CLIP-FH: Fine-Tuning CLIP for Hand-Based Identity Matching

(MSc Artificial Intelligence Final Project)

Babu Pallam P2849288

Supervisor: Dr Nathanael L. Baisa

2nd Marker: Prof. Yingjie Yang

De Montfort University

MSc Artificial Intelligence

Online Viva via MS Teams: Thursday, 05 June 2025

05 June 2025

Outline

Why Hand-Based Biometrics?

- Traditional biometrics (face, fingerprint) present real-world limitations:
  - **Privacy Concerns:** Facial data can be tracked or misused.
  - **Occlusion Issues:** Masks, headgear, or gloves reduce accuracy.
  - **Hygiene Factors:** Fingerprint scanners require physical contact.
- **Hands are a practical alternative:**
  - Easy to capture in public and healthcare settings.
  - Less invasive and socially neutral than face scanning.
- However, hand biometrics are underexplored in deep learning—especially with vision-language models.

Research Gap & Technical Motivation

- Most existing hand recognition systems rely on **CNNs**, which:
  - Struggle to generalise across lighting, pose, and accessories.
  - Are trained with handcrafted modules and domain-specific tuning.
- **CLIP (Contrastive Language-Image Pretraining)** offers a new paradigm:
  - Jointly learns image and text representations.
  - Pre-trained on 400M diverse image-text pairs.
  - Excels in zeroshot generalisation.
- **Question:** Can CLIP be adapted to work effectively in a biometric setting like hand identity matching?

Research Objective

To investigate whether the CLIP vision-language model can be effectively adapted and fine-tuned for hand-based biometric identification.

Research Questions

This project explores whether **CLIP can be adapted** for hand-based identity matching by answering:

- **RQ1:** How well does CLIP perform on hand images without any fine-tuning?
- **RQ2:** Does fine-tuning the image encoder improve hand recognition?
- **RQ3:** Can CLIP-ReID style prompt learning make CLIP more discriminative for hands?
- **RQ4:** Does PromptSG with learnable prompts further improve identity matching?
- **RQ5:** How do different pretrained models differ in hand feature learning and training stability?
- **RQ6:** How does fine-tuned CLIP compare to CNN models like MBA-Net in retrieval performance?

Vision-Language Models for ReID

- **CLIP** (Radford et al. 2021) — Pretrained on 400M image–text pairs using contrastive learning.
  - Enables zeroshot classification, retrieval, and cross-modal understanding.
  - Provides transferable visual features even for unseen domains like biometrics.
- **CLIP-ReID** (S. Li, Sun, and Q. Li 2023) — Two-stage adaptation strategy for person ReID.
  - **Stage 1:** Learns identity-specific pseudo-text tokens with frozen encoders.
  - **Stage 2:** Fine-tunes the image encoder.
  - Achieves strong performance without relying on concrete text descriptions.
- **PromptSG** (Yang et al. 2024) — Semantic prompt-guided ReID with learnable pseudo-tokens.
  - Trains prompts and visual features end-to-end in a unified framework.
  - Uses prompt ensembles (e.g., "a person with [tokens]") for generalisation.
  - Outperforms previous CLIP-based ReID methods without needing extra labels.

System Pipeline Overview

- **Dataset:** 11k Hands dataset(Afifi 2019) (dorsal right only), filtered for accessories and balanced across identity splits.
- **Backbones:** ViT-B/16 and RN50 used as CLIP image encoders.
- **Stage 1:** Zero-shot inference with frozen CLIP.
- **Stage 2:** CLIP-ReID Integration.
- **Stage 3:** PromptSG integration.
- **Evaluation:** 10 Monte Carlo query-gallery splits using cosine similarity, evaluated with mAP and Rank@K metrics.

CLIP-FH Training Stages Overview

Stage 1: Baseline CLIP (Zero-Shot)

- Used pretrained ViT-B/16 and RN50 without any fine-tuning.
- Only image features were extracted using the frozen image encoder.
- Identity matching done using cosine similarity.

Stage 2: CLIP-ReID Integration

- Inspired by CLIP-ReID two-stage training.
- Stage 1 : Learn trainable text tokens per ID - image/text encoders remain frozen.
- Stage 2 : finetune the image encoder to align with the text representations.
- Boosts alignment between image and learned semantic representations.

Stage 3: PromptSG Integration

- Based on PromptSG: dynamic prompts for identity-level supervision.
- Uses an inversion network to generate pseudo-token prompts for each image.
- Applies prompt-driven semantic attention to guide the image encoder.
- image features align better with the learned text tokens

Comparative Summary: Stages 1–3

Stage	Model	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)	mAP (%)
Stage 1	ViT-B/16	71.42	88.53	93.94	78.98
Stage 1	RN50	68.61	83.92	90.43	75.78
Stage 2 (v5)	ViT-B/16	**88.18**	**97.32**	**98.58**	**92.20**
Stage 2 (v1)	RN50	57.64	77.74	85.09	67.07
Stage 3 (v8)	ViT-B/16	85.86	95.32	97.74	89.99
Stage 3 (v3)	RN50	68.99	85.13	90.93	76.28

Performance Insights:

- **ViT-B/16 consistently outperformed RN50** across all stages in both Rank-1 and mAP.
- **Stage 2 (ViT-B/16):** Largest boost (+13.2% mAP) from supervised fine-tuning with ReID enhancements (ArcFace, BNNeck).
- **Stage 3 (PromptSG):** Semantic prompt tuning added moderate yet meaningful gains via joint image-text training.
- **ViT Strength:** Stable across tuning strategies; better generalisation and prompt integration.
- **RN50 Weakness:** Sensitive to configurations and deep tuning; less stable in PromptSG integration.

Comparison with State-of-the-Art

Method	**Rank-1 (%)**	**mAP (%)**
MBA-Net (Nathanael L Baisa et al. 2022a)	**97.45**	**97.98**
ABD-Net (Chen et al. 2019)	95.89	96.76
RGA-Net (Zhang et al. 2020)	94.77	95.67
GPA-Net (Nathanael L. Baisa et al. 2022b)	94.80	95.72
Ours – ViT-B/16 (Stage 3, PromptSG v8)	85.86	89.99
Ours – ViT-B/16 (Stage 2, CLIP-ReID v5)	88.18	92.20
Ours – RN50 (Stage 3, PromptSG v3)	68.99	76.28
Ours – RN50 (Stage 2, CLIP-ReID v1)	57.64	67.07

Key Points:

- ViT-B/16 with CLIP-ReID and PromptSG shows competitive results despite no hand-specific CNN modules.
- All compared SOTA methods use handcrafted attention or multi-branch CNNs tailored for hands.
- Our results validate the potential of general-purpose VLMs like CLIP for scalable biometric matching.

Project Reflection: What Worked and What Didn't

What Went Well

- Staged pipeline (Stages 1–3) allowed systematic evaluation.
- ViT-B/16 performed consistently well throughout.
- CLIP-ReID achived 92% mAP.
- PromptSG improved semantic understanding via prompt learning.
- YAML-based config ensured modularity and reproducibility.
- 10-fold ReID-style setup helped fair benchmarking.

What Didn't Go Well

- RN50 was unstable during prompt-guided training.
- Frozen text encoder limited deeper semantic alignment.
- GPU limits the experiments.

Future Research Directions

- **Dataset Expansion:** Extend to all 11k variants (dorsal/palmar, both hands) and explore HD Hands dataset for broader evaluation.
- **Text Encoder Fine-Tuning:** Unfreeze and jointly train the text encoder to improve semantic alignment and multimodal grounding.
- **Prompt Adaptation:** Use large language models (LLMs) for dynamic prompt generation and ensemble-based robustness.
- **Anatomical Localisation:** Integrate attention maps to highlight biometric cues (e.g., veins, knuckles) for fine-grained matching.

Contributions and Originality

- **Novel Adaptation:** First known staged adaptation of CLIP for hand-based biometrics, extending CLIP-ReID and PromptSG from person to hand identity.
- **ReID Benchmarking:** Introduced a 10-split ReID-style evaluation protocol tailored to hand datasets for fair and repeatable benchmarking.
- **Backbone Comparison:** Empirically showed ViT-B/16 consistently outperforms RN50 in multimodal fine-tuning and stability.
- **Reproducibility Impact:** Delivered a modular, YAML-configurable pipeline contributing to the field of multimodal AI in low-resource biometric domains.

Conclusion & Key Findings

To summarise the key findings and research contributions:

- **CLIP performed well even without fine-tuning**, especially the ViT-B/16 model, which reached over 71% accuracy on hand images.
- **CLIP-ReID integration**—where I fine-tuned the image encoder with ReID losses like ArcFace, Triplet, Center, and SupCon—led to the highest performance gains, achieving 92.20% mAP.
- **PromptSG integration** added semantic supervision using learnable pseudo-prompts and joint image-text training. This improved the robustness of identity retrieval, particularly with the ViT backbone.
- **ViT-B/16 outperformed RN50** across all stages, showing better stability, higher accuracy, and stronger compatibility with prompt learning methods.
- **Compared to MBA-Net:** While MBA-Net achieved the highest accuracy overall, my adapted CLIP models came close—without any handcrafted CNN modules. This shows that general-purpose vision-language models can be effectively adapted for biometric recognition tasks.

References I

- Afifi, Mahmoud (2019). "11K Hands: Gender Recognition and Biometric Identification Using a Large Dataset of Hand Images". In: *Multimedia Tools and Applications* 78.23, pp. 33035–33057. DOI: 10.1007/s11042-019-7424-8. URL: https://doi.org/10.1007/s11042-019-7424-8.
- Baisa, Nathanael L et al. (2022a). "Multi-branch with attention network for hand-based person recognition". In: *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 727–732.
- — (2022b). "GPA-Net: Global and Part-Aware Network for Hand-Based Person Identification Using Deep Feature Learning". In: *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*.
- Chen, Tianlong et al. (2019). "Abd-net: Attentive but diverse person re-identification". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8351–8361.

References II

- Li, Siyuan, Li Sun, and Qingli Li (2023). "CLIP-ReID: Exploiting Vision-Language Model for Image Re-Identification without Concrete Text Labels". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 37. 4, pp. 3919–3927. DOI: 10.1609/aaai.v37i4.25408. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25408.
- Radford, Alec et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *arXiv preprint arXiv:2103.00020*. DOI: 10.48550/arXiv.2103.00020. URL: https://arxiv.org/abs/2103.00020.
- Yang, Zexian et al. (2024). "A pedestrian is worth one prompt: Towards language guidance person re-identification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17343–17353.
- Zhang, Zhizheng et al. (2020). "Relation-aware global attention for person re-identification". In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 3186–3195.

Thank You!

Questions or Comments?