Face Recognition in Historical Documents

KNN - checkpoint report

https://github.com/Knowny/HistoricalFaceRecognition

Bc. Simona Jánošíková (xjanos19)
Bc. Tereza Magerková (xmager00)
Bc. Tomáš Husár (xhusar11)

Date: May 2025

Faculty of Information Technology

1 Introduction

This project focuses on facial recognition in historical archival materials such as books, magazines, and paintings. Modern systems perform well on contemporary images but struggle with historical ones due to degraded quality and differing photographic processes. This domain gap hinders generalization.

Our goal is to build a system that extracts robust facial descriptors from historical images, producing numerical features for each detected face to capture individual identity.

2 Existing Solutions

Modern facial recognition builds on deep learning advances. **Inception** networks [1] introduced parallel convolutional paths to capture multi-scale features, while **ResNets** [2] made training deeper networks feasible.

FaceNet[3] reframed the task as learning embeddings using triplet loss, enabling a unified approach to recognition. This led to improved loss functions like **ArcFace**[4] and **CosFace**, which use angular or margin-based penalties to enhance embedding quality.

Recent approaches emphasize efficiency and scalability. **EfficientNets** optimize depth, width, and resolution, while transformer-based models like **SwinFace**[5] and **TransFace**[6] capture global dependencies, rivaling or surpassing CNN-based methods.

3 Datasets

The choice of dataset plays a critical role in the effective training and evaluation of machine learning models. In the context of this work, a high-quality dataset is particularly important due to the challenges associated with accurate face recognition in historical materials.

3.1 Training Data

The training process requires a dataset containing multiple images for each identity to support the learning of robust facial features. The CASIA dataset was selected for this purpose due to its substantial number of subjects and diverse image conditions. Examples from this dataset are shown in the figure 1.

Figure 1: Examples from the CASIA dataset

To enhance training data and handle the variability of historical images, we use style transfer-based data augmentation to create **stylized_images dataset**. This improves generalization across diverse visual styles in old photos and artworks. Details are provided in subsection 6.1.

3.2 Testing Data

For testing purposes of our model, a WikiFace dataset (provided by our supervisor) has been utilized. This dataset possesses several challenges notably:

- multiple individuals in a single image (multiple face detections)
- erroneous data (where images lack a portrait of a person altogether)
- portraits of varying quality (such as sculptures, less realistic paintings or caricatures, low resolution photos, degraded photos, ...)

Examples from this dataset are shown in the figure 2.

Figure 2: Examples from the WikiFace dataset

The WikiFace dataset, in its initial form, contained:

- 1656 Unique Identities
- 3394 Images Suitable for Detection
- 3778 Detections (including false positives)

3.3 Dataset Pre-processing

Pre-processing is a crucial step to ensure that all face images are consistent and suitable for training and evaluation. This includes aligning facial features, normalizing image dimensions, and more. The process used to standardize facial inputs is described below.

3.3.1 Face Alignment

The model expects a well-aligned image of an individual's face. To achieve this, we use the ArchivalFaces detector, based on YOLO11 [7] and provided by our supervisor, which detects five key facial landmarks: the eyes, nose, and mouth corners.

These landmarks guide a sequence of transformations including rotation, scaling, and cropping, resulting in a normalized face image that matches the model's input requirements. The full alignment process is shown in Figure 3.

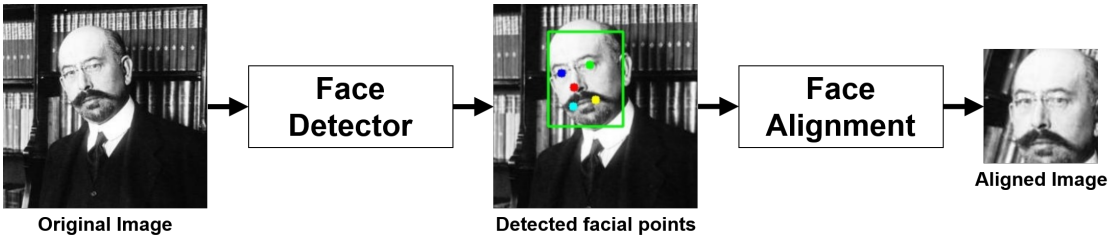

Figure 3: Face Alignment Pipeline

Aligned images produced by this process are then used as input to the feature extraction model. This alignment ensures consistency across the dataset and improves model performance by reducing pose and scale variation.

3.3.2 Removal of False Positives

In this work, false positives refer not only to objects incorrectly identified as faces, but also to cases where the aligned face image does not correspond to the correct identity. This typically occurs when multiple detections are made within a single image, leading to mismatches. These instances were identified through manual inspection of low-confidence scores and duplicate detections, and were subsequently removed.

Additional removals were necessary for:

- Identities with only a single image, which could not be used for comparison.
- A mislabeled identity from the CASIA dataset, where the generated images contained random, unrelated individuals. This appears to be an error in the dataset itself.

After removing the above-mentioned cases, the updated dataset statistics are in the table 1:

Table 1: Basic statistics of stylized_images and WikiFace datasets

	stylized_images	**WikiFace**
Number of Identities	999	1536
Number of Images	14975	3223

4 Evaluation Metrics

Model performance is evaluated using standard binary classification metrics: ROC curve, AUC, and DET curve. Together, these provide a comprehensive view of the model's ability to distinguish between positive and negative classes.

4.1 ROC Curve

The ROC (Receiver Operating Characteristic) curve illustrates the trade-off between the **True Positive Rate (TPR)** and the **False Positive Rate (FPR)** across thresholds:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

An ideal ROC curve approaches the top-left corner, indicating high sensitivity with low false positives. A diagonal curve denotes random guessing, while curves below it suggest poor performance. ROC analysis aids in threshold selection and model comparison.

4.2 AUC

The AUC (Area Under the Curve) summarizes the ROC curve as a single value between 0 and 1. It reflects the probability that a randomly chosen positive instance is ranked higher than a negative one. A value of 1.0 indicates perfect classification, 0.5 indicates chance-level performance, and values below 0.5 imply systematic misclassification. AUC is threshold-independent and useful for assessing ranking quality in facial similarity tasks.

4.3 DET Curve

The DET (Detection Error Tradeoff) curve plots **False Negative Rate (FNR)** against **False Positive Rate (FPR)**:

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

With logarithmic or normal deviate scaling, the DET curve highlights error rates in low-failure regions. Curves closer to the origin indicate better performance. This visualization supports the selection of thresholds that balance missed detections and false alarms.

4.4 Post-Finetuning Metrics

The curves described give a good general-purpose overview to how the model fares on the data. For the finetuned model we followed the same metrics and added some more.

4.4.1 Equal Error Rate (EER)

EER is a single point from the DET (or ROC) curve where false positive rate and false negative rate are equal – where error of both types are balanced. This provides a singular metric that can be minimized in order to minimize error.

4.4.2 True Acceptance Rate (TAR)

In applications where missing a positive instance is more costly than a false alarm (such as biometric authentication) it is important to measure the proportion of actual positive cases that were identified by the model at a specific acceptable level of false alarms. TAR is given at various false positive rates (denoted as `TAR@FPR=0.01`). During finetuning process TAR values at different FPR are maximized.

Its inverse metric is **False Match Rate (FMR)** what follows an inverse concern: *How likely is there going to be a match for two different individuals?*

While FMR represent "security risk" TAR represents the "system reliability". With this in mind this project follows TAR at a fixed FMR rate. The values are measured as: `TAR@FMR=0.1`, `TAR@FMR=0.01`, `TAR@FMR=0.001` and `TAR@FMR=0.0001`.

5 Baseline solution

As a baseline, **FaceNet** by Schroff et al. [3] was used. This model embeds facial images into a Euclidean space where distances reflect similarity. The chosen implementation[1] is based on the **Inception-ResNet-v1** architecture.

This architecture has previously demonstrated solid performance on similar tasks involving historical data [8], where various models were evaluated under comparable conditions. Inception-ResNet-v1 was chosen for further exploration, using the version pretrained on the VGGFace2 dataset [9].

Proposed by Szegedy et al.[10], Inception-ResNet-v1 combines Inception modules with residual connections for improved training efficiency. It was chosen over the v2 variant due to its lower computational cost. The structure consists of repeating Inception-ResNet and Reduction blocks, shown in Figure 4.

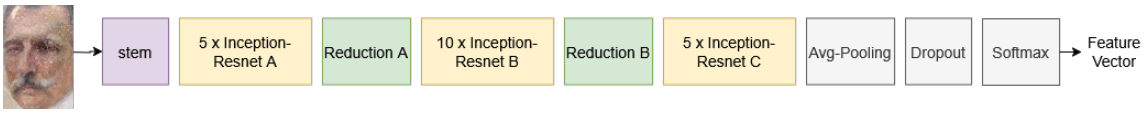

Figure 4: Inception-ResNet-v1 architecture

5.1 Baseline Evaluation

The baseline solution was evaluated on the cleaned and aligned WikiFace data. The model fared relatively well as shown in Figure 5 and Table 2. The AUC shows it has a decent ability to distinguish between pairs of positive and negative pairs. The TAR values suggest that if we desire stricter security from the system it would become unreliable.

[1] https://github.com/timesler/facenet-pytorch.git

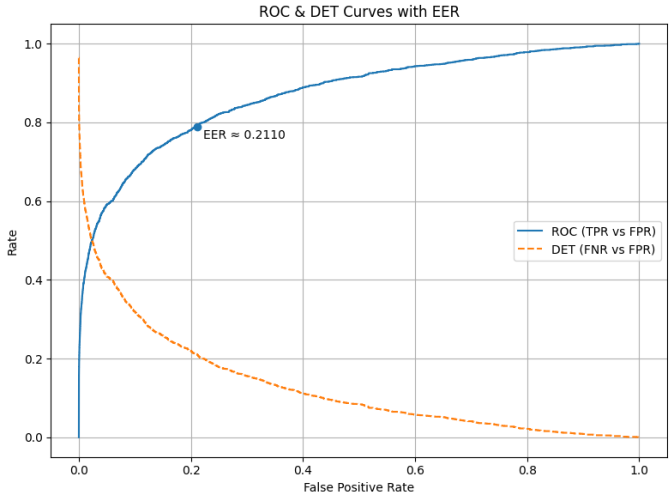

Figure 5: Baseline evaluation on WikiFace dataset

The EER[2] suggests a considerable error of the baseline model. The goal of the finetuning part of this project would be to maximize the AUC value, lower EER and maximize TAR values at defined FMR levels.

6 Our Solution

6.1 Image Generation

Style transfer generates images that retain the original content while adopting a new visual style. While state-of-the-art approaches such as diffusion models achieve impressive stylization, they often struggle to balance content preservation with stylistic changes. For this project, preserving identity, the core content, is prioritized.

The solution uses the diffusion-based model **PhotoMaker**[3], introduced by Zhan et al. [11]. PhotoMaker is a personalized text-to-image generation method that preserves identity embeddings and allows control over stylization strength. This tunable parameter enables balancing between visual fidelity and stylization.

Image generation is guided by a pair of prompts: one positive and one negative. A consistent negative prompt was used across the dataset, discouraging digital artifacts, poor anatomy, and modern elements. Positive prompts were randomly sampled from a curated prompt bank designed to reflect the diversity of the **WikiFace** dataset. Pipeline of this process is shown in the Figure 6.

[2]The EER occurs when FPR = FNR. The ROC curve plots TPR (1 - FNR) vs. FPR, so the EER point appears at (EER, 1 - EER). The DET curve plots FNR vs. FPR, showing EER along y = x. The different y-axes cause the visual separation

[3]https://github.com/TencentARC/PhotoMaker.git

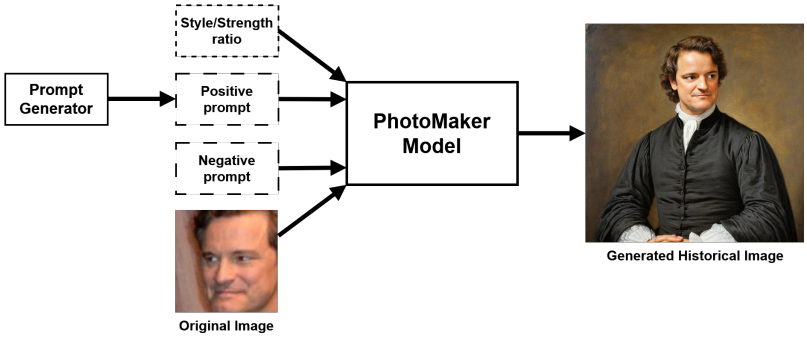

Figure 6: Image Generation pipeline

6.2 Face Recognition

Our approach focused on how the FaceNet was pretrained – triplet loss. The task does not change at all it is still face verfication, therefore the task was to get the network accustomed to the difference in data. This approach was further experimented on different values of hyper-parameters of number of epochs, learning rate, batch size and the value of margin for triplet loss. The results are shown in Table 2 and Figure 8.

7 Experiments & Results

This section provides the approaches taken to reach the presented results. Different experiments and approaches were designed with reference to achieved results and literature review.

7.1 Style Transfer

The generation of the training set was done on CASIA-WebFace subset. The later parts of the pipeline work with the first 1000 personalities from this dataset, where the images were aligned to face and labeled to corresponding identity.

The `style_ratio` described in 6.1 was experimentally set to 15. This setting produced clear preservation of identity while still managing to produce believable atmosphere of a historical piece.

Prompts were manually crafted and applied to the input images. Each personality ended with 15 samples each produced by one script from the prompt bank. The prompts tried to simulate different types of media found in the WikiFace data. The results are shown in Figure 7. The most "good" results were produced by prompts describing oil paintings and cyanotype and tinted photographs. The worst performing were caricatures and sketches (generation of images containing statues was also tested, but the results remained unsatisfactory, so we omitted those from the prompt bank).

The negative prompt was iteratively filled in with non desirable traits such as makeup, bad anatomy (creepy AI fingers are still present) text and (surprising to the authors) nudity.

Figure 7: Historical images generated with PhotoMaker

7.2 Finetuning

The first attempts followed the plan in halfterm report[4] and implemented a classification head that was supposed to be trained on the training data (stylized CASIA). This approach failed as the model failed when the task changed from classification to face verification.

Next attempts consisted of variations of using different learning rates, optimizers, schedulers, loss function, utilizing the logits layer or not using it. No real results were reached here either.

The final experiment was then further experimented on[5] by utilizing a `OneCycleLR` scheduler to achieve better results using super-convergence as described in [12]. It was paired with `AdamW` optimizer. This was only an experiment and this project does not consider it as a final result as it needs more experimenting and tuning of the parameters used.

[4] refer to file: `experiment1.py`
[5] refer to file: `experiment2.py`

8 Conclusion

Table 2 presents a detailed comparison between the baseline FaceNet model and our finetuned variant. Despite being trained specifically on our stylized dataset using triplet loss, the **finetuned model** did not outperform the baseline. In fact, it demonstrated lower overall verification performance across all evaluated metrics, suggesting that the fine-tuning process in its current form was not sufficient to improve upon the pretrained model's capabilities.

This performance gap may be attributed to several factors:

- Suboptimal training hyperparameters (e.g., margin, learning rate, batch size)
- A limited number of training samples per identity
- The training dataset may not have been representative of real-world historical data (e.g., synthetic or stylized variations may not generalize well)

In conclusion, while finetuning showed potential, our current configuration did not yield improved results. Future work should explore more sophisticated triplet mining strategies (e.g., semi-hard negatives), optimize hyperparameters, and consider using larger and more diverse historical datasets to better align training with the intended target domain.

Table 2: Baseline and finetuned model comparison

Model	AUC	EER	TAR@FMR=0.1	TAR@FMR=0.01	TAR@FMR=0.001	TAR@FMR=0.0001
Baseline FaceNet	0.8717	0.2110	0.6808	0.4055	0.2266	0.1350
Finetuned model	0.8216	0.2606	0.5845	0.3140	0.1753	0.1120

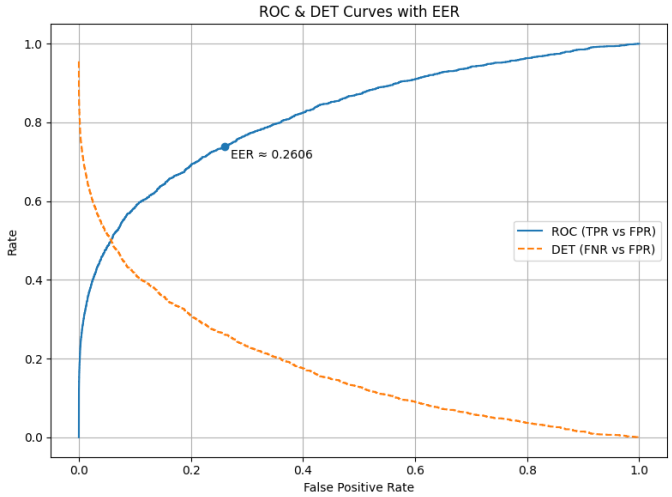

Figure 8: Finetuned model evaluation on WikiFace

References

[1] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[2] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[4] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *CVPR*. 2019.

[5] Lixiong Qin et al. "SwinFace: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation". In: *IEEE Transactions on Circuits and Systems for Video Technology* 34.4 (2023), pp. 2223–2234.

[6] Jun Dan et al. "Transface: Calibrating transformer training for face recognition from a data-centric perspective". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 20642–20653.

[7] Glenn Jocher and Jing Qiu. *Ultralytics YOLO11*. Version 11.0.0. 2024. URL: https://github.com/ultralytics/ultralytics.

[8] Anil Poudel. *Face recognition on historical photographs*. 2021.

[9] Qiong Cao et al. *VGGFace2: A dataset for recognising faces across pose and age*. 2018. arXiv: 1710.08092 [cs.CV]. URL: https://arxiv.org/abs/1710.08092.

[10] Christian Szegedy et al. "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[11] Zhen Li et al. "Photomaker: Customizing realistic human photos via stacked id embedding". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 8640–8650.

[12] Leslie N. Smith and Nicholay Topin. *Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates*. 2018. arXiv: 1708.07120 [cs.LG]. URL: https://arxiv.org/abs/1708.07120.