# Assignment #4: Deep Face Recognition Pipeline

Tjaš Ajdovec
*IBB 2025/26 , FRI, UL*
*ta5946@student.uni-lj.si*

## I. Introduction

Face recognition is a crucial task with many practical applications, such as security and authentication. A typical face recognition pipeline consists of multiple stages, including face detection, feature extraction, and recognition (matching). Approaches range from traditional filter-based methods to deep learning models, such as convolutional neural networks (CNNs). This report presents the use of pretrained deep learning models for face detection and recognition and compares their performance against simpler traditional models.

## II. Methodology

Face detectors produce bounding boxes around faces. The Viola–Jones algorithm, which relies on simple Haar-like features, is widely used for real-time face detection. The **YOLO** detector, presented in *You Only Look Once: Unified, Real-Time Object Detection* [1], is a pretrained lightweight neural network for object detection, with versions fine-tuned for face detection. We also evaluated the detector component of the **InsightFace** end-to-end face recognition model.

Once faces are detected, images can be cropped for feature extraction. We implemented three simple methods: Uniform Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Dense Scale-Invariant Feature Transform (SIFT) using a fixed grid of keypoints. These were compared with deep learning extractors:

- **FaceNet**, introduced in *FaceNet: A Unified Embedding for Face Recognition and Clustering* [2], maps faces into compact 128-dimensional feature vectors.
- **InsightFace** models use ArcFace with additive angular margin loss for highly discriminative embeddings.

Outputs from deep models can be L2-normalized and compared using **cosine similarity**. We used the training part of the dataset to select the best-performing detector. The final recognition evaluation was conducted on the test set, first on full images and then with detected bounding boxes.

## III. Experiments

We used the *CelebA-HQ-Small* dataset, containing around **900 high-resolution** (1024×1024) celebrity face images, with a 55/45 train-test split.

The Viola–Jones detector was tuned via grid search. We also evaluated two YOLO versions, `yolo8n-face` (nano) and `yolov12s-face` (small), and the detector from InsightFace `buffalo_l` model set. Face detection performance, measured by **Intersection over Union** (IoU), is shown in Table I. The InsightFace detector achieved the highest IoU and was selected for the final evaluation.

TABLE I
Face detection performance of different detector models on the training set. Results include IoU and total evaluation time.

| Detector | IoU | Eval Time (s) |
|---|---|---|
| Viola-Jones | 0.673 | 34 |
| YOLOv8n | 0.853 | 26 |
| YOLOv12s | 0.855 | 70 |
| **InsightFace** | **0.861** | **55** |

Simple extractors were implemented using `OpenCV` and `scikit-image`. InsightFace and FaceNet were used via official libraries. The final evaluation included 10 models: each extractor, both with and without the selected detector. Metrics included rank-1 and rank-5 accuracies, indicating the probability that a correct match appears in the top $k$ retrieved faces, and cumulative evaluation time on the test set of 410 faces.

## IV. Results and Discussion

### A. Results

Table II presents the results. Between traditional models, SIFT with a detector achieved the highest performance, with rank-1 accuracy of 43% and rank-5 accuracy of 65%. The best overall performance was InsightFace without a detector, with **rank-1 accuracy of 99.3%** and **rank-5 accuracy of 99.8%**, and the fastest evaluation time of 17 seconds. Deep models were generally more accurate and faster than traditional methods, but their performance decreased when using detected bounding boxes.

IoU scores for detectors on the test set showed no significant differences.

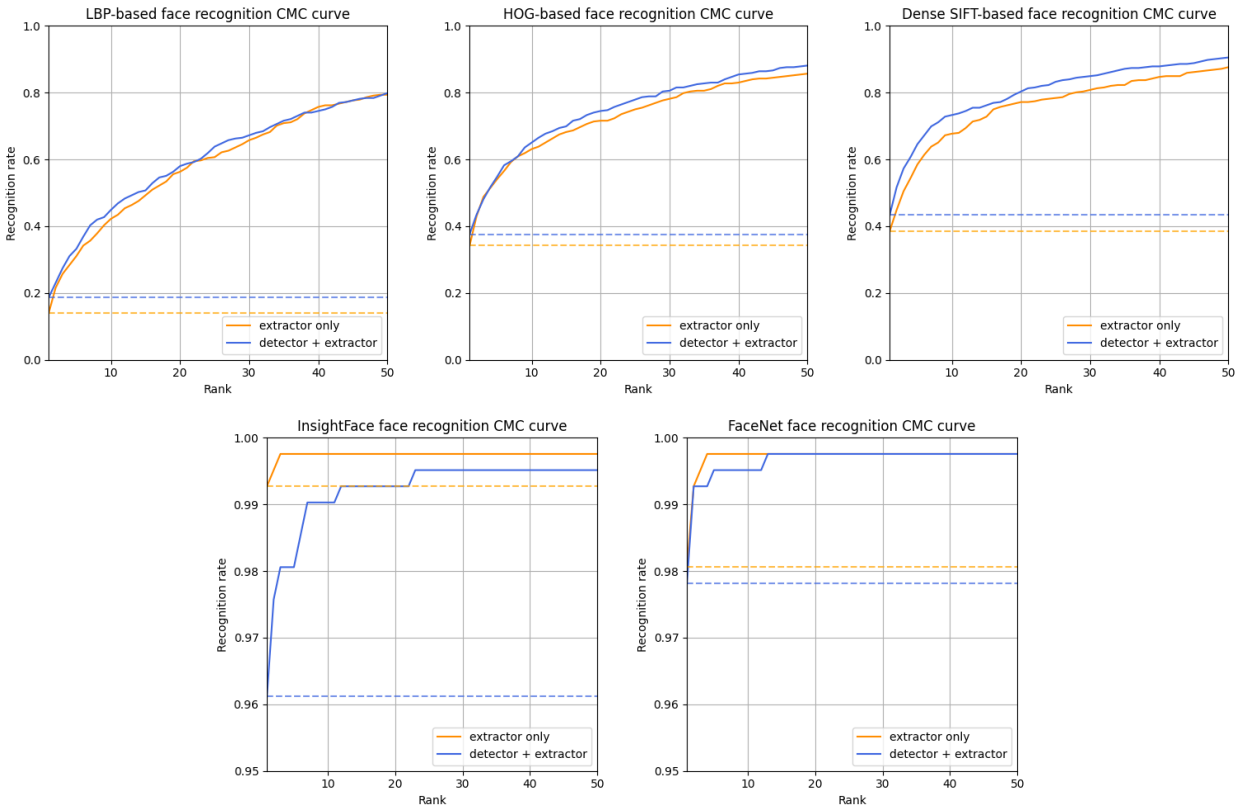


Fig. 1. CMC curves for SIFT, InsightFace, and FaceNet feature extractors, with and without the detector model. The last two plots use a different scale.

TABLE II
FACE RECOGNITION PERFORMANCE ON THE TEST SET OF DIFFERENT FEATURE EXTRACTOR MODELS, WITH AND WITHOUT A DETECTOR. METRICS INCLUDE RANK-1 AND RANK-5 ACCURACIES, AND TOTAL EVALUATION TIME.

| Extractor | Rank-1 Acc | Rank-5 Acc | Time (s) |
|---|---|---|---|
| LBP | 0.141 | 0.311 | 154 |
| Detector + LBP | 0.187 | 0.333 | 112 |
| HOG | 0.342 | 0.541 | 22 |
| Detector + HOG | 0.376 | 0.549 | 73 |
| SIFT | 0.386 | 0.585 | 71 |
| **Detector + SIFT** | **0.434** | **0.646** | **119** |
| **InsightFace** | **0.993** | **0.998** | **17** |
| Detector + InsightFace | 0.961 | 0.981 | 72 |
| FaceNet | 0.981 | 0.998 | 26 |
| Detector + FaceNet | 0.978 | 0.995 | 92 |

### *B. Discussion*

Rank-1 accuracies for traditional models were relatively low, though a detector improved their performance. Deep learned models performed well without preprocessing, but performance slightly degraded on cropped images. This is visible in the **Cumulative Match Characteristic** (CMC) curves in Figure 1. Two possible causes for this are information loss and increased face misalignment. Overall, 99% accuracy on unseen data a highly effective solution. Future work could explore additional CNN architectures, real-time multi-object detection, and handling occlusions.

## V. CONCLUSION

This report evaluated multiple face detection and recognition models on CelebA-HQ-Small. Traditional extractors, including LBP, HOG, and SIFT, achieved limited performance, with Detector + SIFT reaching 43.4% rank-1 and 64.6% rank-5 accuracy. In contrast, FaceNet and InsightFace significantly outperformed these methods. InsightFace achieved 99.3% rank-1 accuracy, 99.8% rank-5 accuracy, and required only 17 seconds for evaluation on 400 images. Using a separate detector slightly improved traditional models but reduced deep model performance due to cropping and misalignment. These results show that pretrained DL models provide an accurate and efficient solution for face recognition, with minimal preprocessing.

## REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[2] F. Schroff, D. Kalenichenko, and J. 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.