An evaluation of lightweight facial recognition models available in the DeepFace library in clustering tasks

Leandro Barbosa da Silva Francisco[1], **Marlon Marcon**[1], **André Roberto Ortoncelli**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Dois Vizinhos - PR - Brasil

`leandrobs1710@gmail.com`

`{marlonmarcon,ortoncelli}@utfpr.edu.br`

Abstract. *This work investigates the use of Deep Learning-based facial recognition models to automatically organize images through unsupervised clustering. We evaluated the FaceNet and ArcFace models using the DeepFace library on a dataset of 100 images of five celebrities. The results show an average accuracy of 92.6%, demonstrating that the choice of model and the clustering algorithm configuration significantly influence performance and enable high accuracy in automatic image organization.*

1. Introduction

Photographers spend a significant amount of time manually organizing photographs after a photoshoot, especially when they need to classify and separate images by the individual portrayed. Although researchers have achieved significant advances in this field using Artificial Intelligence techniques, current tools still present limitations in effectively addressing this specific demand.

Recent advances in facial recognition have prioritized the development of more robust, efficient, and scalable models, especially for application in devices with computational constraints. In the context of lightweight, robust models for facial recognition in uncontrolled environments, one can list various efforts based on Deep Learning for facial recognition [Chen et al. 2018, Xiao et al. 2024].

One notable example is MobileFaceNet, proposed in [Chen et al. 2018], which adapts Google's (MobileNet) principles for facial recognition, prioritizing low computational cost and application on mobile devices. MobileFaceNet is currently one of the most widely used algorithms for facial recognition, and recent work proposes architectural optimizations and training strategies to improve the model's performance and efficiency [Xiao et al. 2024].

Beyond architectural contributions, significant advances in facial recognition have also occurred in the loss functions used for training. Of particular note is ArcFace, proposed by [Deng et al. 2019], which introduces an Additive Angular Margin Loss to increase angular separability between classes in the feature space. Unlike traditional loss functions based solely on softmax, ArcFace imposes an explicit angular margin between embeddings of different individuals, promoting greater interclass discriminability and less intraclass variability. This approach has become a benchmark in facial recognition, widely incorporated into various architectures—including lightweight models—because it delivers consistent performance gains without significantly increasing inference cost.

Other recent works in facial recognition include CFormerFaceNet [He et al. 2023], which combines a CNN with a Transformer-type attention block to capture local and global face representations, reducing parameters and computational cost without sacrificing accuracy, making it suitable for recognition on limited devices. RobFaceNet [Khalifa 2025] uses a lightweight CNN with integrated attention mechanisms at multiple levels to enhance facial feature extraction. This model achieves performance comparable to that of larger architectures at a much lower computational cost, making it stand out for its efficiency. Also noteworthy is FaceLiVT [Setyawan et al. 2025], a hybrid CNN-Transformer model with lightweight linear attention and structural reparameterization, explicitly designed for mobile devices, that offers significant latency reduction and competitive performance on facial recognition benchmarks.

We are exploring opportunities to develop new tools for facial recognition and to support photographers. To that end, this study presents initial results from DeepFace[1] and the DBSCAN algorithm [Ester et al. 1996], an application that makes facial recognition accessible to photographers by dispensing with advanced infrastructure and enabling direct use in real-world professional scenarios.

2. Methodology

We conducted experiments using the Celebrity Face Image Dataset[2]. We selected 100 images from each of five celebrities in the database, for a total of 500 images. The selected celebrities were: i) Brad Pitt; ii) Jennifer Lawrence; iii) Angelina Jolie; iv) Denzel Washington; and v) Hugh Jackman.

In the experiments, we adopted Multi-task Cascaded Convolutional Neural Network (MTCNN) as the backend for facial detection and alignment. We evaluated the ArcFace and FaceNet models for facial embedding generation. We also varied the min_samples parameter of the DBSCAN algorithm, which determines the minimum number of samples required to form a valid cluster, to assess its direct impact on clustering.

In one of the experimental instances, data augmentation techniques were applied, including Mixup, CutMix, Random Erasing, frequency-domain transformations using the DCT (Discrete Cosine Transform), and geometric deformations. These perturbations enable analysis of the models' behavior under different visual conditions and image degradation levels.

Details of the three experiments performed are presented below:

- **Experiment 1:** Evaluated the performance of the FaceNet512 model in the facial clustering task using the DBSCAN algorithm with fixed parameters: *min_samples* = 5 and $\epsilon = 0.65$.
- **Experiment 2:** Investigated the impact of varying the density parameter in DBSCAN, maintaining the FaceNet512 model *min_samples* = 5, but with $\epsilon = 0.75$, allowing analysis of the clustering sensitivity to increasing neighborhood radius.
- **Experiment 3:** With the same configuration as experiment 2, but using the ArcNet model.

[1]https://pypi.org/project/deepface/
[2]https://www.kaggle.com/datasets/vishesh1412/celebrity-face-image-dataset

- **Experiment 4:** Evaluated the impact of the combined application of the Mixup, Elastic Transform, and CutMix data augmentation techniques on the performance of the FaceNet model. Three synthetic variants were generated per sample—one for each technique—totaling 1,500 new images. Thus, each identity had 400 instances (100 original and 300 augmented), resulting in a final set of 2000 images submitted to the DBSCAN algorithm, with ϵ = 0.75 and *min_samples* = 50.

3. Results

Table 1 shows the number of correctly grouped images for each celebrity in each experiment. For the experiment, since the data augmentation techniques increased the dataset from 100 to 400 images per celebrity, we divided the number of correctly grouped images by 4 for a fair comparison of results.

Table 1. Number of images correctly classified in each experiment

Celebrity	**Experiment 1**	**Experiment 2**	**Experiment 3**	**Experiment 4**
Brad Pitt	79	90	49	78.5
Jennifer Lawrence	76	93	88	56
Angelina Jolie	66	91	46	54.25
Denzel Washington	82	94	34	85
Hugh Jackman	68	95	90	61
Average	74.2	92.6	61.4	66.95

Experiment 1, using FaceNet512 with $\epsilon = 0.65$ and *min_samples* = 5, achieved an overall average of 74.2%, indicating that the initial DBSCAN configuration formed reasonably consistent clusters but remained sensitive to intra-class variation.

In Experiment 2, increasing the parameter ϵ to 0.75 resulted in a significant improvement in average performance, reaching 92.6%. This result shows that the previous neighborhood radius was too restrictive, fragmenting groups with the same identity. Increasing ϵ made the algorithm more tolerant of natural variations in facial representations, resulting in more complete and coherent groupings.

Experiment 3, which used the ArcFace model maintaining the same configuration as Experiment 2, showed an average of 61.4%, lower than the performance obtained with FaceNet512. This behavior may indicate that, for the specific dataset used and the adopted DBSCAN configuration, the model-generated embeddings did not exhibit a spatial distribution compatible with the defined parameters. Greater variability is also observed among celebrities, suggesting differences in intra-class dispersion.

In Experiment 4, which incorporated data augmentation techniques, the average was 66.95%. The significant increase in the min_samples parameter made the clustering criterion more rigorous, requiring a higher local density for cluster validation. While this reduces false positives, it may also have increased the rate of points considered noise, negatively impacting the total number of correctly clustered images. Furthermore, the introduction of synthetic variations may have increased the embedding dispersion, making it more challenging to achieve the minimum density required by DBSCAN.

In general, the results indicate that facial clustering performance is susceptible to the DBSCAN configuration, especially the ϵ parameter. Among the configurations evaluated, the FaceNet512 combination with $\epsilon = 0.75$ and min_samples = 5 showed the best balance between intra-class cohesion and inter-class separability. The experiments also show that simply adopting more sophisticated models or applying data augmentation does not guarantee automatic improvement in clustering performance; joint adjustment between the embedding model and the clustering algorithm parameters is fundamental.

4. Conclusion

The results demonstrated that facial clustering performance is strongly dependent on the DBSCAN algorithm configuration, especially the ϵ parameter. The combination of the FaceNet512 model with $\epsilon = 0.75$ and *min_samples* = 5 achieved the best average performance (92.6%), demonstrating that properly adjusting the neighborhood radius is crucial for forming more cohesive and consistent clusters. It was also observed that, in theory, more sophisticated models or data augmentation do not, by themselves, guarantee improved clustering, reinforcing the importance of compatibility between embeddings and the clustering strategy.

Future work will involve expanding the experiments to larger and more heterogeneous databases, including real-world scenarios of photographic events. The use of additional quantitative metrics for cluster evaluation will also be investigated, as will techniques for automatic hyperparameter tuning of DBSCAN. There are also plans to implement a new automatic photo organization tool based on the results, explicitly designed to support photographers in real-world scenarios.

References

Chen, S., Liu, Y., Gao, X., and Han, Z. (2018). Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese conference on biometric recognition*, pages 428–438. Springer.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231.

He, L., He, L., and Peng, L. (2023). Cformerfacenet: Efficient lightweight network merging a cnn and transformer for face recognition. *Applied Sciences*, 13(11):6506.

Khalifa, A. A. A. (2025). *Efficient and robust face recognition in the wild*. PhD thesis.

Setyawan, N., Sun, C.-C., Hsu, M.-H., Kuo, W.-K., and Hsieh, J.-W. (2025). Facelivt: Face recognition using linear vision transformer with structural reparameterization for mobile device. In *IEEE International Conference on Image Processing (ICIP)*, pages 1720–1725. IEEE.

Xiao, J., Wang, W., Zhang, L., and Liu, H. (2024). A mobilefacenet-based face anti-spoofing algorithm for low-quality images. *Electronics*, 13(14):2801.