F. Boutros, N. Damer, M. Fang, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and M. Todisco (Eds.): BIOSIG 2024,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2024

Evaluating the Performance of Facial Recognition Algorithms on Morphed Human Faces

Sandra Gomez Castro[1].

Abstract: This paper evaluates the effectiveness of six face recognition algorithms - DeepID, Facenet, Facenet512, OpenFace, SFace, and VGG-Face - on seven datasets, including those morphed using FaceFusion, OpenCV, FaceMorpher, and UBO techniques. It evaluates the ability of these algorithms to identify individuals from morphed faces compared to standard, differently angled, and illuminated images. Through 585,000 comparisons across 150 reference photos, notable performance differences emerge, highlighting the impact of brightness adjustments in moving images and the superior performance of models such as Facenet and Facenet512. In addition, the study examines the distance-threshold difference and its effect on model accuracy in discriminating morphed images. Finally, the primary errors and image features that lead to misclassifications are analyzed to gain insight into common errors.

Keywords: Facial Recognition Models, Face Morphing Attacks, Performance Evaluation, DeepFace Framework, FRGC dataset

1 Introduction

In the rapidly evolving field of biometrics, facial recognition technology has unique advantages over other modalities. The face serves as a highly distinctive biometric characteristic, offering the potential for high convenience and acceptance among data subjects. In addition, this modality allows for the capture of data without physical contact, taking advantage of the high-resolution cameras found in smartphones, which operate primarily in the visible spectrum, thus eliminating the need for specialized capture devices [Bu24a].

The performance of facial recognition systems relies on the accurate capture and analysis of facial features through several critical processing stages. First, segmentation detects and tracks individuals within the captured scene and locates the face or region of interest. After segmentation, image pre-processing is used to improve the signal quality. Then, feature extraction and comparison analyze the texture of the segmented region and compare probe images to reference feature vectors to verify individuals identity [Bu24b].

When developing facial recognition algorithms, it is essential to prevent face manipulation, which includes a variety of techniques ranging from the creation of

[1] MSc. Human-Centered Artificial Intelligence, DTU Compute Building 303B, ORCID: 0009-0006-6208-7086, s233162@dtu.dk

entirely new faces to the alteration of specific attributes or expressions. When examining the types of face manipulation, it is possible to distinguish four levels of complexity in these techniques [TVFM24].

At the top of the complexity scale is the Entire Face Synthesis technique, which involves the creation of entirely new images of the face. This is typically achieved by using advanced Generative Adversarial Networks (GANs), such as StyleGAN, to generate highly realistic facial images. Such technology offers potential benefits in a number of industries, including gaming and 3D modeling, but also raises concerns about misuse to create fraudulent online profiles [TVFM24].

Next is identity swap, which manipulates existing faces by replacing one person's face with another's in videos. This is done using methods ranging from traditional computer graphics to deep learning techniques, as exemplified by DeepFakes. [TVFM24]

Another level down is attribute manipulation, also known as face editing, which involves changing specific facial features or attributes, such as hair color, skin tone, or age. A level within this category is morphing, which involves changing some of a person's attributes while retaining some of the original features and replacing the entire face [TVFM24].

Finally, the least complex level is expression swap, which changes a person's facial expressions in videos, focusing on techniques such as Face2Face and Neural Textures, which can have serious implications, including misrepresentation [TVFM24].

Regarding the literature on morph detection, it is important to note that several studies have been conducted using computer algorithms versus human ability. A 2019 study found that subjects were highly error-prone when detecting morphed faces, with 50/50 face morphs being accepted as morphed faces 68% of the time in one experiment. After providing basic morph detection instruction and an additional "morph" response option, acceptance as morphed faces dropped to 21%. In contrast, the study showed that a simple computer model outperformed human participants in detecting morphed faces, and that advanced computational techniques may prove more reliable than human training in combating face-morphing attacks, which are a major concern for security agencies [KMF19].

Many models have been developed in the field of morphing detection. Among the most recent, one study presented a detection technique based on high-frequency features and progressive enhancement learning. Specifically, this method first extracts high-frequency information from the three color channels of an image to accurately identify details and texture variations. A progressive enhancement learning framework is then used to integrate high-frequency data with RGB information. This framework includes self-enhancement and interactive enhancement modules that gradually refine features to detect faint morphing signs [JLC23]. The primary goal of this research was to evaluate the effectiveness of high-frequency features and the progressive enhancement learning framework, particularly its self-enhancement and interactive-enhancement modules, in detecting subtle morphing indicators. However, this investigation does not involve any form of enhancement; instead, the database is presented as-is, allowing for the unaltered evaluation of a singularly created model without refinements to focus solely on the evaluation of the model itself.

On the other hand, when it comes to detecting morphing attacks, it is important to have a robust dataset that allows differentiation between individuals. From the amplitude of datasets, it is important to mention the NIST Face Recognition Vendor Test (FRVT) program, which focuses on evaluating technologies for detecting facial morphing in still photographs. This test is designed to evaluate the resistance of facial recognition algorithms to morphing, essentially measuring how well they can distinguish between authentic faces and those that have been morphed together [A24].

This dataset comes from the The Face Recognition Grand Challenge (FRGC) was a comprehensive challenge problem designed to improve the performance of face recognition algorithms. The primary goal was to promote and advance face recognition technology to support existing face recognition efforts in the U.S. Government [JP24].

2 Background

Taking into account the literature and with access to the Face Recognition Vendor Test (FRVT) dataset, facilitated by the supervisors of the 02238 Biometric Systems course several open-source models were selected with the aim of demonstrating their effectiveness within the dataset under consideration, specifically without relying on high-frequency data and self-enhancement techniques.

2.1 DeepFace Framework

To demonstrate its effectiveness the DeepFace framework, a framework that would allow the integration of different models, and was used. This comprehensive framework integrates several state-of-the-art face recognition models. Experiments conducted with this framework have demonstrated the ability to exceed human-level accuracy in face recognition tasks, underscoring the sophisticated capabilities of today's AI algorithms in this area [Se24a]. DeepFace was developed by a team of researchers at Meta and consists of several key components: Face Alignment, Face Representation, Face Recognition, and Face Attribute Analysis. Among these, the decision was made to focus specifically on face recognition [Se24b].

2.2 Description of the models

Among all available models, the decision was made to conduct an experiment with six of them:

- VGG-Face: Developed by the Visual Geometry Group at the University of Oxford, VGG-Face is a variant of the VGG neural network specifically designed for facial recognition [Br24].
- FaceNet: Created by Google, FaceNet stands out for its efficiency and performance in face detection and recognition. Utilizing deep learning, it achieves remarkable accuracy on both the LFW dataset and YouTube Faces DB [SKP15].
- ArcFace: Jointly designed by researchers from Imperial College London and InsightFace, it introduces a novel loss function that significantly improves face recognition accuracy [JGXZ19].
- DeepID: A series of models focused on deep learning-based identification, it aims to identify individuals across varying poses, expressions, and lighting conditions with high precision
- SFace: Also developed by Facebook, DeepFace is a deep neural network-based model trained on a large dataset of faces. Its goal is to bridge the gap between machine and human-level performance in facial recognition tasks [Tt2023].
- OpenFace: This is an open-source face recognition system developed by Facebook, focusing on real-time face recognition and tracking.

Testing these models allowed to observe how each performed in terms of recognition accuracy on such a complex task at first glance.

2.3 Description of the selected face manipulation methods

2.3.1 Reference, Probe and Probe Light

The dataset provided by the course contained a total of 1141 different images from 480 individuals as a reference. However, due to computational issues (discussed in Section 5.1 Limitations and Future Work), 50 different individuals were selected, each with 3 reference images.

In the context of evaluating facial recognition algorithms, the probe dataset consisted of photographs of the same individual under different conditions. These conditions included different lighting, and changes in appearance such as aging or changes in hairstyle. This setup was designed to mimic real-world scenarios where facial recognition may be required, but the reference image (a clear, unaltered photograph) is not available. This database consisted of an additional 150 images, with 3 images per reference individual in different positions and settings.

In addition to the probe dataset, an additional dataset was evaluated. This dataset included the probe dataset, but with a 50% increase in brightness. The main idea behind this was to test whether certain conditions were necessary when analyzing the probe database.

Figure 1 shows an example for the three databases.

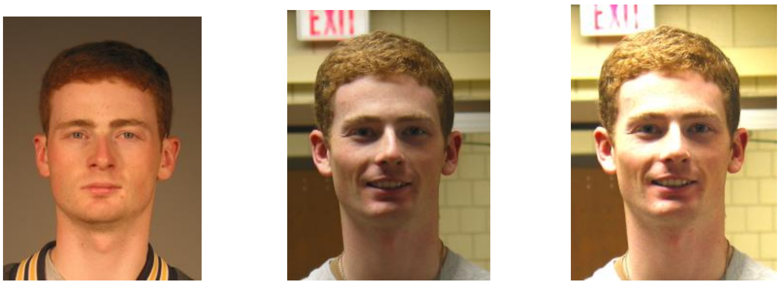

Fig. 1: A subject from the reference dataset, the probe dataset, and the probe light dataset, respectively

2.3.2 Morphed Faces

Similarly, it was decided to evaluate how the reference model performed under four different morphing conditions. These morphings used photographs not included in the dataset; specifically, they involved combining an individual selected from the original pool of 50 with another individual who was not part of the selection process. The methods used on our dataset included several techniques:

The FaceFusion method integrated two or more facial images to create a new image that combined the distinctive features of the original faces. This process involved aligning the faces according to key facial landmarks, followed by blending using various techniques to achieve a natural-looking result.

OpenCV is used as a library to provide face morphing functions. Its procedure requires the definition of control points on both the source and target images, along with the computation of a transformation matrix that effectively maps one set of points to another.

Facemorpher focused on the transformation and interpolation of facial attributes to produce a morphed effect. This technique aims to modify and blend facial features to simulate a composite identity.

Finally, the UBO Morpher tool used triangulation, warping, and blending to create a morphed image. This tool required the identification of specific landmarks for accurate application [IRFDB22].

3 Methodology

Regarding the methodology to be applied, several hypotheses and research questions were developed that allowed us to observe and analyze whether the experiment was correctly performing the recognition at all times, thus allowing a better analysis of its functioning. These included:

- What was the accuracy of each method in identifying reference faces? This analysis included not only assessing the accuracy but also identifying instances of false positives and false negatives
- How are the differences in distance and threshold distributed for each method? In face recognition systems distance refers to the measure of similarity or dissimilarity between two faces based on the features extracted from them. A threshold, on the other hand, is a predefined value that determines whether or not two faces are recognized as belonging to the same person. It serves as a decision boundary in the feature space. The calculation of the difference between the threshold and the distance was performed to determine how closely a face is accepted as being the same.
- Which faces are most often misidentified? Is there a pattern in the faces that are most often misidentified when some faces are misidentified? This also included the detection of similar features between these misidentified individuals.

3.1 Roadmap

To carry out this methodology, the model shown in Figure 2 was implemented. This figure illustrates how the initial datasets consisting of 1441 images were analyzed. From these, 50 of the 1139 recognized images were selected. Then, 50 morphed faces were selected for each of the morphing techniques. It was decided to get 50 morphed faces so that an image from the reference database was not morphed with another image from the reference database, but with an image that was not part of the 50 chonen individuals. That is, when A and B were morphed, only A was in the reference database.

Similarly, faces from the probe dataset were collected and tested with the algorithm. In addition, a light filter was applied to these probe images, making them 50% brighter, as many were darker. This adjustment was made to assess the effect of lighting on the model's performance. A data frame was then created for each of our seven experiments, with six models per experiment, for a total of 585,000 comparisons and 42 models. Finally, a script was developed to evaluate and address the research questions posed.

3.2 Scripts and Running Time

Once all datasets were selected, a program was run that performed a loop through the command:

```
DeepFace.find(img_path=full_img_path, db_path=db_path,
    model_name=model, distance_metric="cosine")[0]
```

This command extracted photos identified as identical from each dataset, storing them in a new dataset as the loop processed our entire dataset. Subsequently, an additional command was added to verify whether the identified person was indeed the same individual.

As explained in section 5.1 Limitations and Future Work, the experiment was initially conducted with all 1441 photos in the dataset; however, due to the large number of photos that the algorithm would need to detect and store in 7 different datasets, managing this became complex. Therefore, it was decided to select a representative sample, taking on average about 30 minutes per experiment.

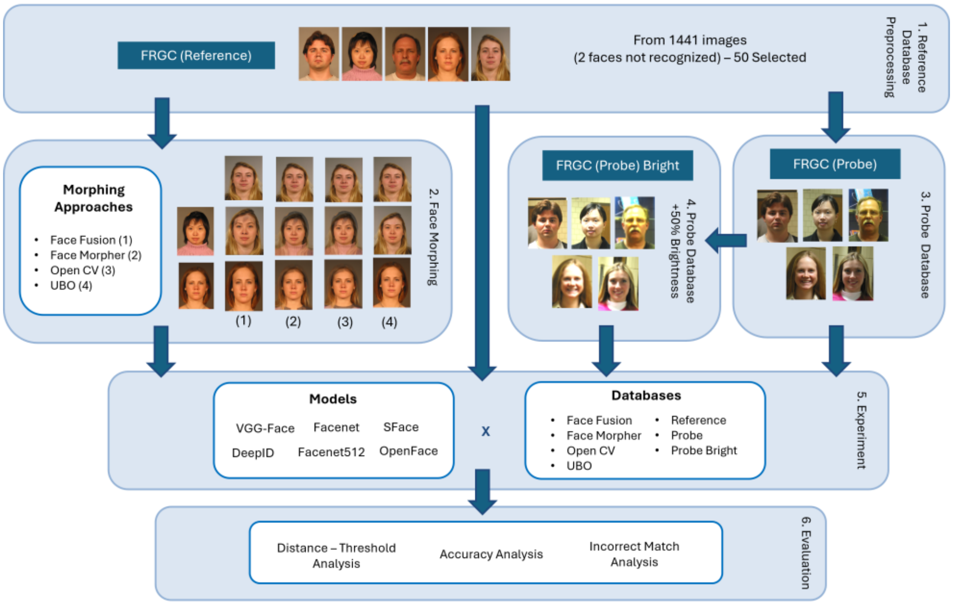

Fig. 2: Roadmap and methodology followed divided in 6 steps

4 Results

Regarding the outcomes, it is crucial to categorize them according to the three questions and hypotheses proposed.

4.1 Accuracy Implementation

To measure accuracy, all experiments were evaluated in a way that the amount of true positives were divided by the actual number of matches that there had to be for the model to be 100% correct. The results are depicted in Figure 3.

From this experiment, it is important to highlight that the reference model, which is the one that had been used to train the models, did not achieve 100% accuracy in all cases. Additionally, it should be noted that the probe with brightness performed better than the standard probe in most cases, indicating that the clearer the image appears, the better the model can distinguish it.

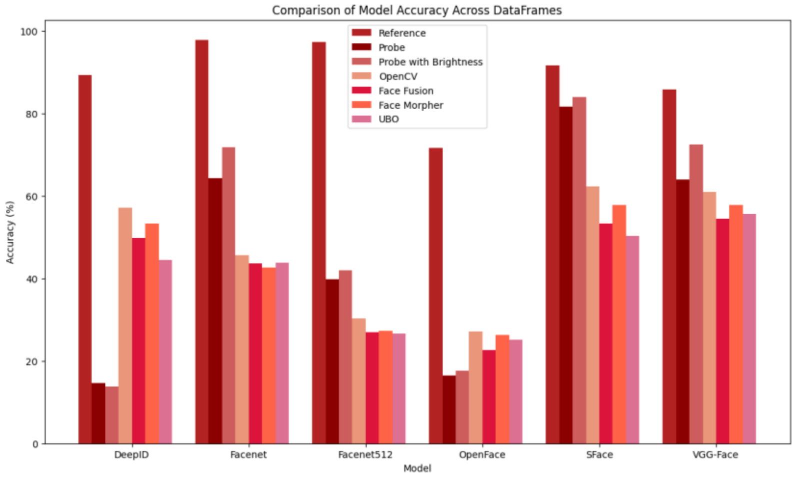

Fig. 3: Model Accuracy across all 7 databases for each one of the 6 models

It is also important to note that, with the exception of the DeepID and OpenFace models, the recognition of morphed faces did not achieve the same level of accuracy in most cases as the Probe dataset and the Probe dataset with Light. This showed that the model was able to detect when the presented face was not the same as the morphed face. In other words, the generally low accuracy in these four scenarios indicated that the model was more resistant to accepting morphed faces.

After analyzing accuracy, it is equally important to examine the number of errors the model made in interpreting faces. This involves looking at the number of false positives that were incorrectly passed. To illustrate this, Figure 4 was plotted.

Figure 4 shows that in all seven databases, the DeepID model had an unusually high acceptance rate, meaning that it accepted faces as identical when they were not, and therefore had a very high number of false positives. However, models such as FaceNet 512 and FaceNet had very low error rates, as their classifications were mostly accurate. The rest of the models performed better than DeepID, but in the case of the probe database and the probe light database, had very high inaccuracies.

4.2 Distance – Threshold Difference

Among the questions to be developed, it was proposed how the difference between distance and threshold of models was distributed, specifically the false positives. The distribution of the difference between distance and threshold is critical because it allowed to evaluate how models handle different levels of confidence when identifying faces. In models with a high number of errors, the distribution of the difference between distance and threshold could be wider, indicating a lower ability to correctly distinguish between similar and different faces.

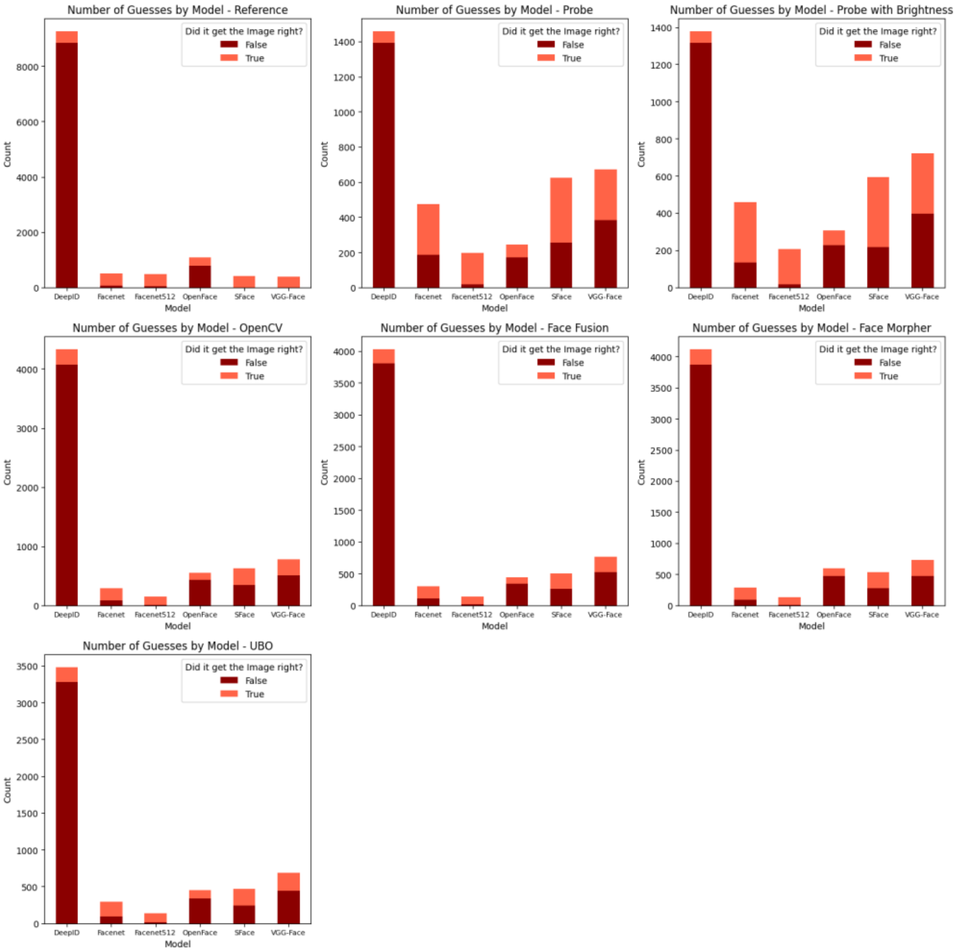

Fig. 4: Number of false positives and true positives by model and by database

On the other hand, in more accurate models, this distribution would be narrower, indicating a better ability to set a clear threshold that adequately separates relevant facial features from those that are irrelevant for identification. This analysis helps identify areas for improvement in the design and training of facial recognition models, allowing specific adjustments to be made to optimize their performance.

This analytical approach also highlighted the importance of adaptability in facial recognition models. This means a detailed understanding of the distribution of distance and threshold could allow future researchers and developers to dynamically adjust these parameters. Figure 5 shows the results of the experiment.

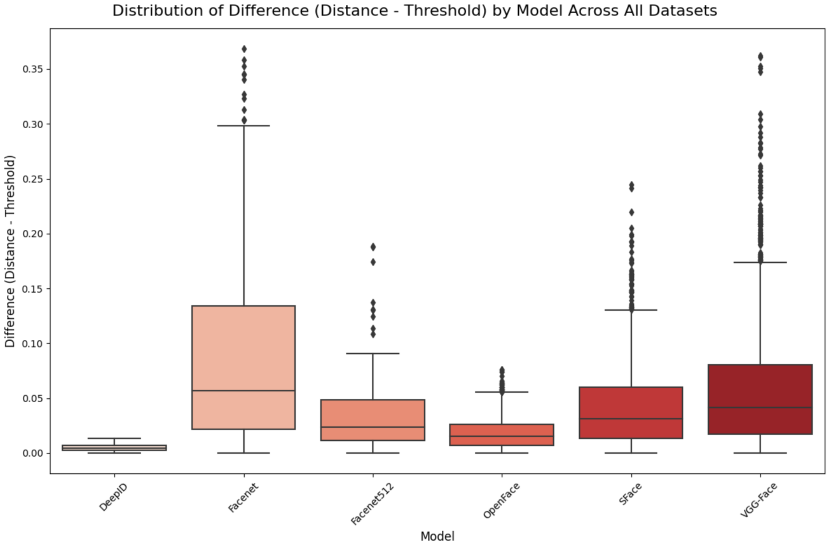

Fig. 5: Distribution in Difference of Distance - Threshold by model across all datasets

The results of this figure were revealing, as the poorer performing models were distributed over smaller distances between the distance difference and the threshold. However, contrary to what was hypothesized, the best performing model had a larger error range for false positives.

This led to propose a new hypothesis for future studies. This hypothesis was that when this distance and threshold was discarded for those with an accuracy below 0.3, the accuracy was significantly higher. This can be seen in Figure 6 and shows another line of research for future work.

4.3 Misidentified faces

Since each image was tested 3900 times (650x6 models), it was important to study the distribution of false positives (where the algorithm erroneously determined that it was the same face when in fact it was not) to determine if there were certain images that were misidentified more often. Of the 150 reference images, Figure 7 shows the images that were most frequently misidentified, with error rates of 11.36%, 10.44%, 10.33%, 10.15%, and 10.05%, respectively.

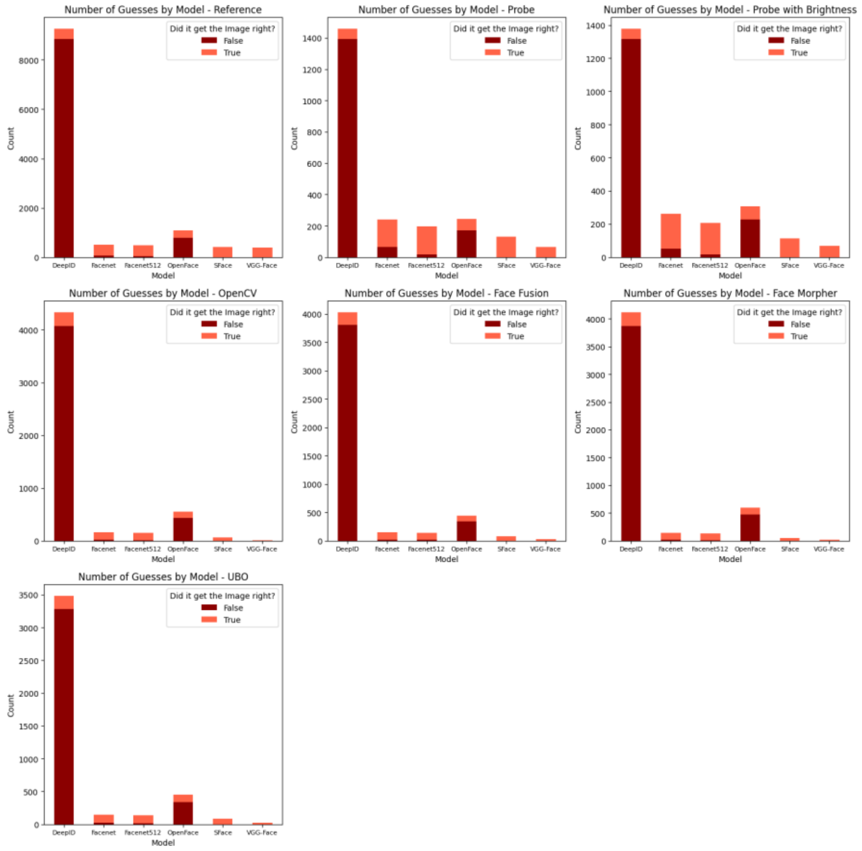

Fig. 6: Number of false positives and true positives by model and by database when discarded difference of Distance-Threshold larger than 0.3

From this experiments, not many conclusions could be drawn since, despite all having skin white and perhaps eyes that were more pronounced compared to the rest of individuals, it was not a mistake large enough to determine a pattern in faces that were always misidentified.

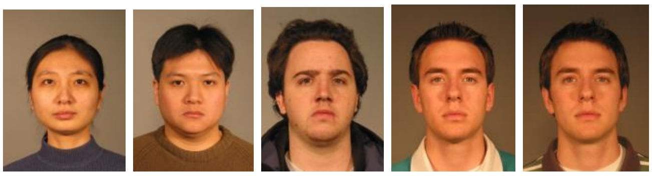

Fig. 7: Misidentified images in 11.36%, 10.44%, 10.33%, 10.15%, and 10.05% of cases, respectively

Nonetheless, it was important to see if there were pairs of images that actually produced errors, i.e., morphs or people that were so identical in our dataset that it caused the models to fail. These results are shown in Figure 8.

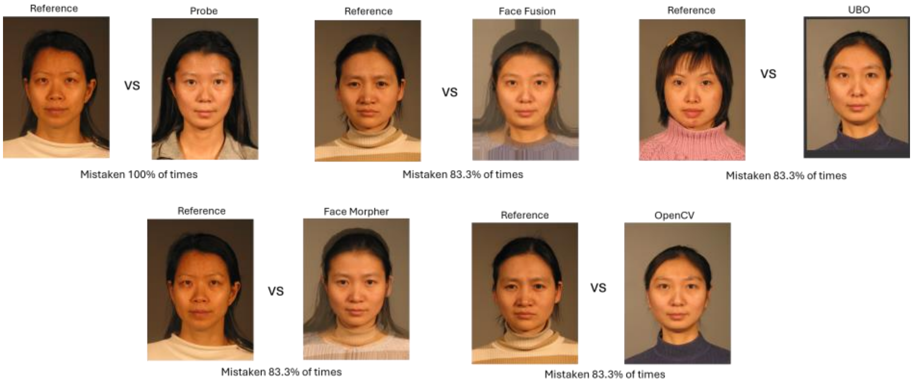

Fig. 8: Misidentified image pairs and their probability of being misidentified

As can be observed in this case, the first comparison had a 100% error rate despite not involving the same individual. Generally speaking, there was a noticeable pattern here since, in these instances, pairs of individuals who experienced the highest misidentification errors were Asian women. In most cases, this involved 5 out of 6 models making the mistake (83.3%).

5 Discussion

The results showed that it was particularly noteworthy that the DeepID model had an excessively high acceptance rate in all seven experiments, meaning that it accepted faces as identical when they were not. This figure also shows that models such as FaceNet 512 and FaceNet had very low error rates, as the classifications made by these models were correct in most cases.

It is also important to note that, with the exception of the DeepID and OpenFace models, the recognition of morphed faces did not achieve the same level of accuracy as the Probe and Probe with Light datasets in most cases.

Similarly, regarding the hypothesis of filtering out those with a distance and threshold less than 0.3, it was observed that the accuracy was significantly higher. This can be seen in Figure 6 and shows another line of research for future work. This highlights the importance of performing a final analysis before determining whether one face is the same as another.

In terms of misclassified images, it is important to highlight that there were pairs of images that were so identical in our dataset that it caused the algorithms to make mistakes. These results are shown in Figure 8.

5.1 Limitations and Future Work

In terms of future work, it's important to note that as the project progressed, it would have been beneficial to use a larger dataset. However, due to limitations and runtime constraints, running 1441 images on a personal computer would have been computationally expensive, and although we attempted to do so initially, our code was shown to exceed 10,000,000 comparisons and more than 2,000,000 rows in a dataframe.

Regarding future work, it would be key to focus on the line of research where, after an image has been classified as belonging to the same person or not, performing an analysis of the difference in the threshold distance as mentioned throughout the study.

6 Conclusion

In summary, this paper has examined the performance of six facial recognition algorithms on both standard and morphed faces. The study highlights several important findings. First, algorithms such as FaceNet and FaceNet512 exhibit superior performance in recognizing faces under a variety of conditions, highlighting their robustness in dealing with changes in brightness and morphing techniques.

Second, the analysis of the distance-threshold relationship between different models provided insightful data on model sensitivity and error rates. This aspect of the research could lead to more sophisticated approaches to algorithm tuning in order to improve accuracy, especially under different operating conditions.

Finally, the study's examination of common misclassification errors and the specific conditions that lead to such errors provides valuable insights for improving facial recognition systems. By understanding the nuances of where these algorithms fail, programmers can better address these weaknesses, potentially leading to safer and more reliable systems in real-world applications. Morphed face recognition is critical for mitigating illegal activity.

References

[Bu24a] Busch, C.; 2D Face Recognition, 02238 Biometric Systems, Lecture F, Slide 4,2024.

[Bu24b] Busch, C.; 2D Face Recognition, 02238 Biometric Systems, Lecture F, Slide 34, 2024.

[TVFM24] Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. ResearchGate, 2020.

[KMF19] Kramer, R.S.S., Mireku, M.O., Flack, T.R. et al. Face morphing attacks:

Investigating detection with humans and computers. Cogn. Research 4, 28, 2019.

[JLC23] Jia CK.; Liu YC.; Chen YL.; Face morphing attack detection based on high-frequency features and progressive enhancement learning, Front Neurorobot, 2023.

[A24] Anonymous; Introduction to NIST FRVT: background, tests and results, https://www.paravision.ai/news/introduction-to-nist-frvt/, 15-06-2024.

[JP24] Jonathon Phillips P.; Face Recognition Grand Challenge (FRGC): Information about the primary goal of the FRGC and description, https://www.nist.gov/programs-projects/face-recognition-grand-challenge-frgc, 15-06-2024.

[Se24a] Serengil S.; DeepFace: A Popular Open Source Facial Recognition Library, What is DeepFace and how to use the library, https://viso.ai/computer-vision/deepface/ , 18-06-2024.

[Se24b] Serengil S.; Github repository on DeepFace, https://github.com/serengil/deepface, 18-06-2024.

[Br24] Brownlee J.; How to Perform Face Recognition With VGGFace2 in Keras, https://machinelearningmastery.com/how-to-perform-face-recognition-with-vggface2-convolutional-neural-network-in-keras/ , 22-06-2024.

[SKP15] Schroff, F..; Kalenichenko, D.; Philbin, J.; FaceNet: A Unified Embedding for Face Recognition and Clustering, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015

[JGXZ19] Jiankang D.; Guo J.; Xue N.; Zafeiriou S.; ArcFace: Additive Angular Margin Loss for Deep Face Recognition, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019

[Tt2023] Tran-Thanh T.; Information about the SFace library and its accuracy, https://trungtranthanh.medium.com/sface-the-fastest-also-powerful-deep-learning-face-recognition-model-in-the-world-8c56e7d489bc, 22-06-2024.

[IRFDB22] Ibsen, M.; Rathgeb, C.; Fischer, D.; Drozdowski, P.; Busch, C.; Digital Face Manipulation in Biometric Systems. In: Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Busch, C. (eds) Handbook of Digital Face Manipulation and Detection. Advances in Computer Vision and Pattern Recognition. Springer, Cham, 2022