Facial Marks: detection and integration into a facial recognition system

Sjoerd van den Belt
University of Twente
Enschede, Netherlands
s.p.vandenbelt@student.utwente.nl

***Abstract*—Facial marks can be used as a complement to facial recognition systems, and have been used for this purpose before. New advances in convolutional neural network (CNN) architectures have enabled more accurate detection of facial marks. In this paper state-of-the-art CNN models are trained on the FRGCv2 dataset for the recognition and detection of facial marks. The resulting systems are combined with a FaceNet facial recognition system in order to improve facial recognition performance. In particular, the improved ability to differentiate between twins will be studied. Due to their similarities, twins are exceptionally difficult for facial recognition systems to distinguish. The Twins Days dataset is used in order to investigate the ability of the combined system to differentiate between twins. This paper demonstrates significant improvements in facial recognition and twin differentiation performance when facial mark detection is used.**

***Index Terms*—Facial Marks, Facial Recognition, Twins, CNN**

I. Introduction

State-of-the-art facial recognition systems (FRS), such as FaceNet, manage to attain highly accurate results [1]. Nevertheless, the accuracy of such models decreases when posed with a particularly difficult tasks. Discriminating between twins is one such task, as FRS performance has been demonstrated to decrease when tested on a population of twins, in various previous works [2]–[5].

When humans are presented with the task of discriminating between faces of identical twins they tend to make use of subtle facial features, such as moles, scars and freckles [6]. Figure 1 shows a pair of identical twins with annotated facial marks. The figure clearly shows that, whilst their facial features may be difficult to distinguish, their facial mark patterns differ vastly. Incorporating the detection of such facial features into an FRS may increase the accuracy of facial recognition. In particular, the benefits may be valuable for facial recognition on twins.

This work will implement modern convolutional neural networks (CNN) to recognize and detect facial marks. CNNs are a type of neural network based on convolutional layers. These convolutional layers pass images through a number convolutional filters with trained weights, called kernels, in order to extract key features from an image [7]. Figure 2 illustrates a single convolutional filter. The complexity of CNN models differs vastly between architectures. Shallow models can be implemented using only a few convolutional layers, whilst state-of-the-art architectures may incorporate over one hundred layers.

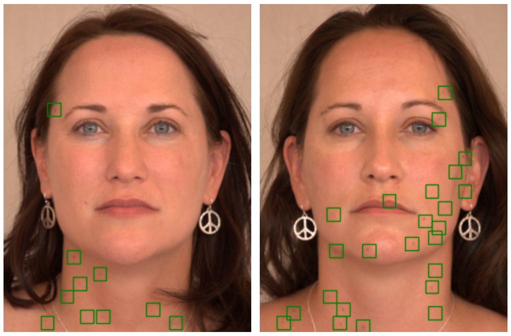


Fig. 1. Identical twins with clearly distinguishable facial mark patterns

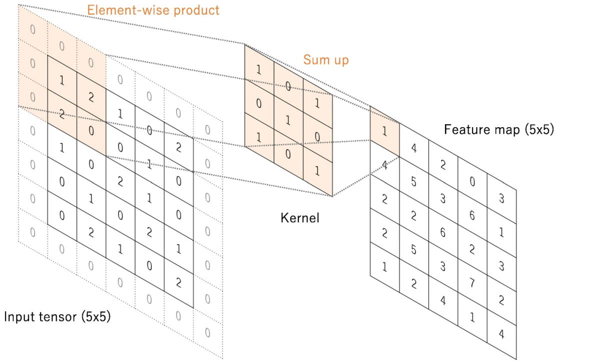


Fig. 2. Illustration of a single convolutional filter. The kernel slides over the tensor in order to extract a feature map. Taken from [7]

In the context of this paper, recognition and detection of facial marks refers to two distinct processes. Facial mark recognition (FMR) refers to the ability of a system to determine whether a small skin patch taken from a facial image contains a facial mark. Facial mark detection (FMD) refers to the ability of a systems to localize facial marks given a full facial image. Detection will produce a list of coordinates mapping detected facial marks on the image.

Using a dataset of facial images with annotated facial marks, skin patches featuring facial marks are isolated [8]. Various FMR models are trained on those patches, and the effectiveness of these models is compared.

Three different systems will be tried for the implementation of FMD. One system will be an FMD implementation of the previously mentioned FMR models. The other systems implement state-of-the-art object detection architectures, trained for facial mark detection. The effectiveness of the different FMD systems will be compared.

The FMD systems will be implemented for facial recognition by determining the similarity between detected facial mark patterns. The FMD-based facial recognition will be tested on subsets of the FRGCv2 dataset, and of the Twins Days dataset [9]. The ability to discriminate between identities using only FMD-based facial recognition will be analyzed.

Finally, the FMD systems will be integrated with a FaceNet FRS, and the performance of fused facial recognition will be tested on both datasets. In particular, the ability to discriminate between the faces of twins will be analyzed, since this is a task on which a standalone FRS performs poorly.

Concisely, this work will attempts to answer the following research questions.

1) How effectively can skin patches featuring facial marks be recognized using modern CNN architectures?
2) To what extend can facial marks be localized on a face using object detection based on CNNs?
3) To what extend can facial mark detection based on CNNs be used to improve facial recognition?

The rest of the paper is organized as follows. Section II will go over previous work related to the posed research questions. Section III will identify and clarify the systems that will be implemented for FMR, FMD, and facial recognition. Section IV. will specify the experiments that will be executed and further expands on the datasets used.

II. Related Work

As discussed in the introduction, differentiating twins, through biometric systems, has been a subject of interest in previous research. Sun *et al.* [2] investigated the capabilities of discriminating between identical twins using multiple biometric systems, under which facial identification. Phillips *et al.* [3] was the first to do in-depth research on facial recognition on twins, introducing the Twins Days dataset. Consequently, multiple studies have researched the application of facial recognition systems on the Twins Days dataset, taking into account a large number of covariates [4] [5]. These studies have concluded that there is an increased difficulty of differentiating identical twins, especially under non-ideal conditions. Images from different days, under different lighting conditions and of subjects wearing glasses pose such difficulties. In their work, Biswas *et al.* [6] investigated the ability of humans to discriminate between faces of identical twins. In this work it has been shown that humans pay attention to various biometric indicators, including moles, scars and freckles, when differentiating identical twins.

Facial marks inhibit potential use as forensic evidence, and the discriminatory power of facial marks has been analyzed thoroughly [8]. Park and Jain have created a facial mark detection system, making use of Laplacian of Gaussian (LoG) blob detection [10]. Integrating this system into a facial recognition system, Park and Jain have demonstrated to improve facial recognition on a limited dataset. A more recent study implementing facial mark detection with facial recognition has shown the increased accuracy on a larger dataset as well [11].

The implementations described in [10] and [11] both make use of the LoG method for blob detection. The LoG filter is vulnerable to false positives, due to the non-uniform structure of a face. In order to mitigate the false positives, the primary facial features on the faces are masked. Masking may, however, mask potential facial marks as well, and manual masking is labour intensive.

Current state-of-the-art object detection and image classification systems are dominated by convolutional neural networks (CNN). In [12] Shallow CNNs have been trained and used for the recognition of facial marks. Zeinstra and Haasnoot [12] design and train shallow CNNs to recognize facial marks on patches from facial images. Facial marks can be recognized with high accuracy using these models.

This work will further detail on facial mark recognition using CNNs, and in particular broaden the scope to deeper CNN architectures. In addition, it will focus on facial mark detection and facial recognition. Facial recognition will be analyzed using facial marks as a biometric modality on its own, as well as fused with a facial recognition system.

III. Methods

A. *Recognition of facial marks*

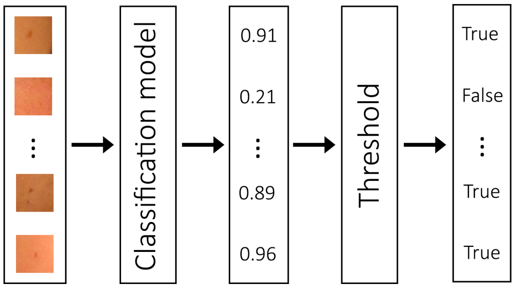
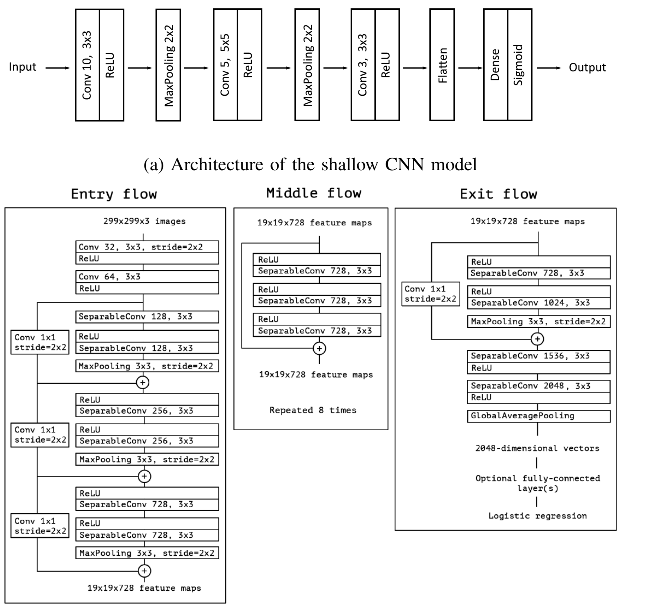


Fig. 3. Pipeline for facial mark recognition. The model determines a confidence value for each skin patch. Using a threshold, models are classified as featuring a facial mark (True), or not (False)

Facial mark recognition (FMR) is the process of classifying whether a skin patch does, or does not, contains a facial mark. An FMR model will, given a skin patch, output a measure of confidence, indicating whether the patch does or does not feature a facial mark. FMR models will be trained and tested on patches. These patches are images cropped from a facial image. Figure 3 shows the process of FMR, Figure 5 and Figure 6 show variations on a patch with a facial mark.

To determine the increase in FMR performance using modern, complex, CNN architectures, first a baseline model is established. This baseline model will be a shallow CNN



(a) Architecture of the shallow CNN model

(b) Architecture of Xception model. Taken from [13]

Fig. 4. Architectures of a shallow CNN model (a) and a state-of-the-art CNN model (b)

model. To demonstrate the contrast between the two different models, Figure 4(a) shows the architecture of the shallow baseline model, and Figure 4(b) shows the architecture of a state-of-the-art CNN, Xception [13]. The shallow CNN will be tested using both RGB as well as grayscale patches, examples shown in Figure 6(a) and Figure 6(b) respectively.

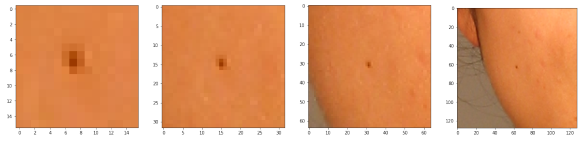
With the shallow CNN model as a baseline, three different modern CNN architectures are trained and tested for FMR:

1) MobileNetV2
2) ResNet50V2
3) Xception

ResNet50V2, Xception and MobileNetV2 are all state-of-the-art CNNs for image classification [13]–[15]. MobileNetV2, in particular, is chosen due to its focus on computational efficiency. ResNet50V2 is chosen for its performance as well as the fact that it can process images of down to 32 px in size. Xception is chosen based on its superior performance [13].

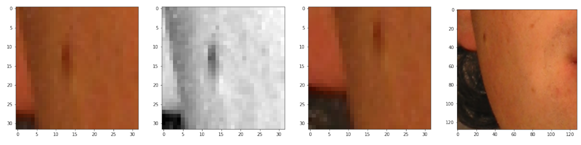
Next to the variety of models that is being tried, different sizes of patches will be considered. Due to the considered implementation of facial mark detection (FMD), which needs to recognize regions of interest by scanning large patches, and needs to recognize facial marks on small patches, a wide range of sizes is considered. The sizes will range between 16 px and 128 px. Figure 5(a) through 5(d) shows facial mark patches of various sizes.

When applying FMR models in an FMD system, facial marks are not always located in the center of a skin patch. To account for this, a subset of FMR models will be trained and tested with off-center facial mark patches. These off-center facial marks are distributed randomly within a margin around the center of the patch. Figure 6(c) and 6(d) show examples of off-center facial mark patches at 32 px and 128 px respectively.



(a) 16 px (b) 32 px (c) 64 px (d) 128 px

Fig. 5. Patch featuring a facial mark in various sizes



(a) Default (b) grayscale (c) off-center (d) off-center

Fig. 6. Patch featuring a facial mark without augmentation (a), in grayscale (b), and randomly off-center at 32 px (c) and 128 px (d)

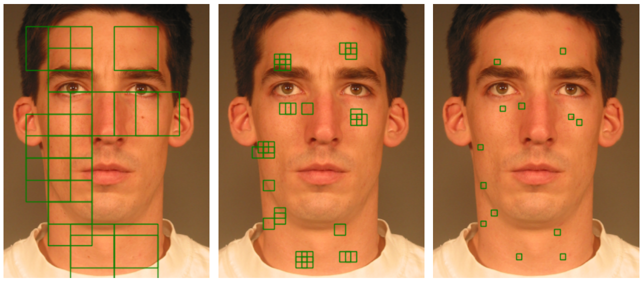
Since training modern CNN models requires tuning millions of parameters, it can be computationally expensive and time consuming. Furthermore, an extensive dataset is required to fully train such models. To train models of existing architectures more efficiently and with limited data, transfer learning can be applied [16]. When applying transfer learning, the parameters of models that have previously been trained on a foreign dataset are transferred and used as a starting point for a new model. To train the new model for its desired task, the layers at the head (output) of the model are retrained on the task-specific training dataset. After this, the model is fine-tuned. During fine-tuning, the rest of the layers are retrained as well, using a low learning rate as not to destroy the existing network. To investigate the effectiveness of transfer learning when training an FMR model, models are trained with and without the use of transfer learning. The transferred parameters used in these models have been trained on the ImageNet database [17].

B. *Detection of facial marks*

Facial mark detection (FMD) is the process where, given a facial image, the FMD system computes the locations of the facial marks on that image. In this paper, two distinct approaches are tried for implementing an FMD system. Firstly, an implementation of an FMD system will be designed based on the previously discussed facial mark recognition (FMR) models. Secondly, two existing object detection architectures will be trained and implemented for the facial mark detection.

1) Using FMR models: Various approaches can be taken to implement the FMR models for FMD. One of the most trivial is a sliding window system. A window will scan over the image using a fixed stride. Each window that is scanned will be processed by the FMR model. The main downside to this approach is the large number of sub-images that need to be scanned. To illustrate, an image of size (800, 600), scanned with a window of size (40, 40), using a stride of 20, will already yield 1131 sub-images to be scanned. At a false positive rate of only 0.01 this will, on average, result in 11 false positives. This is a significant amount, considering the FRGCv2 subset annotates an average of 5.4 facial marks per image.

To reduce the amount of images that need to be scanned by the FMR model, regions of interest (ROI) will first be detected, by scanning with a larger window. The regions detected in the ROI stage can subsequently be scanned by a model with a smaller window, for finer detection. Multiple stages of ROI scanning can be implemented to filter out more potential false positives. Adding stages will, however, be potentially more computationally expensive. An FMD system using three FMR models, for three layers of detection, will be implemented. Figure 7 shows the process of detection stage-by-stage.



(a) Stage 1

(b) Stage 2

(c) Stage 3

Fig. 7. Stages of sliding window facial mark detection. 128 px window (a), 32 px window (b), 16 px window (c)

Each stage in the sliding window system produces a feature vector and a corresponding vector of confidences. Each feature corresponds to the position of one window, scanned at the stage, and each confidence being a measure of likelihood that the window features a facial mark. For each stage, a threshold is chosen, and the feature vector of each stage is filtered by confidence. Filtering implies that each feature, or window, with a confidence greater than the threshold, is kept, while the other features are discarded. This results in a feature vector of variable length where each feature has a confidence greater than the threshold value. The output of the system is the filtered feature vector of the last stage, where each feature is represented by the center of the window.

2) Using object detection architectures: In addition to the implementation based on the FMR models, existing state-of-the-art CNN-based object detection models are trained and tested. The facial mark annotations that have been used to extract the skin patches can also be used to train object detection systems. Instead of extracting patches, the models will be trained on facial images where the facial marks have been annotated by bounding boxes.

To compare performance, two different object detection models will be implemented. EfficientDet is a highly efficient object detection system, which is largely based on the Single Shot Detector (SSD) [18], [19]. It is chosen for its fast performance. The other system that will be implemented uses the Faster R-CNN (FRCNN) architecture [20]. This architecture is less efficient than the SSD architecture, but may be more accurate.

Similarly to the sliding window approach, the EfficientDet and FRCNN systems produce a feature vector, and a vector of corresponding confidences. These features correspond to bounding boxes around detected facial marks. A single threshold is chosen and the feature vector is filtered by confidence. From the filtered feature vector, the center of each bounding box is determined, and a feature vector of box-centers is constructed. The output of the systems is a variable length feature vector, where each feature is the center of a bounding box.

C. Facial recognition

Using the aforementioned FMD systems, facial mark patterns can be extracted from facial images. A score can be assigned to the similarity between the facial mark patterns from a pair of facial images. The score is a measure of confidence that the facial images correspond to the same subject. Conventional facial recognition systems (FRS), such as FaceNet, extract key facial features in order to identify a subject. The FMD-based system and the FaceNet FRS can be combined for improved facial recognition. Figure 8 shows the pipeline for a combined implementation. The systems will label a pair of faces as being from the same subject (True), or being from different subjects (False).

1) Facial marks score: The feature vector from the FMD system contains the location of the center of each detection. In order to determine the similarity between a pair of facial mark patterns, a score based on overlapping facial marks is calculated. To determine overlap, each facial mark is represented by a circular region around the center of the detection. Two facial marks are said to overlap if their regions overlap. In other words, two facial marks overlap if the distance between the marks is less than the diameter of the region. Each mark can only be counted to overlap once. Consider p, the amount of overlapping pairs. N the sum of the amount of marks from each image

$$N = N_1 + N_2$$

. then the overlap score s is defined as

$$s = \frac{2 \cdot p}{N}$$

If $N = 0$, then the score is defined as $s = 1$. This way faces with no facial marks also have a positive correlation. Using this method of scoring, $s = 1$ corresponds to a perfect match, and $s = 0$ corresponds to a perfect mismatch.

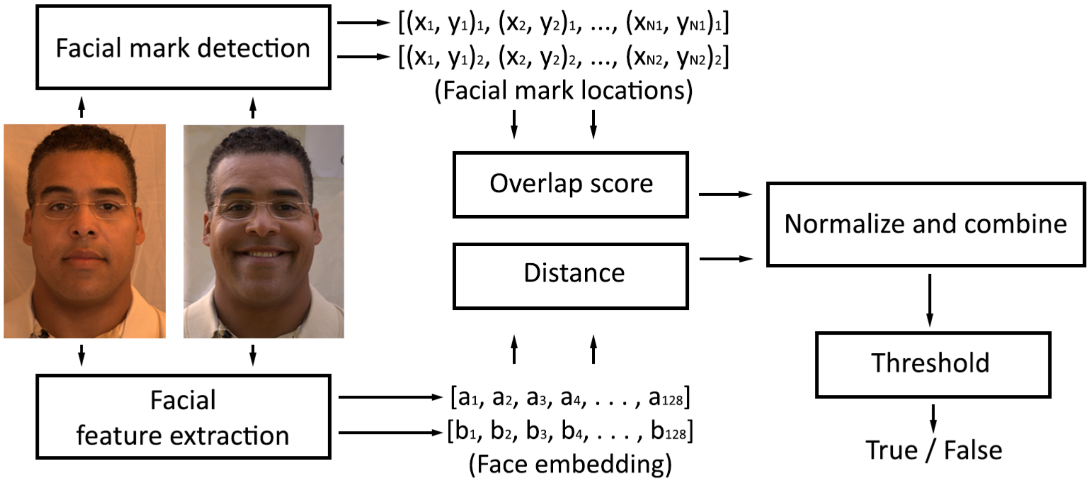

Facial mark detection
$[(x_1, y_1)_1, (x_2, y_2)_1, ..., (x_{N1}, y_{N1})_1]$
$[(x_1, y_1)_2, (x_2, y_2)_2, ..., (x_{N2}, y_{N2})_2]$
(Facial mark locations)
Overlap score
Distance
Normalize and combine
Threshold
True / False
Facial feature extraction
$[a_1, a_2, a_3, a_4, . . . , a_{128}]$
$[b_1, b_2, b_3, b_4, . . . , b_{128}]$
(Face embedding)

Fig. 8. Facial recognition combining facial mark detection (FMD-based system) and facial feature extraction (FaceNet FRS). From the two facial mark patterns, an overlap score, indicating similarity, is determined. From the FaceNet embeddings, the euclidean distance is calculated. The two scores are normalized and fused

2) FaceNet distance: When using FaceNet for facial recognition, all faces will be mapped to an embedding of 128 features. To determine the similarity between the faces, the euclidean distance between two embeddings is calculated. A lower distance corresponds to greater similarity.

3) Fusing the scores: To normalize a set of overlap scores and distances, adaptive z-score normalization is applied to each set [21]. Thereafter, the scores are fused using a weighted sum. A set of normalized scores is defined as

$$Z = \frac{X - \mu}{\sigma}$$

Where X is the set of scores or distances, μ is the mean of the set and σ is the standard deviation of the set. Consider a set of overlap scores S and FaceNet distances D. If Z_S denotes the standard scores of S and Z_D denotes the standard scores of D, then the fused score S_C is defined as

$$S_C = Z_S - \alpha \cdot Z_D \quad \text{where} \quad \alpha \geq 0$$

Where α is a factor weighing the influences of the two sets.

Since the performance of the FaceNet FRS varies among different sets of subjects (such as twin and non-twin populations), the weights to fuse the scores are chosen specific to the set [22]. This implies that for a test set where FaceNet performance is high, a greater value for α is chosen, such that Z_D has a greater weight. In cases where FaceNet is less accurate α is lowered for better results. It is worth noting that choosing a non-ideal value for α may result in an accuracy of the combined system that is lower than the accuracy of a standalone system.

IV. EXPERIMENTS

A. Datasets and preprocessing

For training both the facial mark recognition (FMR) and facial mark detection (FMD) models, a subset of the FRGCv2 dataset with annotated facial marks is used [8]. This dataset contains 12306 images of 568 different subjects. The dataset was split up into training, testing and validation subsets. The training set contains 7925 images of the first 279 subjects, the validation set contains 1982 images of next 115 subjects and the test set contains 2400 images of the last 174 subjects. The Twins Days dataset is only used for testing, no models are trained on this dataset. Only the images in the Twins Days dataset with forward-facing subjects have been used. The resulting Twins Days subset contains a total of 7422 images featuring 435 different subjects.

1) Preprocessing for FMR: Using the facial mark annotations, patches with facial marks are extracted from the FRGCv2 images. To train the FMR models, the models need to be trained on patches that are positive (featuring a facial mark), as well as negative (featuring no facial mark). Patches featuring no facial marks must therefor be extracted as well. To create a balanced dataset, the amount of patches with facial marks must be similar to the amount of patches without facial marks. Since, on average, an FRGCv2 image contains 5.4 annotated facial marks, 5 patches without facial marks were sourced from each image. These patches were selected by picking 5 semi-random locations on each image, and extracting patches around each location. The locations are picked such that the semi-random patches do not overlap, and do not include any annotated facial marks.

2) Preprocessing for FMD: The images in the FRGCv2 dataset are transformed such that the right and left pupils are mapped to fixed locations at (200, 250) and (400, 250) respectively. The images are cropped to the size of (800, 600). The images in the Twins Days subset are transformed using a dlib face alignment implementation [23] and are also cropped to size (800, 600).

The EfficientDet and FRCNN training processes implement an additional number of random data augmentation methods. For both object detection models the data augmentation methods used are as follows.

- Horizontal flip (over the width of the image)
- Adjusted hue
- Adjusted contrast
- Adjusted saturation
- Scaling (factor between 0.6 and 1.3)

These methods are implemented because the detection models train on a limited amount of data. The number of facial images is less than the number of facial mark patches. If no data augmentation is used, a model may overfit to the training data, causing it to only be effective on the training subset.

B. Recognition of facial marks

Instances of the shallow CNN model are trained on RGB and grayscale patches at 23, 32 and 64 px. Instances of the modern CNN models are trained at 16, 32, 64 and 128 px, varying the use of transfer learning and off-center facial marks. Patches of 16 px will be up-scaled to 32 px using bilinear interpolation, since 32 px is the minimum size MobileNetV2 and ResNet50V2 can process.

ROC curves will be used in order to visualize and analyze the trained FMR models. An ROC curve plots the true positive rate (TPR) of a system, against the false positive rate (FPR). On the ROC curve, the equal error rate (EER) can be found. On this point the rate of false positives equals the rate of false negatives, a lower EER indicates a better performing system.

C. Detection of facial marks

For the sliding window FMD implementation, FMR models of the following three sizes are used:

- 128 px
- 32 px
- 16 px

The stride of the models is 64, 16 and 8 px respectively. The first two models are trained using off-center facial mark patches. The third model is trained using centered facial mark patches. The models implement the MobileNetV2 architecture.

For filtering the output feature vector of the EfficientDet and FRCNN systems, a value of 0.05 is chosen. For the sliding window system 0.15, 0.90 and 0.80 are chosen for the first, second and final stage respectively. The threshold values were chosen based on empirical results, the values resulted in a satisfactory high amount of true positives, and an acceptable amount of false positives on each system.

The sliding window, EfficientDet and FRCNN system will each be used for facial mark detection on all facial images in the FRGCv2 test subset and the Twins Days subset.

D. Facial recognition

For the FaceNet facial recognition system (FRS), two pre-trained models based on the Inception-ResNetv1 architecture are implemented [24]. One of those models is trained on the CASIA-Webface dataset, and the other is trained on the VGGFace2 dataset [25], [26]. Each models will be used to determine the embeddings of the faces in the FRGCv2 test subset and the Twins Days subset. The euclidean distance between each embedding in a subset will be calculated.

Using the results from the FMD systems, the overlap score of each pair of facial mark patterns in a subset will be determined. The size of the regions representing the facial marks is set to be 32 px.

From the embedding distances and overlap scores, the fused scores will be calculated. For this calculation, a value for weighing factor α must be picked. Through empirical observation, a factor of $\alpha = 5$ was deemed most effective, and will be used.

To analyze the ability of the each system to differentiate between twins, a separate set of embedding distances, overlap scores and fused scores is constructed. In this set, only the subjects from the Twins Days dataset with a twin will be considered. This set will only compare facial pairs corresponding to the same subject (matching pair) and corresponding to a subject and the subject's twin sibling (non-matching pair). Since the FaceNet FRS performs poorly on this task, another value for α is chosen. Through empirical observation, a factor of $\alpha = 1$ is found to be effective, and will be used.

V. Results and Discussion

A. Recognition of facial marks

In Table I and Figure 9 the results on facial mark recognition (FMR) performance using the shallow CNN model and modern CNN models are shown by their equal error rate (EER) and ROC curves. Form the data it becomes clear that the complex modern CNN models have a significantly lower EER than the shallow CNN models. This behaviour was observed to be consistent among patches of all sizes. Furthermore Table I shows the difference between using grayscale and RGB patches for shallow CNN models. No consistent difference in performance was observed when comparing the RGB models to grayscale models.

TABLE I. Comparing modern CNN and shallow CNN models for facial mark recognition

Model	**EER**	
	RGB	**Grayscale**
Shallow (23 px)	0.030	0.035
Shallow (32 px)	0.031	0.028
Shallow (64 px)	0.029	0.028
MobileNetV2 (32 px)	**0.019**	N/A
ResNet50V2 (32 px)	**0.019**	N/A

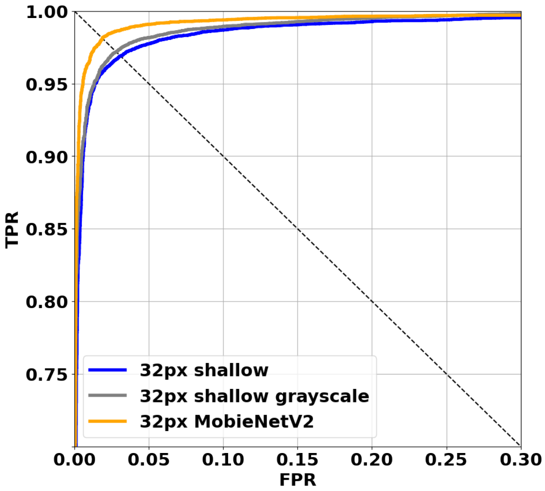


Fig. 9. Comparing modern CNN and shallow CNN models for facial mark recognition

The effect of using transfer learning to train the modern CNN models is plotted in Figure 10. No significant difference in performance between the models is observed. However, when training the models using transfer learning, the models tended to converge to their best accuracy with less training than the models that did not use transfer learning. Moreover, the models that were trained using transfer learning more consistently converged to their best accuracy. The models trained from scratch were more likely to plateau at lower accuracies.

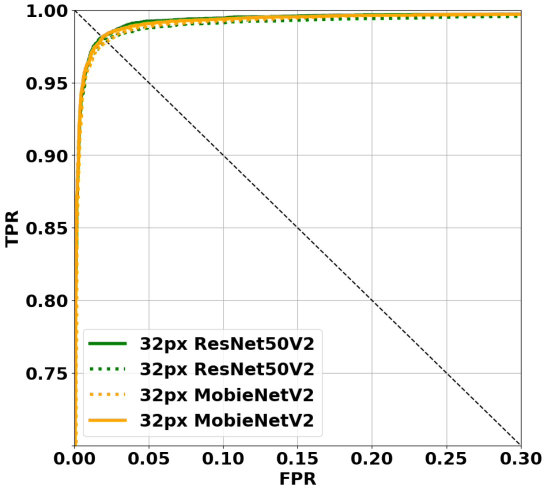
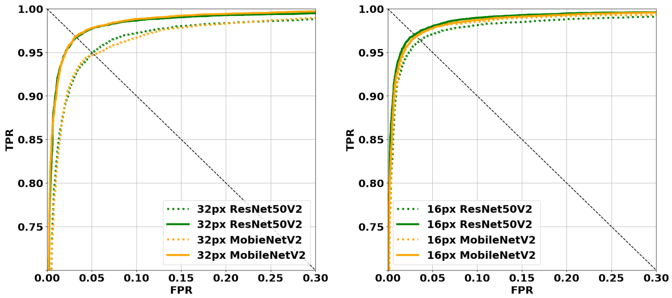


Fig. 10. Modern CNN models trained With transfer learning (solid) and without transfer learning (dotted)

In Figure 11 the results of various FMR models tested on patches with off-center marks are shown. Figure 11(a) and Figure 11(b) show the results at a patch size of 32 px and 16 px respectively. In Figure 11(a) a significant difference can be observed, the models trained with off-center marks perform better than the models trained on centered marks. On the contrary, in Figure 11(b), no significant difference is observed.

These results indicate that training using off-center facial mark patches can be beneficial to a model, depending on the patch size. For patches of at least 32 px, there is a visible benefit. For smaller patch sizes there seems to be little benefit. The difference in effect can be attributed to, firstly, the limited space for randomization on a 16 px image. Secondly, the fact that the annotations of the marks do not, in all cases, point to the pixels at the center of the facial mark. On small patches this results in inherent off-center mark locations.



(a) 32 px Models

(b) 16 px Models

Fig. 11. Modern CNN model performance on off-center facial marks. Models trained on off-center facial marks (solid) and models trained on centered facial marks (dotted)

Figure 12 shows the performance of the modern CNN models on a relatively large patch size. No significant difference in performance can be made out between the different models. With this result in mind, the MobileNetV2 model can be considered the best option for FMR as it has the most lightweight architecture. This also indicates that there may be potential for even more lightweight architectures to attain similar performance.

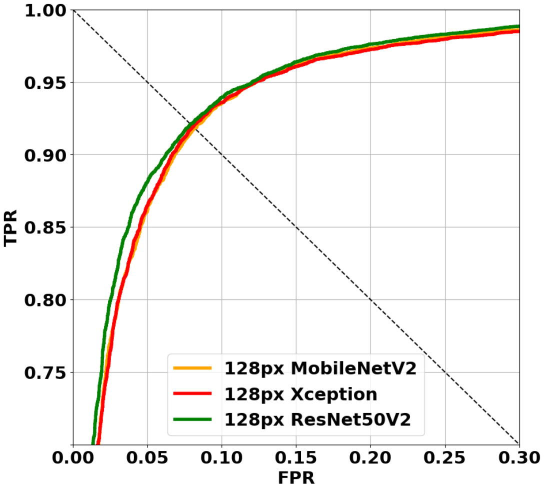


Fig. 12. Comparing modern CNN model performance for facial mark recognition on larger patches

Conclusively, from the results on FMR models, it is clear that the modern CNNs outperform the shallow CNN. It is noted that the shallow model tested in this paper performed poorer than the shallow model used in the work by Zeinstra and Haasnoot, in which the best result was an EER of 0.279 for 23 px grayscale patches [12]. Nevertheless the modern CNN models in this paper consistently outperform those results as well, at an EER of under 0.020. FMR using modern CNN architectures attains state-of-the-art accuracy.

B. Detection of facial marks

Figure 13 shows images from the FRGCv2 subset where facial mark detection (FMD) has been performed on, by the sliding window system. In 13(a) many facial marks are detected, and in 13(b) few facial marks are detected. The figure illustrates the ability of the system to detect facial mark patterns that are trivially distinguishable and specific to the subjects. Furthermore, figure 14 shows detections by each individual FMD system. Although differences can be observed among the systems, all systems generally agree on prominent facial marks.

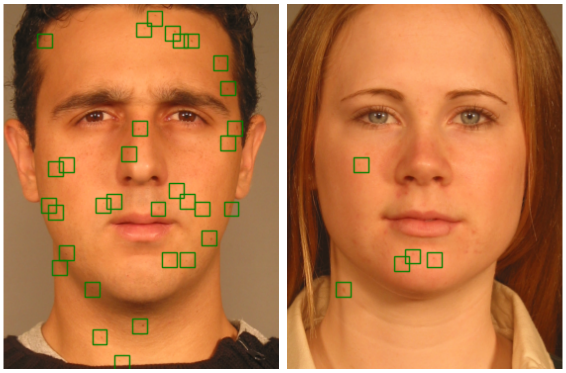

(a) Many detected marks (b) Few detected marks

Fig. 13. Facial images with annotated detected facial marks. Subject (a) has many facial marks, subject (b) has fewer facial marks

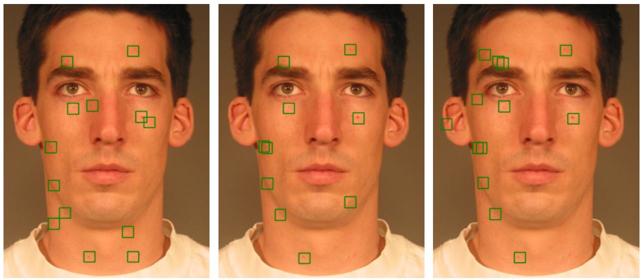

(a) Sliding window (b) FRCNN (c) EfficientDet

Fig. 14. Detected facial marks using different architectures

The difference between the three systems becomes apparent when the detection time is taken into account. Table II compares the average time each system takes to process one face. The averages have been taken from the detection time of each images in the FRGCv2 test subset. The models were run on an Intel Core I7-7700HQ CPU. No GPU acceleration has been implemented.

It should be noted that that the speed of FRCNN and EfficientDet was observed to be relatively consistent, the processing time per face did not vary by more than a tenth of a second on any face. In contrast, the speed of the sliding window system was more variable, due to the nature of the 3-stage detection. The detection time for the sliding window model has been observed to vary between 1.0 and 2.5 seconds, with the average time being 1.3 seconds.

TABLE II. Comparing speed of different facial mark detection architectures

Architecture	**Average time [s]**
Sliding window	1.3
FRCNN	1.8
EfficientDet	**0.4**

The FMD results on the FRGCv2 images are observed to be more consistent and effective than the detections on the Twins Days subset. This is in part because the FRGCv2 subset contains images taken in more consistent lighting conditions than the Twins Days images. To illustrate, Figure 15 shows detections on the same subject from the Twins Days subset, under different lighting conditions. It can be seen that the difference in lighting affects the color of the face notably. This complicates consistent detection of facial marks. Furthermore, the FMD systems have been trained on FRGCv2 training data, which benefits the FRGCv2 test subset. Nevertheless, the systems still successfully detect most prominent facial marks on the Twins Days subset.

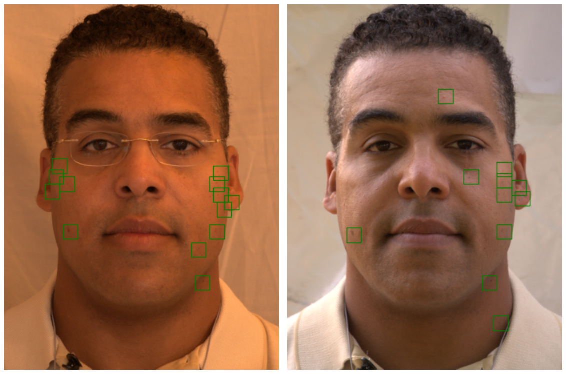


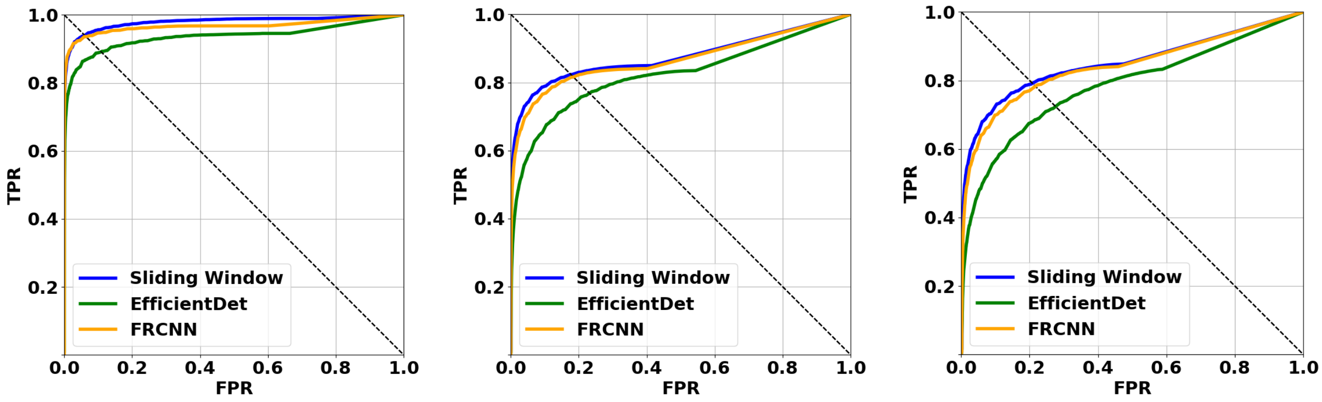
Fig. 15. Varying lighting conditions on the Twins Days dataset hinders facial mark detection

C. Facial recognition

1) Using facial marks: Table III shows the EERs for FMD-based facial recognition, using only overlap scores of detected facial mark patterns. The corresponding ROC curves are plotted in Figure 16. Figure 16(a) shows the results on the FRGCv2 test subset. Figure 16(b) shows the results for the Twins Days subset. Finally, Figure 16(c) shows the results for twin differentiation.

TABLE III. Facial recognition performance using only facial mark detection-based overlap score

Facial mark detection architecture	**EER**		
	FRGCv2	**Twins Days**	**Twin Differentiation**
Sliding window	**0.059**	**0.175**	**0.203**
FRCNN	0.066	0.182	0.223
EfficientDet	0.106	0.239	0.268



(a) FRGCv2

(b) Twins Days

(c) Twin Differentiation

Fig. 16. Facial recognition performance using only facial mark detection-based overlap scores

Figure 16 and Table III show that FMD-based facial recognition performs significantly better on the FRGCv2 subset than on the Twins Days subset. This is consistent with the conclusion drawn regarding FMD performance. The performance on the Twins Days subset is similar to the performance for twin differentiation. This suggests that twin subjects can be distinguished by detected facial mark patterns with the same effectiveness as non-twin subjects.

The results also indicate that the EfficientDet system performs worse than the FRCNN and sliding window system. This could indicate that the EfficientDet architecture is less effective for FMD than the other architectures. However, it could also be a consequence of suboptimal parameters for the system, such as the feature vector threshold. Since the systems have not been thoroughly optimized, no definitive conclusion on this can be drawn.

Figure 16 shows that on all plots, and especially visible in Figure 16(b) and 16(c), the ROC curves initially converge to a true positive rate (TPR) value less than 1.0. After plateauing, yet before reaching a false positive rate (FPR) of 1.0, the curves linearly approach $(1.0, 1.0)$. The linear approach is an interpolated line between the two final data points. To examine why this behaviour is observed Figure 17 shows a histogram visualizing the overlap scores for the Twins Days dataset.

The histogram shows that the bin containing a score of 0 contains a large number of matching and non-matching pairs. These are instances where facial pairs do not have any overlapping marks. In the case of matching pairs, this error may be due to poor FMD performance, a change in pose, changing lighting conditions, or misalignment of the faces. When the threshold of the system reaches 0 on the ROC curve, all pairs with score 0 will be labelled positive at once. This is the cause for the remarkable ROC curves encountered in Figure 16.

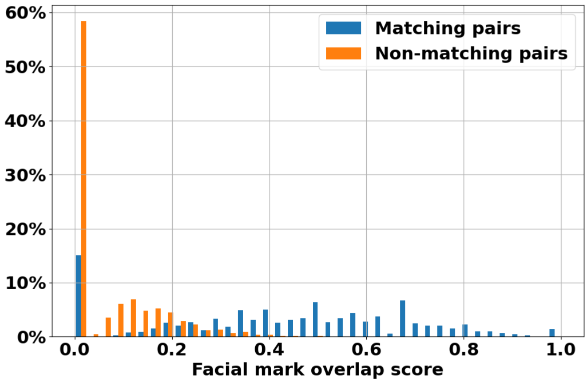


Fig. 17. Facial mark overlap scores for matching and non-matching facial pairs from Twins Days subset

2) *Combining FMD and FaceNet:* The effects of fusing the FMD-based overlap scores with FaceNet's embedding distances are shown in Table IV, Table V and Table VI. In these results model A denotes the FaceNet model trained on the VGGFace2 dataset, and model B denotes the FaceNet model trained on the CASIA-WebFace dataset.

TABLE IV. Effect of score fusion on facial recognition performance for FRGCv2 images

Facial mark detecton architecture	**EER**	
	Model A	**Model B**
Sliding window	0.0033	0.0014
FRCNN	**0.0032**	**0.0012**
EfficientDet	0.0038	0.0019
Without FMD	0.0065	0.0034

Results in Table IV show that on the FRGCv2 test subset, the EER of facial recognition can be decreased by up to 58%, in the case of the sliding window method. This is a remarkable decrease, especially considering the low EER the standalone FaceNet FRS already features. Although not as effective as the other two FMD systems, the decrease in EER is still significant when using the EfficientDet model, at 42% and 44% for model

A and model B respectively. These results do confirm the ability to improve FRS performance using FMD-based facial recognition.

TABLE V. Effect of score fusion on facial recognition performance for Twins Days images

Facial mark detection architecture	**EER**	
	Model A	**Model B**
Sliding window	**0.0093**	**0.0099**
FRCNN	**0.0093**	0.0101
EfficientDet	0.0103	0.0111
Without FMD	0.0115	0.0123

The increase in facial recognition performance for the Twins Days subset is more subtle, yet not insignificant. In this case the EER is decreased by up to 19% using the sliding window system. In the results where the faster EfficientDet system is implemented, the EER for the Twins Days subsets decreased by 10% for both FaceNet models. The decreased improvement on the Twins Days dataset, relative to the FRGCv2 dataset, is in line with the results in Figure 16.

TABLE VI. Effect of score fusion on facial recognition performance for twin differentiation

Facial mark detection architecture	**EER**	
	Model A	**Model B**
Sliding window	**0.196**	**0.193**
FRCNN	**0.196**	0.194
EfficientDet	0.237	0.236
Without FMD	0.299	0.300

The increase in performance for twin differentiation can be seen in Table VI and in Figure 18. In Figure 18 model B has been used as FaceNet model. The baseline system, using FaceNet as a standalone FRS, is labelled ”No FM”. In both the table and the figure a significant increase in performance is observed. As indicated in other works, the standalone FRS struggles to distinguish twins [2]–[5]. Since the FMD-based systems do not appear to suffer from this issue, score fusion contributes a significant increase in performance. The improved performance decreases the EER by 35% and 20% using the sliding window and EfficientDet system respectively. This confirms that FMD-based facial recognition can be used effectively for improved twin differentiation.

In all of the results on facial recognition, it has been observed that combining the FaceNet FRS with the FMD-based systems leads to improved facial recognition performance. The sliding window and FRCNN systems consistently show greater improvements than the EfficientDet system, and improvements on the FRGCv2 subset is the most drastic. This is in line with the difference in performance as seen in Figure 16. The results give a clear indication that FMD-based facial recognition performs as well on twins as it does on unrelated subjects. Because of this, the FMD-based systems can significantly improve twin differentiation performance.

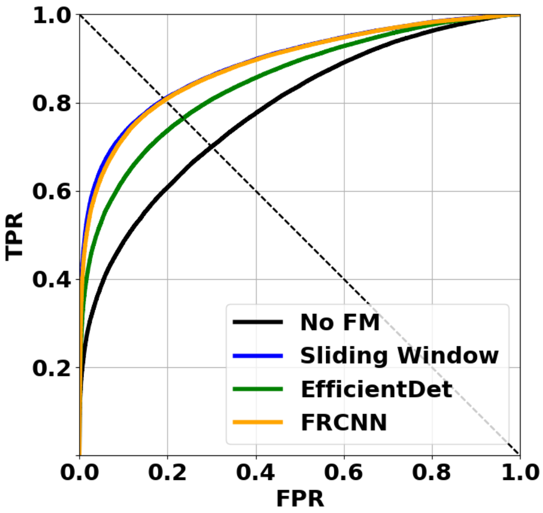


Fig. 18. Twin differentiation with, and without, score fusion. The No FM curve uses a standalone FRS, the other curves combine the FRS with an FMD-based system

VI. Conclusion

Using modern CNN architectures, the accuracy of facial mark recognition has been further advanced. The more complex CNNs outperform shallow CNNs when it comes to facial mark recognition. Using MobileNetV2, an EER of 0.019 for patches of 32 px was demonstrated.

Using a sliding window method, as well as other object detection architectures, facial marks could consistently and accurately be detected. Using these systems, the false positive issue, which hindered Laplacian of Gaussian facial mark detection, was successfully mitigated. Using CNNs for facial mark detection appears to be the state-of-the-art approach.

This is confirmed by the performance of FMD-based facial recognition, when combined with an existing facial recognition system. The EER for facial recognition has been reduced by up to 58% on the FRGCv2 test subset, and up to 19% on the Twins Days subset.

An equally remarkable result is the improved performance on twin differentiation. Facial recognition performance of an FMD-based system was not affected by the similarity between twin siblings. A decrease in EER of up to 35% shows that an FMD-based system is an effective addition to a facial recognition system, especially in the case of twin differentiation.

VII. Future work

In the implementation of the FMD-based facial recognition systems, certain parameters have been decided upon through empirical results. This paper has not looked into optimizing said parameters, in order to optimize the performance of the FMD systems. Such parameters include:

- Feature vector thresholds
- Facial mark pattern matching method
- Score fusion method

Further research into the optimization of these, and perhaps other, parameters, will further improve facial recognition through facial mark detection.

This paper has implemented facial mark detection on high-resolution facial images, using general object detection systems. Research into the feasibility of implementing facial mark detection on low-resolution images, and using optimized detection models, could result in faster and more widely applicable facial mark detection.

REFERENCES

[1] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[2] Zhenan Sun, Alessandra A. Paulino, Jianjiang Feng, Zhenhua Chai, Tieniu Tan, and Anil K. Jain. A study of multibiometric traits of identical twins. In B. V. K. Vijaya Kumar, Salil Prabhakar, and Arun A. Ross, editors, *Biometric Technology for Human Identification VII*, volume 7667, pages 283 – 294. International Society for Optics and Photonics, SPIE, 2010.

[3] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt. Distinguishing identical twins by face recognition. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 185–192, 2011.

[4] J. R. Paone, P. J. Flynn, P. J. Philips, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, M. T. Pruitt, and J. M. Grant. Double trouble: Differentiating identical twins by face recognition. *IEEE Transactions on Information Forensics and Security*, 9(2):285–295, 2014.

[5] M. T. Pruitt, J. M. Grant, J. R. Paone, P. J. Flynn, and R. W. V. Bruegge. Facial recognition of identical twins. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2011.

[6] S. Biswas, K. W. Bowyer, and P. J. Flynn. A study of face recognition of identical twins by humans. In *2011 IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2011.

[7] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, Aug 2018.

[8] C. Zeinstra, R. Veldhuis, and L. Spreeuwers. Grid-based likelihood ratio classifiers for the comparison of facial marks. *IEEE Transactions on Information Forensics and Security*, 13(1):253–264, 2018.

[9] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt. Distinguishing identical twins by face recognition. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 185–192, 2011.

[10] U. Park and A. K. Jain. Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security*, 5(3):406–415, 2010.

[11] Fabiola Becerra-Riera, Annette Morales-González, and Heydi Méndez-Vázquez. Facial marks for improving face recognition. *Pattern Recognition Letters*, 113:3 – 9, 2018. Integrating Biometrics and Forensics.

[12] C. Zeinstra and E. Haasnoot. Shallow cnns for the reliable detection of facial marks. In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2018.

[13] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 630–645, Cham, 2016. Springer International Publishing.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[16] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[17] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[18] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787, 2020.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 91–99. Curran Associates, Inc., 2015.

[21] Arun A. Ross, Karthik Nandakumar, and Anil K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*, pages 114–115. Springer US, Boston, MA, 2006.

[22] Arun A. Ross, Karthik Nandakumar, and Anil K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*, pages 131–137. Springer US, Boston, MA, 2006.

[23] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[24] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4278–4284. AAAI Press, 2017.

[25] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.

[26] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017.