EPiC Series in Computing

Volume 85, 2022, Pages 48–61

Proceedings of 43rd Conference of the South African Institute of Computer Scientists and Information Technologists

Robust Facial Recognition for Occlusions using Facial Landmarks

Kyle Johnston[1] and MKhuseli Ngxande[2]

[1] Stellenbosch University, Stellenbosch, Western Cape, South Africa
Kylejohnston1999@gmail.com
[2] Stellenbosch University, Stellenbosch, Western Cape, South Africa
ngxandem@sun.ac.za

Abstract

Convolutional neural networks have proven to be very powerful for image classification problems, but still has its shortcomings in the presence of non-ideal data. Recently, facial recognition has become popular with usages such as surveillance and automatic tagging of individuals on social media sites. This paper explores a facial recognition solution that utilizes a feature masking strategy focused on facial landmarks with the goal of developing a solution capable of facial recognition in the presence of occlusions. The main driving factor behind this paper is based on the idea that the most commonly found occlusions in the wild are found in the regions of the facial landmarks and that these landmarks play a crucial role during the recognition process. It is found that using a masking strategy based on facial landmarks can be beneficial if the network is trained adequately and the dataset contains mostly well aligned faces, offering improved performance in comparison to using an arbitrary grid layout for all the tested occlusions. Furthermore, it is discovered that masks are not precise at removing the targeted features, causing the masking strategy to also harm recognition process in some cases by accidentally removing critical features.

1 Introduction

Deep convolutional neural networks (CNNs) have recently seen great strides in both popularity and performance. CNN architecture submissions to competitions such as the ImageNet large scale visual recognition challenge (ILSVRC) where image classification and detection is tested on 1000 different object categories have also shown great performance gains in recent years, moving from a 28% classification error in 2010, to a 2.3% classification error in 2017 [12]. CNNs have many applications ranging from detecting various types of cancer, such as melanoma cancer [5, 17] and prostate cancer [14] to enhancing the performance of hearing aids [10]. In many cases, these CNNs are able to outperform humans, such as in [5] where the network was able to outperform many of the dermatologists in detecting melanoma cancer. It is therefore no shock that researchers are constantly researching ways of improving the performance of the CNNs [7, 3]. In the area of facial recognition, CNNs have shown great promise for the constrained recognition problem. Unfortunately, for unconstrained problems performance is negatively impacted by variations in illumination, occlusions, facial expression,

and misalignment [6]. This paper focuses specifically on improving the performance of CNNs for the partially occluded facial recognition problem. The approach followed in this paper is similar to the implementation of Song et al. [13]. Song et al. took inspiration from the human visual system that only focuses on the non-occluded areas of the face. The idea is thus to find the correspondence between the occluded facial regions and the features that are affected by the occlusions. A pairwise differential Siamese network (PDSN) is developed to train a mask generator module. The idea is that the mask generator should be capable of identifying which features have been affected by the occluded area and ultimately be used to remove those features from the recognition process. In [13], a mask generator is trained for a set of predefined blocks in the form of a grid. This paper proposes to rather learn mask generators for regions covering each of the facial landmarks i.e., the eyes, nose, mouth, chin, eyebrows. Two approaches are explored in this paper, both centered around the use of facial landmarks. The idea is inspired by the fact that these landmarks have the largest influence on the recognition process and that most occlusions are found in the regions of these landmarks. The main contribution of this paper is thus the following: it is proposed that a mask dictionary is explicitly formed based on regions corresponding to the facial landmarks in an attempt to prevent the unnecessary removal of facial landmark features and to decrease the number of required masks within the mask dictionary.

2 Background

2.1 Related work

Partial occlusions have long been a challenge for facial recognition systems. Given that partial occlusions are commonly found on faces in real-world scenarios, such as people wearing sunglasses and face masks, it is clearly an area of interest. Previous work has investigated the use of local descriptor vectors that are identified to be non-occluded to be used for classification. Following this approach has achieved some success, as demonstrated by Rui Min et al. [9] where Gabor wavelets, principal component analysis (PCA) and support vector machines (SVM) are used to detect the occluded area while using block-based binary patterns to classify the non-occluded areas. In [8], scale-invariant feature transform (SIFT) features are obtained and used to produce two new local feature descriptors, namely sparse histogram of gradients (HOG) and sparse local binary patterns (LBP) with the goal of extracting more discriminative features from the occluded face for recognition. In [2], the occlusion is detected by using PCA and SVMs after which the area identified as non-occluded is processed by blocked-based weighted local binary patterns for recognition. Another common approach to this problem is to attempt a reconstruction of the occluded area. In [18], the discriminative nature of sparse representation is exploited for classification. This is achieved by finding the minimal number of training samples needed to represent a test image, thus reconstructing the test image using the images from the training set.

More recently, the use of deep learning is a very popular choice in the field facial recognition. Using a convolutional neural network, there is no longer a need to extract the features prior to sending the data to the model. The convolutional layers are capable of automatically finding those features that are crucial to the recognition process. Unfortunately, these deep learning models are also negatively impacted by partial occlusions. Some attempts have been made to diminish the effect that these partial occlusions have on the recognition process such as in [15] where the network is trained with augmented data, to ensure that the model learns discriminative features more equally across the whole face. Although this approach managed to

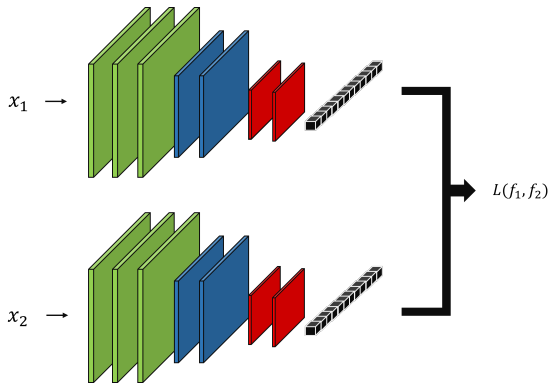

Figure 1: Illustration of a Siamese neural network.

improve performance, the network still utilizes the occluded features during recognition. Song et al. [13] aims at learning a mask dictionary that is capable of removing those features that are corrupted by the occlusions through the use of a Siamese network. This is done by selecting a set number of predefined areas and finding those features that are affected by occluding these selected areas. This method has shown to be successful, managing to diminish the effect of the occlusions on the recognition process.

2.2 Siamese neural networks

Siamese neural networks, also commonly referred to as a twin neural network, is a form of artificial neural networks during which two input vectors are passed through the same network to produce comparable output vectors. These output vectors are then used to calculate the loss and update the weights accordingly. Usually, Siamese networks are visually represented as two identical networks running in parallel to compute the outputs of two different input vectors, as can be seen in figure 1. In reality, one could just use a single network and pass through both input vectors before calculating the loss and updating the weights. Siamese networks therefore do not output probabilities, but rather a measure of similarity between the two input vectors. In this paper, the Siamese network is used to compare the feature maps of an occluded image to that of the non-occluded image and train a mask generator such that when the output from the mask generator is multiplied with the feature maps, the difference between the two masked feature maps is minimized while retaining as much information for successful recognition as possible.

3 Methodology

3.1 Overview

The proposed implementation follows a similar approach to that of [13]. This entails a four-stage process. The first stage involves the training of the base CNN using an unconstrained facial recognition dataset. The second stage trains the mask generators using a pairwise differential Siamese network. The idea behind the mask generator is to find the correspondence between the occluded areas and the feature elements that are affected by these occluded areas. The third stage forms the mask dictionary from the mask generators and the fourth stage combines

the masks of those occluded areas to form the feature discarding mask which is then used to remove those affected features from the recognition process. This paper proposes that the masks responsible for discarding features be targeted for the facial landmarks, which in effect should decrease the number of necessary feature discarding masks (FDMs) and combat the possibly unnecessary removal of features corresponding to these facial landmarks.

3.2 The mask generator

The process starts off with training a CNN with any suitable architecture on a dataset of clean, non-occluded images. This trained model is referred to as the base CNN. The base CNN is then used to train each mask generator using a Siamese network. During the training of a mask generator, the base CNN weights are fixed. Two images that are identical except for one having an occlusion present at some specified location, which will be discussed in section 4, are sent through the PDSN. The absolute difference between the feature maps obtained from passing the two images through the convolutional layers of the base CNN are then used as input to the mask generator. The mask generator consists of a convolutional layer, a PReLU activation layer, a batch normalization layer and finally a sigmoidal activation layer. The output from the mask generator is multiplied elementwise with the feature maps before being sent to the fully connected layers for classification. As discussed in [13], a combination of two losses is used to train the PDSN. The first is formulated as the mean absolute error (MAE) between the masked feature maps, mathematically expressed as

$$l_{diff}(\tilde{f}(x_j^i), \tilde{f}(x^i)) = \frac{1}{n}\|M_\theta(\cdot)f(x^i) - M_\theta(\cdot)f(x_j^i)\|_1 \tag{1}$$

with $f(x^i)$ and $f(x_j^i)$ being the feature maps obtained from the non-occluded image x^i and the occluded image x_j^i that is occluded on block j respectively. $M_\theta(\cdot)$ is the output from the mask generator, $\|\cdot\|_1$ is the L1 norm and n is the absolute size of the vector. $\tilde{f}(x_j^i)$ and $\tilde{f}(x^i)$ are thus the masked feature maps. The second loss is the mean classification loss of the two outputs, meaning that the outputs from both the occluded and non-occluded image should produce accurate classification predictions. Mathematically, this combined classification loss is expressed as

$$\begin{aligned} l_{cls}(\theta; \tilde{f}(x_j^i), \tilde{f}(x^i)y^i) = &-\frac{1}{2}log(p_{y^i}(F(\tilde{f}(x_j^i)))+ \\ &-\frac{1}{2}log(p_{y^i}(F(\tilde{f}(x^i))) \end{aligned} \tag{2}$$

where y^i is the target label and $F(\tilde{f}(x_j^i)$ and $F(\tilde{f}(x^i))$ are the outputs from the masked feature maps. The resulting loss can thus be expressed as

$$l_{final} = l_{cls} + \lambda l_{diff} \tag{3}$$

where λ is a scalar value that is problem dependent and used to balance the MAE loss and classification loss.

3.3 The mask dictionary

Once the mask generator is trained, the mask dictionary can be formed. A mask is created for each of the predefined areas and will correspond to one of the facial landmarks as discussed in section 4. A mask is formed by sending numerous images occluded on one of the facial landmarks through the mask generator trained for that landmark. Each of the outputs from

the mask generator is min-max normalized and the average of all the outputs is then used to obtain a mean mask. This mean mask is converted to a binary mask by applying a thresholding technique such that all values below a certain percentile is set to zero and the rest to 1. The reasoning behind this is that single features can not partially occluded, so those features with high values are likely not occluded and should not be affected by the masks whereas those lower values from the mean mask is likely occluded and should be completely removed from the recognition process. Mathematically, a mask corresponding to block i can be denoted as

$$M_i = \begin{cases} 1 & \text{where } \overline{m_i} > \delta \\ 0 & \text{elsewhere} \end{cases} \tag{4}$$

where $\overline{m_i}$ is the mean mask for block i and δ is the value corresponding to a certain percentile of all the values in the mean mask. This percentile is discussed later on. The mask dictionary is established once all the masks are formed and can then be used to retrieve masks that correspond to the blocks that are occluded.

3.4 The use of facial landmarks

Determining suitable predefined areas that are used to form the mask dictionary is important as it has an impact on the performance of recognition. In [13], the aligned faces are split into a 3x3 grid of equally sized blocks. This paper suggests forming the predefined areas such that it covers the main facial landmarks, such as the nose, eyebrows, eyes and mouth and chin.

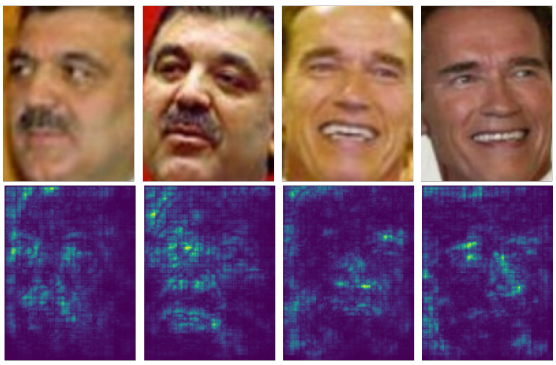

Figure 2: Saliency maps of four individuals taken from the altered Labeled Faces in the Wild (LFW) dataset [4] .

One of the main driving factors behind the proposed implementation is the idea that the facial landmarks have a large influence on the recognition process. Figure 2 contains saliency maps of four different images. The above saliency maps are found by computing the gradients of the output with respect to the input image. The brighter areas indicate areas that have a larger influence on the output. In layman's terms, it is the areas that the model pays attention to. In figure 2, it is evident that the general trend among the randomly selected sample of images is that the facial landmarks make a large contribution towards the classification process with lighter areas predominantly covering the regions of the facial landmarks. A predefined area is seen as occluded if the occlusion overlaps with more than 50% of the predefined landmark region, thus if a standard 3x3 grid is used it might be the case that a block with a facial landmark in it is classified as occluded even though the landmark is still visible. Furthermore, having

a predefined block for an area that does not really contribute to the output is also not ideal. Targeting the facial landmarks aims at preventing the unnecessary removal of facial landmarks, while also reducing the number of FDMs required. Figure 3 contains a plot of the frequencies of the respective facial landmark distances relative to their mean positions in the augmented LFW dataset [4]. These plots indicate, as expected, that most of the landmark occurrences are within a nearby range of the mean. With this in mind, it is proposed that the regions be set up in such a manner that each facial landmark falls within some predefined region. This paper explores two different approaches. The first approach consists of four regions. These regions correspond to the left eye, right eye, nose and mouth. Each region's width and height spans a distance of $4\sigma_1$ and $4\sigma_2$ respectively and is centered around the mean of its corresponding landmark, with σ_1 and σ_2 being the standard deviation of the horizontal and vertical positions of its corresponding landmark respectively. It is found that each landmark falls within their respective region for roughly 95% of the training set. The second approach consists of five regions. Each region will cover the whole width of the image. The reasoning behind this is that most common occlusions that cover a facial landmark end up spanning the whole width of an aligned face, such as face masks, sunglasses, and headbands. The first region will cover the section above the eyes, such as forehead and eyebrows. The second region will cover the eyes themselves. The height of this region will range from 3σ in the upwards direction to 2σ in the downwards direction relative to the mean, where σ is the standard deviation of the vertical positions of the eyes in the training set. The third region will cover the section between the eyes and the mouth, thus corresponding to the nose landmark. The fourth region follows a similar approach to the second region, but with a range of 2σ in each vertical direction relative to the mean and is focused on the mouth rather than the eyes. Lastly, the fifth region covers the section below the mouth, targeting the chin. It is found that using the above-mentioned regions would result in roughly 90% of the training data being sectioned correctly i.e., each landmark falls within their respective region. Looking at each region individually, it is found that each region covers their respective landmark for roughly 95% of the training set. For the remainder of this paper, LM1 would refer to the first masking strategy and LM2 would refer to the second masking strategy. Figure 4 contains an illustration of where the predefined areas would be relative to a 128x96 image. It is thus expected that each of those regions illustrated in figure 4 overlaps with their respective facial landmarks for a large majority of the images given to the model. Figure 4 also illustrates one of the advantages of utilizing the facial landmarks as opposed to using the 3x3 grid. In this illustration, one can witness how using a 3x3 grid could result in losing information about both eyes even though they are visible, due to the large overlap with the occlusion and the grids in which the lower half of the eyes occur. This is not the case when using facial landmarks and the information regarding the eyes are kept, since the overlap between the occlusion and the predefined regions is much smaller. It is thus clear that these two approaches would assist in retaining important information for commonly found occlusions such as face masks.

3.5 Forming and applying the feature discarding mask

The feature discarding mask is formed by combining all the masks corresponding to those areas that are occluded. Since the masks are binary masks with 1 corresponding to the features to be kept and 0 corresponding with the corrupted features, the masks are combined by simple elementwise multiplication. The occluded areas are detected by using a segmentation network. If the output of the segmentation network thus has a 50% overlap with any of the landmark regions, the masks corresponding to those landmarks form part of the feature discarding mask.

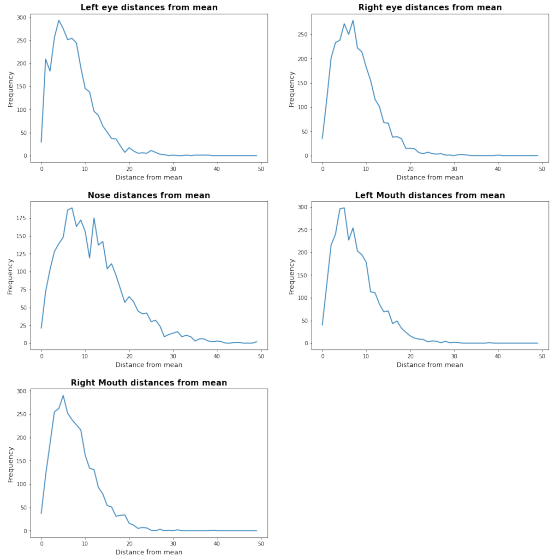

Figure 3: The frequency of the respective facial landmarks relative to their distances from the mean position of the landmarks. The data is retrieved from an augmented LFW dataset [4] dataset

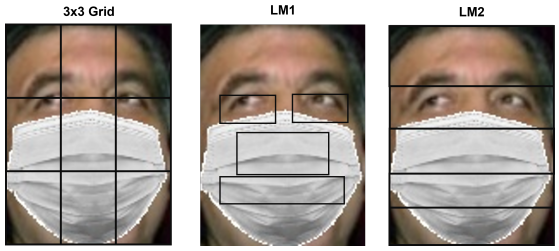

Figure 4: Illustration of the calculated predefined areas displayed over an image from the augmented LFW dataset [4].

Obtaining the feature map of the occluded image and multiplying it with the feature discarding mask calculated for that image then results in the masked feature map which can be sent to the fully connected layer for classification.

4 Empirical Procedure

4.1 Data pre-processing

All images are processed by the MTCNN face detector [19] to obtain the positions of the facial landmarks, as well as the regions bounding the faces. Most of the background information within the images from the LFW dataset is then removed, leaving only the regions bounding the faces to remain. This area is then resized to the dimensions of 128x96. The positions of the facial landmarks are then used to compute the bounds of predefined areas that each facial landmark

corresponds to. An augmented dataset is then created, which has the same targets as the main

Figure 5: The segmentation network training labels before and after applying morphological image processing. Illustrated image is an augmented image from the LFW dataset [4].

dataset but utilizes MaskTheFace [1] which adds synthetic masks to the input images. To obtain labels for the segmentation network, the absolute difference between the original images and the augmented images are calculated and those values that deviated from the original corresponds to the occluded areas and will be labelled as such for the segmentation network. These labels tend to be noisy, so morphological opening and closing procedures are applied to the labels to remove some of the noise. Figure 5 illustrates an augmented image, a binary image obtained from those values that deviated from the original, as well as the binary image after applying morphological opening and closing. It is evident that applying morphological processing to the labels offers a better target label for the segmentation network. Furthermore, a dataset with

Figure 6: The before and after images from adding the synthetic sunglasses to an image from the altered LFW dataset [4]

.

synthetic sunglasses as occlusions is created, also having the same targets as the LFW dataset. The synthetic occlusions are added by using the information regarding the positions of the eyes from the MTCNN face detector [19]. The sunglasses are resized and orientated such that it reasonably covers the appropriate region. Figure 6 illustrates images from the LFW dataset before and after adding the occlusion.

4.2 Training the models

The base CNN utilizes the ResNet34 architecture with a large margin cosine loss (LMCL) [16] classifier. This model is trained with the altered LFW dataset consisting of 2959 training images from 158 classes and is trained until the validation accuracy stagnates or declines. The dataset as a whole consists of 4324 images and has a training, validation, and testing split of 70%, 15%, and 15% respectively. The LFW dataset [4] consists mostly of middle-aged men, with a small portion of the images being women. The conditions (lighting, pose and occlusions) in which

the images were taken are also fairly consistent. The mask generator consists of a convolutional layer, followed by the PReLU activation and batch normalization with sigmoidal output. Each mask generator is trained until the validation accuracy and MAE between the masked features stagnate. A mask is formed for each of the predefined areas by sending numerous image pairs from the altered LFW dataset through the Siamese network. The image pairs consist of the clean image, as well as the synthetically occluded image. The synthetically occluded image is formed by adding blocks of uniform intensity to the predefined area of focus on the clean image. The value of δ which determines the threshold for the binary mask does not seem to impact performance drastically if chosen to be in the range of the 20th to 30th percentile. For the traditional grids, δ is chosen to be the value corresponding to the 25th percentile. For the LM1 masking strategy, δ is set to correspond to the 20th percentile and for the LM2 masking strategy, δ is set to the 30th percentile. The reasoning for this is that the larger the predefined blocks are, the more features would be impacted by occluding that block. Larger δ values are thus assigned to larger blocks. Occlusion detection is achieved using a U-Net segmentation model [11] with a VGG16 base model. Training data for the segmentation model is an augmented LFW dataset.

4.3 Testing the models

The models are tested and compared using the augmented LFW dataset. The models are compared on rank 1 classification accuracy. The testing procedure follows a nearest neighbour approach, during which the output of each of the masked feature maps are compared to each of the outputs from the masked feature maps of the training set. The reasoning behind this is that those features that are affected by the occlusions in the test image, should be removed from both the clean training set data and the occluded test data before their similarity is measured, to allow for a more accurate comparison.

5 Research Results

5.1 Base CNN

The ResNet34 architecture with the LMCL classifier manages to achieve a 82.51% rank 1 test accuracy and a 88.87% rank 5 test accuracy on the LFW dataset. This test accuracy is obtained through a nearest neighbour implementation using cosine similarity.

5.2 Segmentation network

The U-Net segmentation network manages to obtain a 90.49% accuracy, which is adequate for the purposes of this implementation, especially when considering that the test labels are imperfect. Figure 7 provides an illustration of what to expect as output from the segmentation network.

5.3 Test results

Table 1 displays the results obtained for various occluded areas using the different masking techniques. More precisely, this table contains the test accuracy values obtained for the various masking strategies when presented with an occlusion as described in the left-most column. Landmark occlusions are formed synthetically for the 5 facial landmark regions by setting the

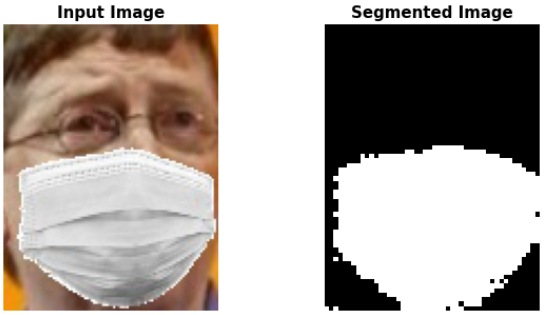

Figure 7: Output from the segmentation network given the occluded face image from the augmented LFW dataset [4].

respective region to be a white block stretching the width of the image. For the face mask, the MaskTheFace [1] library is used and the resulting image is a mask covering the chin, mouth and nose of the individual. The sunglasses are dealt with as described in the section 4.1 and occludes the region of the eyes. From table 1, it is evident that using a no masking strategy in the case of the synthetic landmark occlusions offers poor performance. Furthermore, it can also be observed that occluding the regions in the lower half of the face has a much smaller impact on the recognition process in comparison to occluding regions in the top half. This implies that the network relies heavily on the features present within the upper half of the face. Comparing the three masking strategies, it is clear that LM2 performs the best among the synthetic occlusions. This is to be expected, because the synthetic occlusions are targeted at the facial landmarks and these are exactly the regions this masking strategy is designed for. Figure 8 illustrates the shift in focus from the model when using the FDM from the LM2 masking strategy, moving away from the occluded area and focusing on the non-occluded regions.

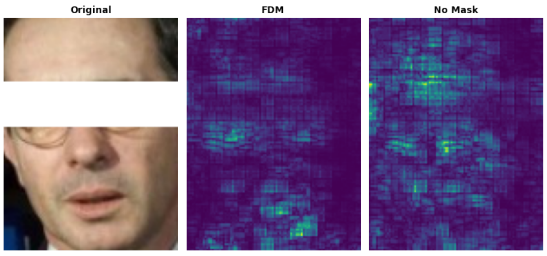

Figure 8: Saliency map with and without applying the FDM. The original image is an augmented image from the LFW dataset [4].

LM1's poor performance across the various synthetic landmark occlusions is understandable, because even after masking many of the features would still be occluded due to the small areas covered using this masking strategy. In some cases, such as the forehead and chin, there is no mask trained for those areas, so it would in effect be the same as having no masking strategy. The 3x3 grid strategy has good performance with some of the occlusions and poor performance for others. Looking at the face mask occlusion, the 3x3 grid performed the worst of the masking strategies. Due to the large overlap that the face masks might have with the middle row of the grid, features correlating to the eyes might be removed which then impacts performance. On

the other hand, occluding the forehead still offered good performance, as the top row of cells within the grid is mostly aligned with the forehead occlusion. With reference to the 3x3 grid's performance for the synthetic nose occlusion, it is evident that its performance is poor. This is once again due to the overlap between the occlusion and the middle row, causing important features regarding the regions of the eyes to be removed. Both the proposed masking strategies aims at avoiding the unnecessary removal of landmark features as is the case with the 3x3 grid in the above-mentioned scenario.

Table 1: A Comparison between the Masking Strategies

Augmented Labelled Faces in the Wild Test Accuracy				
Occlusion	No FDMs	LM1 FDMs	LM2 FDMs	3x3 Grid FDMs
Forehead	06.18	06.18	23.32	18.37
Eyes	02.83	02.47	08.13	03.00
Nose	14.49	15.37	20.49	13.25
Mouth	24.20	26.86	41.34	19.26
Chin	55.12	55.12	65.37	53.00
Face Mask	24.20	20.26	19.89	18.39
Sunglasses	08.30	07.77	07.24	05.83

With regards to the face mask results in table 1, it can be observed that using no masking strategy outperforms all of the other masking strategies. This result may seem strange, but because the base CNN relies very heavily on the features present within the top part of the images and all the masking strategies have some chance of removing non-occluded features i.e., features in the region of the eyes and forehead, it is understandable that using no masking strategy would perform better. Due to the large focus placed by the network on the region of the eyes and forehead, the network is more affected by the accidental removal of some features in those areas than the occluded areas in the regions of the nose, mouth and chin. Unfortunately, the training process for the mask generators is quite complex, with the optimizer struggling to find optimal weights to focus on only those regions affected by the occlusion. This then leads to the behaviour seen in figure 9, where it can be seen that section underneath the occlusion that is crucial for recognition is also removed from the area of focus. Improvements might be seen by further tuning the value of δ, as it is a possibility that the combination of several masks leads to too much information being lost.

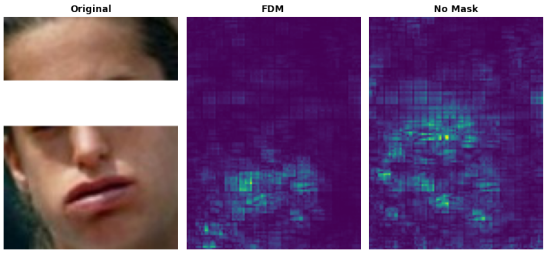

Figure 9: Saliency maps with and without applying the feature discarding mask. The original image is an augmented image from the LFW dataset [4].

Furthermore, since the base CNN is not perfect it is possible that background sections surrounding the face also carries some weight during the recognition process. If the model does indeed focus on areas around the mask, then using an FDM that stretches across the whole width of the image could perhaps harm the recognition process since it might remove those features of interest to the model. This, however, is more a flaw relating to the training of the base CNN than it is a flaw of the proposed masking strategies. To test this hypothesis, a synthetic occlusion comprised of the chin, mouth and nose synthetic landmark occlusions is created and tested against the face mask, because these two occlusions should mostly occlude the same facial features. It is found that using the synthetic landmark occlusion produces a test accuracy of 11.13% when using no FDM. This is quite a large drop from the 24.20% that is obtained from the face masks, meaning it is likely that the base CNN also focuses on background information. Of course, it is also a possibility that the face is orientated such that other features like the ears become more prominent and in such a case, the LM2 approach would still discard these features if the face mask overlap is large enough, which could clearly harm the recognition process. Focusing on the sunglasses, these results are somewhat unexpected following the results obtained from the synthetic eye landmark occlusion. These results stem from the variation in face alignment across the dataset. In a scenario where someone's face might be tilted to the left or right, the sunglasses would occlude much less of the image than its synthetic eye landmark occlusion counterpart. The LM2 masking strategy would then still remove those features on the sides of the face, regardless of whether it is occluded. The LM1 masking strategy would be impacted slightly less, but its alignment would be off. Due to the layout of the 3x3 grid, the sunglasses would in some cases not overlap enough with any of the cells for a mask to be included in the FDM. In other cases, a lot of non-occluded information would be removed due to the poor alignment between the grid and sunglasses, hence the poor performance. The grid layout is, however, beneficial in a case where the occlusion only occurs on one side of the face and triggers one of the cells, such as sunglasses when the face is tilted. It is possible to alter the LM2 approach such that each horizontally stretched cell is split into two to cater for occlusions occurring on only one side of the face, but this comes at the expense of training more mask generators. Once again, it is a possibility that those background regions surrounding the sunglasses are also used by the model for recognition, which would lead to improved performance when not using a masking strategy. The results are therefore mixed. It is clear from the synthetic landmark occlusions that the masking strategies definitely has the ability to improve the performance of the models, but this result does not translate too well when using the synthetic face masks and sunglasses. It appears that the base model has placed some focus on background regions and the FDMs are designed such that these features are removed if the section of the face near it is occluded. Furthermore, the alignment of the face is also an issue and causes masking strategies such as the LM2 strategy to remove non-occluded features. To fully understand the impact that the facial landmark sectioning has on the recognition process, the base CNN would have to be trained such that the focus is fully placed on the face itself and the faces would have to be well aligned. Considering that the whole idea behind tackling occlusions is to improve performance on the unconstrained facial recognition problem, one does not want to add the constraint of the face being well aligned. Taking all of the above information into consideration, it would be best to form a masking strategy that is targeted at facial landmarks, but also caters for variation in horizontal alignment of the face to prevent the removal of non-occluded features in the case of imperfect facial alignment.

6 Conclusion

This paper proposes a facial recognition method that utilizes a pairwise differential Siamese network to find the correspondence between the areas relating to the facial landmarks and the feature elements affected by those regions. When presented with certain commonly found occlusions, it is found that forming a masking strategy that is focused on facial landmarks is more capable of retaining important features in comparison to arbitrary grids. The ability to retain important features result in better test accuracies, thus allowing the landmark masking strategy to outperform arbitrary grids for commonly found occlusions. It is also found that the FDMs are not precise at removing the targeted features, which could lead to important information being lost, harming the recognition process. Furthermore, it is clear that the model also pays attention to regions in the background. This can also cause the FDMs to harm the recognition process, as this background information is available regardless of the occlusion worn in the case of using a no masking strategy, but will likely be removed when using the proposed masking strategies. This, however, is more a flaw of the base CNN than it is a flaw in the masking strategies. In conclusion, targeting the facial landmarks does have merit, but it is necessary to also cater for variation in horizontal alignment in order to get good results for occlusions found in the wild. In future, it is worth exploring a combination of the technique proposed by Trigueros et al. [15] and the technique proposed in this paper, such that the features are extracted more equally across the various facial landmarks, which should result in more stable performance for all occlusions. It is also worth exploring a masking strategy that caters for horizontal variations while still utilizing the positions of the facial landmarks. Furthermore, training the base model such that background information is not utilized during classification would also result in more fruitful results that might be closer to the true impact that the sectioning of the masks has on the facial recognition process when presented with everyday occlusions.

References

[1] Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication. *arXiv preprint arXiv:2008.11104*, 2020.

[2] Zhaohua Chen, Tingrong Xu, and Zhiyuan Han. Occluded face recognition based on the improved svm and block weighted lbp. In *2011 International Conference on Image Analysis and Signal Processing*, pages 118–122. IEEE, 2011.

[3] Dan Claudiu Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence*, 2011.

[4] B Huang Gary, Ramesh Manu, Berg Tamara, L Erik, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Technical Report 07-49, University of Massachusetts*, volume 1. 2007.

[5] Holger A Haenssle, Christine Fink, Roland Schneiderbauer, Ferdinand Toberer, Timo Buhl, Andreas Blum, A Kalloo, A Ben Hadj Hassen, Luc Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology*, 29(8):1836–1842, 2018.

[6] Samil Karahan, Merve Kilinc Yildirum, Kadir Kirtac, Ferhat Sukru Rende, Gultekin Butun, and Hazim Kemal Ekenel. How image degradations affect deep cnn-based face recognition? In *2016 international conference of the biometrics special interest group (BIOSIG)*, pages 1–5. IEEE, 2016.

[7] Jungkyu Lee, Taeryun Won, Tae Kwan Lee, Hyemin Lee, Geonmo Gu, and Kiho Hong. Compounding the performance improvements of assembled techniques in a convolutional neural network. *arXiv preprint arXiv:2001.06268*, 2020.

[8] Na Liu, Jianhuang Lai, and Huining Qiu. Robust face recognition by sparse local features from a single image under occlusion. In *2011 Sixth International Conference on Image and Graphics*, pages 500–505. IEEE, 2011.

[9] Rui Min, Abdenour Hadid, and Jean-Luc Dugelay. Improving the recognition of faces occluded by facial accessories. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 442–447. IEEE, 2011.

[10] Se Rim Park and Jinwon Lee. A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*, 2016.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[13] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 773–782, 2019.

[14] Zaneta Swiderska, Thomas Bel, Lionel Blanchet, Alexi Baidoshvili, Dirk Vossen, Jeroen Laak, and Geert Litjens. Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. *Scientific Reports*, 10, 09 2020.

[15] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing*, 79:99–108, 2018.

[16] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition, 2018.

[17] Julia K Winkler, Katharina Sies, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, et al. Melanoma recognition by a deep learning convolutional neural network—performance in different melanoma subtypes and localisations. *European Journal of Cancer*, 127:21–29, 2020.

[18] John Wright, Allen Yang, Arvind Ganesh, Shankar Sastry, and Lei Yu. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31:210 – 227, 03 2009.

[19] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.