OCCLUSION ROBUST FACE RECOGNITION BASED ON MASK LEARNING

Weitao Wan, Jiansheng Chen

Department of Electronic Engineering, Tsinghua University, Beijing 100084, P.R. China
wwt16@mails.tsinghua.edu.cn jschenthu@mail.tsinghua.edu.cn

ABSTRACT

Face occlusion has been a long standing challenging issue in face recognition. In the state-of-the-art deep Convolutional Neural Network (CNN) face recognition models, occluded facial parts are generally embedded into the learned features together with the non-occluded parts in an equivalent manner. As such, the discriminative power of the generated face representation may be weakened for occluded face images. To address this problem, we propose the MaskNet, a trainable module which can be included in existing CNN architectures. With end-to-end training supervised by only the personal identity labels, MaskNet learns a proper way of adaptively generating different feature map masks for different occluded face images. Intuitively, MaskNet automatically assigns higher weights to the hidden units activated by the non-occluded facial parts and lower weights to those that are activated by the occluded facial parts. Experiments on datasets consisting of real-life and synthetic occluded faces demonstrate that MaskNet can effectively improve the robustness of CNN models towards occlusions in face recognition.

Index Terms— Face recognition, occlusion, Convolutional Neural Network, mask

1. INTRODUCTION

Recently, face recognition techniques have achieved dramatic performance improvements on the public databases [1, 2, 3]. However, in real-life application, face recognition systems often suffer a degradation in accuracy compared to their performances on experimental test datasets. This is largely due to the fact that real-life face images often contain uncontrollable illumination variations, facial expressions and occlusions. Among all these variations, the face occlusion has been considered highly challenging.

In the occluded face images, the occluded parts are irrelavant to the personal identity. For instance, people of different identities may wear the same kind of sunglasses while the same person may be occluded by different objects such as sunglasses or a scarf. Current state-of-the-art methods for face recognition are mainly based on the deep Convolutional Neural Network (CNN) [4, 5, 6, 7]. Input images are processed through deep CNN to generate the discriminative face representations with the property of relatively low intra-class distance and high inter-class distance. When using current CNN models for face recognition with occlusion, the occluded parts are embedded into the face representation in an equivalent manner as the non-occluded parts, which will increase the intra-class variation and decrease the inter-class variation. This usually leads to a degradation in accuracy.

We believe that a more reasonable way is to enable the CNN to discriminate the occluded parts from the non-occluded parts adaptively. However, it is difficult for the conventional deep CNN to have this ability. For a convolutional layer, the convolution operation is local and the kernel is shared at every spatial location. While for the fully connected layer, the effective receptive field of its input may become very large when the CNN goes deeper. However, for face recognition, the depth of a CNN has already been proved to be critical for achieving high recognition accuracy. These factors will be discussed in detail in Sect. 3. To address this problem, we propose a neural network module namely MaskNet, aiming at learning to focus on the image features with high fidelity and ignore those distorted by occlusions.

Our main contributions are summarized as follows: (i) we propose the MaskNet module for learning feature masks to improve the performance of face recognition with occlusion; (ii) we show that MaskNet can be included in different CNNs and optimized with end-to-end training; (iii) we demonstrates the effectiveness of MaskNet with real-life as well as synthetic occluded face images.

2. RELATED WORK

Many different approaches have been proposed for solving the occlusion problem. The method in [8] is based on Principal Component Analysis (PCA) which projects the face image into a low dimensional subspace. This method can handle only horizontal occlusions because it relies on row-based occlusion detection.

Since the occlusion can corrupt the features of the whole image, local representation methods are proposed to address this problem. The method [9] divides the face image into several local regions and independently analyzes them to solve the occlusion problem. The method [10] estimates the probability of occlusion based on the Local Gabor Binary Patterns

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

(LGBP). And the probability is used as the weight for local regions. The discriminative power of this approach is restricted because the local operators are handcrafted.

Several methods have been proposed to directly remove the occlusion and explicitly reconstruct the occluded face areas [11, 12, 13]. Generally speaking, the recognition performances of these methods rely heavily on the realistic degree of the image reconstruction result. [14] seeks a sparse solution based on an occlusion dictionary to recover the faces. The generalization ability of the method is restricted because the occlusion patterns which can be handled are limited to the occlusions in the training set.

The CNN-based method of learning a mask for occlusion robust face recognition has not been specifically discussed. However, in other vision tasks, such as the image caption, the method [15] which focuses on salient regions has inspired us. Our proposed model is also inspired by the architecture proposed in [16] which includes a CNN module into other networks to learn the spatial transformation of the feature map.

3. MASKNET

3.1. Architecture

The proposed MaskNet is a differentiable module which can be included in a CNN network and optimized with end-to-end training as shown in Fig. 1. The MaskNet takes an image with fixed size as input, followed by a relatively shallow convolutional network. A regression layer is fully connected to the last convolution output in MaskNet. This regression layer outputs a weighting coefficient for each spatial location of the feature map U in the recognition network. The matrix M consisting of these weighting coefficients is the learned mask. The mask and feature map U have the same width and height. One weight in the mask is assigned to the features in the same spatial location in U. The intersection point of the MaskNet and the recognition network can be selected accordingly for different architecture designs.

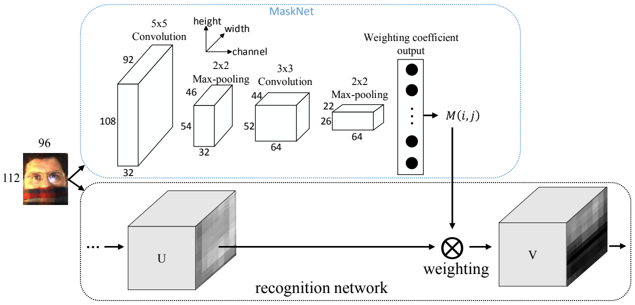

Fig. 1. The architecture of the MaskNet included into the recognition network. All activation layers are omitted for simplicity. The output of the fully connected layer of MaskNet is reshaped to the spatial size of U . Then the mask M is computed with an activation function. The weighting unit multiplies each weight in the mask M with the features of U at the same spatial location. V is the feature map after weighting. The whole network is optimized with end-to-end training using only the label of person identities.

Table 1. The spatial size of feature maps and the RF in ResNet. The size of the input image is 112×96.

Layer	Pool1	Conv2	Pool2	Conv3	Conv4_2
RF	6	18	20	60	100
Feature	54x46	52x44	26x22	24x20	12x10

3.2. The necessity of an explicit mask

It is possible for a fully connected layer in the conventional CNN to learn weighting coefficients with respect to different locations in the input feature map. Nevertheless, the problem is that the effective receptive field (RF) for each hidden unit becomes larger and larger as the network goes deeper. Table 1 shows the spatial size of feature maps and the effective RF of several different layers in the ResNet [17]. The RF size has already reached 100×100 in the fourth Residual Unit. Practical instances of ResNet usually contain even more residual units, indicating that the effective RF of the input hidden units of the fully connected layer may have already spanned the entire input image. As such, the occlusion in the input image may be strongly correlated to all the input hidden units of the fully connected layer. This makes it difficult to generate reasonable weighting coefficients for hidden units at different spatial locations. For face recognition with occlusion, the occluded parts can corrupt the extracted image features and thus lead to inaccurate recognition results. Explicitly learning a mask can put more weights on useful features while relatively ignore those corrupted by the occlusion.

3.3. Optimization

We address the problem in Sect. 3.2 by including the MaskNet in the recognition network at an appropriate depth. For the convergence of the training, the intersection point of the MaskNet and the recognition network should be neither too deep nor too shallow. Because deep neurons have too large effective RFs as described in Sect. 3.2 while the feature maps at shallow layers are short of descriptive capability, leading to difficulties in learning reasonable masks. For example, when the ResNet is used as the recognition network, we suggest to choose the output of the second pooling layer to be the feature map U, of which the spatial size is 26×22. As such, the fully connected layer in MaskNet has the output of the size $26 \times 22 = 572$.

Unlike most fully connected layers in other CNNs, an activation function is applied to the output of the fully connected layer in the MaskNet to constrain the numerical range of the generated mask. Actually, we have found through experiments that such a design also benefits the convergence of end-to-end training. The sigmoid function is utilized instead of the commonly used ReLU because the sigmoid function maps the output into the interval of $[0, 1]$, which fits the property of a normalized weighting coefficient. The training of the proposed network requires only the labels of person identities. Extra annotation such as the information about the occlusion is not needed. We use joint supervision of the softmax loss and the center loss[5] in the training.

4. EXPERIMENTS

4.1. Mask visualization

We first use LFW [1] for a qualitative test of mask generation. The test images are roughly aligned [18], which means that the target face may not be strictly centered in the aligned image and there may exist background objects or even faces of other people. For the recognition network, we use the Maxout network [19] with 8 convolution layers and 3 pooling layers. The MaskNet is applied to the feature map of the 3rd pooling layer of the recognition network and the mask is of the size 9×9. The model is trained on the CASIA-Webface [20] dataset which contains 10,575 subjects and 494,414 face images. The training images are aligned using MTCNN [21] and the aligned images are resized to 100×100. Each aligned image is horizontally flipped for data augmentation.

To visualize the generated mask, we map the weight coefficients back to each pixel of the input image. Some sample mapped masks are visualized in Fig. 2 together with the input aligned face images. A higher value in the mapped mask indicates a larger weight on the image at the corresponding spatial location. The visualized masks are composited of square blocks simply because the overall stride in the 3rd pooling layer of the Maxout network is 8. The results show that the MaskNet can effectively learn where the useful information is in the input image for face recognition. While unrelated backgrounds or objects are efficiently ignored by assigning very low weights by the MaskNet.

A similar experiment is also performed on face images with different types of occlusions. Following the data protocol used in [22], we fine-tune the model on AR dataset [3] using the first 21 configurations out of 26 for each of the 126 individuals. We visualize sample mapped masks for some AR images as well as some Internet images in Fig. 3. It can be observed that for test images with real-life occlusions, the MaskNet successfully generate reasonable masks to place more weights on non-occluded face areas. Further experiments will demonstrate how the mask improves face recognition accuracy on face images with both synthetic and real-life occlusions in Sect. 4.2 and Sect. 4.3.

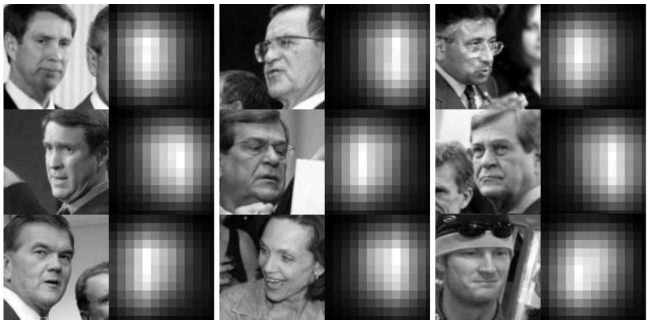

Fig. 2. LFW test images and the mapped masks.

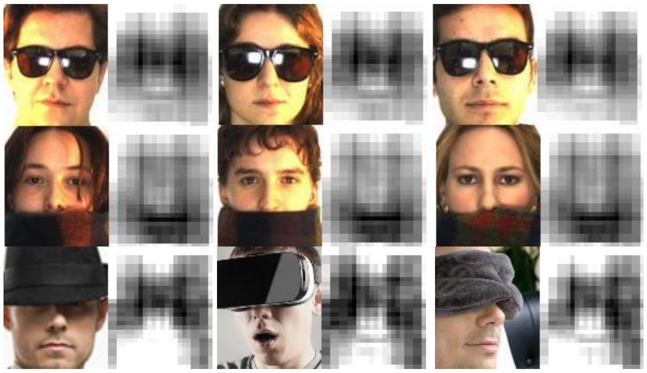

Fig. 3. Test images and the mapped masks, with the first two rows for AR dataset and the last row for Internet images.

4.2. Face verification on the occluded LFW

In this section we evaluate the face verification performance of the proposed model on dataset with synthetic occlusions. The test dataset is generated from LFW by randomly placing a $n \times n$ black square on each image. Fig. 4 shows examples of the synthetic test data for different n values.

Two different CNN models are used as recognition networks: a Maxout network [19] with 4 maxout units (8 convolution layers), and a 28-layer ResNet [17]. The MaskNet is included into each of the the two recognition networks respectively to get Mask-Maxout and Mask-ResNet. The MaskNet is applied to the third and the second pooling layer of the two recognition networks respectively. Joint softmax loss and center loss supervision is adopted for the optimization.

For fair comparison, all the 4 models, namely Maxout, ResNet, Mask-Maxout and Mask-ResNet, share the same training configuration. The training set is CASIA-Webface dataset and its synthetic counterpart in which 50% images contain synthesized sunglasses and 50% images contain synthesized scarves. All images are first aligned by MTCNN [21] and then resized to 112×96 before training.

The face verification results on the occluded LFW dataset

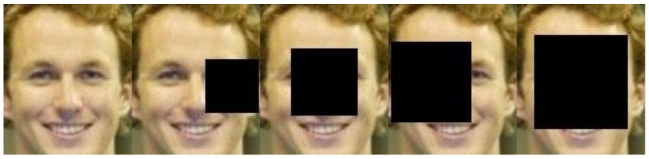

Fig. 4. Example LFW test images with synthetic occlusion. The left-most one is the original image followed by images with random block sizes n = 40, 50, 60, 70 respectively.

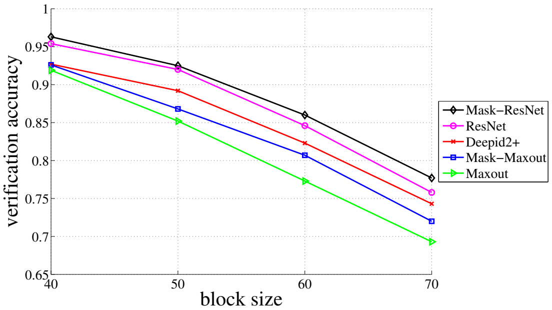

Fig. 5. Face verification accuracy on the occluded LFW with square block occlusion at random position. All results are for a single model. The result for Deepid2+ is quoted from [6].

are illustrated in Fig. 5. We have the following observations from the results. First, Mask-Maxout and Mask-ResNet outperform Maxout and ResNet respectively, improving the verification accuracy by 3.4% and 1.4% when the block size is 60 × 60. It shows that MaskNet can be included in different CNN models for improving the recognition performance on synthetic occluded faces with end-to-end training. Second, the MaskNet improves the performance of ResNet obviously. Mask-ResNet not only beats the single-model deepid2+ but also achieves comparable accuracy as 25 deepid2+ models combined [6] as shown in Table 2. While our single model needs much less running time (1.2ms vs. 35ms per image for feature extraction) and simpler training process.

4.3. Face identification on AR dataset

We further explore the effectiveness of the proposed model through face identification experiments on the AR dataset with real-life occlusions. Following [14], we randomly choose 90 individuals out of 126 for the test. For each test individual, 7 images without occlusion are included in the gallery and 12 images (6 with sunglasses, 6 with a scarf) are in the probe set. In the training set, 7 images without occlusion for each individual, and 6 images with sunglasses or scarf for each of the 36 individuals outside the test set are used. We fine-tune and test the 4 models introduced in Sect. 4.2 for the face identification task.

Table 2. Comparison of verification accuracies of a single-model Mask-ResNet and 25 Deepid2+ models combined.

block size	40	50	60	70
Mask-ResNet	96.3%	**92.5%**	86.0%	**77.7%**
25 Deepid2+	**96.4%**	92.4%	**86.5%**	77.6%

Table 3. Face identification accuracy on the AR dataset. 'All' stands for the whole test set. And the last two columns are for faces with sunglasses and scarves respectively.

Method	All	sunglasses	scarf
MPCA-LDA[22]	91.9%	-	-
SOC-struct[14]	90.5%	-	-
F-LR-IRNNLS[23]	84.3%	89.8%	78.8%
KLD-LGBP[10]	90.0%	82.0%	**98.0%**
ResNet	89.0%	83.9%	94.1%
Maxout	89.8%	85.3%	94.2%
Mask-ResNet	91.6%	87.2%	96.0%
Mask-Maxout	**93.8%**	**90.9%**	96.7%

The identification accuracy on the test set is shown in Table 3. Comparing Mask-Maxout with Maxout, and Mask-ResNet with ResNet, the overall identification accuracy is improved by 2.6% and 3.0% respectively. Mask-Maxout achieves the best result under this train/test protocol for AR dataset. It even outperforms the MPCA-LDA [22] (91.9%) which uses 21 images from each individual for training. Considering only the faces with sunglasses, Mask-Maskout improves the accuracy significantly from 85.3% to 90.9%. KLD-LGBP only performs well on faces with scarves while the accuracy degrades dramatically on faces with sunglasses (82.0%). This experiment demonstrates that the MaskNet improves face identification accuracies on the AR dataset.

5. CONCLUSIONS

In this paper we propose a trainable module called MaskNet for occlusion robust face recognition. The module can be easily optimized with end-to-end training when included into existing CNN architectures. Qualitative experiments show that the MaskNet can well distinguish the useful facial areas from the occluded parts. Further quantitative experiments on datasets with synthetic and real-life occlusion demonstrate that MaskNet can effectively improve the face recognition performance for occluded faces. It is worth noticing that MaskNet needs only personal identity labels and that it requires only small amount of extra computation compared to existing CNN based recognition models.

This work was supported by the National Natural Science Foundation of China under Grant 61673234.

6. REFERENCES

[1] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," Tech. Rep. UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[2] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.

[3] A. M. Martinez, "The ar face database," *CVC Technical Report*, vol. 24, 1998.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[6] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2892–2900.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, et al., "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[8] T. Y. Kim, K. M. Lee, S. U. Lee, and C.-H. Yim, "Occlusion invariant face recognition using two-dimensional pca," in *Advances in Computer Graphics and Computer Vision*, pp. 305–315. Springer, 2007.

[9] A. M. Martínez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 24, no. 6, pp. 748–763, 2002.

[10] W. Zhang, S. Shan, X. Chen, and W. Gao, "Local gabor binary patterns based on kullback–leibler divergence for partially occluded face recognition," *IEEE signal processing letters*, vol. 14, no. 11, pp. 875–878, 2007.

[11] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust lstm-autoencoders for face de-occlusion in the wild," *arXiv preprint arXiv:1612.08534*, 2016.

[12] J.-S. Park, Y. H. Oh, S. C. Ahn, and S.-W. Lee, "Glasses removal from facial image using recursive error compensation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 805–811, 2005.

[13] L. Cheng, J. Wang, Y. Gong, and Q. Hou, "Robust deep auto-encoder for occluded face recognition," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1099–1102.

[14] Y. Wen, W. Liu, M. Yang, Y. Fu, Y. Xiang, and R. Hu, "Structured occlusion coding for robust face recognition," *Neurocomputing*, vol. 178, pp. 11–24, 2016.

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, et al., "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044*, vol. 2, no. 3, pp. 5, 2015.

[16] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[18] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 10, pp. 1978–1990, 2011.

[19] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio, "Maxout networks.," *ICML (3)*, vol. 28, pp. 1319–1327, 2013.

[20] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[22] J. Harguess and J. Aggarwal, "A case for the average-half-face in 2d and 3d for face recognition," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2009, pp. 7–12.

[23] M. Iliadis, H. Wang, R. Molina, and A. K. Katsaggelos, "Robust and low-rank representation for fast face identification with occlusions," *arXiv preprint arXiv:1605.02266*, 2016.