Deep Difference Analysis in Similar-looking Face recognition

Yaoyao Zhong
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Email: zhongyaoyao@bupt.edu.cn

Weihong Deng
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Email: whdeng@bupt.edu.cn

***Abstract*—Deep convolutional neural networks (DCNNs) have recently demonstrated impressive performance in face recognition. However, there is no clear understanding of what difference they find between two similar-looking faces. In this paper, we propose a visualization method that gives insight into difference of similar-looking faces found by DCNNs. This method, used as an assistant role, could help human to identify people who try to invade the biometric system using a similar-looking face. We design a crowdsourcing task to evaluate our method. With assistance of our method, accuracy of participants is greatly increased by 8%, which is also better than the accuracy of network, while participants get little improvement with assistance of Deconvolutional network or Gradient Back-propagation. The experiment result suggests that our method makes a difference in human-machine cooperation.**

I. Introduction

Labeled Faces in the Wild (LFW) [1] provides a same/not-same benchmark which addresses the face recognition problem as a similarity problem, which requires that methods learn to evaluate the similarity of unseen face pairs. It has become the de-facto standard regarding to unconstrained face-recognition evaluation in recent years. With the recent deep learning technique developing and underlying large training dataset accumulating, DCNNs have reported extremely high accuracy rates on LFW. There is an intense debate on whether unconstrained face verification problem has already been solved.

Database SLLFW [2] just makes a little change to LFW by selecting similar-looking face pairs in the original LFW database as negative pairs. As a result, accuracy of several state-of-the-art methods drops about 10% - 20% compared to the corresponding LFW performance. Images in the real world are not easy as either LFW or SLLFW. As they said in [3], there exists a gap between people in daily life and images captured in LFW which are smiling, make-up, young, and beautiful. We also should factor in the possible large pose variation, heavy make-up, or occlusions in real-world application.

So I guess we have a long road ahead. On one hand, people fight for more accurate models and larger amounts of data. On the other hand, we try to interpret and visualize the model, for the purpose of, when the machine outperforms humans, convincing the user, and when machine and human have equal but complementary ability, making it possible for user to select the reasonable decisions of machine by showing them why machines make those decisions.

Fortunately there is a strong complementary relationship between machine and human operators. In the human survey of work [4], the controlled human survey yields 99.85% accuracy on LFW but their accuracy drops to 92.03% on SLLFW, which suggests that it may be difficult for human operators to detect deliberate imposters. According to our experiment, on the dataset where human operators have made wrong decisions, the accuracy of DCNNs network only drops about 3%. At the same time, we find the network could make stupid mistakes in the view of human. By taking advantages of the complementarity, it is possible for DCNNs to help human operators.

Currently DCNNs based method could only provide human with a two-valued result, showing whether a pair of face belong to a person or not. DCNNs could not get entirely right, in this case, it is necessary to provide more information about DCNNs to human.

This motivated us to explore what difference DCNNs find between two similar-looking faces and show it to human operators so that they could get more specific help by combining their opinion with the reasonable judgement of machines. In this work, we propose a patch-by-patch occlusion method which could apply to any DCNNs. This patch-by-patch occlusion method allows us to observe the two-value result of DCNNs and different regions of two similar-looking faces at the same time.

II. Related Work

DCNNs have led to impressive performance on a variety of visual tasks. However, it is difficult to understand DCNNs exactly because we could not use functions to describe a particular, trained DCNNs due to the large number of interacting, non-linear parts. A number of previous works have visualized DCNNs predictions in the literature. A common thought of understanding the decisions of DCNNs is to find regions of an image that were particularly important to classification.

Some methods get influential regions by calculating a weighted sum of features of the last convolutional layer, which could be used in weakly-supervised localization. Zhou et al. [5] proposed a technique called Class Activation Mapping (CAM) that replace fully-connected layers with an average

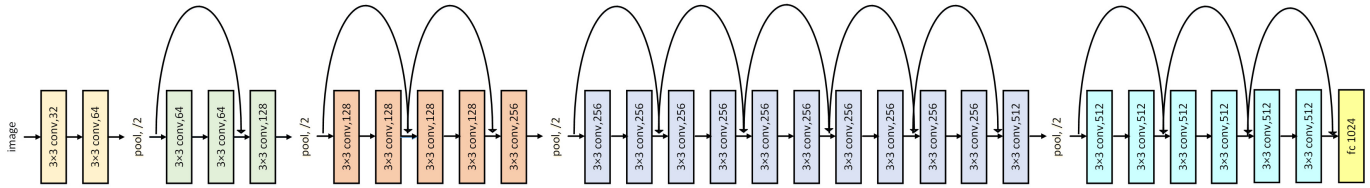


Fig. 1: The modified ResNet is the network architecture we used in this paper.

pooling layer that could help highlight task relevant regions. Grad-CAM [6] and Grad-CAM++ [7] make modifications to the weight coefficient of features of the last convolutional layer.

Some methods calculate gradients to assign a value that reflects the influence of an individual pixel on the final classification. Zeiler et al. [8] use Deconvolutional networks to visualize which patterns from the training set activate the feature map. Simonyan et al. [9] back propagate a class score to the input data layer, visualizing image-specific class saliency map. Guided Backpropagation [10] make modifications to raw gradients that result in improvements.

Other methods use occlusion techniques to find regions that are most useful to predict the classification score. Zeiler et al. [8] perturb input by occluding patches and monitoring classification scores which will be lower when the relevant objects are occluded. Bazzani et al. [11] propose a technique for self-taught object localization by analyzing the change in the recognition scores when artificially masking out different regions of the image.

The methods we mentioned above conduct class-discriminative analysis. While there are no effective methods for non-class-specific similarity analysis in face recognition. We dedicate to finding the corresponding regions of face pairs that were particularly important for verifying the face pairs, that is, to find difference of a pair of faces. Face recognition involves two faces, twice forward pass of DCNNs, meaning that conventional class-specific discriminative analysis methods appropriate for a single image would not help. Taking into account all these factors we could still deploy a deep difference analysis by occlusion, but in a pairwise manner.

III. Approach

The difference analysis is based on DCNNs and our pairwise occlusion method could be used with a variety of networks. In this paper, the network we used is a modified ResNet [12], inspired by [13], as shown in Figure 1. The network maps a color image composed of three 2D arrays containing pixel intensities in the three colour channels, via a series of layers, to a probability vector over different classes. The middle layers of the network are composed of convolutional layers with shortcut connections which turn the network into its counterpart residual version. Softmax loss and Center loss [14] are used for training.

The trained network yields 99.18% accuracy on LFW and 95.55% accuracy on SLLFW, indicating that the network has a relatively good identification ability. Besides there is a strong complementary relationship between the network and human. We pick a dataset of 478 pairs where human participants have made mistakes, containing 193 negative pairs and 265 positive pairs in SLLFW. Surprisingly, the network yields 92.36% accuracy on this dataset. The accuracy on this dataset only drops about 3% compared with that on the whole SLLFW database. It means that our network has a certain degree of ability to identify face images that are difficult for human operators.

A. Pairwise occlusion

In the process of face verification, we import an aligned input face pair with a generic shape $\{x_1, x_2\}$ to the network, extract their features from a selected layer of CNN, and then compute their Cosine distance. The judgement standard is the distance metric of face pairs. If their distance is bigger than the threshold, images are judged as images of the same person, and vice versa.

We attempt to find regions of two face images that were important to the similarity score by a manner of pairwise occlusion. Since the face pair is aligned, deep difference analysis is deployed by systematically occluding different regions of the input face pair, and monitoring the fluctuation of the distance between the deep activation features relative to the original distance. Each time, the position of a patch p is taken, and the value of a face pair $\{x_1, x_2\}$ in position p is set to the mean value of the whole SLLFW database. After occlusion, we get a new pair $\{x_1^{'}, x_2^{'}\}$ and evaluate their new distance. The difference value of the distance measures the importance of this region to the similarity of this face pair.

The bigger the difference value is, the more sensitive and more important this region is, and vice versa. We take regions patch-by-patch, traversing a pair of images, and repeat the process. Eventually, we assign the difference value from all the position, to a 2D measure matrix Δ of the same size as the input face images, which reflects relative importance of all positions. The procedure is illustrated in Figure 2.

B. Normalization

We conduct a statistical analysis on the measure matrix Δ of 300 face pairs in SLLFW. As we mentioned before, the region is not important if Δ is negative. What we concern on is somewhere influential, therefore, we set the negative value to zero for convenience. The distribution information of their Δ at all locations is illustrated in Figure 3a, taking a form of histogram, about 3 million difference values ($300\times104\times96$ = 2995200 pixels, size of input image is 104×96) from value 0 to 0.157 with the interval of 0.001. We plot them

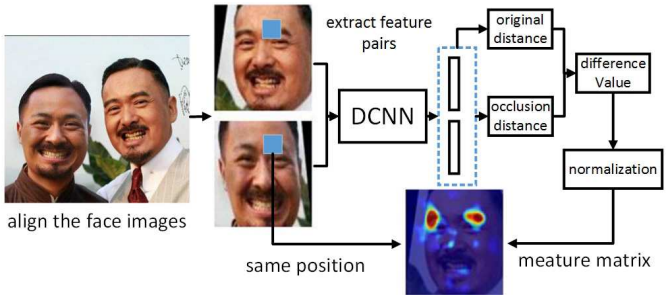
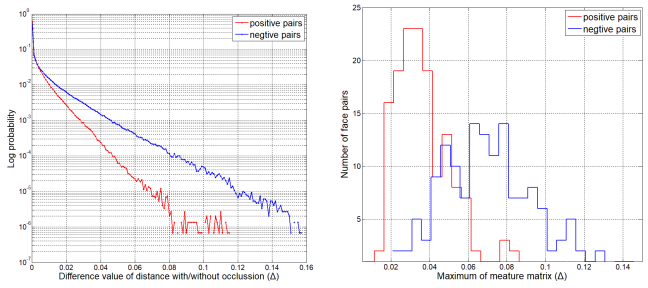


Fig. 2: The procedure of patch-by-patch pairwise occlusion. A face pair is aligned first. The deep difference analysis method attempts to systematically occlude different regions of the aligned face pair, and monitor the difference value of the distance between the deep activation features relative to the original distance. Then the normalization is applied to the difference value, obtaining a measure matrix which reflects the degree of difference between the face pair at all positions.



(a) (b)

Fig. 3: Statistical analysis. (a): The distribution information of their Δ at all locations, taking a form of histogram, on logarithmic coordinate, about 3 million difference values ($300\times104\times96$ = 2995200 pixels, size of input image is 104×96) from value 0 to 0.157 with the interval of 0.001. (b): Maximum of Δ of 150 negative pairs and 150 positive pairs

on logarithmic coordinate to see them clearly. As Figure 3a shows, there is a higher fluctuation in Δ of the negative pairs. It is reasonable that negative pairs have more bigger Δ, indicating that these pairs have more different regions and higher degree of difference.

Figure 3b shows Maximum of Δ of 150 negative pairs and 150 positive pairs. As expected, there is an evident distributional difference between Δ of negative and positive pairs. Usually negative pairs have bigger maximum of Δ.

The statistical characters make it possible to visualize the difference of negative/positive pairs discriminatively. Our intuitive idea is that heatmaps of positive pairs show nowhere highlighted and heatmaps of negative pairs show the most discriminative regions, which could help human operators to distinguish. Guided by this idea, we proposed a min-max normalization function with threshold that maps the measure matrix Δ to the range of [0,1].

$$\Delta = max\{0, min(1, \frac{\Delta}{\Theta})\} \quad (1)$$

In function (1) the threshold Θ is counted from a reference set. According to the distributional difference shown in Figure 3b, Θ is set by a principle that maximums of Δ of negative pairs are bigger than Θ and those of positive pairs are less than Θ. With the normalization method, the most discriminative regions of negative pairs would be highlighted.

Figure 2 and Algorithm 1 illustrate how the pairwise occlusion method can be implemented, incorporating the proposed normalization.

Algorithm 1 Deep difference analysis by pairwise occlusion

Input: aligned face pairs $\{x_1, x_2\}$, patch size k, DCNN feature extractor feature(), $(n1, n2)$ = sizeof(x_1)
for $\{x_1, x_2\}_i$ in face pairs **in** SLLFW **do**
  $d1$ = Distance(feature(x_1),feature(x_2))
  Δ = zeros($n1, n2$), counts = zeros($n1, n2$)
  for every patch p of size $k \times k$ **in** Δ **do**
    $x_1^{'}$ = copy(x_1)
    $x_2^{'}$ = copy(x_2)
    $x_1^{'}$(coordinates of p) = mean value of SLLFW
    $x_2^{'}$(coordinates of p) = mean value of SLLFW
    $d2$ = Distance(feature($x_1^{'}$),feature($x_2^{'}$))
    Δ(coordinates of p) += $d1$ - $d2$
    counts(coordinates of p) += 1
  end for
  Δ /= counts //*point-wise division*
end for
Output: Δ = max$\{0,$ min$(1, \frac{\Delta}{\Theta})\}$

IV. Visualization Result of SLLFW

In this section, we analyze and visualize discriminative regions in face pairs of SLLFW, using our pairwise occlusion method and other two methods which we have mentioned in the related work.

Using our pairwise occlusion method, we map the normalized 2D measure matrixes Δ to heatmaps. Figure 4 shows 9 pairs of them. There are 6 negative pairs in the first two rows and 3 positive pairs in the third row. The superimposition of original images and the same heatmap is shown on the right of the original face pairs. In a superimposition, the intensity of the color and the size of highlight regions represent the difference degree of a pair. From Figure 4, we observe that the most discriminative area of negative pairs is highlighted and there is little regions of positive pairs highlighted. For example, in row 1, col 1, the most discriminative regions for this pair are eyebrows and nose. In row 3, the heatmaps show discriminative area rarely. We will conduct an experiment in the next section to explore whether these heatmaps, which explain the discriminative area for DCNNs-based decisions, is possible to help human in face recognition.

To visualize the discriminative area, we also try other class-discriminative method expecting to deploy difference analysis by observing difference of two discriminative heatmaps of a face pair.

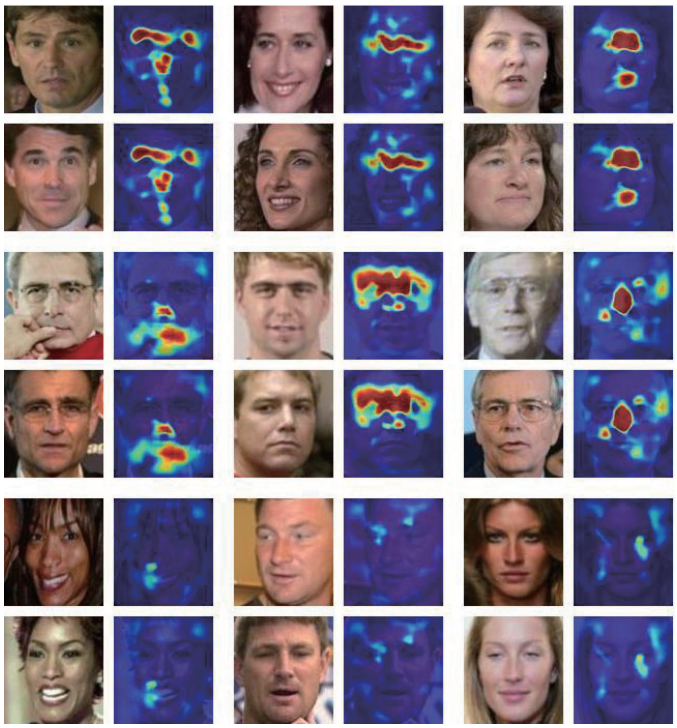


Fig. 4: Results of 9 pairs of our pairwise occlusion method. There are 6 negative pairs in the first two rows and 3 positive pairs in the third row. Each pair is shown in the form of two original images lie on the left and on the right are the same heatmap combined with two original images Both the intensity of the color and the size of highlight area represent the different degree of a pair.

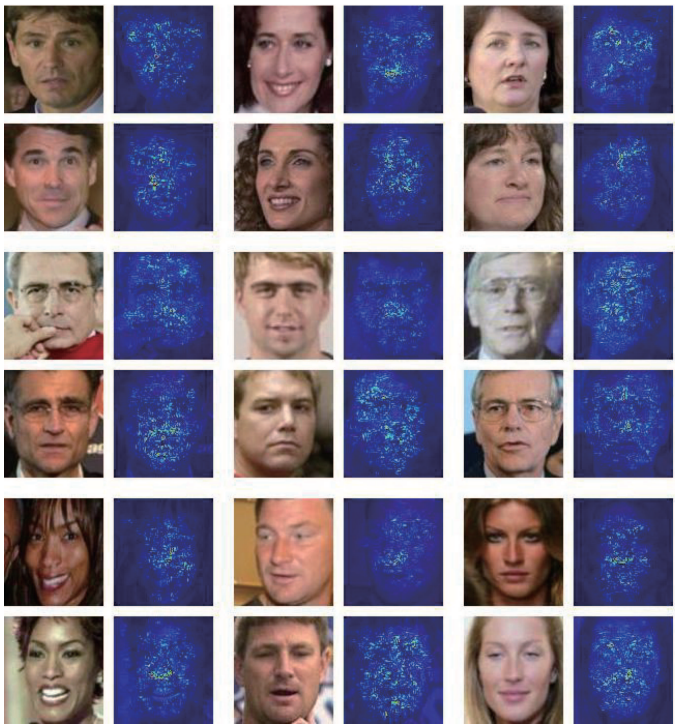


Fig. 5: Discriminative areas of 9 pairs using the Gradient Back-propagation. Each pair is shown in the form of original images lying on the left and on the right are original images combined with their heatmap. Difference of a face pair may be observed by the difference of two discriminative heatmaps.

Work of Simonyan et al. [9] proposes a method for computing image-specific class saliency map using a single back-propagation pass through a classification DCNNs. Figure 5 shows the discriminative regions of 9 pairs, face pairs and their orders are same as before. We try the multi-layered Deconvolutional network proposed by Zeiler et al. [8] that allows to visualize the most influential parts of a face image for classification. Figure 6 shows the discriminative regions of 9 pairs.

V. The Crowdsourcing Task

In order to evaluate the technique, we design a crowdsourcing task that compares humans' performance on face recognition in different settings: with no help of machine, with help of visual cues of one of three visualization methods we have mentioned.

A. Experimental design

We would like to observe how much the reasonable decisions of network could help human, so we select a dataset which is relatively fair for both machine and human. In this case, participants performance would not improve if they follow the cues of machine completely. The dataset, selected from SLLFW, consists of 200 face pairs, where the network

Fig. 6: Discriminative areas of 9 pairs using the Deconvolutional network. We observe that the most discriminative regions of a face is displayed. Difference of a face pair may be observed by the difference of two discriminative maps.

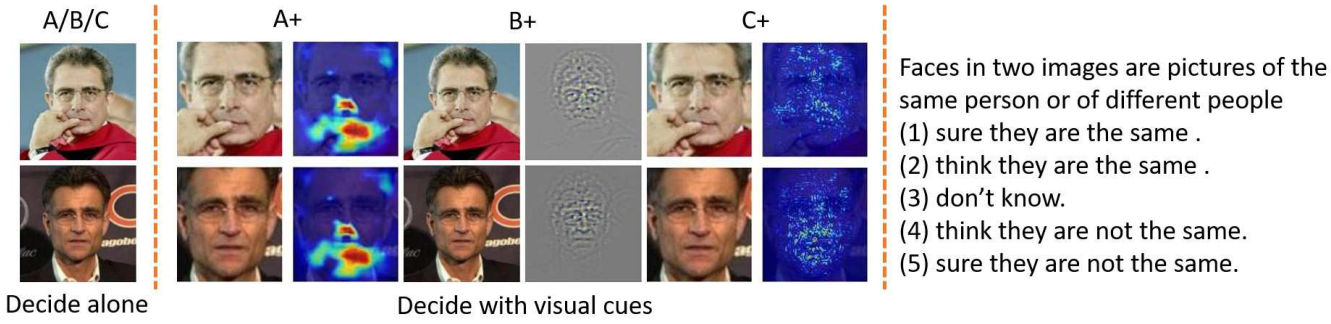


Fig. 7: Experiment demo shown to participants in the crowdsourcing experiment.

yields 75 % accuracy. According to the human survey result of work [4], human also yields 75 % accuracy on the 200 pairs approximately.

Participants need to make decision on 200 face pairs twice. The first time they decide alone without any cues and the second time they get help from one of the three types of visual cues. They are always asked to describe the similarity of 200 face pairs using one of the five choices as Figure 7 shows.

Undergraduate students from the School of Information and Communication Engineering at the Beijing University of Posts and Telecommunication volunteered to participate in this experiment in exchange for a research credit. In the experiment, participants receive no training or practice before and they did not know the celebrities in SLLFW previously. They are allowed to make their choices independently on the dataset for the first time. Right after that they are allowed to judge another time with the assistance of visualization created by one of the three methods we have mentioned above. In the whole process, participants have no idea about if their choices are correct. They have unlimited time to make choices for each pair, with images remaining on the screen until a choice was entered.

A total of 277 voluntary students paricipated in the experiment. Actually, participants differ in understanding of the experiments, ability to distinguish faces and attitude towards experiments. To eliminate these irrelevant factors, we select 90 participants which are serious, responsible and good at face verification (their original accuracy on the dataset is around 75% as we expect). Among them, 30 participants first made choices on their own and then with the assistance of visualization created by our pairwise occlusion (This before-and-after result is refer to as A and A+); 30 participants first made choices on their own and then with the assistance of visualization created by Deconvolutional networks (This before-and-after result is refer to as B and B+); 30 participants first made choices on their own and then with the assistance of visualization created by the Gradient Back-propagation (This before-and-after result is refer to as C and C+).

B. Results and discussion

We deal with the choices as [15]. Choices were transformed into "same" or "different" judgments for individual pairs of faces. Choices 1 and 2 were deemed "same" judgments and choices 3, 4, and 5 were deemed "different" judgments. We compute the mean value and standard deviation of their accuracy to generate a histogram. Accuracy of every group is shown in Figure 8, with more detailed data shown in Table I. As Figure 8 shows, our network yeilds 75% accuracy on the dataset and accuracy of 90 participants we selected is around 76%, which is accord with our hypothesis. With the help of our technique, the mean value of A+ is increased by 8 % compared with that of A. While, compared with B and C, the mean value of B+ and C+ remains unchanged. Note that, the standard deviation of A+ declines obviously, implying that

TABLE I: The mean value and standard deviation of verification accuracy of every 30 participants on 200 face pairs in the experiment.

	The verification accuracy (%)
our network	75.00%
A (original accuracy)	76.19%±4.95%
A+ (with help of pairwise occlusion)	84.14%±1.13%
B (original accuracy)	76.05%±5.71%
B+ (with help of Deconvolutional network)	76.33%±6.45%
C (original accuracy)	76.47%±4.29
C+ (with help of Gradient Back-propagation)	76.26%±3.20%

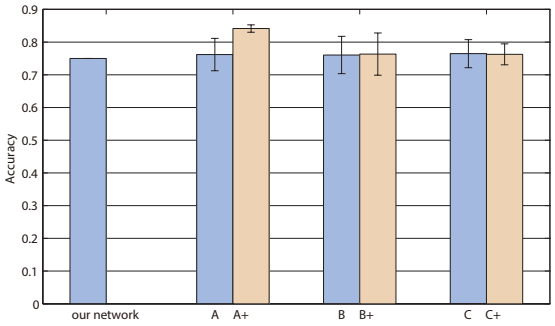


Fig. 8: The mean value and standard deviation of participants' accuracy. In the histogram, bars show the mean value and error bars show the standard deviation.

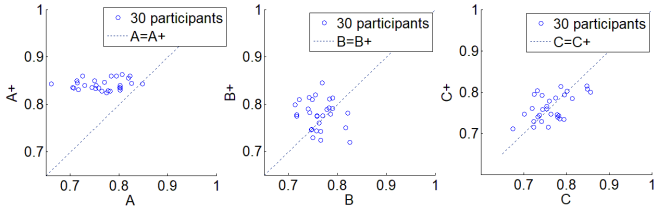


Fig. 9: Behavior of each participant. Each circle represents a participant, plotted by accuracy of his first time label (e.g. A) as the horizontal axis and second time label (e.g. A+) as the vertical. Left: Behavior of participants whose results are A and A+ (with assistance of our pairwise occlusion method). Middle: Behavior of participants whose results are B and B+ (with assistance of Deconvolutional networks). Right: Behavior of participants whose results are C and C+ (with assistance of Gradient Back-propagation).

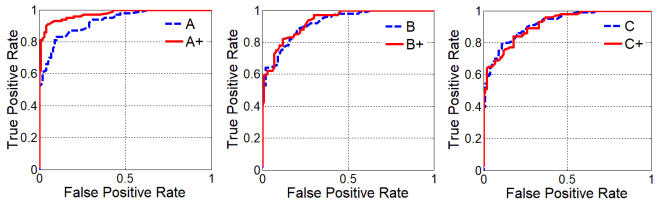


Fig. 10: Left: The ROC curves of the 30 participants whose results are A and A+ (with assistance of our technique). Middle: The ROC curves of the 30 participants whose results are B and B+ (with assistance of deconvolutional networks). Right: The ROC curves of the 30 participants whose results are C and C+ (with assistance of Gradient Back-propagation).

in some way human have reached the highest level on this dataset.

For the purpose of observing behavior of each participant, we have three scatterplots as shown in Figure 9. Each circle represents a participant, plotted by accuracy of his first label stage (e.g. A) as the horizontal axis and accuracy of second label stage (e.g. A+) as the vertical axis. The diagonal line is shown in each scatterplot. A circle in the area above diagonal represents a participant whose accuracy increased with the assistance of visualization. A circle in the area below diagonal represents a participant whose accuracy decreased in the second label stage. A circle on diagonal represents a participant whose accuracy remain unchanged.

From the first scatterplot (where participants' recognition results are A and A+) in figure 9, we could observe that almost all of circles lie in the area above diagonal, implying that with assistance of our technique, the recognition results of almost all of the 30 participants get better. At the same time, circles in the second scatterplot (where participants' recognition results are B and B+) and the third scatterplot(where participants' recognition results are C and C+) distribute in the area near diagonal, which implies that visual cues of Deconvolutional networks or those of Gradient Back-propagation offer little or no aid to these participants.

Each positive/negative judgement from the full range of response data could be assigned a certain value to compute the ROC curves. As Figure 10 shows, we would reach the same conclusion as before.

VI. Conclusion

In this work, we propose a patch-by-patch pairwise occlusion method to visualize the difference of a face pair found by DCNNs. With assistance of the pairwise occlusion, the mean value of human's accuracy is increased by 8%. Eventually, we make a conclusion that the pairwise occlusion could help in helping human to distinguish similar-looking faces, that is, it makes a difference in human-machine cooperation.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 61573068, and 61375031, and Beijing Nova Program under Grant No. Z161100004916088.

References

[1] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," *Month*, 2008.

[2] W. Deng, J. Hu, N. Zhang, B. Chen, and J. Guo, "Similar-looking labeled faces in the wild (sllfw)." [Online]. Available: http://www.whdeng.cn/SLLFW/index.html

[3] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of LFW benchmark or not?" [Online]. Available: http://arxiv.org/abs/1501.04690

[4] W. Deng, J. Hu, N. Zhang, B. Chen, and J. Guo, "Fine-grained face verification: Fglfw database, baselines, and human-dcmn partnership," *Pattern Recognition*, 2016.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," pp. 2921–2929, 2016.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *See https://arxiv. org/abs/1610.02391 v3*, 2016.

[7] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," *http://arxiv.org/abs/1710.11063*, 2017.

[8] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," vol. 8689, pp. 818–833, 2013.

[9] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *Computer Science*, 2014.

[10] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[11] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani, "Self-taught object localization with deep networks," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–9.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2015.

[13] B. Chen, W. Deng, and D. Junping, "Noisy softmax: Improving the generelizationg ability of dcnn via postponing the early softmax saturation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[14] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," vol. 47, no. 9, pp. 11–26, 2016.

[15] A. J. O'Toole, P. P. Jonathon, F. Jiang, J. Ayyad, N. Penard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 29, no. 9, pp. 1642–1646, 2007.