A STUDY OF CNN OUTSIDE OF TRAINING CONDITIONS

Gabriel Dahia, Matheus Santos, Maurício Pamplona Segundo

Intelligent Vision Research Lab, Department of Computer Science, Federal University of Bahia

ABSTRACT

Convolution neural networks (CNN) are the main development in face recognition in recent years. However, their description capacities have been somewhat understudied. In this paper, we show that training CNN only with color images is enough to properly describe depth and near infrared face images by assessing the performance of three publicly available CNN models on these other modalities. Furthermore, we find that, despite displaying results comparable to the human performance on LFW, not all CNN behave like humans recognizing faces in other scenarios.

Index Terms**—** Face Recognition, Deep Learning, CNNs

1. INTRODUCTION

Face recognition of color images in the wild has seen major improvements recently and has closed the gap to human performance on that same task [1], mainly due to the development of Deep Learning techniques. Convolutional Neural Networks (CNNs) achieved 1% EER on the challenging Labeled Faces in the Wild (LFW) [2] dataset, while traditional descriptors such as LBP, SIFT and HOG, even when combined with other techniques as Joint Bayesian, are only able to achieve 4.83% EER [3]. However, all published CNNs that tackle face description were trained for color images; other modalities were ignored so far.

The interest for these other modalities arise from their well documented advantages over color for face recognition. Depth based face recognition is invariant to both pose [4] and ambient illumination variations and NIR images are only mildly affected by illumination changes [5]. While there are reported results for deep networks performing heterogeneous NIR-color face recognition [6, 7], no work, to the best of our knowledge, has used Deep Learning for homogeneous NIR face recognition or any depth recognition whatsoever. The main reason for that is the inadequacy of the available facial datasets for other modalities to the currently known Deep Learning techniques. The key to the CNNs' performance is the size of the training dataset (*e.g.* 200 millions of images for FaceNet [8]) – no publicly available datasets (and we suspect no privately owned as well) for depth or NIR have the necessary depth (*i.e.* number of images per subject) or width (*i.e.* number of subjects overall) that meet the Deep Learning training standards.

In this paper, we investigate if CNNs trained for color images can describe NIR and depth images efficiently without further training or fine tuning. There is some intuition that corroborates these claims: NIR and color images share some texture information, even if illumination patterns make them consistently different, and depth, while not encoding any texture, represent the 3D shape that is also inferable from color images.

The same intuitions that lead our first questioning are the ones behind our second. We intuitively believe them because they are true for human beings analyzing faces, so if they hold for some CNN, other characteristics from human recognizing faces must also do. Therefore, we compare the performance of some CNNs for face recognition to humans performing that same task in an array of situations. First, we compare the results across modalities with our assumed human performance on that task. Second, we assess the effects of pose variation in recognition.

In order to verify our claims, we tested three publicly available CNNs, with results comparable to the state-of-the-art methods, in three different face datasets, each with a different investigative purpose.

2. FACE RECOGNITION USING STATE-OF-THE-ART DEEP REPRESENTATION

In this work we use three deep public CNNs: OpenFace [9], Wu *et al.* deep representation (hereon called Wu-C) [7] and VGG-Face [10]. These differ in terms of face normalization, description size and matching scheme, as described in Sections 2.1 and 2.2.

2.1. Normalization

In the normalization for OpenFace, an affine transformation is employed to standardize outer eye corners and nose tip locations in an image with 96×96 pixels [9]. For Wu-C, a 128×128 gray-scale face crop is obtained after setting the distance between eye and mouth centers to 48 pixels and the

The authors would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Akyiama Soluções Tecnológicas for their financial support.

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

distance between eyes and image border in the y-axis to 40 pixels [7]. VGG-Face uses 224×224 images but does not provide specific instructions for normalization, despite explicitly mentioning a transformation *"to map the face to a canonical position"* [10]. Therefore, we evaluated its performance without normalization, with OpenFace's normalization and with Wu-C's normalization, and the latter one gave the best results. Similarly to OpenFace, we used Dlib's [11] face and facial landmarks detectors to perform these normalization processes, and an example of the obtained results is shown in Figure 1(a). VGG-Face originally used Mathias *et al.*'s [12] and Everingham *et al.*'s [13] works to locate the face and its facial landmarks, and Wu-C used Sun *et al.*'s work [14] for this task, but our implementations using Dlib achieved comparable recognition results for both representations. To adapt Dlib's landmarks for Wu-C's normalization, we estimated the center of the eyes and mouth as the average of their contours.

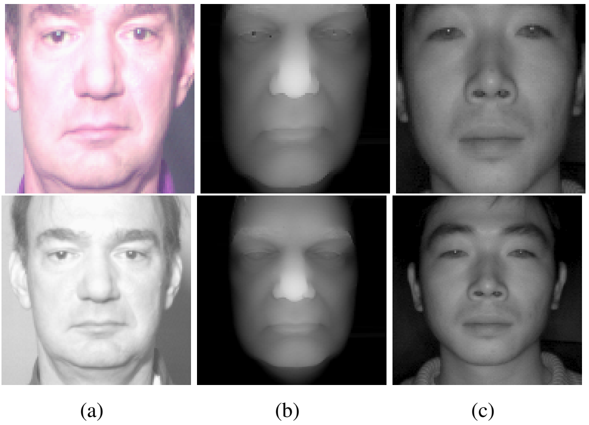


Fig. 1. Face normalization results following (top) Openface [9] and (bottom) Wu-C [7] specifications for (a) color, (b) depth and (c) NIR images.

Since NIR images are similar to their color counterparts, we were able to apply the same procedures described above to normalize them (see Figure 1(c)). However, depth images are very different from color ones, causing Dlib's detectors to fail on several of them. Since these images can be registered, we use the landmarks from the color image to normalize its registered depth image. In order to homogenize pixel intensities, we align the input image to an average face using Iterative Closest Points [15], using only points in the eyes and nose area to avoid problems with facial expressions [16]. Finally, we fill holes by propagating the value of border pixels. A resulting image may be seen in Figure 1(b).

2.2. Description and matching

OpenFace [9] is an open source library for face recognition based on FaceNet [8]. Its neural network was trained using approximately 500K images and 10K identities from CASIA-WebFace [17] and FaceScrub [18] datasets. The resulting descriptor has 128 dimensions and comparisons are performed using L2 distance. The reported EER on the LFW database for the model used in this article, *nn4.small2.v1*, is 7.08%.

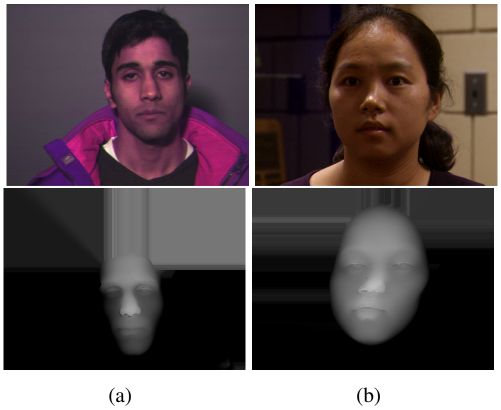


Fig. 2. Examples of problems encountered on FRGC.

Wu-C [7] was trained using CASIA-WebFace and MS-Celeb-1M [19] databases, totaling more than 5M images. Its descriptors have 256 dimensions and are compared through cosine distance, reaching 1.20% EER on the LFW database for *model C*, the same one used in this article.

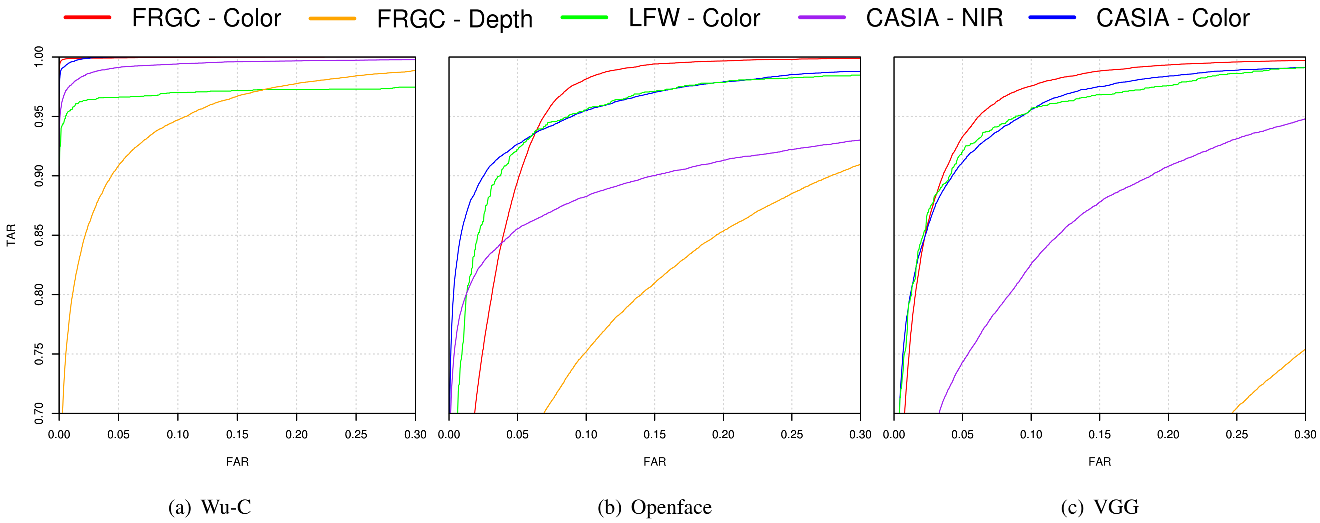
VGG-Face [10] was trained using a private database with about 1M images and 2.6K identities. Although its face representations were supposed to be compared through L2 distance, we found, experimentally, that inner product yields better results. It achieves 1.05% EER on the LFW database with a 4096-dimensional descriptor. In this work, we used the available CNN trained using the softmax method.

In our experiments, we found different EER for the three descriptors on LFW: Openface achieved 7.28%, Wu-C achieved 2.25% and VGG achieved 6,48%. The difference in OpenFace's and Wu-C's results is due to our discarding of faces that could not be normalized and the minor differences from our normalization to theirs, as described in section 2.1. The reason for the large discrepancy between ours and VGG-Face's results is that, while they crop 10 patches, center with horizontal flip and average the feature vectors from each patch, we just pass the face image once, to do justice to the other methods and to save experimental time.

3. EXPERIMENTS

3.1. Datasets

We chose the Labeled Faces in the Wild (LFW) [2], the Face Recognition Grand Challenge (FRGC) [20] and the CASIA NIR-VIS 2.0 (CASIA) dataset [21] because they are the most used in the literature to evaluate face recognition based on color, depth and NIR images, respectively. This way, we are able to evaluate the chosen CNNs in these three modalities. Coincidentally, we can also use them evaluate the performance of the chosen CNNs for color images in uncon-



(a) Wu-C (b) Openface (c) VGG

Fig. 3. ROC curves for different descriptors in different modalities and databases.

strained (LFW), in mildly constrained (CASIA) and in controlled (FRGC) settings.

The LFW is a dataset for unconstrained face recognition with more than 13,000 color images of 1,680 subjects. The View 2 of LFW, as specified by its maintainers, was used in the experiments, since it is regularly used as a benchmark in the literature.

FRGC provides 4,950 registered depth and color images of 556 subjects, mostly with frontal faces and various degrees of illumination variations and facial expressions. These images are divided in training and testing sets, and all training images were used to create the average face model required by the normalization in Section 2.1.

FRGC's testing set has 4,007 images of 466 subjects, from which five images were not properly normalized and four images were found to be defective – these nine images were the only discarded. The perceived problems were deformed depth shapes (see Figure 2(b)) and misalignment between color and depth shapes (see Figure 2(a)).

CASIA provides 17,580 images of 725 subjects captured in both color and NIR. It contains face images of children, adults and seniors with pose and facial expression variations. The subjects in this dataset do not have the same number of NIR and color images. Therefore, in order to perform a fair comparison between NIR and color modalities, we randomly selected the maximum number of images so that all subjects have the same number of images in both NIR and color. After that we got 5,025 correctly normalized images of 715 subjects for each modality.

3.2. Results

Our first experiment is a performance comparison between the chosen CNNs across different modalities. We describe each test image through a forward pass in each CNN and then perform an all-vs-all comparison among representations of the same CNN and modality. This experiment has two different objectives: (1) assess if CNNs trained for color face images can be transfered to different domains without fine tuning; and (2) determine if CNNs have human-like behavior in recognizing faces (*i.e.* fare better in controlled scenarios).

Figure 3 shows the obtained ROC curves for this experiment. As may be seen, all evaluated CNNs, even if only trained for color images, are able to properly describe and recognize NIR and depth face images. Although Wu *et al.* [7] already determined Wu-C's potential for heterogeneous face recognition from color to NIR images, to the best of our knowledge this is the first report on how color-only CNNs perform on recognizing faces in NIR and depth images. Their capacity of describing faces across modalities, even if with varying success, shows that these CNNs learned intrinsic facial characteristics, such as shape and texture.

From Figures 3(b) and 3(c), we can see that OpenFace's and VGG-Face's EER do not significantly alter for color images in constrained conditions. At low FAR, counterintuitively, both perform better for more unconstrained datasets. This may represent an overfitting to wild scenarios, since it diverges from what is expected from a face recognition system. Unlike OpenFace and VGG-Face, for Wu-C we can see that the more controlled the dataset, the higher the performance in terms of EER (see Figure 3(a)). Wu-C also presents the greater generalization power, with the best results for both NIR and depth modalities. This CNN's behavior resembles what we presume to be the human performance on this task: it fares better on constrained NIR face images than on the unconstrained LFW color images (even if color is the modality for which it was trained), and it is worse for depth images than for color images, constrained or not. For these reasons, we only use Wu-C in the following experiment.

In the second experiment, for each normalized image in CASIA and FRGC, we created two images by mirroring the left side (see Figure 4(a)) and right side (see Figure 4(b)) of the original image. We then compared these new images in three settings. The first setting compares them to frontal images (Frontal x Side), which simulates having a frontal gallery image and a profile probe image frontalized. This scenario is common in applications like authentication in unconstrained conditions. The second setting uses mirrored images of the same side (Side x Side) and the third setting images from opposite sides (Side x Other Side) in their comparisons, which is the worst case scenario of face recognition on the wild after frontalization, due to asymmetries in human faces. For these comparisons we also computed the ROC curve and compared them to previously obtained LFW, FRGC and CASIA results.

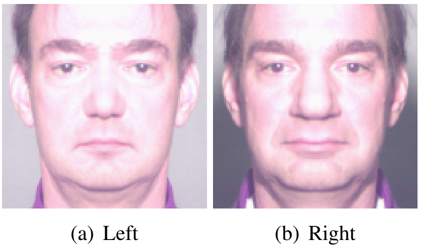

(a) Left (b) Right

Fig. 4. Mirrored images using left and right sides of the face.

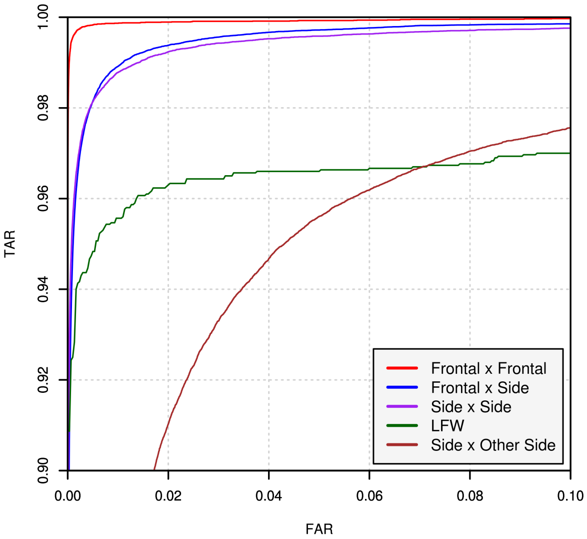


Fig. 5. Wu-C's ROC curves for FRGC.

Figures 5 and 6 show the results of the second experiment. In both cases, the original Frontal x Frontal experiment outperformed other results, but in less critical scenarios (Frontal x Side and Side x Side) the EER increase less than 1% for FRGC and 0.5% for CASIA. When considering Side x Other Side, however, this increment raises to nearly 5% in both datasets, being outperformed by LFW results. From these results, we can see that, despite the high performance achieved when comparing faces without pose correction, it can be improved by comparing a frontal gallery face with a profile probe or comparing two profile faces of the same side. This claim holds even in a dataset that does not feature strict pose constraint, like CASIA, on which mirroring half of the face does not correspond to perfect frontalization. These results present a behavior previously reported for humans recognizing faces [22], that is, recognition is consistently better for frontal (even if partial) images.

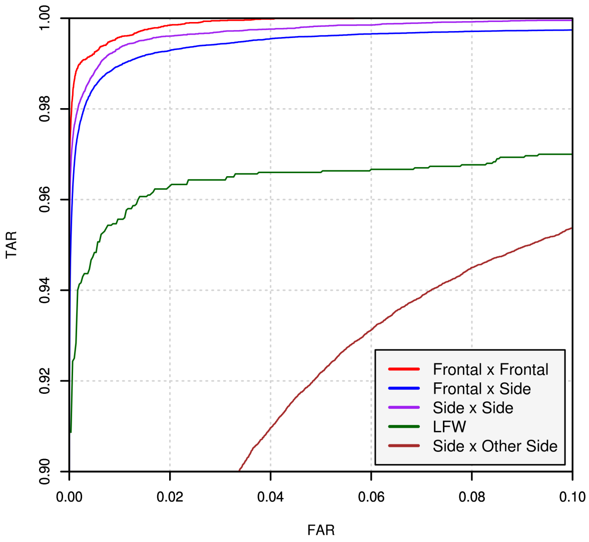


Fig. 6. Wu-C's ROC curves for CASIA.

4. CONCLUSIONS

In this paper, we evaluated three CNN descriptors and their performance describing color, NIR and depth face images over the LFW, CASIA and FRGC datasets. The best descriptor, Wu-C, can achieve better results on controlled scenarios and is able to describe NIR and depth faces, even if not trained to do so. We show that combining frontalization methods with the Wu-C descriptor, even for profile face images as probes, yields better results than using traditional 2D normalization.

These experimental results show the potential of using depth information for pose normalization and combining both NIR and depth with color to improve the current state-of-the-art face recognition methods. We intend to follow this work with an in-depth study investigating these results.

Our results also present interesting questions which we plan to investigate. One of them is why VGG-Face and OpenFace did not perform better, as expected, on more constrained scenarios. The other is correlating the perceived human behavior of Wu-C not only to its better results on controlled acquisition conditions, but also to human performance on other face recognition modalities.

5. REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1701–1708.

[2] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

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

[4] G. Medioni and R. Waupotitsch, "Face modeling and recognition in 3-d," in *2003 IEEE International SOI Conference. Proceedings (Cat. No.03CH37443)*, Oct 2003, pp. 232–233.

[5] Stan Z Li, RuFeng Chu, ShengCai Liao, and Lun Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 4, pp. 627–639, 2007.

[6] Xiaoxiang Liu, Lingxiao Song, X. Wu, and T. Tan, "Transferring deep representation for nir-vis heterogeneous face recognition," in *2016 International Conference on Biometrics (ICB)*, June 2016, pp. 1–8.

[7] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, "A light cnn for deep face representation with noisy labels," *arXiv preprint arXiv:1511.02683*, 2015.

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

[9] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," Tech. Rep., CMU-CS-16-118, CMU School of Computer Science, 2016.

[10] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[11] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[12] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *ECCV*, 2014, pp. 720–735.

[13] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in tv video," *Image Vision Computing*, vol. 27, no. 5, pp. 545–559, 2009.

[14] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *CVPR*, 2013, pp. 3476–3483.

[15] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb 1992.

[16] Felix Juefei-Xu and Marios Savvides, "An augmented linear discriminant analysis approach for identifying identical twins with the aid of facial asymmetry features," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Washington, DC, USA, 2013, CVPRW '13, pp. 56–63, IEEE Computer Society.

[17] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[18] Hong-Wei Ng and Stefan Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 343–347.

[19] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *14th European Conference on Computer Vision*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., 2016, pp. 87–102.

[20] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek, "Overview of the face recognition grand challenge," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, June 2005, vol. 1, pp. 947–954 vol. 1.

[21] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 348–353.

[22] G. Hole and V. Bourne, *Face Processing: Psychological, Neuropsychological, and Applied Perspectives*, Oxford, 1 edition, 2010.