CNN Transfer Learning for Robust Face Recognition in NAO Humanoid Robot

D. Bussey[1], A. Glandon[2], L. Vidyaratne[2], M. Alam[2] and K. M. Iftekharuddin[2]
[1] is with Department of Electrical Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114 and [2] are with the Vision Lab in Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA 23529 (email: busseyd@my.erau.edu, {aglan001, lvidy001, malam001, kiftekha}@odu.edu)

***Abstract*—Application of transfer learning for convolutional neural networks (CNNs) has shown to be an efficient alternative for solving recognition tasks rather than designing and training a new neural network from scratch. However, there exists several popular CNN architectures available for various recognition tasks. Therefore, choosing an appropriate network for a specific recognition task, specifically designed for a humanoid robotic platform, is often challenging. This study evaluates the performance of two well-known CNN architectures; AlexNet, and VGG-Face for a face recognition task. This is accomplished by applying the transfer learning concept to the networks pre-trained for different recognition tasks. The proposed face recognition framework is then implemented on a humanoid robot known as NAO to demonstrate the practicality and flexibility of the algorithm. The results suggest that the proposed pipeline shows excellent performance in recognizing a new person from a single example image under varying distance and resolution conditions usually applicable to a mobile humanoid robotic platform.**

Keywords*—*Transfer Learning, Convolutional Neural Network (CNN), Facial Recognition, Humanoid Robot.

I. INTRODUCTION

In recent years, transfer learning has shown to be an effective method in solving machine learning tasks. The idea of modifying an existing solution to a specific problem to create a new solution for a related problem can reduce the amount of time and effort to solve machine learning based problems. One particular area of study where transfer learning can be applied is retraining convolutional neural networks (CNNs) to accomplish recognition tasks. One of the most desirable, yet complex recognition task is face recognition. Conditions such as image angle and brightness can pose problems for any face recognition algorithm. Most face recognition frameworks also require multiple images of a single person to yield accurate recognition results.

The architecture of CNNs is inspired by biological processes [1] as they are designed to mimic the vision system of humans. Because of this complex architecture and the connectivity pattern of neurons within the neural network, CNNs have shown to perform well in complex recognition tasks such as face recognition [2], text recognition [3], and natural language processing [4]. CNNs excel in the mentioned recognition tasks due to their ability to learn features from images rather than hand-crafted, as in the case of most classical recognition algorithms. However, one of the major drawbacks of CNNs is the often prohibitive amount of time it takes to be properly trained for a task. High performance hardware such as Graphics Processing Units (GPUs) are essential to complete the training process in a reasonable amount of time due to the complex computational nature of the training process [5].

Several studies have been conducted using CNNs for face recognition task that show excellent accuracy such as DeepFace [6], FaceNet [7], and DeepID3 [2]. DeepFace contains 120 million parameters and 8 layers. The network is trained on a portion of the Social Face Classification (SFC) [6] dataset. DeepFace shows professional levels of performance achieving a true positive rate of around 97% and a false positive rate of less than 2% [6]. FaceNet utilizes a more complex architecture which consists of 22 layers and 140 million parameters. The network achieves approximately 99.6% recognition accuracy when tested on the Labeled Face in the Wild (LFW) [8] dataset. DeepID3 consists of 14 layers and is trained on a combination of the CelebFaces+ [9] and WDRef [10] datasets. When tested on the LFW dataset, DeepID3 achieves close to 99.5% accuracy. Even though these methods successfully handle face recognition tasks, they typically require multiple images of a single person in the database for comparison in order to achieve accurate results in a real world setting.

This paper analyzes the performance of two existing CNN architectures popularly known as AlexNet [5] and VGG-Face [11] to perform face recognition tasks. In this study transfer learning is applied to AlexNet while VGG-Face remains unaltered. The face recognition framework utilized in this study requires only one example image per person to achieve accurate face recognition. More importantly, we utilize a humanoid robotic platform known as NAO to evaluate the face recognition performance of the above mentioned CNNs in a practical environment. NAO's low resolution camera and a separate high-resolution camera are utilized to obtain the experimental results. Our results suggest that VGG-Face achieves better face recognition performance for both low resolution and high-resolution case compared to the AlexNet with transfer learning.

This paper is organized as follows. Section II provides a background on transfer learning, the architecture of AlexNet and VGG-Face and the humanoid robotic platform NAO. Section III provides our transfer learning approach for the CNNs, the distance based algorithm for classifying the faces and the detailed pipeline of our proposed CNN based face recognition technique using NAO. The experimental results and related discussions are provided in Section IV. Finally, section V concludes the paper.

II. BACKGROUND REVIEW

A. Transfer Learning

The idea behind transfer learning is taking the solution of one problem and modifying that solution to solve a related problem, saving time and effort [12]. Transfer learning is typically accomplished by retraining a CNN on a large labeled dataset [13]. The first layers of a CNN typically extract the basic features of an image. The basic feature extraction can be applied to any generic image; it is not until the final layers of the CNN until the highly abstract features of an image are extracted [14]. Because of the shared features in the initial layers of the CNN, fine-tuning an optimally-trained CNN on a new dataset can result in high level of accuracy with less effort [13].

Torrey and Shavlik [15] discuss how the relatedness between the source task and the target task has a correlation with the performance of transfer learning. They claim that as the relatedness between the source and target tasks increase, transfer learning becomes a more and more viable solution. Transfer learning is currently being used for complex recognition tasks in several new fields of research. A recent study shows transfer learning to be used for classifying non-Gaussian noises in LIGO detectors [13]. Transfer learning may not yield the best possible results for recognition tasks, but it is a favorable alternative to developing and training a new neural network entirely. This study utilizes the transfer learning concept for retraining the AlexNet using a face dataset.

B. AlexNet

AlexNet is a CNN designed for object recognition and is trained on a subset of the ImageNet [16] database. ImageNet contains millions of images consisting of several thousand categories and is used in the ImageNet Large Scale Visual Recognition Challenge. AlexNet contains 11 different layers containing 5 convolution layers, 3 max pooling layers, 3 fully connected layers and a Softmax layer. The input layer of AlexNet expects a resolution of 227x227. Table I shows the detailed architecture of the AlexNet including filter numbers and input size at each layer.

TABLE I. THE ARCHITECTURE OF ALEXNET

Layer #	**Layer Type**	**# filters**	**Input**
1	Conv.	96	11x11x3
2	Max-Pool	-	3x3
3	Conv.	256	5x5x48
4	Max-Pool	-	3x3
5	Conv.	384	3x3x256
6	Conv.	384	3x3x192
7	Conv.	256	3x3x192
8	Max-Pool	-	3x3
9	FC	-	4096
10	FC	-	4096
11	FC	-	1000
12	SoftMax	-	-

C. VGG-Face

VGG-Face is a CNN based off the VGG-16 architecture and is trained on a face data set acquired by the Visual Geometry Group that consists of over 2.5 million images and 2622 unique identities [11]. VGG-face contains 22 layers that consist of 13 convolutional layers, 5 max pooling layers, 3 fully connected layers and a Softmax layer. The expected image resolution of the input layer is 224x224. Table II shows the detailed architecture of the VGG-face including filter numbers and input size at each layer.

TABLE II. THE ARCHITECTURE OF VGG-FACE

Layer #	Layer Type	# filters	Input
1	Conv.	64	3x3x3
2	Conv.	64	3x3x64
3	Max-Pool	-	2x2
4	Conv.	128	3x3x64
5	Conv.	128	3x3x128
6	Max-Pool	-	2x2
7	Conv.	256	3x3x128
8	Conv.	256	3x3x256
9	Conv.	256	3x3x256
10	Max-Pool	-	2x2
11	Conv.	512	3x3x256
12	Conv.	512	3x3x512
13	Conv.	512	3x3x512
14	Max-Pool	-	-
15	Conv.	512	3x3x512
16	Conv.	512	3x3x512
17	Conv.	512	3x3x512
18	Max-Pool	-	2x2
19	FC	-	4096
20	FC	-	4096
21	FC	-	1000
22	Softmax	-	-

D. NAO Robot

NAO is a fully programmable humanoid robot developed by Aldebaran. NAO is equipped with an Intel Atom Z530 and multiple sensors to analyze the environment. For this study, the most relevant feature of NAO is the integrated camera. NAO's camera is capable of taking photos at a resolution of 640x480 and recording video at 30 frames per second. NAO is used in this study to demonstrate how the facial recognition frameworks using AlexNet and VGG-Face can be implemented into a practical environment to simulate real world situations.

III. METHODS AND SETUP PARAMETERS

A. Transfer learning for AlexNet

To apply transfer learning to AlexNet, the CNN is retrained on the CASIA-WebFace [17] database. This database includes just under half a million images of celebrity faces in a total of 10575 unique identities. This dataset is chosen for transfer learning because it is the largest publicly available face dataset. The images in the data set are initially resized to match the required input layers of the CNN.

AlexNet is trained using the stochastic gradient descent (SGD) with momentum with an initial learning rate of 0.001. However, for this study, VGG-Face remains trained on the data set mentioned in section II. Note that the VGG-Face is already trained for the face recognition task and may not require fine-tuning in the first place. After the CNNs have been retrained they are used to extract features from images for facial recognition. First, the Viola-Jones algorithm [18] is used to crop out the face in the image. The image of the cropped face is then resized to the appropriate resolution for the CNNs to extract the features.

B. Feature Classification

Feature classification task is performed by computing the distance between the features of two images. For each input image, features are extracted by the CNNs. Both AlexNet and VGG-Face extract 4096 features for one image. Once the features of an image is extracted, the recognition task begins.

To determine whether a face match exists depends on a measure of feature distance between the features of the input image and the images in the database of people. The feature distance measure is obtained by computing the Euclidian distance between feature vectors, as shown in (1)

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \; ; \qquad (1)$$

where p_i is the i^{th} feature of the input image, q_i is the ith feature of the image in a person's database folder and d indicates the Euclidian distance. This technique is chosen for distance calculation since it requires low computation that results in a quick run time while still being a good method for feature comparison. The distance between the features of the input image and the images in the database are then saved to a new vector for analysis purposes. By taking the difference between the minimum value and the second minimum value of the new vector one can determine if a match is detected. If the difference between the two minimums is less than 15 percent of the average feature distance between the input image and everyone in the database, the system determines that the person shown in the input image does not exist in the database, otherwise the system detects a match. The 15% threshold used is determined experimentally during the development phase of the face recognition framework as it seems to yield the best results. Our experiments show that the above procedure only require one image per person in the database for the facial recognition to work with sufficient accuracy in a practical environment.

C. Incorporating NAO

NAO robot is utilized in this study to demonstrate the feasibility of implementing the proposed facial recognition framework in a practical environment. NAO's role in this experiment is to capture images and to provide instructions for the user on how to proceed. The experiment begins telling the user to take a seat and provides a countdown warning to indicate the user when the picture is about to be taken. After taking the picture, NAO notifies the user if the face is not able to be cropped out of the original photo and prompts the user to take another photo. After the photo has been processed, NAO uses its speech functions to either greet the user by name if a match is detected or prompt the user to enter either his or her name on the computer so he or she can be recognized in the future. The full pipeline is illustrated in Fig. 1.

D. Preparing the Testing Dataset

The dataset used for testing the CNN face recognition performance in this study is an in-house set of face images. The face images are obtained using NAO's camera and a high-resolution camera of 3264x2448. To create the in-house database, 9 different people are used and only one image of each person is obtained to store in the database for recognition purposes. To test the face recognition performance of AlexNet and VGG-Face, nine different images of each person are taken at different distances. Three images are taken from 2 feet, 4 feet, and 6 feet, respectively to get a better understanding of the proposed recognition framework when the distance of the photo is altered. The images of the cropped faces are resized to 227x227 for AlexNet and 224x224 for VGG-Face. The same process is then repeated using a high-resolution camera to analyze the effect of resolution on the performance for each CNN

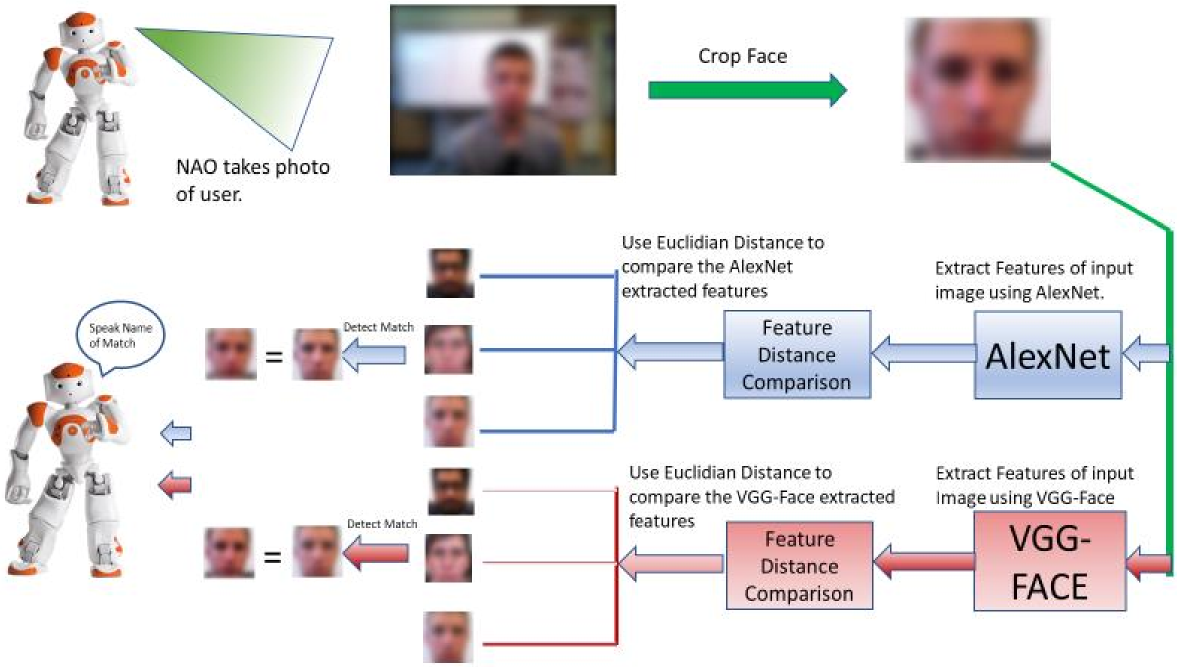
NAO takes photo of user.
Crop Face
Use Euclidian Distance to compare the AlexNet extracted features
Extract Features of input image using AlexNet.
Speak Name of Match
Detect Match
=
Feature Distance Comparison
AlexNet
Use Euclidian Distance to compare the VGG-Face extracted features
Extract Features of input Image using VGG-Face
Detect Match
=
Feature Distance Comparison
VGG-FACE

Fig 1. CNN based face recognition pipeline for NAO. The images are blurred due to possible privacy issues.

IV. RESULTS AND DISCUSSION

The performance of the facial recognition for each CNN is evaluated by determining the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The overall performance of each network's facial recognition abilities are determined by true positive rate (TPR), false positive rate (FPR) and an overall accuracy score.

The evaluation methods for these performance measures are shown below,

$$TPR = \frac{TP}{TP+FN}, \tag{2}$$

$$FPR = \frac{FP}{FP+TP}, \tag{3}$$

$$Accuracy = \frac{TP}{TP+FN+FP}. \tag{4}$$

For testing the recognition framework on low-resolution images, a single image of 9 different people is collected to establish the database of people. 9 images of each person are used in the testing dataset resulting in a total of 81 images. The same process is then repeated using a high-resolution camera to obtain another set of 81 images. The experimental results for the true positive rate is shown in Fig. 2.

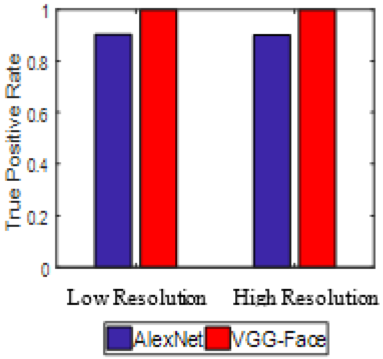

Fig 2: Performance comparison for both AlexNet (Blue) and VGG-Face (Red) using low-resolution (640x480) and high-resolution images (3264x2448), in terms of TPR.

As shown in Fig. 2, VGG-Face achieves a higher TPR when testing both high and low-resolution images. VGG-Face performed excellently demonstrating 100% recognition accuracy in the testing for both high and low-resolution images where AlexNet managed approximately 90%. The fact that neither network showed any performance difference for each resolution demonstrates the robustness of the models. The exact values for TPR and FPR are shown in Table III.

TABLE III. TPR AND FPR USING ALEXNET AND VGG-FACE FOR HIGH AND LOW RESOLUTION IMAGES

Network	**Image Resolution**	**TPR**	**FPR**
AlexNet	640x480	0.9012	0
Vgg-Face	640x480	1	0
AlexNet	3264x2448	0.9012	0
Vgg-Face	3264x2448	1	0

Table III shows that the resolution of the image had no impact on the performance of either CNN. Both CNNs also had a FPR of 0. For every instance where AlexNet failed to classify the correct person the result was a FN. A zero FPR is essential for security applications such as granting access to facilities, computers, and smartphones. The fact that both AlexNet and VGG-Face demonstrate a 0 FPR shows that either network is suitable for practical security applications.

The face recognition performance with respect to distance is also analyzed for each CNN. The results of the test are shown in Table IV and Table V. The accuracy calculated in the tables utilizes equation 4.

TABLE IV. FACE RECOGNITION ACCURACY OF ALEXNET AND VGG-FACE WITH RESPECT TO DISTANCE USING LOW RESOLUTION IMAGES.

Distance of Face from Camera	**Accuracy using AlexNet**	**Accuracy using VGG-Face**
2 Feet	91.67%	100%
4 Feet	91.67%	100%
6 Feet	87.50%	100%

Table IV shows that the distance from a low resolution camera does not have a significant impact on the performance of either CNN. AlexNet's performance remains the same when tested at 2 and 4 feet while the performance slightly dips at 6 feet. VGG-Face managed to have 100% match recognition accuracy at all distances. This is most likely due to the more complex nature of VGG-Face's model. However, the complexity of VGG-Face leads to a significantly longer run-time when extracting the features of an image compared to AlexNet. VGG-Face (240ms) typically takes approximately 8 times longer than AlexNet (30ms) to extract the features of an image. Despite the better performance of VGG-Face, this slower run-time may result in VGG-Face not being the best choice for certain applications where execution speed is a top priority. The face recognition accuracy with respect to distance using high resolution images is shown below.

TABLE V. FACE RECOGNITION ACCURACY OF ALEXNET AND VGG-FACE WITH RESPECT TO DISTANCE USING HIGH RESOLUTION IMAGES.

Distance of Face from Camera	**Accuracy using AlexNet**	**Accuracy using VGG-Face**
2 Feet	100%	100%
4 Feet	88.89%	100%
6 Feet	81.48%	100%

Table V shows that the even though performance of AlexNet at 2 feet is 100%, the performance decreases as distance from the high resolution camera increases. However, VGG-Face maintains 100% recognition accuracy at the three tested distances.

V. CONCLUSION

This study explores the use of transfer learning on the CNN known as AlexNet to accomplish a real-time face recognition task. The performance of this model is then compared to VGG-Face, a much deeper CNN that is specifically trained for face recognition. Overall, the proposed face recognition pipeline demonstrates that only one image per person is adquate to achieve a high level of accuracy. The study shows that VGG-Face yields better performance compared to AlexNet for face recognition tasks. However, VGG-Face yields a significantly longer execution time during feature extraction compared to AlexNet. The study also shows that the resolution of the image does not appear to have a significant impact on the performance. However, the results further suggest that the distance of the face from the camera has an impact on recognition accuracy when using AlexNet.

The study further implements the proposed face recognition framework on the humanoid robot NAO. This demonstrates the flexibility of the proposed method for utilization in a practical environment.

In the future, we plan to extend the proposed framework for solving more challenging recognition tasks multiple face recognition, and generalized object recognition. Further, we plan to perform extensive comparison between various deep learning models for recognition tasks.

VI. ACKNOWLEDGEMENT

This research is supported in part by NSF under grant CNS-1659795.

REFERENCES

[1] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks,* vol. 16, no. 5-6, pp. 555-559, 2003.

[2] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873,* 2015.

[3] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 3304-3308.

[4] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," presented at the Proceedings of the 25th international conference on Machine learning, Helsinki, Finland, 2008.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[6] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701-1708.

[7] 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.

[8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst2007.

[9] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891-1898.

[10] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Transactions on Information Forensics and Security,* vol. 9, no. 12, pp. 2144-2157, 2014.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *BMVC*, 2015, vol. 1, no. 3, p. 6.

[12] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," presented at the Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014.

[13] D. George, H. Shen, and E. Huerta, "Deep Transfer Learning: A new deep learning glitch classification method for advanced LIGO," *arXiv preprint arXiv:1706.07446,* 2017.

[14] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806-813.

[15] L. Torrey, T. Walker, J. Shavlik, and R. Maclin, "Using advice to transfer knowledge acquired in one reinforcement learning task to another," presented at the Proceedings of the 16th European conference on Machine Learning, Porto, Portugal, 2005.

[16] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision,* journal article vol. 115, no. 3, pp. 211-252, December 01 2015.

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

[18] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision,* vol. 57, no. 2, pp. 137-154, 2004.