AcFR: Active Face Recognition Using Convolutional Neural Networks

Masaki Nakada, Han Wang, Demetri Terzopoulos
University of California, Los Angeles
nakada@cs.ucla.edu, hannahoo@g.ucla.edu, dt@cs.ucla.edu

Abstract

We propose AcFR, an active face recognition system that employs a convolutional neural network and acts consistently with human behaviors in common face recognition scenarios. AcFR comprises two main components—a recognition module and a controller module. The recognition module uses a pre-trained VGG-Face CNN to extract facial image features, along with a nearest-neighbor identity recognition criterion. The controller module can make three different decisions based on the results—greet a recognized individual, disregard an unknown individual, or acquire a different viewpoint from which to reassess the subject, which are natural reactions when people observe passers-by. Evaluated on the CMU PIE face database, our recognition module yields higher accuracy on images acquired at angles more similar to those saved in memory. The view-dependent accuracy provides evidence for the proper design of the controller module.

1. Introduction

Face recognition is a classic computer vision problem. A big challenge is improving recognition accuracy given limited information. In this context, the active perception approach in computer vision [3, 1, 4, 14] seems promising. For example, when observing the profile of someone who looks like a friend, before presuming to greet the individual, one might move to a better viewpoint from which to see more of the face of interest in order to ascertain the person's identity. The idea of viewpoint-dependent recognition has been studied in neurophysiology [2], and Wang et al. [15] found face-selective neurons tuned to specific viewpoints. These considerations have motivated us to develop an active face recognition (AcFR) system that models human behaviors in common face recognition scenarios. Unlike most existing face recognition systems, it does not rely solely on a single input image; instead, like people in real-world scenarios, it acquires additional images from different viewpoints to improve recognition accuracy.

AcFR comprises two main components, a face recognition module and a behavior controller module. The face recognition module, which employs VGG-Face, a popular convolutional neural network (CNN), in conjunction with a nearest-neighbor identity recognition criterion, evaluates the input image and provides the data needed to make decisions. The controller module uses these data to drive its follow-up behavior, including whether to greet the (known) observed individual, disregard the (unknown) individual, or obtain a different viewpoint from which to reassess the subject.

The main contributions of this study are twofold: First, we propose the AcFR approach, which actively recognizes faces by mimicking human behaviors. Second, the use of a CNN makes the visual processing more biologically plausible, which is relevant in future applications of AcFR to biomechanical virtual human models.

The remainder of the paper is organized as follows: In Section 2, we briefly review recent related work. Section 3 describes the methodology of AcFR in detail and Section 4 demonstrates its effectiveness through experiments. Finally, in Section 5, we present our conclusions and discuss future work.

2. Related Work

Our work is relevant to human modeling and simulation. Early efforts on this topic focused on human face and body modeling. Over time, technologies from computer graphics, computer vision, machine learning, psychology, etc., have been converging on this area. Some recent work, including ours, considers brain functionalities such as visual recognition, learning, and communication abilities.

With regard to recognition, much work has been done on object recognition, especially face recognition. Learning-based face recognition has recently made major strides. Traditional approaches represent faces using hand-crafted features extracted from the image, such as SIFT, LBP, and HOG [12, 16, 7], which are then employed in the classification step. So-called deep learning approaches employ artificial neural networks that learn appropriate features automatically via training on massive quantities of image data. Among neural networks, CNNs are preferred by the com-

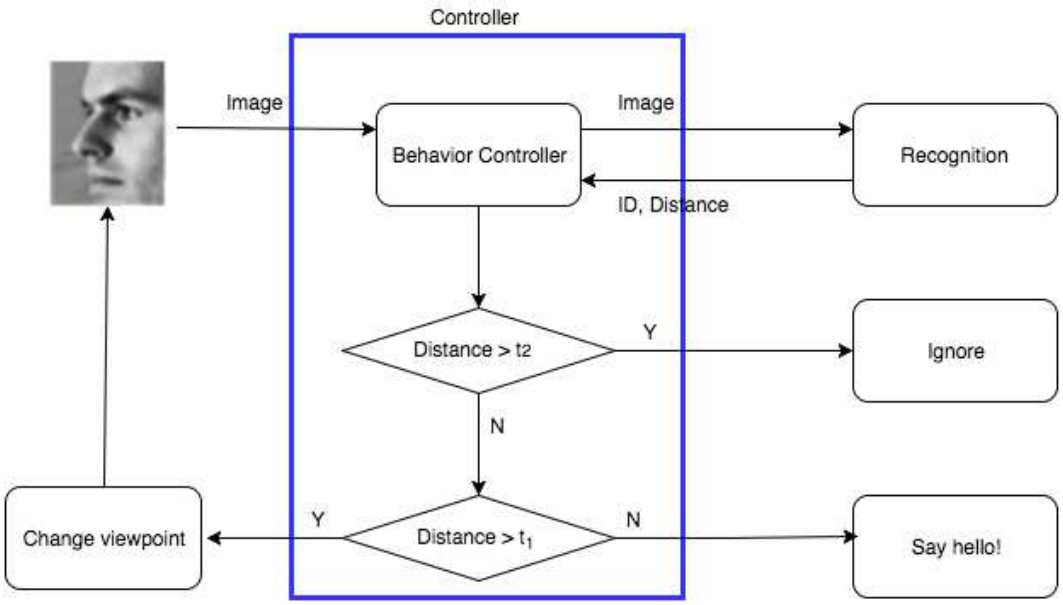

Figure 1. System architecture

puter vision community in part because the mechanisms underlying their architectural design are suggestive of cortical mechanisms in biological vision systems.

Although CNNs have been applied to face recognition as far back as 1997 [6], the recent availability of massive image datasets have revealed their power. A representative work of this class of approaches is Deep-Face [13], which uses a 9-layer CNN trained on 4.4 million labeled facial images including over 4,000 identities. It has achieved outstanding performance in both the Labeled Faces in the Wild (LFW) [7] and YouTube Faces (YTF) [16] benchmarks. Subsequently, Parkhi et al. [10] proposed the VGG-Face network, which we have adopted in our work. It uses a 16-layer CNN trained on 2.6 million images and achieves even better accuracy in these benchmarks.

3. Methodology

As illustrated in Fig. 1, our AcFR system consists of two main modules—a behavior controller module, which is served by a face recognition module. Given a facial image, the recognition module will attempt to determine the subject's identity and will provide its results to the controller. The controller has two functionalities: to re-evaluate the face whenever the viewpoint changes and to model human behaviors based on the results of recognition.

3.1. Face Recognition

The modern face recognition pipeline usually consists of four stages: detection, alignment, representation, and classification. Detection and alignment are often included as preprocessing steps. Given a good facial representation, the system can predict identity through classification algorithms.

3.1.1 Preprocessing

Face detection and alignment algorithms are often employed because many face recognition algorithms require the input images to be carefully positioned into a canonical pose. Sometimes the assumption is made that the detection step has provided rough alignment. In this project, faces are detected using the algorithm by Mathias et al. [8].

3.1.2 Face Representation

Face representation heavily influences the performance of the face recognition system and is also the focus of current recognition-related research. In this project, we employ the VGG-Face network (Fig. 2), a 16-layer CNN that is trained on over 2 million celebrity images. In addition to its outstanding performance in benchmarks, we chose it over other CNNs because (i) the image dataset that VGG-Face used is similar to ours, which makes its performance more reliable in our application, and (ii) the pre-trained VGG-Face model is available in the Caffe [5] Zoo Library, which makes it easy to use. Using VGG-Face, we are able to extract suitable image features from the output of the fc-6 layer and use them in our subsequent classification stage. In this way, each 224×224 image is represented by a 4,096-dimensional feature vector.

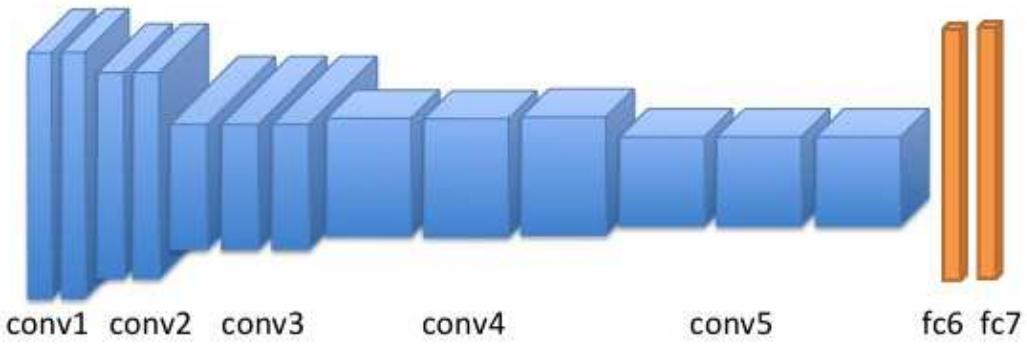

Figure 2. VGG-Face network architecture

3.1.3 Classification

Given the extracted features, we experimented with various classification algorithms. With half the dataset used for training, Support Vector Machines and Linear Regression yielded poor accuracy (below 20%), whereas K-Nearest-Neighbor (KNN) classification achieved 90% accuracy with $K = 30$. We determined that the performance gap is attributable to the "curse of dimensionality". Furthermore, maintaining only a single frontal image per person in the gallery G results in an improvement over KNN. Accordingly, we decided to use the Nearest-Neighbor (NN) classifier. Given the feature vector θ associated with an unknown image, NN will compute its Euclidean distance from each of the feature vectors θ_i stored in the gallery G and output the identification of the image as

$$\text{ID} = \arg\min_{\theta_i \in G} \parallel \theta - \theta_i \parallel . \tag{1}$$

We also used Euclidean distance in the subsequent behavioral model.

3.2. Behavior Modeling

In active face recognition scenarios, human behaviors can be roughly categorized into three types: change viewpoint, greet, and ignore. If a person of interest is likely to be someone we know, but we are uncertain, we will seek a better position from which to observe the subject's face until we can confidently recognize the subject or relegate the subject a stranger, at which point we would choose to greet or ignore the subject, respectively. Hence, our controller module is designed to model such behaviors based on the results of facial image recognition.

The controller module is initialized with two distance thresholds, t_1 and t_2. Given the output of the recognition module, the controller can model the aforementioned behaviors in the following simple way:

$$\text{Behavior} = \begin{cases} \textit{Greet} & \text{if } d \leq t_1, \\ \textit{Ignore} & \text{if } d \geq t_2, \\ \textit{ChangeView} & \text{if } t_1 < d < t_2, \end{cases} \tag{2}$$

where $d = \min_{\theta_i \in G} \parallel \theta - \theta_i \parallel$. In the first and second cases, our system is confident that the subject is a friend or a stranger, respectively, while in the third case it must acquire more information via a change in viewpoint.

4. Experiments

4.1. Experimental Setup

We used Caffe [5] to implement our recognition module. It is one of most popular neural network frameworks and is widely used by many large-scale computer vision applications. Additionally, Caffe provides various popular pretrained neural networks in its Model Zoo. Considering the high computational costs of CNNs and the complexity of installing Caffe, we decided to develop this project on Elastic Compute Cloud (EC2), which provides a scalable computing capability on Amazon Web Service (AWS), and we used the g2.2xlarge GPU instance, which has Caffe already installed, that takes advantage of the high-performance parallel processing capabilities of NVIDIA GPUs.

4.2. Dataset

The PIE facial image database used in this project was collected by Carnegie Mellon University's Robotics Institute [11]. It contains 41,368 images of 68 people taken under 43 different illumination conditions, 4 different facial expressions, and 13 imaging viewpoints (from 13 cameras) ranging from -90° to 90°. Among the 13 cameras, 9 are positioned at head height in an arc from approximately a full left profile to a full right profile. As shown in Fig. 3, images acquired from these 9 viewpoints may be used to model a sequence of faces seen by an active observer.

Figure 3. A sequence of 9 images of the same person

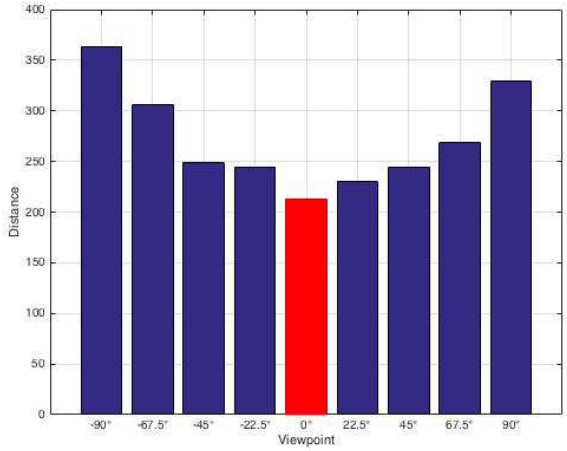

Figure 4. Distance for different viewpoints

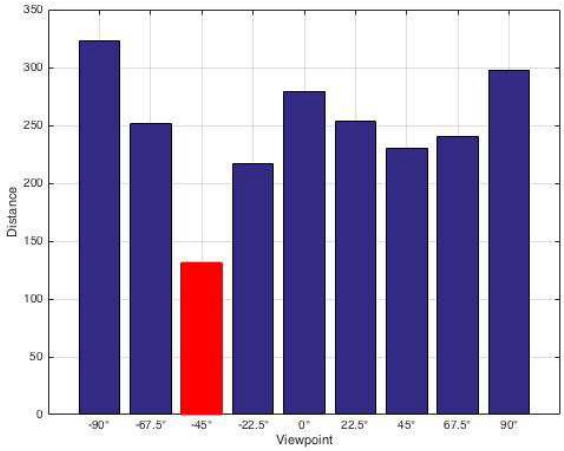

Figure 6. Distance for different viewpoints for side-view gallery

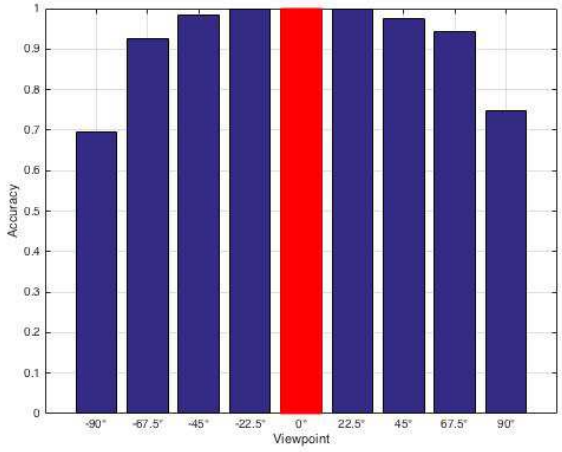

Figure 5. Accuracy for different viewpoints

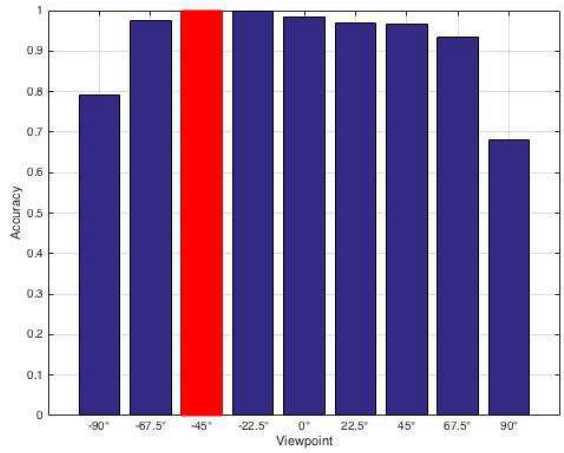

Figure 7. Accuracy for different viewpoints for side-view gallery

4.3. Face Recognition Module Performance

To evaluate the performance of our recognition module, we reserved one frontal image of each person, saved its VGG-Face feature vector in the gallery and used the remaining images for testing. Fig. 4 and Fig. 5 show that the accuracy and distance are view-dependent, as expected. For images with angles closer to the frontal view, the accuracy can reach 100% and the distance is also minimized. This shows that views can be represented by the features extracted by VGG-Face and also provides evidence for the view-dependent design of our behavior model.

To investigate further the influence of the gallery on the results, we changed the stored feature vectors in the gallery from frontal views to -45° side views. Fig. 6 and Fig. 7 show that the optimal view (the view that results in the highest accuracy and minimal distance) changes accordingly. Also, due to the similarity between the left and right side views, a distance local minimum occurs at 45° and the accuracy remains roughly the same for angles near -45° and 45°. This is reasonable because the images most familiar

to the active observer are the images (feature vectors) in memory—the gallery.

To test how our AcFR system works when observing strangers, we removed 10 people randomly from the gallery to relegate them strangers. For these strangers, the average distances under different views range from 286 to 350, which are close to the distance of the worst views at -90° and 90°. This shows that our system is able to distinguish strangers from friends. Thus, for properly set thresholds in the behavior controller module, it exhibits appropriate behavioral modeling.

4.4. Behavior Controller Module Performance

For each of the 68 subjects in the PIE database, we choose a sequence of images, one per viewpoint angle, under one specific illumination condition (Fig. 3), albeit sometimes with different expressions. This is because viewpoint changes usually happen quickly, which means that illumination is likely to remain unchanged, although expression may change. Then our AcFR system plays the role of an active observer with the feature vectors associated with the frontal images of all 68 subjects in memory. We tested the active observer's first reaction when acquiring images from different viewpoint angles.

Modifying the thresholds in our behavior model changes the "personality" of our AcFR system. As expected, the system greets the subject more often for larger t_1, and ignores the subject more often for smaller t_2. By default, we set the thresholds to $t_1 = 250$ and $t_2 = 325$.

Fig. 8 shows the behaviors when the initial observation is at different angles. Obviously, the frontal view is the best from which the observer can immediately recognize the subject. By contrast, the full left profile and full right profile are the most difficult views to recognize and they sometimes even lead to incorrect results. This mimics human active face recognition in the real world.

4.5. Time Complexity

Computational efficiency is an important concern for us because large latencies undermine real-time behavioral modeling. The initialization of our AcFR system takes approximately 2.2 seconds and the recognition of each image takes only about 0.067 seconds. This implies that our system is capable of performing real-time recognition, which makes it suitable for virtual human modeling applications.

5. Summary, Limitations, and Future Work

We have proposed an approach to simulating human behaviors in facial recognition. Motivated by real-life face recognition scenarios and related psychological findings, we assumed that the recognition strategy should be active; therefore, the controller module in our prototype Active

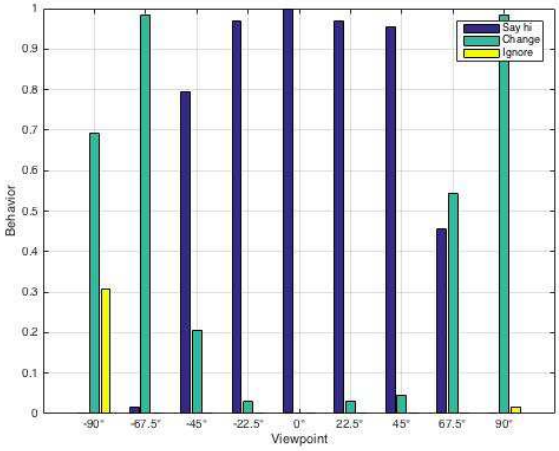

Figure 8. Behaviors for different initial viewpoints

Face Recognition (AcFR) system was designed to perform facial recognition in a view-driven sequential manner. Our use of a convolutional neural network makes the recognition module of the AcFR system more biomimetic and more powerful compared to alternative approaches. The experimental results support our design decisions.

The direction of movement when our AcFR system decides to change its viewpoint was not carefully investigated. This can be a problem because the system may decide to move from a side view to look at the subject from behind, whereas people normally move towards the front of a subject in order to see the face more clearly. Therefore, the direction of the face must be estimated for more realistic active face recognition.

In future work, we will incorporate our AcFR system into a biomechanical human model of the face-head-neck complex [9]. To this end, the behavioral repertoire of the system's controller module will need to be expanded.

References

[1] J. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988. 1

[2] T. J. Andrews and M. P. Ewbank. Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *Neuroimage*, 23(3):905–913, 2004. 1

[3] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. 1

[4] D. H. Ballard. Animate vision. *Artificial Intelligence*, 48(1):57–86, 1991. 1

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe:

Convolutional architecture for fast feature embedding. In *Proc. ACM International Conference on Multimedia*, pages 675–678, 2014. 2, 3

[6] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997. 2

[7] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with GaussianFace. *arXiv Preprint arXiv:1404.3840*, 2014. 1, 2

[8] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *Proc. European Conference on Computer Vision (ECCV)*, pages 720–735, 2014. 2

[9] M. Nakada and D. Terzopoulos. Deep learning of neuromuscular control for biomechanical human animation. In *Proc. International Symposium on Visual Computing*, pages 339–348, December 2015. 5

[10] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. British Machine Vision Conference (BMVC)*, pages 41.1–41.12, September 2015. 2

[11] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003. 3

[12] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: Video shot retrieval for face sets. In *Proc. International Conference on Image and Video Retrieval*, pages 226–236, 2005. 1

[13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014. 2

[14] D. Terzopoulos and T. F. Rabie. Animat vision: Active vision in artificial animals. *Videre: Journal of Computer Vision Research*, 1(1):2–19, September 1997. 1

[15] G. Wang, M. Tanifuji, and K. Tanaka. Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience Research*, 32(1):33–46, 1998. 1

[16] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 529–534, 2011. 1, 2