DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman Ming Yang Marc'Aurelio Ranzato Lior Wolf

Facebook AI Research
Menlo Park, CA, USA
{yaniv, mingyang, ranzato}@fb.com

Tel Aviv University
Tel Aviv, Israel
wolf@cs.tau.ac.il

Abstract

*In modern face recognition, the conventional pipeline consists of four stages: detect ⇒ align ⇒ represent ⇒ classify. We revisit both the alignment step and the representation step by employing explicit 3D face modeling in order to apply a piecewise affine transformation, and derive a face representation from a nine-layer deep neural network. This deep network involves more than 120 million parameters using several locally connected layers without weight sharing, rather than the standard convolutional layers. Thus we trained it on the largest facial dataset to-date, an identity labeled dataset of four million facial images belonging to more than 4,000 identities. The learned representations coupling the accurate model-based alignment with the large facial database generalize remarkably well to faces in unconstrained environments, even with a simple classifier. Our method reaches an accuracy of 97.35% on the Labeled Faces in the Wild (*LFW*) dataset, reducing the error of the current state of the art by more than 27%, closely approaching human-level performance.*

1. Introduction

Face recognition in unconstrained images is at the forefront of the algorithmic perception revolution. The social and cultural implications of face recognition technologies are far reaching, yet the current performance gap in this domain between machines and the human visual system serves as a buffer from having to deal with these implications.

We present a system (*DeepFace*) that has closed the majority of the remaining gap in the most popular benchmark in unconstrained face recognition, and is now at the brink of human level accuracy. It is trained on a large dataset of faces acquired from a population vastly different than the one used to construct the evaluation benchmarks, and it is able to outperform existing systems with only very minimal adaptation. Moreover, the system produces an extremely compact face representation, in sheer contrast to the shift toward tens of thousands of appearance features in other recent systems [5, 7, 2].

The proposed system differs from the majority of contributions in the field in that it uses the deep learning (DL) framework [3, 21] in lieu of well engineered features. DL is especially suitable for dealing with large training sets, with many recent successes in diverse domains such as vision, speech and language modeling. Specifically with faces, the success of the learned net in capturing facial appearance in a robust manner is highly dependent on a very rapid 3D alignment step. The network architecture is based on the assumption that once the alignment is completed, the location of each facial region is fixed at the pixel level. It is therefore possible to learn from the raw pixel RGB values, without any need to apply several layers of convolutions as is done in many other networks [19, 21].

In summary, we make the following contributions : (i) The development of an effective deep neural net (DNN) architecture and learning method that leverage a very large labeled dataset of faces in order to obtain a face representation that generalizes well to other datasets; (ii) An effective facial alignment system based on explicit 3D modeling of faces; and (iii) Advance the state of the art significantly in (1) the Labeled Faces in the Wild benchmark (*LFW*) [18], reaching near human-performance; and (2) the YouTube Faces dataset (*YTF*) [30], decreasing the error rate there by more than 50%.

1.1. Related Work

Big data and deep learning In recent years, a large number of photos have been crawled by search engines, and uploaded to social networks, which include a variety of unconstrained material, such as objects, faces and scenes.

This large volume of data and the increase in computational resources have enabled the use of more powerful statistical models. These models have drastically improved the robustness of vision systems to several important variations, such as non-rigid deformations, clutter, occlusion and illumination, all problems that are at the core of many computer vision applications. While conventional machine