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

Report on the paper

Artem Komarichev

February 7, 2016

Outline

- New alignment technique
- New DNN architecture
- New large dataset with labeled face images
- Experiments and Results on benchmarks dataset LFW and YTF

Introduction

The goal of *face verification* is to define either two face images belong to the same person or not.

There exist other face recognition tasks such as: *face identification*, *face clustering* and *face detection*.

In their paper they are considering problem in unconstrained environment.

Benchmark datasets

1. LFW (Labeled Faces in the Wild) consists of 13,323 web photos of famous people. 5,749 unique identities. 6,000 face pairs.

2. YTF (Youtube Face Database) contains 3,425 youtube videos of 1,595 unique identities. They splitted all this videos on 5,000 video pairs.

New collected dataset

Their DeepFace model was trained on a new huge dataset extracted from Facebook photos, so called *Social Face Classification (SFC)* dataset.

SFC dataset contains **4.4M** labeled face images belonging to **4030** unique identities.

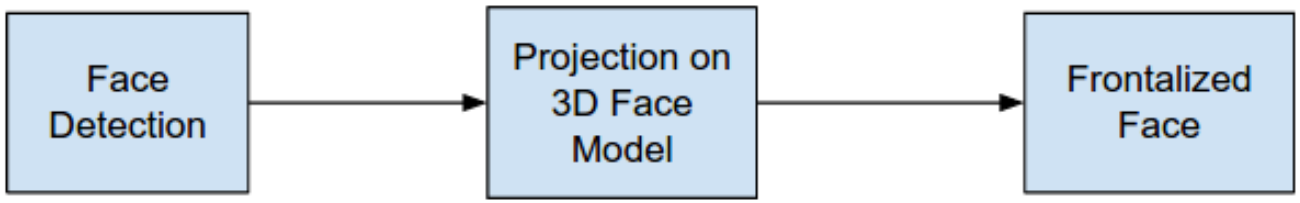
5% of this dataset was used for testing purposes.

Face alignment

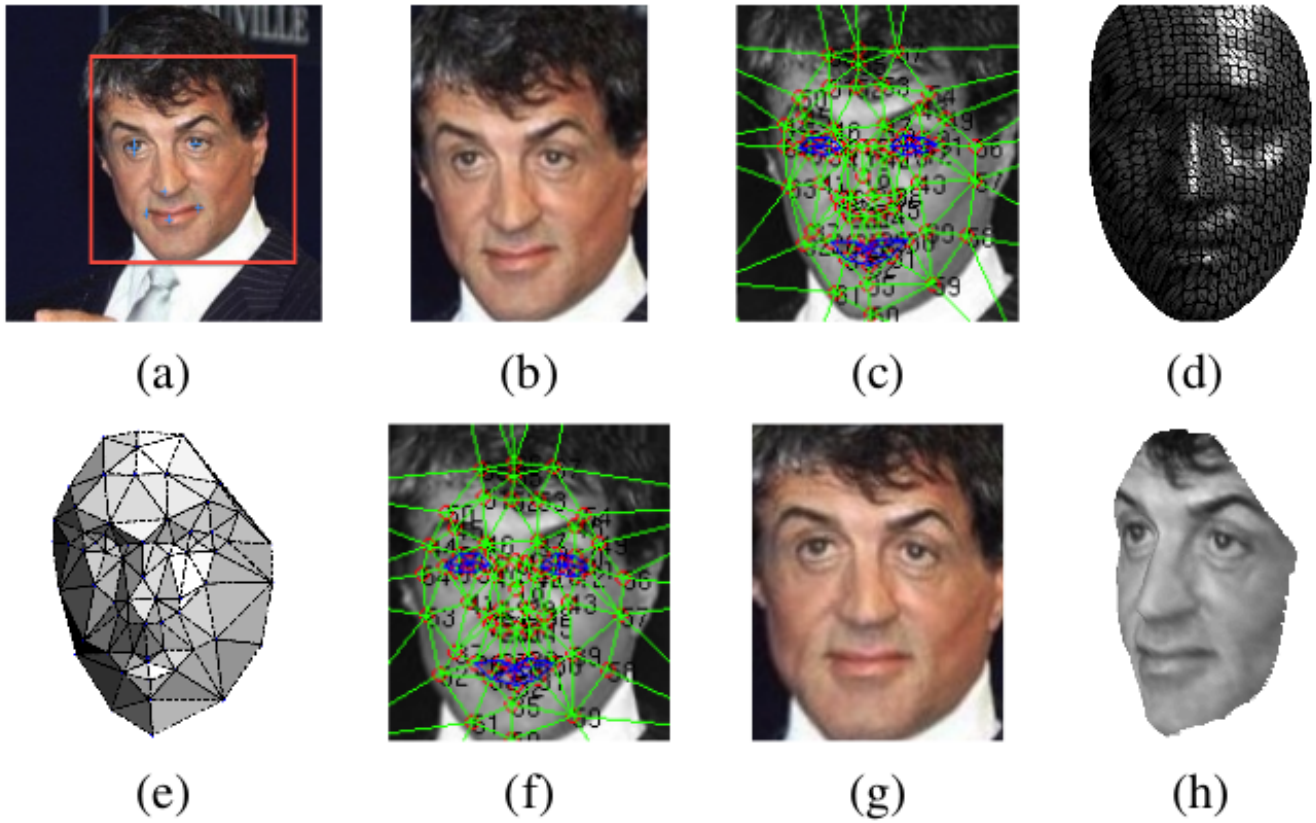
Alignment is still considered to be difficult problem to solve especially in the unconstrained environment.

In the recent few years 3D modeling was used extensively.

Frontalization procedure are illustrated below:

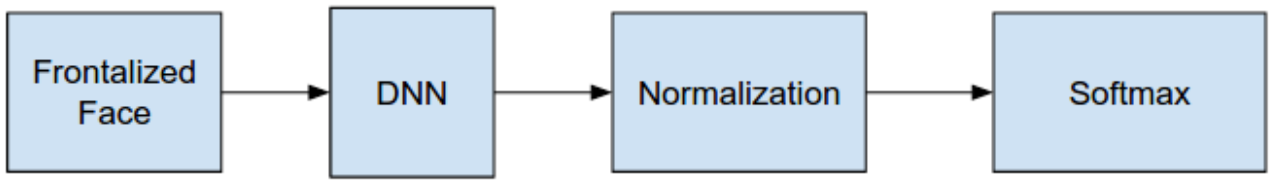


Face alignment (example from paper)

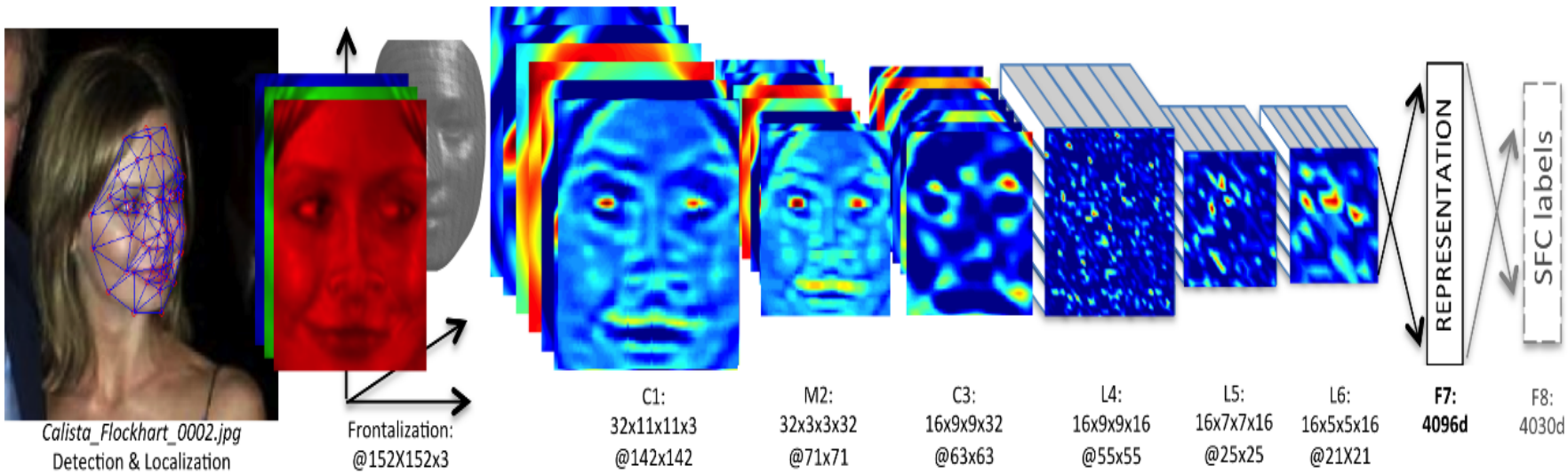


DeepFace Architecture

The big picture of DeepFace model is represented below:



DeepFace Architecture (DNN from paper)



DeepFace Architecture

As *loss function* was chosen cross-entropy loss:

$$L = \sum_{i}^{N} -log(p_i) \tag{1}$$

where p_i is a prediction that this image belongs to class i, and N is a number of classes.

To update parameters was used standard BackPropagation algorithm and Stochastic Gradient Descent (SGD). The size of mini-batch was equal to *128*.

Non-linear activation function: *rectifier linear units* - $max(0, x)$.

DeepFace Architecture

The number of parameters was close to **120M**, where *95%* was from LC and FC layers.

Sparse architecture: 75% of gradients were equal to zero in the last five layers. Sparsity was encorouged in the last time (dropout, maxout).

Training process took **three days** to run model on the whole SFC dataset for **15 epochs**.

Learning rate was decreased during the training procedure up to **0.0001** starting from **0.01**.

Verification Metrics

- Unsupervised: inner product of feature vectors
- Siammese network: $d(f_1, f_2) = \sum_i^N w_i |f_1^{(i)} - f_2^{(i)}|$,
  where N is a number of learned features,
  f_1 and f_2 - feature vectors for first face image and for second one, and
  w_i - trainable weights.
- Chi squared distance: $\chi^2(f_1, f_2) = \sum_i^N w_i \frac{(f_1^{(i)} - f_2^{(i)})^2}{(f_1^{(i)} - f_2^{(i)})}$

Experiments and Results

Models	Datasets	
	LFW	Youtube Faces DB
FaceNet	**99.63%**±0.15	**95.12%**±0.39
DeepFace	**97.35%**±0.25	**91.4%**±1.1
Parkhi's approach	**98.65%**	**97.3%**
DeepID2	...	...
DeepID2+	...	...
DeepID3	...	...
...	...	...

Table: State-of-the-art in face verification.

Experiments and Results (size of training dataset)

They provided several experiments with the size of the training dataset, different number of unique identities.

- When whole *4K* identities was used they got the highest accuracy. It is not suprising fact, because the more diverse our training dataset, the more better features we can learn.
- The origianl SFC was reduced up to 10% with the same 4K people showed twice less accuracy than 50% dataset.

Experiments and Results (different architectures)

Comparing different reduced architectures:

- without second conovolutional layer
- without first two locally connected layers
- without three these layers

with the original one, they showed that the deep of the architecture is a critical issue.

Deeper DNN can learn better feature representations.

Therefore, deeper architectures shows higher final accuracy than shallower ones does.

Conclusion

Main contributions:

- They achieved the new state-of-the-art on the LFW and YTF datasets by the time when the paper was published
- They proposed a new alignment techniques based on 3D modeling of the face images in the unconstrained settings
- They came up with a new DNN architecture which was trained on their own huge dataset