Web-Scale Training for Face Identification

1 Ming Yang, 1 Marc’Aurelio Ranzato, 2 Lior Wolf, 1 Yaniv Taigman

1 Facebook AI Research, 2 Tel Aviv University
Why faces?

1. One class. Billions of unique instances.

2. Plays an important role in our social interactions, conveying people’s identity; The most frequent entity in the media by far: e.g. ~1.2 faces / Photo by avg

3. Enables many applications in Man-Machine interaction
Applications
Face Recognition main objective

Find a **representation** & **similarity measure** such that:

- Intra-subject similarity is high
- Inter-subject similarity is low
Milestones in Face Recognition

1964 Bledsoe Face Recognition
1973 Kanade’s Thesis
1991 Turk & Pentland Eigenfaces
1997 Belhumeur Fisherfaces
1999 Blanz & Vetter Morphable faces
1999 Wiskott EBGM
2001 Viola & Jones Boosting
2006 Ahonen LBP

Slightly modified version of Anil Jain’s timeline
Problem solved?

NIST’s best-performer’s on:

1. Its internal dataset with **1.6 million** identities: 95.9%

2. On LFW (public) with ‘only’ **4,249** identities: 56.7%

→ Answer: No.
Types of Face Recognition

- ‘Constraining’ – Mainly for traditional purposes
- ‘Unconstraining’ – General purpose

NIST’s FR Vendor Test (FRVT) 2006

In the wild
Challenges in **Unconstrained** Face Recognition

1. Pose
2. Illumination
3. Expression
4. Aging
5. Occlusion

Probes for example

Gallery
LFW: Progress over the recent 7 years

- Descriptor methods in the Wild, ECCV-W 2008
- Attribute and simile classifiers for face verification, ICCV 2009.
- Multiple one-shots for utilizing class label information, BMVC 2009.
- Large scale strongly supervised ensemble metric learning, with applications to face verification and retrieval, NEC Labs TR, 2012.
- Bayesian face revisited: A joint formulation, ECCV 2012.
- Tom-vs-pete classifiers and identity preserving alignment for face verification, BMVC 2012.
- Probabilistic elastic matching for pose variant face verification, CVPR 2013.
- Fusing robust face region descriptors via multiple metric learning for face recognition in the wild, CVPR 2013.
- Fisher vector faces in the wild, BMVC 2013.
- Hybrid deep learning for computing face similarities, ICCV 2013.
- A practical transfer learning algorithm for face verification, ICCV 2013.
Verification

= 

≠

= 

≠
LFW: Progress over the recent 7 years

Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments (results page), Gary B. Huang, Manu Ramesh, Tamara Berg and Erik Learned-Miller.
Verification Impacts Recognition

Recognition Performance

Same/Not-Same Performances

- 97.53%
- 96.33%
- 95.17%
DeepFace: Closing the Gap to Human-Level Performance in Face Verification; Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato and Lior Wolf (CVPR 2014)
Face Recognition Pipeline

Detect
Align
Represent
Classify
Face Recognition Pipeline

Detect
Align
Represent
Classify

Yaniv
Lubomir
Marc’Aurelio
Faces are 3D objects
Texture vs. Shape

Bornstein et al. 2007
Face alignment
(‘Frontalization’)

Detect  2D-Aligned  3D-Aligned
2D alignment

\[ p^0 = A_{n \times d} \cdot f \]

\[ p^t = s_t[R_t|t_t] \cdot p^{t-1} \]
3D alignment

\[
\text{loss}(\tilde{P}) = r^T \Sigma^{-1} r
\]

\[
r = (x_{2d} - X_{3d}\tilde{P})
\]

\[
\widetilde{x}_{3d}(x, y) := x_{3d}(x, y) + r(x, y)
\]

Piecewise Affine Warping
Next: Representation Learning

- **2004 – 2013**: Feature engineering monopoly, mostly LBP.
  - Contributions mainly in Classification.

- **2012**: The resurrection of LeCun’s Deep Convolutional Neural Networks (CNNs) by Krizhevsky, Sutskever and Hinton.

---

• 'Multi-Shots'; Taigman, Hassner, Wolf

LBP; Ahonen 2004

High-Dim LBP; Chen, Cao, Wen, Sun
CNNs for: Image Classification vs. Face Recognition

1. We mostly care about feature learning
   - We do not know the number of identities before-hand
   - Transfer Learning

→ **Last layer** can be removed or replaced
→ We still need to think about the Classification stage (later)
2. Geometry is **physically** relaxed:
   - Translation, scale and 2D-rotation due to Detection and 2D Alignment
   - Out-of-plane rotation due to 3D Alignment.

Aligned pixels → Enables **Untying** the weights → ‘**Locally-connected**’ layers.

→ Greater focus in training on what’s not solved already.
3. Several levels of (max-) **pooling** would cause the network to lose information about the precise position of detailed facial structure and micro-textures.
DeepFace Architecture

Localization

Frontalization

Front-End ConvNet

Local (Untied) Convolutions

Globally Connected

Calista_Flockhart_0002.jpg

Detection & Localization

Frontalization

DeepFace: Closing the Gap to Human-Level Performance in Face Verification; Taigman, Yang, Ranzato, Wolf
SFC Training dataset
(pre-cropping)

4.4 million photos blindly sampled, containing more than 4,000 identities (permission granted)
(a) Cosine angle

\[ S'(f_1, f_2) = \frac{\langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|} \]

(b) Kernel Methods

\[ S_{\chi^2}(f_1, f_2) = \sum w_i \frac{(f_1[i] - f_2[i])^2}{f_1[i] + f_2[i]} \]

(c) Siamese Network\(^1\)

\[ S_{Siam}(I_1, I_2) = \frac{1}{1 + e^{-(W|f(I_1)-f(I_2)|+b)}} \]

Deep Siamese Architecture [1]

Binary Label

Face A
Network / Feature Extractor → \( f_1 \)

Face B
Network / Feature Extractor → \( f_2 \)

\[ |f_1 - f_2| \]

Logistic

Binomial Cross Entropy Loss

\[
p = \frac{1}{1 + e^{-(W \cdot |f_1 - f_2| + b)}}
\]

\[
E = -y \log(p) - (1 - y) \log(1 - p)
\]

Results on LFW
‘Explaining’ the False Negatives pairs (1.65%)
False Positive pairs (1.00%)
Results on YouTube Faces (Video)

False negatives

False positives
Unaccounted challenges in verification:
I. Reliability
II. Large confusion (P x G)
III. Different distributions
IV. Unknown class
LFW Identification (1:N) Protocols

1. **Close Set**
   - #Gallery\(^1\): 4,249
   - #Probes: 3,143
   Measured\(^3\) by Rank-1 rate.

2. **Open Set**
   - #Gallery\(^1\): 596
   - #Probes: 596
   - #Impostors: 9,491 (‘unknown class’)
   Measured\(^3\) by Rank-1 rate @ 1% False Alarm Rate.

---

\(^1\) Each identity with a single example

\(^2\) Unconstrained Face Recognition: Identifying a Person of Interest from a Media Collection

Best-Rowden, Han, Otto, Klare and Jain (Technical Report MSU-CSE-2014-1)

\(^3\) Training is not permitted on LFW (‘unsupervised’)
Cosine similarity measure ('unsupervised'):

Confusion Matrix = $G^T \cdot P$

$G$ is 256 x 4249

$P$ is 256 x 3143
(part of the) t-SNE visualization of LFW faces
Why does it work so well?

1. **Coupling** alignment with Locally-Connected layers

2. Large capacity model that *actually* enjoy large data

But can we understand more with respect to the roles of:

- What each layer is actually doing
- Is alignment necessary?
- Is regularization needed?
- Dimensionality & Sparsity
- Will more data help?
Localization is needed

Original + ImageNet

Original
2D-Aligned
3D-Aligned

75%
89%
94%
97%
Localization is needed but insufficient.

- Alignment – DNN + LBP → Accuracy drops to 91.5% (-6%)
Local Patches are Insufficient

False Positive
→ Fully-Connected Layer is the holistic representation

Projects input ‘features’ Into the representation.

1. **Correlates** between different local parts
2. Can exploit **symmetries** in faces
3. **High-Level templates**, a-la Eigenfaces (PCA)
The RELU := max(0,x) encourage sparsity.

Weights can be ‘thought of’ as weak template classifiers:

\[ \text{Output} := \max(0, W\times\text{input} + b) \]

Bias ‘b’ is a trainable threshold / filter:

IF : \( W\times\text{input} < -b \) THEN
\[ \text{Output} := 0 \]
ELSE
\[ \text{Output} := W\times\text{input} + b \]

80% of the dims are zero by avg.
Most of the information is encoded in whether a unit is fired or not.

\( X := (X > 0) \rightarrow \) Performance drops only a bit.
The norm of the representation is a measure of signal acquisition.

For faces: $\| F(I) \|$ is a measure of feed-forward confidence.

Smallest norm's in LFW:

Largest norm's in LFW:
Understanding feature response

C1: 32 filters 11x11
M2: 3x3
C3: 16 filters 9x9
L4: 16 x 9 x 9 x 16
L5: 16 x 7 x 7 x 16
L6: 16 x 5 x 5 x 16
F7: 4096d

Occlusion

Failed Alignment
Correlation between norm & accuracy confidence

- True Positive Or True Negatives
- False Positives or False Negatives
Bottleneck is an important Regularizer in Transfer Learning

The network **overfits less** on the **SOURCE** training set, and performs better on the **TARGET** when reducing the representation layer (F7) from 4K dims to **256 dims**.
Bottleneck regularizes Transfer Learning

Web-Scale Training for Face Identification; Taigman, Yang, Ranzato, Wolf
CNN’s (can) saturate

“Results can be improved simply by waiting for faster GPUs and bigger datasets to become available” -- Krizhevsky et al.

What happens when the network is fixed & the number of training grows from 4m \(\rightarrow\) 0.5b ?

Answer: our findings reveal that this holds to a certain degree only.
Scaling up

DeepFace: 4.4 million images / 4,030 identities
Random 108k: 6 million images / 108,000 identities
Random 250k: 10 million images / 250,000 identities
(yes : 250K softmax)

→ Saturation
Scaling up: Semantic Bootstrapping

- 0.5B images $\rightarrow$ 10M hyperplanes
- Lookalike hyperplanes $\rightarrow$ DB2
- Training on DB2 with more capacity.

Web-Scale Training for Face Identification; Taigman, Yang, Ranzato, Wolf
Second round results

Results on LFW

- New (98.4%)
- Human cropped (97.5%)
- DeepFace-ensemble (97.35%)
- DeepFace-single (97.00%)
- TL Joint Baysian (96.33%)
- High-dimensional LBP (95.17%)
- Tom-vs-Pete + Attribute (93.30%)
- combined Joint Baysian (92.42%)

true positive rate
false positive rate
Comparison to NIST’s State Of The Art

Second-round DeepFace

Same system that achieved 92% Rank-1 accuracy on a table of 1.6 million identities. (NIST’s State-Of-The-Art, Constrained)
Additional works


- 200 ConvNets from 400 patches ← 2D Aligned (no 3D)
- With Joint Bayesian source / target adaptation
  → 99.15% on the verification (1:1) task.
Additional works

New free public large face dataset from SMU:

**WLFDB : Weakly Labeled Faces on the Web**

Wang, Dayong, Hoi, Steven C. H., He, Ying, Zhu, Jianke, Mei, Tao and Luo, Jiebo, 
Retrieval-Based Face Annotation by Weak Label Regularized Local Coordinate Coding

714,454 facial images / 6,025 identities
Conclusion:

• Coupling 3D alignment with Large locally-connected networks

• Two-stage 3D alignment system

• Regularization in Transfer Learning

• Scaling up through bootstrapping

• At the brink of human-level performance
Thank you!

1 Ming Yang  1 Marc’Aurelio Ranzato  2 Lior Wolf

1 Facebook AI Research  2 Tel Aviv University

References:
1. DeepFace: Closing the Gap to Human-Level Performance in Face Verification; Taigman, Yang, Ranzato, Wolf
2. Web-Scale Training for Face Identification; Taigman, Yang, Ranzato, Wolf
3. Multi-GPU Training of ConvNets; Yadan, Adams, Taigman, Ranzato