Deep face recognition

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman

Key Questions

- Can large scale datasets be built with minimal human intervention?

Yes!

- Can we propose a convolutional neural network which can compete with that of internet giants like Google and Facebook?

Yes!

Achievements

- New face dataset of 2.6 Million Faces
- **State of the art** results on YouTube faces in the wild dataset
- Comparable to the state of the art results on the Labeled faces in the wild dataset

Can large scale datasets be built with minimal human intervention?

Dataset Collection

1. **Candidate list generation**: **Finding names of celebrities**

   - Tap the knowledge on the web
   - 5000 identities

Robert Downey Jr. Ashley Hamilton Barack Obama

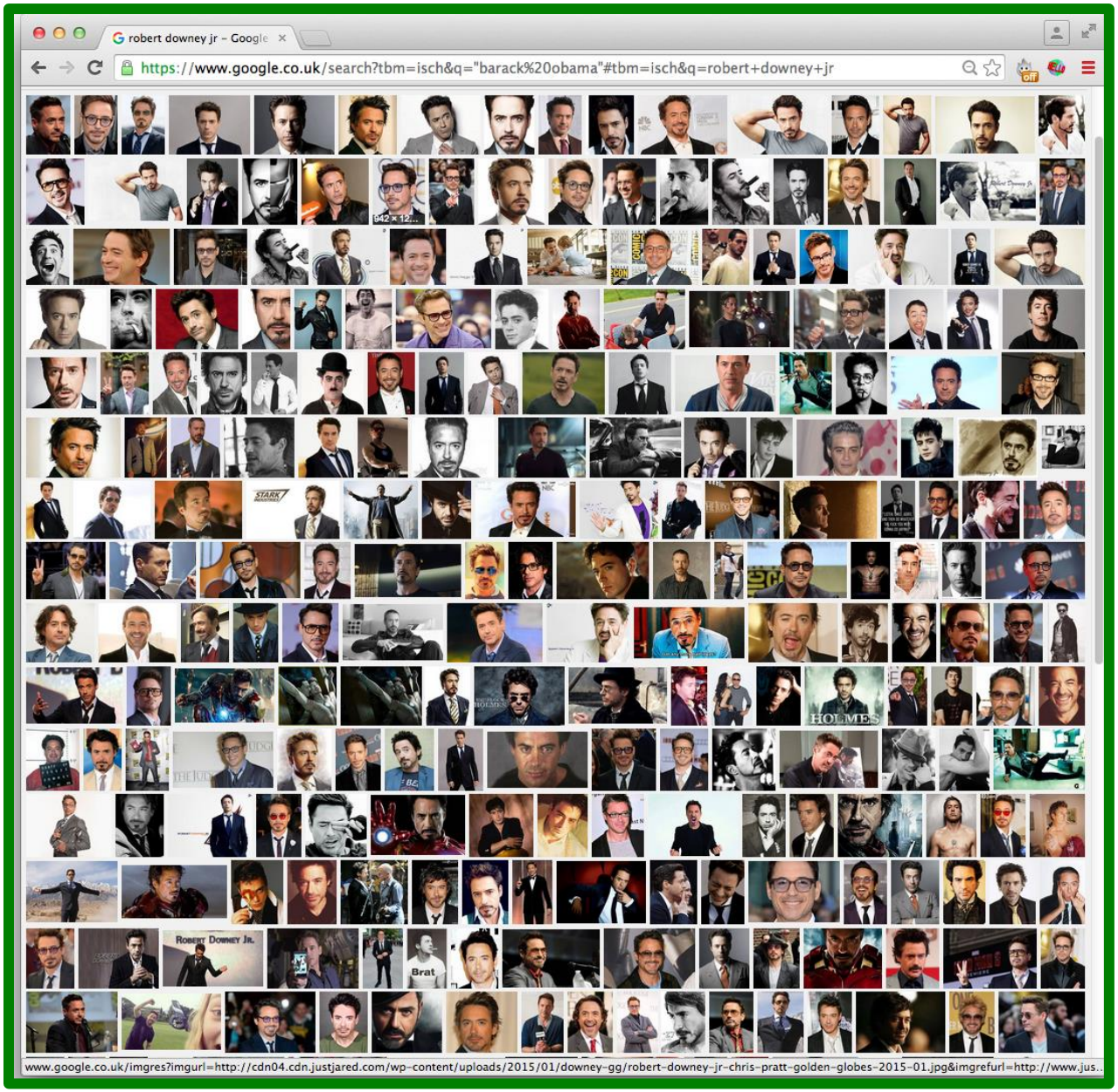
Allison Hannigan Amitabh Bacchan Vladimir Putin

Dataset Collection

2. **Manual verification of celebrities: Finding Popular Celebrities**

- Collect representative images for each celebrity
- 200 images/identity
- Remove people with low representation on Google.
- Remove overlap with public benchmarks
- 2622 celebrities for the final dataset

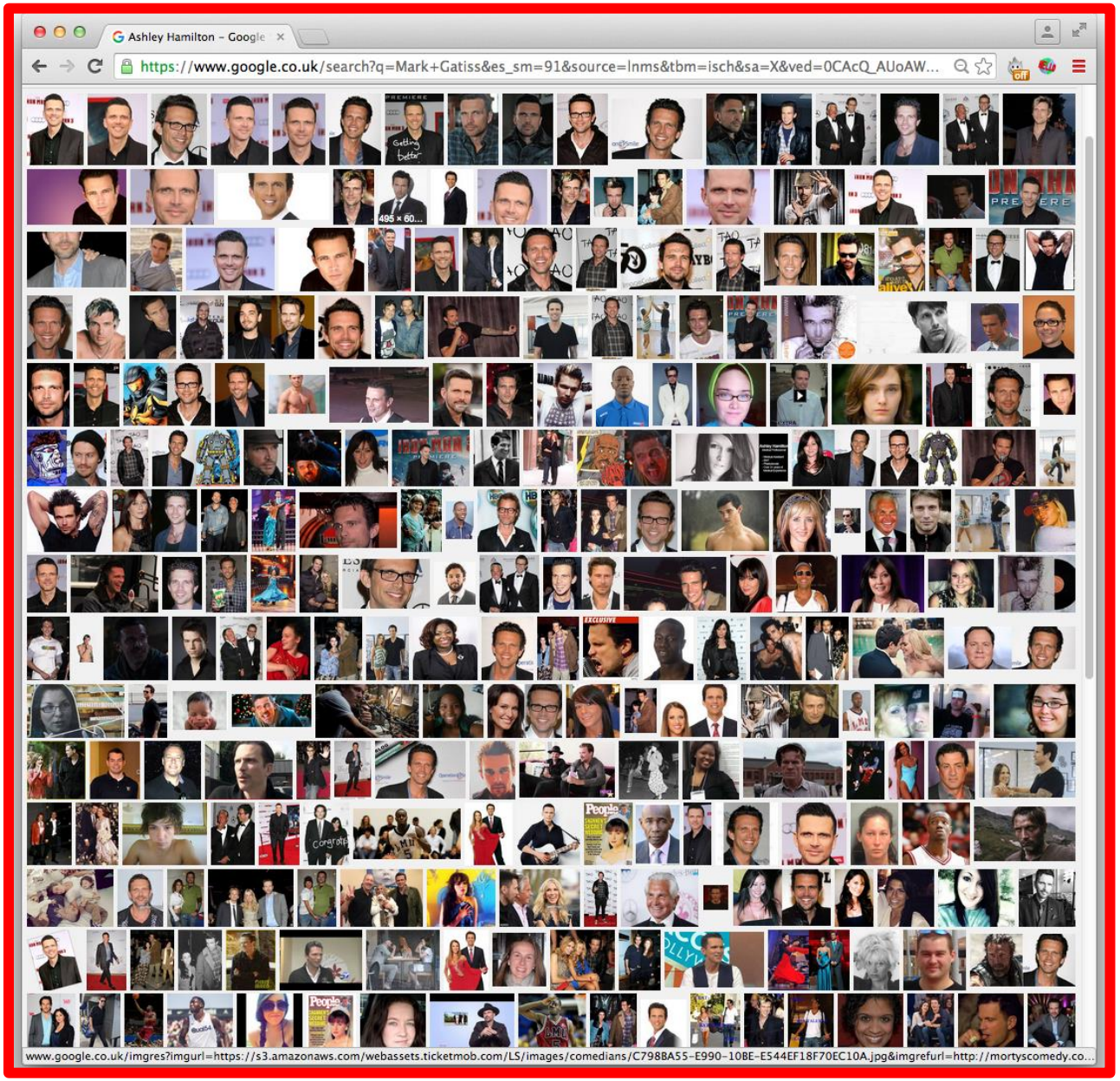
Dataset Collection



Popular Celebrity

Robert Downey Jr.

Dataset Collection



Not so popular celebrity

Ashley Hamilton

Dataset Collection

2. **Manual verification of celebrities: Finding Popular Celebrities**

- Collect representative images for each celebrity
- 200 images/identity
- Remove people with low representation on Google.
- Remove overlap with public benchmarks
- 2622 celebrities for the final dataset

Dataset Collection

3. **Rank image sets**

- 2000 images per identity
- Searching by appending keyword “actor”

- Learning classifier using data obtained the previous step. c
- Ranking 2000 images and selecting top 1000 images
- Approx. 2.6 Million images of 2622 celebrities

Dataset Collection

4. **Near duplicate removal**
   - VLAD descriptor based near duplicate removal

5. **manual filtering**
   - Curating the dataset further using manual checks

No.	Aim	Mode	# Persons	# images /person	Total # images	Anno. effort
1	Candidate list generation	Auto	5000	200	1,000,000	-
2	Candidate list filtering	Manual	2622	-	-	4 days
3	Rank image sets	Auto	2622	1000	2,622,000	-
4	Near duplicate removal	Auto	2622	623	1,635,159	-
5	Manual filtering	Manual	2622	375	982,803	10 days

Dataset Collection

Dataset Comparison

No.	Aim	# Persons	Total # images
1	Labeled Faces In the Wild	5,749	13,233
2	WDRef	2995	99,773
3	Celeb Faces	10177	202,599
4	Ours	2622	1,635,159
5	Facebook	4030	4.4M
6	Google	8M	200M

Dataset Collection

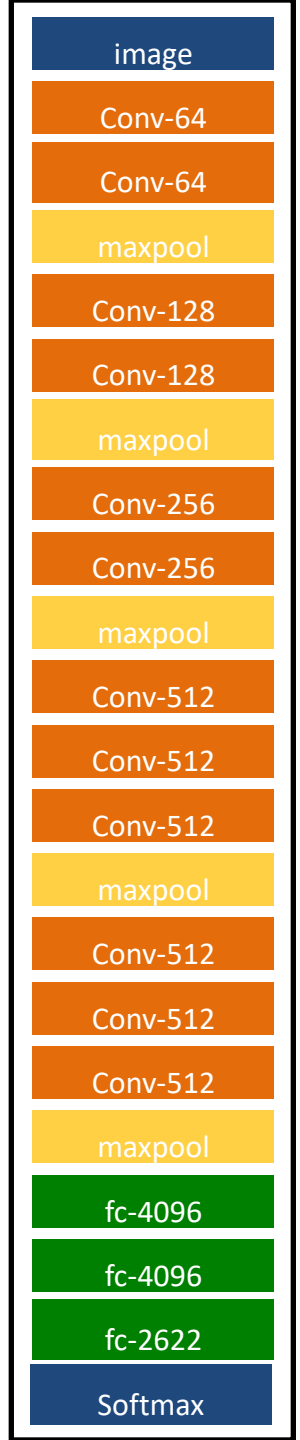
Example Images From Our Dataset

Can we propose a convolutional neural network which can compete with that of internet giants like Google and Facebook etc.?

Convolutional Neural Network

- The “Very Deep” Architecture
  - Different from previous architectures proposed for face recognition:
    - locally connected layers (Facebook)
    - inception (Google)

- Network Details:
  - 3 x 3 Convolution Kernels (Very small)
  - Conv. Stride 1 px.
  - Relu non-linearity
  - No local contrast normalisation
  - 3 Fully connected layers



Very Deep Convolutional Networks for large-Scale Image Recognition.
K. Simonyan and A. Zisserman.

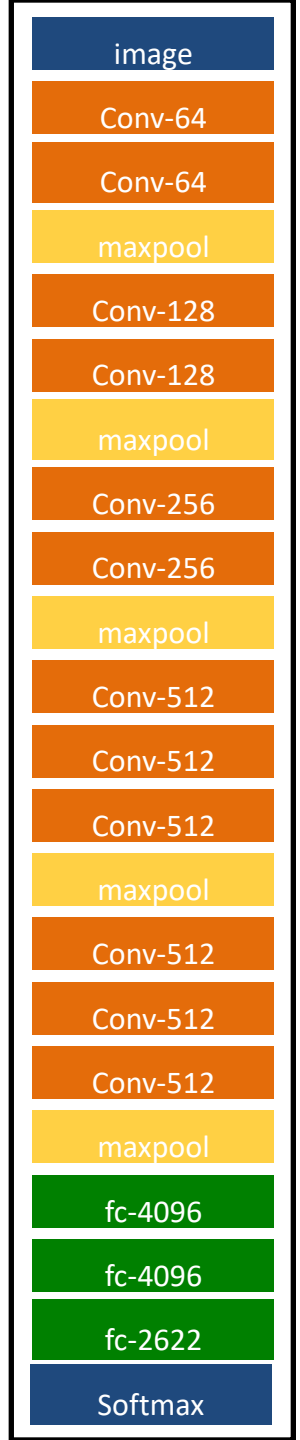
Convolutional Neural Network

Training

- MatConvNet Tootlbox
  - Nvidia CuDNN bindings
  - Multi GPU Training (approx 3.5x speedup)
  - Nvidia Titan Black
  - 7 days of training

- Random Gaussian Initialization
- Stochastic Gradient Descent with back prop.
  - Accumulator Descent for large batch sizes
- Batch Size: 256
- Incremental FC layer training

- 2622 way multi class criterion (soft max)



Matconvnet – convolutional neural networks for matlab.
A Vedaldi and K. Lenc. Arxiv - 2014.

Convolutional Neural Network

Training: Learning Task Specific Embedding

- Learning embedding by minimizing triplet loss

$$\sum_{(a,p,n)\in T} \max\{0, \alpha - \|\mathbf{x}_a - \mathbf{x}_n\|_2^2 + \|\mathbf{x}_a - \mathbf{x}_p\|_2^2\}$$

- Learning a projection from 4096 to 1024 dimensions
- On line triplet formation at the beginning of each iteration
- Fine tuned on target datasets
- Only the projection layers learnt

Design Choices

1. Network configuration
   - Does increasing the depth improves performance?

1. Face alignment
   - Can the network be invariant to changes in the faces?

2. Task specific learning
   - What are the effects of task specific embedding learning on performance?

Labeled Faces In the Wild Dataset (LFW)

- Face Verification: Given a pair of images specify whether they belong to the same person
- 13K images, 5.7K people
- Standard benchmark in the community
- Several test protocols depending upon availability of training data within and outside the dataset.

[G. Huang, M. Ramesh, T. Berg and E. Learned-Miller - Tech Report 07]

Effects of design choices (LFW Unrestricted Protocol)

No.	Network Config.	Dataset	Face Align Training	Face Align Testing	Embedding	100%-EER
1	A	Curated	No	No	No	92.83
2	A	Full	No	No	No	95.80
3	A	Full	No	Yes	No	96.70
4	B	Full	No	Yes	No	97.72
5	B	Full	Yes	Yes	No	97.07
6	D	Full	No	Yes	No	96.60
7	**B**	**Full**	**No**	**Yes**	**Yes**	**99.13**

Comparison with the State of the Art (LFW Unrestricted Protocol)

No.	Method	# Training Images	# Networks	Accuracy
1	Fisher Vector Faces	-	-	93.10
2	DeepFace	4 M	3	97.35
3	DeepFace Fusion	500 M	5	98.37
4	DeepID-2,3	Full	200	99.47
5	FaceNet	200 M	1	98.87
6	FaceNet+ Alignment	200 M	1	99.63
7	VGG Face	2.6 M	1	98.95

YouTube Faces Dataset (YTF)

same **different**

- Video Face Verification: Given a pair of videos specify whether they belong to the same person
- 3425 videos, 1595 people
- Standard benchmark in the community
- Wide pose, expression and illumination variation

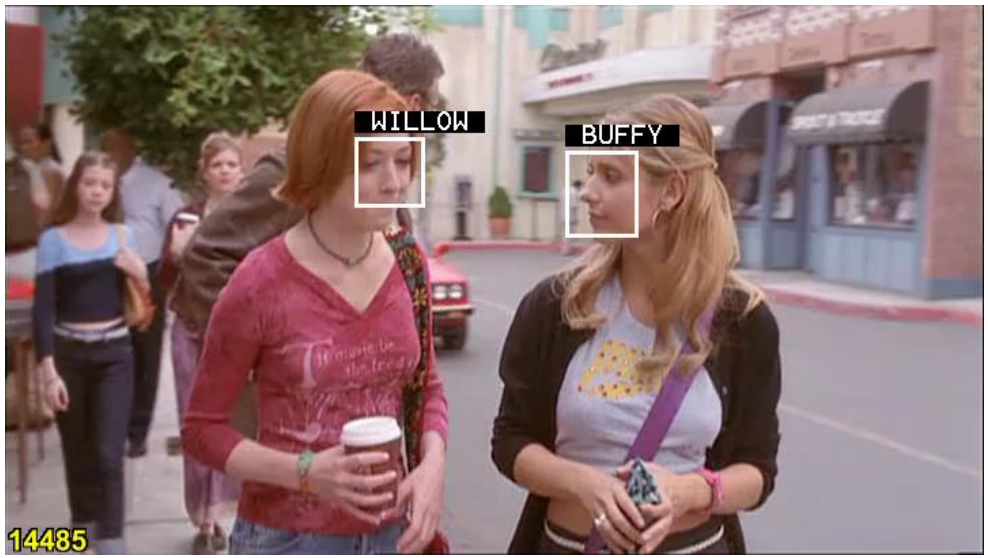
[Wolf, Hassner, Moaz CVPR 2011]

Comparison with the State of the Art (YTF Unrestricted Protocol)

No.	Method	# Training Images	# Networks	100%-EER	Accuracy
1	Video Fisher Vector Faces	-	-	87.7	93.10
2	DeepFace	4 M	1	91.4	91.4
4	DeepID-2,2+,3		200	-	93.2
5	FaceNet + Alignment	200 M	1	-	95.1
7	**VGG Face**	**2.6 M**	**1**	**97.4**	**97.3**

Oxford Buffy Dataset

Weakly supervised face classification



- **“Buffy The Vampire Slayer”**
  - Face tracks from 7 episodes of season 5.
  - Both frontal and profile detections
  - Weak supervision from transcript and subtitles
  - Multi Class classification for every episode

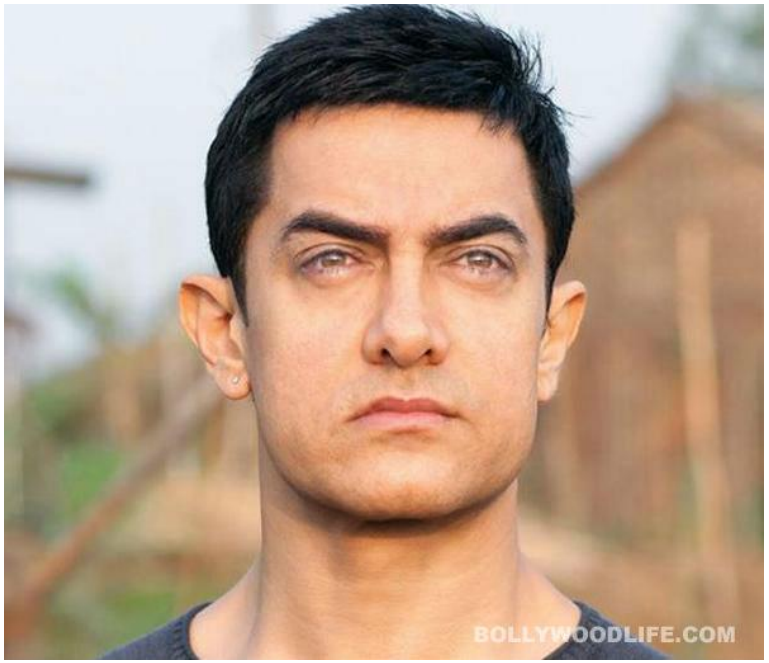
[Everingham et al. IVC 2009, Sivic et al. CVPR 2009]

Comparison with the State of the Art (Oxford Buffy Dataset)

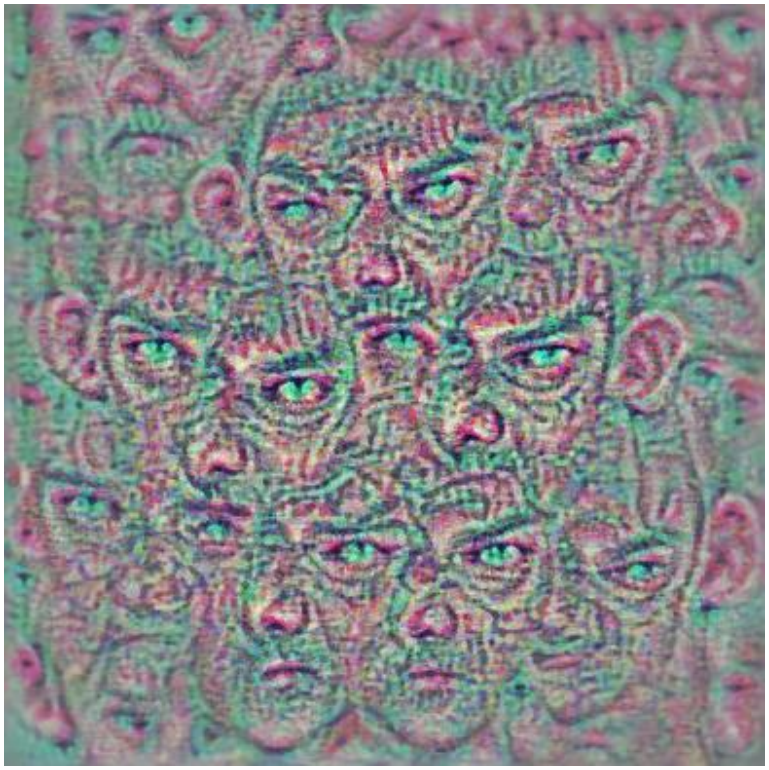
No.	Method	Mean AP
1	Sivic et. al (HOG + RBF SVM)	0.81
2	Video Fisher Vector Faces (FV + Lin SVM)	0.86
3	**Ours (CNN + MIL SVM)**	**0.95**

Deep Face Dreams

Inversion by maximizing class specific neuron responses



Representative Image



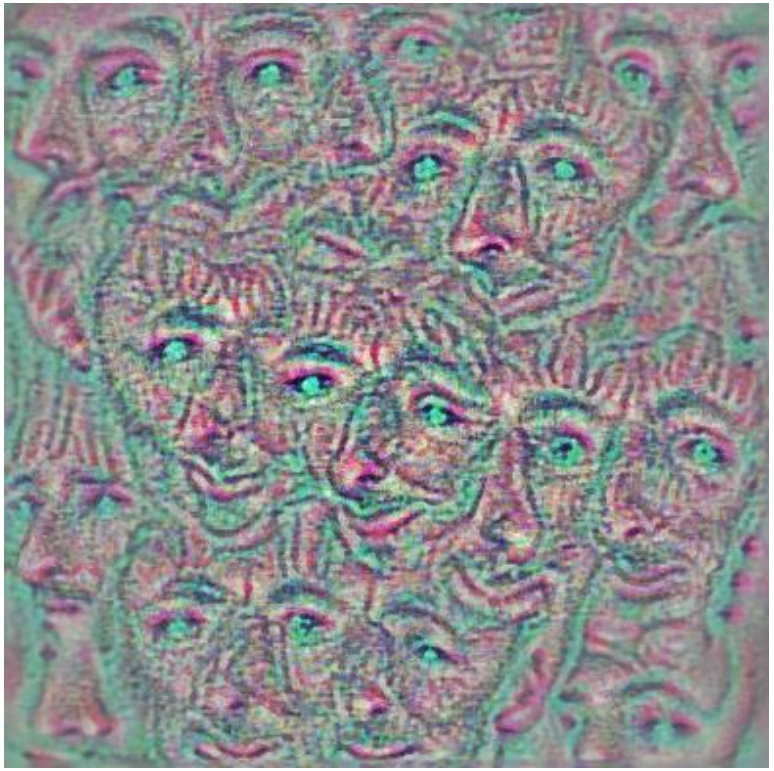
Neuron Inversion

[Simonyan et al. NIPS 2014, Mahendran et al. CVPR 2015]

Deep Face Dreams

Inversion by maximizing class specific neuron responses

Representative Image



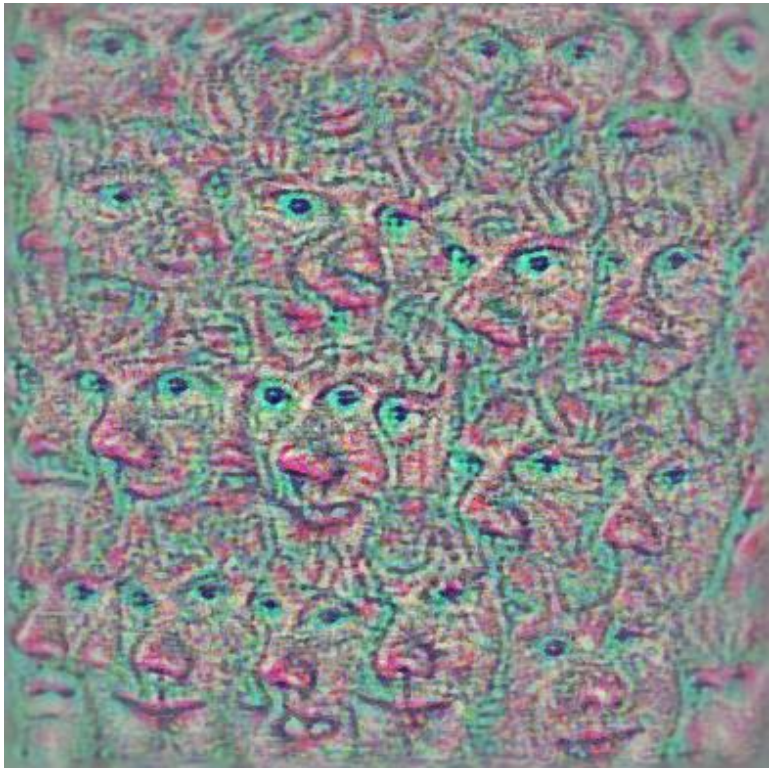
Neuron Inversion

[Simonyan et al. NIPS 2014, Mahendran et al. CVPR 2015]

Deep Face Dreams

Inversion by maximizing class specific neuron responses

Representative Image



Neuron Inversion

[Simonyan et al. NIPS 2014, Mahendran et al. CVPR 2015]

Deep Face Dreams

Inversion by maximizing class specific neuron responses

Representative Image

Neuron Inversion

[Simonyan et al. NIPS 2014, Mahendran et al. CVPR 2015]

Thank You!