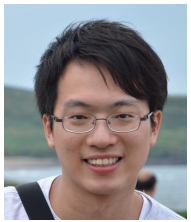
One-to-many Face Recognition using Bilinear CNNs

Aruni RoyChowdhury Tsung-Yu Lin
Subhransu Maji Erik Learned-Miller

College of Information and Computer Sciences
University of Massachusetts, Amherst

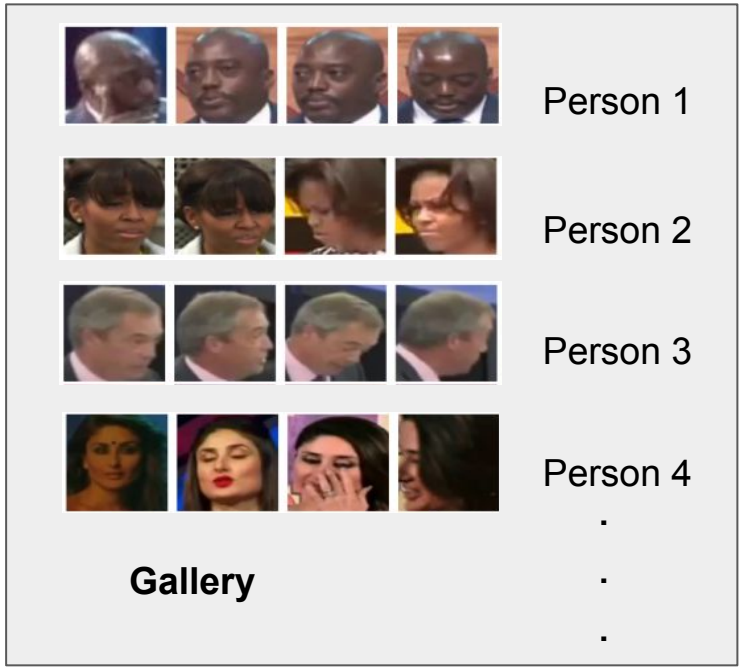



Motivation

- How do **high-performing** fine-grained recognition models[1] translate to face recognition?
- How well can we do with publicly-available moderate-sized **training datasets**?
- Do deeper architectures in CNNs matter in this setting?

[1] ***Bilinear CNN Models for Fine-grained Visual Recognition,***
Tsung-Yu Lin, Aruni RoyChowdhury, Subhransu Maji, ICCV '15

Face Recognition



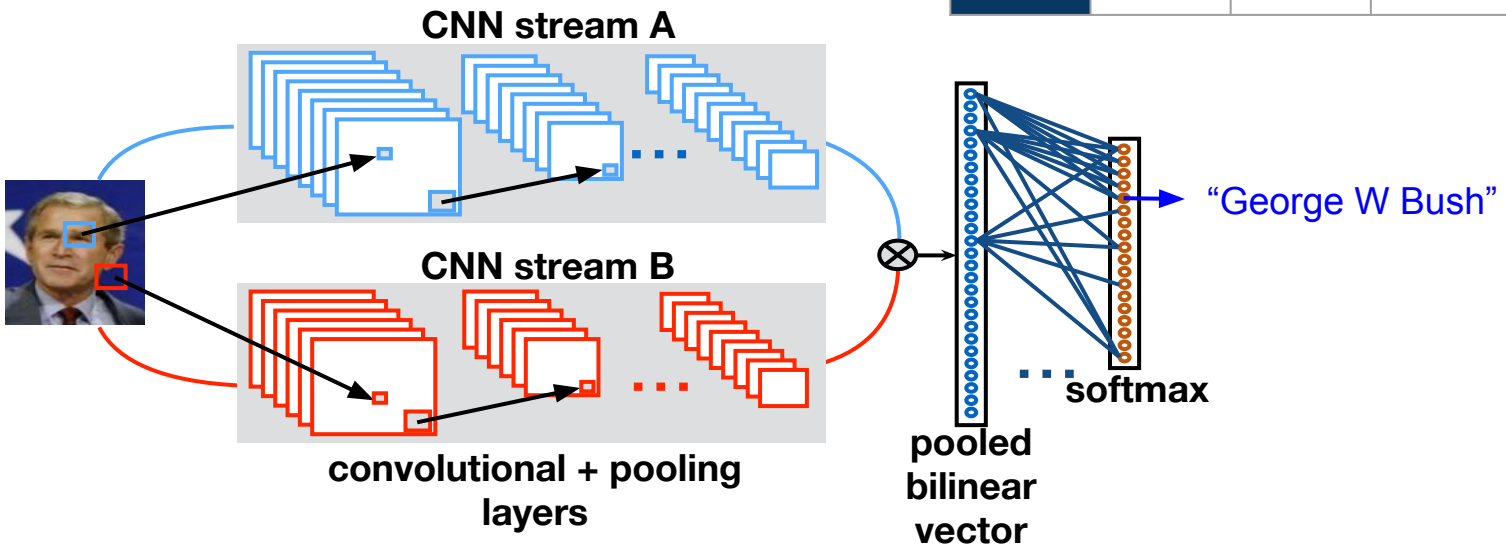
Verification: is it the same person?

- Verification (e.g. LFW) has saturated
- Moving on to recognition: actually naming the person
- The challenging IARPA Janus A (IJB-A) protocol **[Klare et al., CVPR '15]**

Bilinear CNN Model

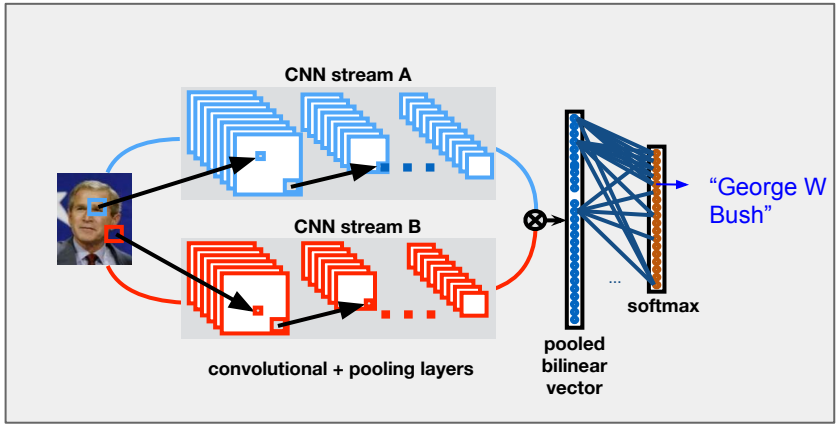
- Achieves state-of-the-art results on fine-grained recognition datasets.
- Models **co-occurrence statistics** of features, e.g. "*brown eyes*"
- Translation invariant

	hair	eye	beard	eye brow
black				
brown				
blue				
green				



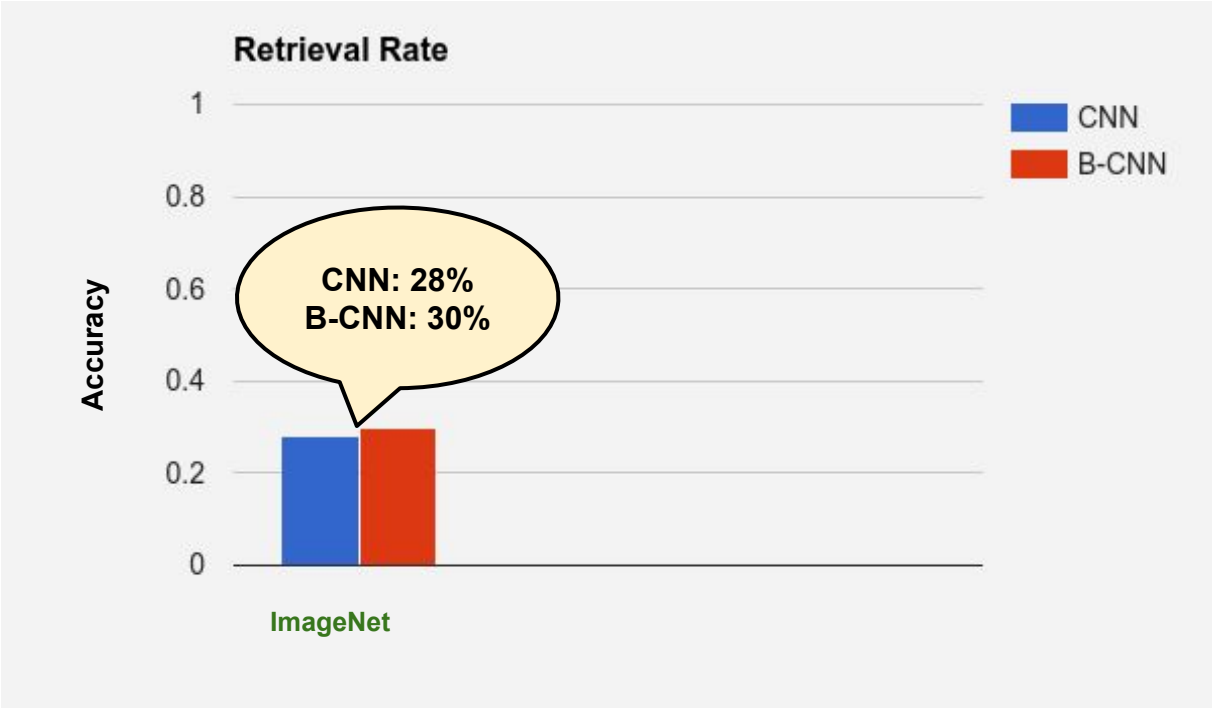
Experiments

- Dataset:
  - **FaceScrub** (513 / 89045)
  - **IJB-A Train** set (~333 / ~16900)
- Model:
  - **Stream A** and **Stream B** are identical
  - ***VGG-M*** model
  - uses '`conv5+ReLU`' layer features



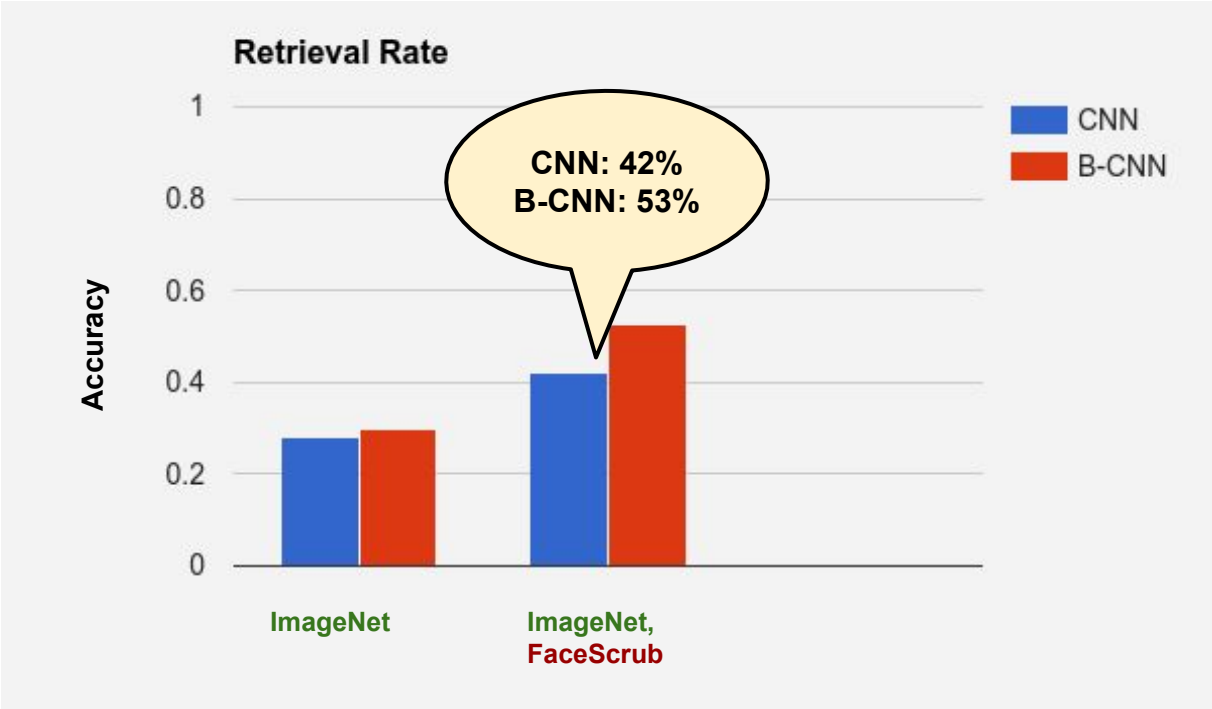
- Linear SVM classifier learned for each person in Gallery
- Max-pooling features or classifier scores to aggregate multiple media

Results: Fine-tuning CNN and B-CNN



Low initial accuracy with ImageNet pre-trained models

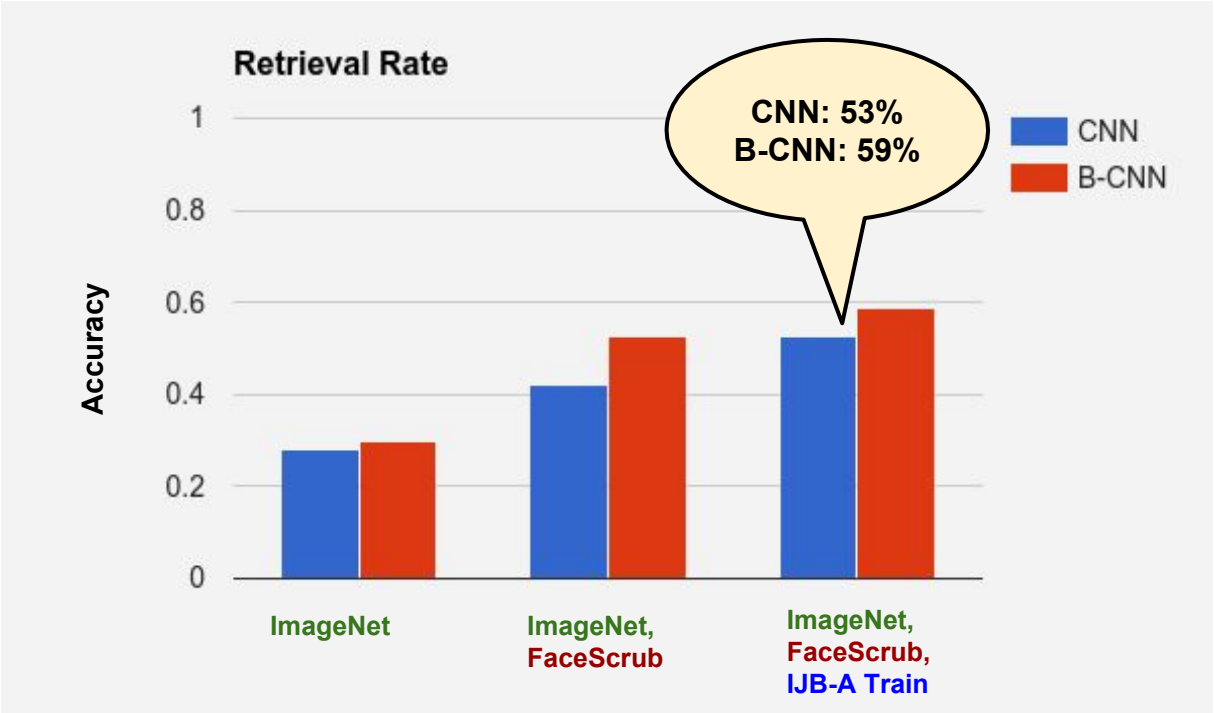
Results: Fine-tuning CNN and B-CNN



Low initial accuracy with ImageNet pre-trained models

Accuracy increases with domain-specific fine-tuning

Results: Fine-tuning CNN and B-CNN

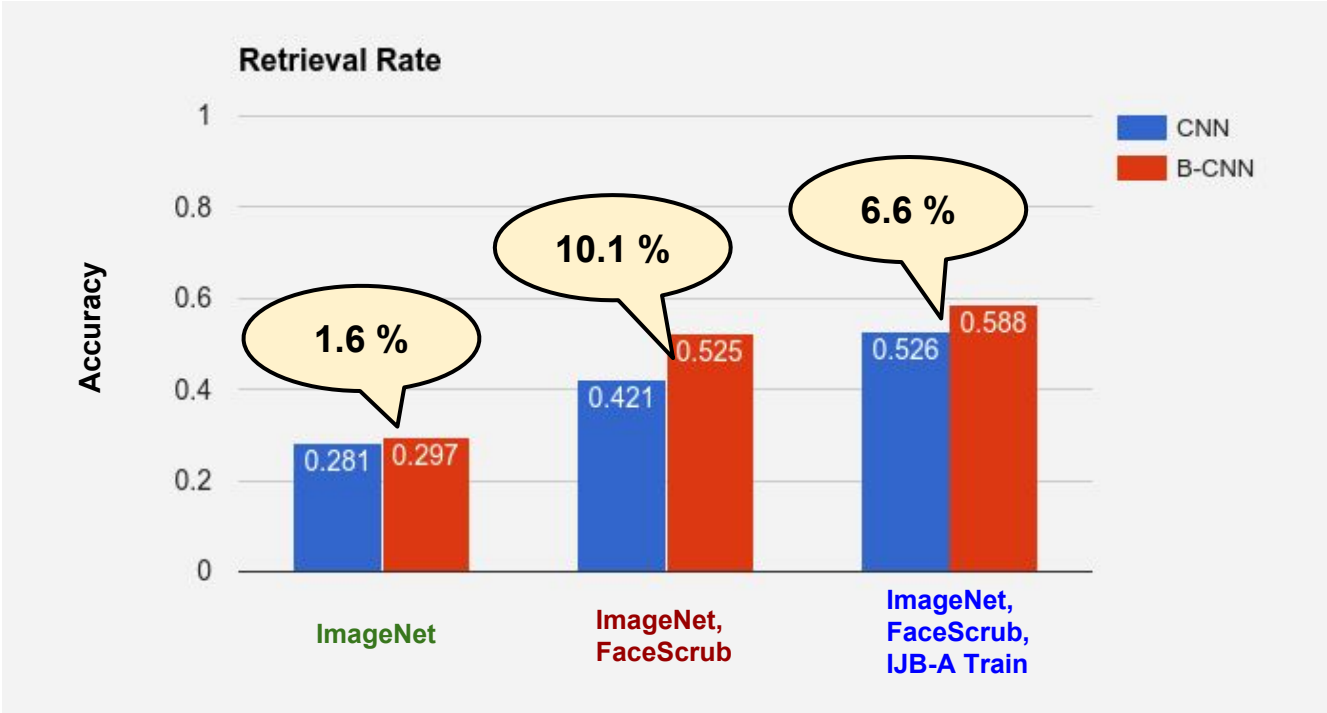


Low initial accuracy with ImageNet pre-trained models

Accuracy increases with domain-specific fine-tuning

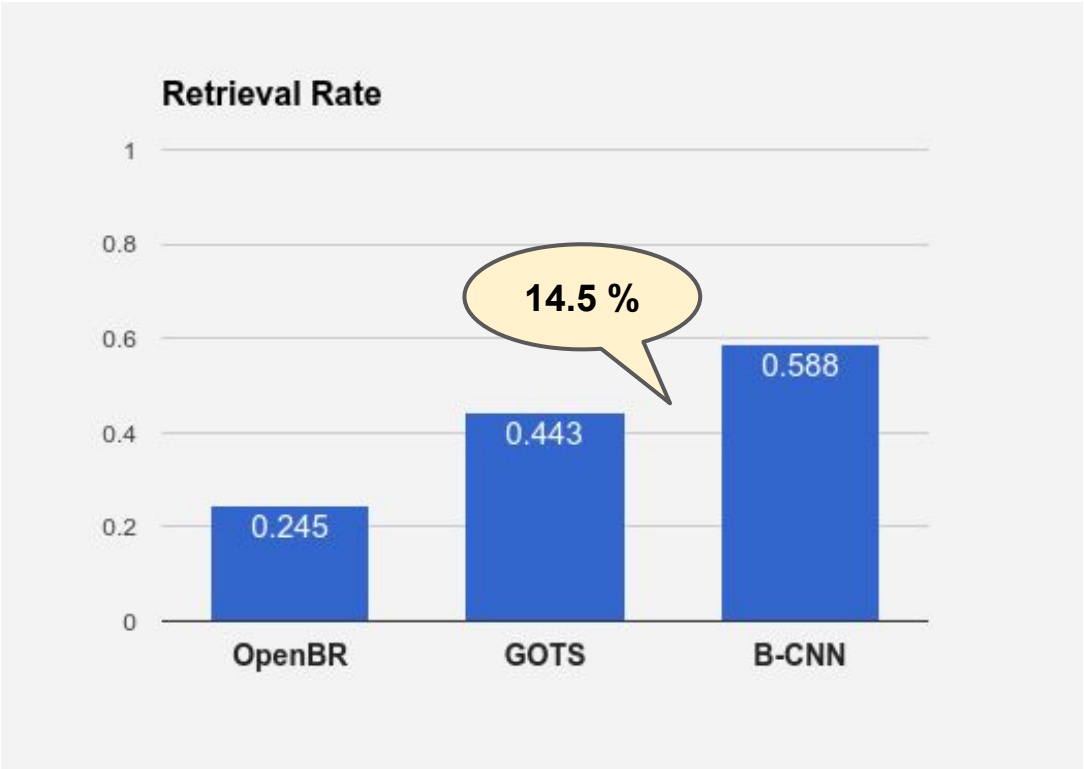
Highest accuracy when fine-tuning on IJB-A Train set

Results: CNN versus B-CNN



B-CNN consistently outperforms the regular **CNN**

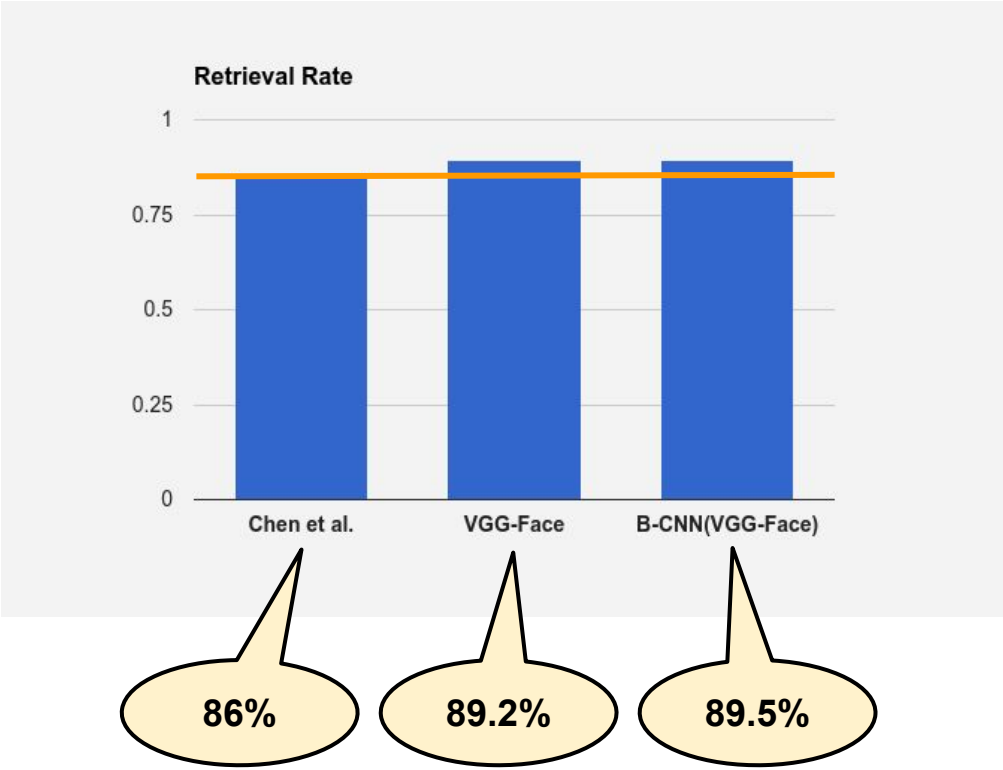
Results: Comparison with IJB-A Baselines



B-CNN exceeds best baseline - **"GOTS"** - by a large margin

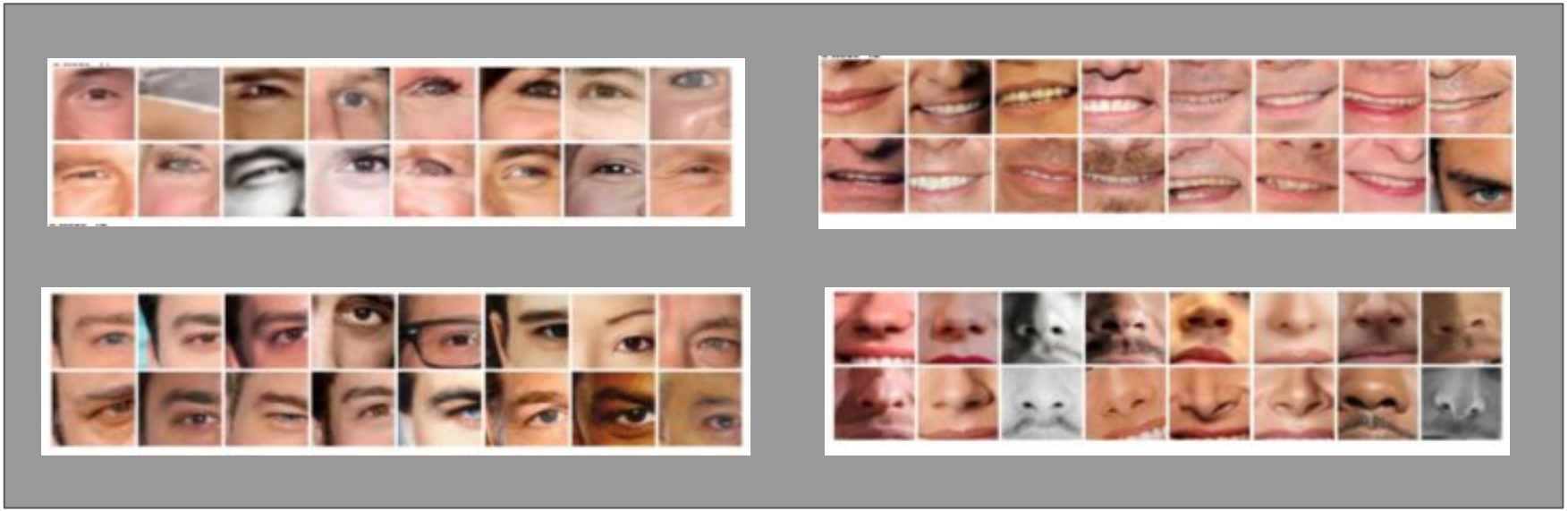
- **58.8%** versus 44.3%

Results: Pre-trained Networks



- CNN pre-trained on **massive external datasets**
  - VGG-Face **[Parkhi et al.]**
  - **[Chen et al.]**
- Bilinearization:
  - start with a **pre-trained network**
  - **B-CNN features:** sum-pooled outer product of penultimate layer outputs
  - Train **identity-specific SVMs** on new B-CNN features
- B-CNN gives a slight improvement over VGG-Face CNN

Results: Filter Visualizations



Traditional facial features such as

- eyes,
- eyes+eyebrows
- partially open mouth
- noses.

Results: Filter Visualizations

Filters correlated with hair, both facial and on the head.

Filters correlated with accessories such as eyeglasses and earrings.

Thank You