Face Representations in Deep Convolutional Neural Networks

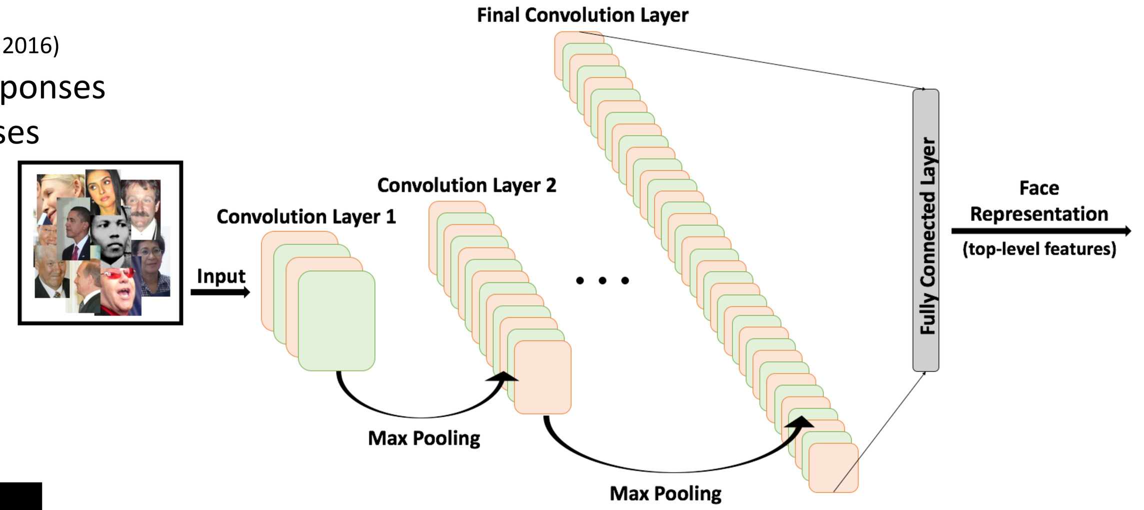
Connor J. Parde[1], Carlos Castillo[2], Matthew Q. Hill[1], Y. Ivette Colon[1], Swami Sankaranarayanan[2], Jun-Cheng Chen[2], and Alice J. O'Toole[1]

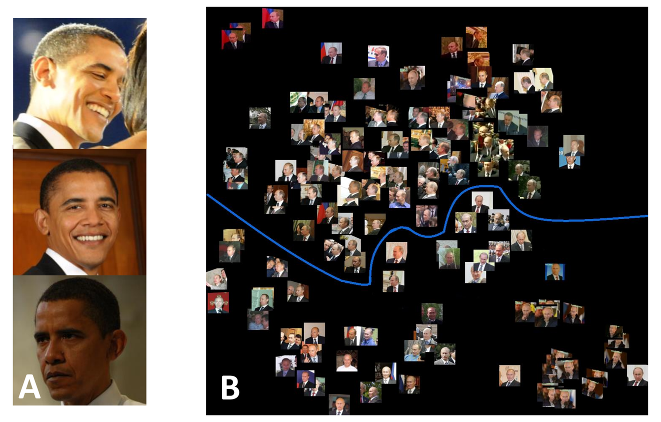
[1]The University of Texas at Dallas, [2]University of Maryland

Introduction

- State-of-the-art face recognition → DCNN models
  - (Taigman et al., 2014; Schroff et al. 2015; Chen et al., 2016; Sankaranrayanan et al., 2016; Ranjan et al., 2017)
- DCNNs designed to model primate visual system (Krizhevsky, et al., 2012)
  - neural network with multiple layers that convolve and pool image data
  - representations expand in intermediary layers
  - highly compressed final representation of image emerges at top layer

- primate vision for objects (Yamins & DiCarlo, 2016)
  - *early network layers* → V1-V3 responses
  - *intermediate layers* → V4 responses
  - *top levels* → IT responses
    - category orthogonal information (e.g. viewpoint, size) represented in top-level features





(A) Examples of variation in PIE. (B) t-SNE visualization of a single identity. Hand-drawn blue line shows distinct grouping by view.

Representations for faces in DCNN?

- similar coding between DCNNs & humans?
- performance maintained across pose, illumination, expression (PIE) and image quality
- *visualizations show image information remains in the top-level*
  - from Parde et al. (2016)
  - *t*-SNE compresses multidimensional data for visualization while preserving relative point distances (Maaten & Hinton, 2008)

Goal

Explore nature of face representations in top-level DCNN feature codes:

1: Retention of image data in DCNN representation?
- Predict yaw, pitch, and media type from top-level features using linear classifiers

2: Robustness of DCNN features to image change
- determine view and media robustness of top-level features
- analyze impact of feature invariance on face recognition performance

3: Image quality codes in representation
- Find indications of image quality in the top-level feature space

Approach

- Analyzed top-level features produced by two state-of-the-art DCNNs:
  - Network A (Chen et al., 2015) and Network B (Sankaranarayanan et al., 2015)
  - developed for IARPA Janus Competition
  - trained on CASIA Webface database (490,000+ images, 10,000+ identities)
  - top-level feature descriptor length: Network A–320 features, Network B–512 features
- Test set: 25,787 images of 500 identities

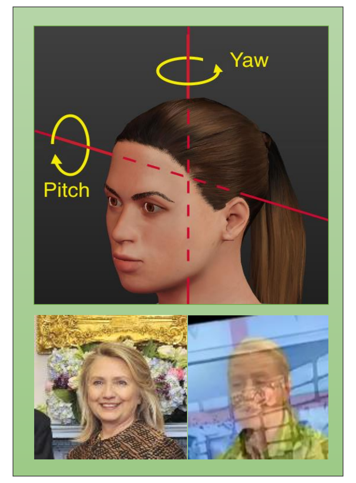
Acknowledgements

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Do DCNNs for Faces Retain Image Information in Top-level Features?

top-level features from Networks A and B as input to LDA classifier
- to predict yaw, pitch, and media type (still image vs. video frame)
- ground truth:
  - yaw and pitch scores assigned by Hyperface (Ranjan, Patel & Chellappa, 2016)
  - media type provided in dataset
- tested with 20 bootstrap iterations

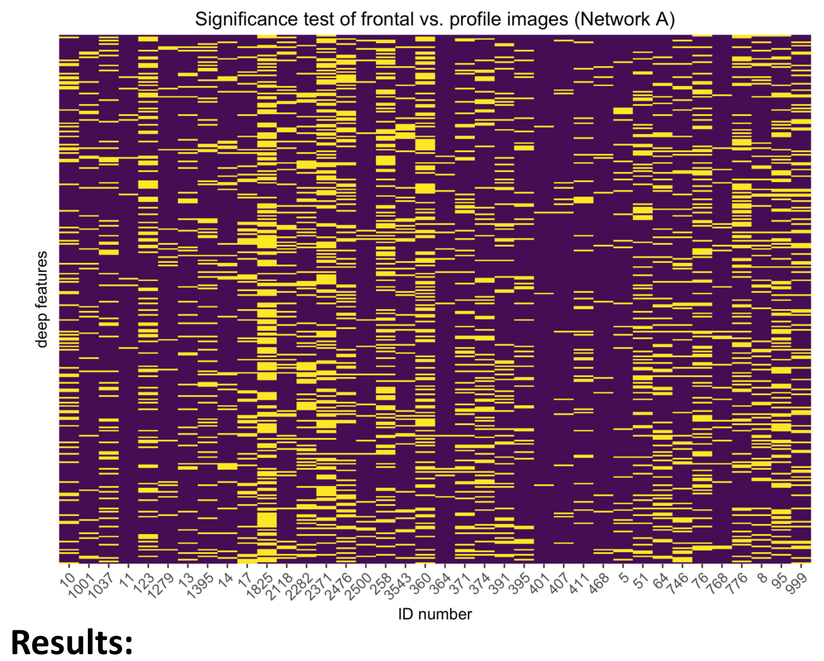
Results: yaw, pitch, & media type accurately predicted
- consistent with object recognition findings in IT (Hong et al., 2016)

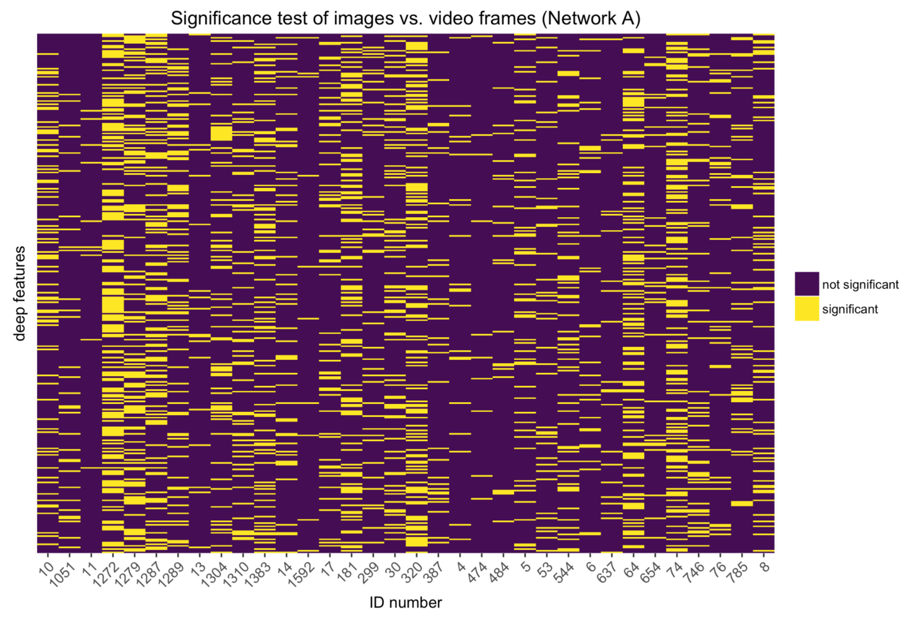


Network	Yaw	Pitch	Media Type
A	+/- 8.06˚ (*SD* = 0.078)	77% correct	87.1 % correct
B	+/- 8.59˚ (*SD* = 0.071)	71% correct	93.3 % correct

Are Feature Values Stable Across Viewpoint/Media Type?

developed feature robustness index: 1) across frontal and profile; 2) across still images and video frames
- analyzed identities with 20+ images in each condition (profile vs. frontal; still images vs. video frames)
- computed *t*-tests to indicate statistically significant differences for top-level features across conditions
  - alpha level Bonferroni corrected (p = .000156)
  - significance acts as an index of feature robustness across conditions

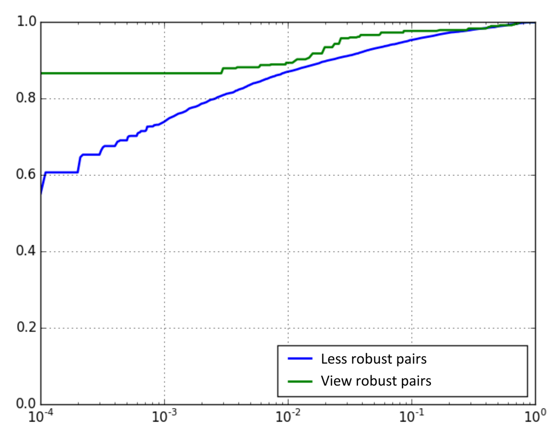




Results:
- robustness to view or media type is *identity-specific* rather than *feature-specific*
  - some identities robustly coded across features—others not

Identity Robustness & Algorithm Performance

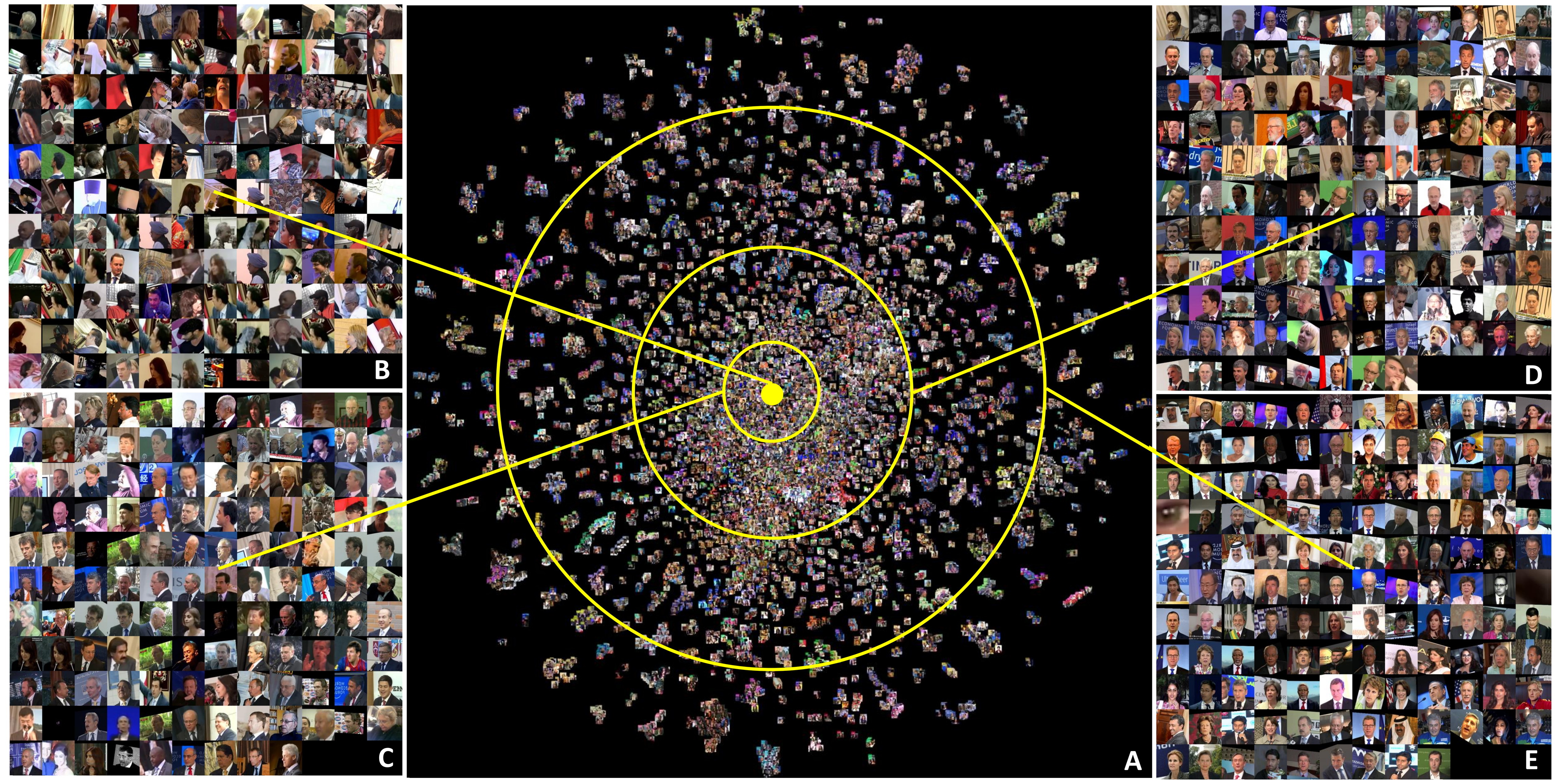
does identity robustness across view affect algorithm performance?
- compared Network A performance in 2 subgroups:
  - 7 most view-robust subjects
  - all other identity pairings
- **Results:**
  - strong face recognition advantage for identities coded robustly



Pictures of identity with most robust coding across views

Pictures of identity with most view-dependent coding

Global Organization of the DCNN Face Space



(A) 2-dimensional t-SNE visualization of full feature space (Network A). Each point represents the image from which the features were computed. (B) Images closest to the center of the full space. (C) More distant 20th, (D) 50th, and (E) 90th percentiles of distances from the center

Image "quality" improves monotonically as distance from origin of top-level feature space increases
- "high quality" images located along periphery → e.g. frontal view, well-lit, little occlusion
- "low quality" images located near origin → e.g. extreme viewpoints, harshly lit, blurry, heavily occluded

Conclusions

1. Image information (pose and media type) preserved in top-level DCNN features trained for face recognition
2. No top-level feature consistently codes view or media type, however some identities are more robust to these changes than others
3. Distance from origin of raw feature space related to "quality"—low quality images close to origin and high quality images close to perimeter

References

- J.-C. Chen, "Unconstrained face verification using deep cnn features," *arXiv preprint arXiv:1508.01722*, 2016.
- Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, *19*(4), 613-622.
- B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus bench- mark a," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1931–1939.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
- Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579-2605.
- Parde, C. J., Castillo, C., Hill, M. Q., Colon, Y. I., Sankaranarayanan, S., Chen, J. C., & O'Toole, A. J. (2016). Deep Convolutional Neural Network Features and the Original Image. *arXiv preprint arXiv:1611.01751*. (to be published 2017 in *Proceedings of the IEEE Workshop on Automatic Face and Gesture Processing*)
- R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi- task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *arXiv preprint arXiv:1603.01249*, 2016
- Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. "An All-In-One Convolutional Neural Network for FaceAnalysis" *arXiv preprint arXiv:1611.00851v1*, 2017
- Sankaranarayanan, S. "Triplet probabilistic embedding for face verifi- cation and clustering," *arXiv preprint arXiv:1604.05417*, 2016.
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815-823).
- Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1701-1708).
- Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356-365.

Scan this QR code to download our conference proceeding from the 2017 IEEE Face and Gesture conference.

For information:
connor.parde@utdallas.edu