2017 IEEE 12th International Conference on Automatic Face & Gesture Recognition

Face and Image Representation in Deep CNN Features

Connor J. Parde[1] and Carlos Castillo[2] and Matthew Q. Hill[1] and Y. Ivette Colon[1] and Swami Sankaranarayanan[2] and Jun-Cheng Chen[2] and Alice J. O'Toole[1]
[1] School of Behavioral and Brain Sciences, The University of Texas at Dallas, USA
[2] Department of Electrical Engineering, University of Maryland, College Park, USA

***Abstract*—Face recognition algorithms based on deep convolutional neural networks (DCNNs) have made progress on the task of recognizing faces in unconstrained viewing conditions. These networks operate with compact feature-based face representations derived from learning a very large number of face images. Although the learned feature sets produced by DCNNs can be highly robust to changes in viewpoint, illumination, and appearance, little is known about the nature of the face code that emerges at the top level of these networks. We analyzed the DCNN features produced by two recent face recognition algorithms. In the first set of experiments, we used the top-level features from the DCNNs as input into linear classifiers aimed at predicting metadata about the images. The results showed that the DCNN features contained surprisingly accurate information about the yaw and pitch of a face, and about whether the input face came from a still image or a video frame. In the second set of experiments, we measured the extent to which individual DCNN features operated in a view-dependent or view-invariant manner for different identities. We found that view-dependent coding was a characteristic of the *identities* rather than the DCNN features– with some identities coded consistently in a view-dependent way and others in a view-independent way. In our third analysis, we visualized the DCNN feature space for 24,000+ images of 500 identities. Images in the center of the space were uniformly of low quality (e.g., extreme views, face occlusion, poor contrast, low resolution). Image quality increased monotonically as a function of distance from the origin. This result suggests that image quality information is available in the DCNN features, such that consistently average feature values reflect coding failures that reliably indicate poor or unusable images. Combined, the results offer insight into the coding mechanisms that support robust representation of faces in DCNNs.**

I. INTRODUCTION

Face recognition algorithms based on convolutional neural networks and deep learning show considerable robustness to changes in imaging parameters (e.g., pose, illumination, and resolution) and facial appearance (e.g., expression, eyewear). This robustness accounts for the impressive gains made by CNNs on the problem of unconstrained face recognition [1], [2], [3], [4], [5], [6]. Performance on datasets such as LFW [7], [8], IJB-A [9], [10], and Mega-Face [11] offer evidence that face recognition by machines can, in some cases, approach human performance [1]. Indeed, human recognition of familiar faces (e.g., friends, family) operates in highly unconstrained environments and over changes in appearance and age that can span decades. This kind of performance remains a goal of automated face recognition systems.

Although humans remain a proof-of-principle that highly invariant face recognition is possible, the underlying nature of the face representation that supports invariance in humans is poorly understood. The nature of the representation captured in DCNN features is similarly elusive. The goal of this paper is to characterize the features that emerge in a DCNN trained for face recognition so as to better understand why they are robust to yaw, pitch, and media type (still image or video frame). The approach we take is to first examine the extent to which the "robust" feature sets that emerge in a CNN *retain* information about the original images. As we will see, DCNNs that show considerable robustness to pose and media type retain detailed information about the images they encode, even at the deepest and most compact level of the network. Second, we explore the view-dependency and media-dependency characteristics of DCNN features. Third, we examine cues within the structure of a DCNN feature space that can provide information pertaining directly to an image's quality.

II. BACKGROUND AND PROBLEM

The problem of image-invariant face perception has been studied for decades in both computer vision [12] and psychology. Traditionally, two classes of models have been considered: a.) representations that capture 3D facial structure and b.) representations based on collections of 2D, image-based views of faces. The former can enable specification of appearance across arbitrary affine and non-affine transformations. The latter can show invariance in any given instance via interpolation to image representations taken in conditions similar to that of the probe image. Notably, this requires "experience" with enough diverse views to be successful across a range of possible probes. Active appearance models [13] comprise an intermediary class, which relies on class-based knowledge of faces, including 3D structure and reflectance-map information for many examples. Although these models can achieve impressive performance in computer graphics representations made from single images, they are not practical for face recognition as they are computationally intense and require high quality, 3D data on diverse classes of faces.

The recent gains made in face recognition can be tied both to the computational power of DCNNs and to the quality and quantity of the training data now available from web-scraping. In theory, the goal of a DCNN is to develop an invariant representation of an individual's face through exposure to a wide variety of images showing that person in different settings, with different poses, and in images that vary in quality. Given enough data, it is expected that the network will be able to learn a representation of an individual that does not rely on these non-static, image-level attributes. Instead,

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.85

the intent is that the learned features represent the invariant information in a face that makes the face unique.

The fact that DCNNs support robust recognition across image transformation does not preclude the possibility that the features used to code faces in these networks also retain information about the image properties. Rather, DCNNs may succeed across appearance-related and image-related variation by incorporating both identity and image parameters into the face codes. This code may support the separation of image and identity for identity verification. This separation may ultimately be achieved at a post-DCNN stage via another type of classifier that operates on image or person representations extracted from the deepest, most compact layer of the DCNN.

The motivation for the present work came from visualizing the way *single identities* cluster in a low-dimensional space derived from the top-level features produced by two recent DCNNs [9], [10]. These DCNNs were developed to work on the JANUS CS2 dataset, an expanded version of the IJB-A dataset [14]. We describe the architecture of the two DCNNs in detail in the methods section. For present purposes, this visualization was done by first applying t-Distributed Stochastic Neighbor Embedding (t-SNE) [15] to the top level features of each network to produce a flattened, 2-dimensional space. t-SNE is a dimensionality reduction technique that uses stochastic probability methods to preserve the high-dimensional Euclidean distances between data points while embedding them in a low-dimensional space. We then used the 2-dimensional t-SNE output for each image as a pair of coordinates, allowing us to plot each image on a 2-dimensional map such that every image was surrounded by its nearest neighbors from the learned feature face. We visualized single identities that had large numbers of images available in the Janus CS2 dataset. Figure 1 shows the t-SNE space for the top level features of 140 CS2 images of Vladimir Putin, extracted from the two DCNNs. Both plots exhibit roughly separable clusters of profile and frontal images of the subject. The blue curves were hand-drawn onto the visualizations to indicate the position of an approximate border.

The clustering patterns seen in Figure 1 suggest that the top-level features produced by both of these DCNN networks preserve salient, view-related information captured in the original image, while still clustering by identity. More generally, this suggests that DCNNs contain a deeper-than-expected representation of the original image in their top-level features. See [16] for a similar finding in a DCNN for object recognition. Notably, the clustered images of Putin still varied substantially in other appearance- and image-based attributes (e.g., age, illumination).

In what follows, we quantify the clustering behavior of image-based attributes in these two DCNNs. This paper is organized as follows. In Section III, we present the networks and the datasets analyzed. In Section IV we use the top-level features of the DCNNs as input into linear classifiers aimed at predicting metadata about the images including yaw, pitch, and media type (still image or video). In Section V, we analyze the extent to which top-level features operate invariantly across

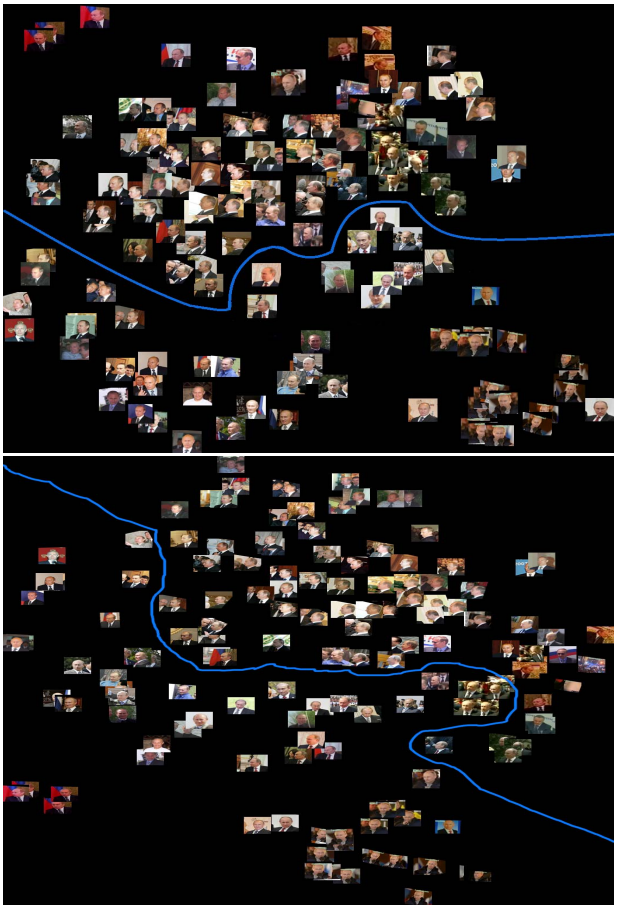

Fig. 1. These figures show the t-SNE visualization of the top level DCNN features for 140 images of Vladimir Putin from the Janus CS2 dataset. The visualizations are based on the 320 top-level DCNN features from Network A [9] (top) and the 512 top-level DCNN features from Network B [10].

viewpoint and media type. In Section VI, we examine the top-level feature space's organization in the context of image quality.

The contributions of this study are as follows. We show that:

- Image quality can be "read out" from the top-level features. This readout can be used to eliminate or attenuate the contributions of low quality image data.
- Individual identities are coded either robustly across viewpoint (invariantly) or with feature dissociations for different viewpoints (variantly). These different types of codes can be detected from the top-level features and be used to predict network accuracy on an individual item basis.
- Image properties such as yaw and pitch are available "for free' from a network trained for identity. No special training is necessary. These attributes can be obtained from an image's top-level feature descriptor.
- In scientific terms, the nature of the representation that

TABLE I
NETWORK A

Name	Filter Size/Stride	Output	Parameters
conv11	3x3x1/1	100x100x32	.28K
conv12	3x3x32/1	100x100x64	18K
pool1	2x2/2	50x50x64	
conv21	3x3x64/1	50x50x64	36K
conv22	3x3x64/1	50x50x128	72K
pool2	2x2/2	25x25x128	
conv31	3x3x128/1	25x25x96	108K
conv32	3x3x96/1	25x25x192	162K
pool3	2x2/2	13x13x192	
conv41	3x3x192/1	13x13x128	216K
conv42	3x3x128/1	13x13x256	288K
pool4	2x2/2	7x7x256	
conv51	3x3x256/1	7x7x160	360K
conv52	3x3x160/1	7x7x320	450K
pool5	7x7/1	1x1x320	
dropout (40%)		1x1x320	
fc6		10548	3296K
softmax cost		10548	
total			5006K

TABLE II
NETWORK B

Layer	Kernel Size/Stride	Parameters
conv1	11 x 11/4	35K
pool1	3 x 3/2	
conv2	5 x 5/2	614K
pool1	3 x 3/2	
conv3	3 x 3/2	885K
conv4	3 x 3/2	1.3M
conv5	3 x 3/1	2.3M
conv6	3 x 3/1	2.3M
conv7	3 x 3/1	2.3M
pool7	6 x 6/2	
fc6	1024	18.8M
fc7	512	524K
fc8	10548	10.8M
Softmax Loss		Total 39.8M

emerges in DCNNs is inconsistent with the assumption that DCNNs eliminate information about the image to directly code identity.

III. General Methods

A. Description of DCNN's

We analyzed the feature-space produced by two DCNNs (Network A, [9]; Network B, [10]) using the JANUS CS2 dataset. Both networks approach the problem by constructing a feature-based representation of all input images using a DCNN. For present purposes, we discuss only the training of these two networks since our analysis focuses only on the top-level features produced by either network. Full details on these networks are available elsewhere.

The base architectures of the DCNNs appear in Tables I and II. In both networks, parametric ReLU (PReLU) were used as the activation function. In Network A, a learned feature space was developed from scratch and produced a 320-dimensional feature vector for each input image. The second network (Network B) builds upon the AlexNet model [17] and assigns each input image a 512-dimensional feature vector. At its lower levels, Network B initially assigns weights based on the values generated by AlexNet and then trains its higher layers using the CASIA-Webface database. Network A also uses CASIA-Webface for training and does so for all layers in the network.

B. CS2 Dataset

The images used for the analyses discussed in this paper were sourced from the JANUS CS2 dataset. This dataset includes approximately 25,800 images of 500 subjects. CS2 is an expanded version of the IARPA Janus Benchmark A (IJB-A) [14], a publicly available "media in the wild" dataset. Some key features of the IJB-A dataset include: full pose variation, a mix of images and videos, and a wider demographic variation of subjects than is available in the LFW dataset. The dataset was developed using 1,501,267 crowd sourced annotations. Baseline accuracies for both face detection and face recognition from commercial and open source algorithms are available in [14].

The original IJB-A dataset included metadata from crowd-sourcing. Here we used metadata provided by the Hyperface system described in [18]. The Hyperface system provides key-point locations to aid in face detection, as well as estimated measurements of face pose (yaw, pitch, and roll).

Of the 25,800 items in the CS2 dataset, we omitted 1,298 items from our analysis. This was due to either Network A's or Network B's inability to compute features for one of these images, or Hyperface's inability to compute the pose of the subject within an image. This left us with 24,502 items that could be considered when training classifiers to predict each metadata attribute of interest.

IV. Predicting Image-Related Metadata from the DCNN Features

For each experiment described in this section we used a bootstrap method to predict different image-related metadata attributes from the top-level features produced by Network A and Network B. Predictions were computed using a linear discriminant analysis (LDA) classifier. We performed 20 iterations of the bootstrap test for each metadata attribute. For each iteration, we randomly selected 18,000 image descriptors to use as training data. We tested the classifier on the remaining 6,502 items. The reported results display the average accuracy of the classifier across the 20 bootstrap iterations.

A. Predicting Yaw

The yaw values provided by the Hyperface system for the CS2 dataset describe the yaw angle of the face in an image. Yaw values are measured in degrees and can vary from -90 (left profile) to +90 (right profile). A yaw value of 0 indicates a frontal pose. Both Network A and Network B employ pre-processing steps to produce a mirrored version of all left-facing images, so that the range of yaw scores is limited to include only positive values. Therefore, we used the absolute value of the yaw scores provided by Hyperface as output

TABLE III
YAW AND PITCH PREDICATION ACCURACY

Network	Yaw	Pitch
A	+/-8.06 degs. (*sd.* 0.078)	77.0% correct
B	+/-8.59 degs. (*sd.* 0.071)	71.5% correct

TABLE IV
MEDIA TYPE

Network	Media Type
A	87.1% (sd. 0.004)
B	93.3 % (sd. 0.002)

for the classifier. In each bootstrap iteration, a classifier was trained to predict the Hyperface yaw values from the DCNN features. Prediction accuracies for both Networks A and B appear in Table III and are surprisingly high. Both networks predict the yaw to within 9 degrees, and are consistent across bootstrap iterations.

B. Predicting Pitch

Pitch estimates for the CS2 dataset were provided by Hyperface and are measured in degrees. Positive pitch scores indicate an upward-looking face, and negative scores indicate a downward-looking face. A score of 0 indicates that a face is looking directly at the camera. The majority of images in the CS2 dataset depict faces with a relatively centered pitch. Given the low frequency of data points with particularly high or low values, we chose to categorically code the pitch scores in this experiment as either *centered* or *deviating*. Centered pitch was defined to include all values between -8 and +8 degrees. Deviating pitch was defined to include all values outside of the centered range.

Using the top-level DCNN features as input, we predicted whether each image in the CS2 data set showed a face with either centered pitch or deviating pitch. The pitch-prediction scores output by the LDA classifier were continuous values from 0 (centered) to 1 (deviating). These values were rounded to the nearest integer (0 or 1) to obtain the final prediction values. The results appear in Table III and are reported as percent correct. As with yaw, the ability of the top-level DCNN features to predict the category of pitch (centered or deviating) of the face in an image was unexpectedly accurate (77.0% and 71.5% correct for Networks A and B, respectively).

C. Predicting Media Type

The media type is provided for all images in the CS2 dataset. Each image originated as either a still photograph or a video frame. An image's media type might be considered a proxy-measure for some aspects of image quality. In the CS2 dataset, the images that originated as still photographs often have better illumination or higher resolution. The images that originated as video frames often come from lower-quality data sources (such as CCTV footage).

We assigned a score of 1 to all images in the CS2 dataset that originated as still photographs, and a score of 0 to all images that originated as video frames. We then applied the bootstrapped classification method to predict media type from the top-level DCNN features produced by both Network A and Network B. The predictions for our test data were continuous values from 0 to 1. These were rounded to the nearest integer (0 or 1) to obtain the prediction values. The results appear in Table IV and are reported as percent correct. Predictions using the DCNN features were highly accurate and consistent for both networks.

Fig. 2. Heat map illustration of view-dependent DCNN features for Network A displayed for each identity in the database with at least 20 frontal and 20 profile images (top). Heat map illustration of the quality-dependent top-level DCNN features for each identity in the database with at least 20 still images and 20 video frames (bottom).

D. Interim summary

The classification experiments showed that metadata from individual images, including yaw, pitch, and media type, was available in the top level DCNN features of both Network A and Network B.

In the next section the goal was to analyze the extent to which individual features operate invariantly, or at least robustly, across pose and media type.

Fig. 3. Image clusters of two individuals (Bono and G. Bush) who were both coded with a majority of view-independent features (312 and 289 of 320 respectively). These clusters show mixed viewpoints aligned closely, which may correspond to distinctive features (e.g. Bono's sunglasses) that are easy to detect across variable views.

V. CNN Features and Invariance: Are features invariant or are people invariant?

A. View (In)variance Coding

We developed an index of feature robustness to examine whether specific features in the learned feature space varied across frontal and profile poses. First, we sub-selected identities in the database (n = 38) for which there were at least 20 frontal images and 20 profile images. Second, within each of these identities, for each of the 320 DCNN features in Network A, we computed a t-test to determine whether the feature's values differed significantly between frontal and profile images of that particular individual. We set the alpha level for statistical significance at 0.000156[1]. The resultant p-values act as an of index of feature invariance for an individual. The results of this analysis are displayed in the top-panel heat map in Figure 2 and are surprising. In the figure, individual identities are displayed across columns and individual features are displayed across rows. We anticipated that individual features would consistently code identities in either a view-dependent or view-invariant way. This would have produced horizontal bands in the heat map, suggesting the consistency of a feature across identities. Instead we found the inverse. Individual *identities* were coded in either a view-dependent or view-invariant way across features. This is indicated by the vertical, banded lines evident in the heat map. More formally, the percentage of features that differentiated faces by viewpoint for an individual was as high as 55.31%. Individual features did not consistently code in a view-dependent or view-independent manner.

To further interpret the results shown in these heat maps, we visualized the most- and least-differentiated identities by selecting the most strongly banded columns from the heat map. Two examples of the most invariant identities appear in Figure 3, showing Bono and Pres. George W. Bush. For Bono, 90.31% of the 320 features were undifferentiated by viewpoint; for Bush, 97.5 % were undifferentiated. These clusters show mixed viewpoints aligned closely–possibly reflecting the presence of distinctive identity features that are

[1]This is a two-tailed alpha level of 0.05, Bonferroni corrected for 320 multiple comparisons.

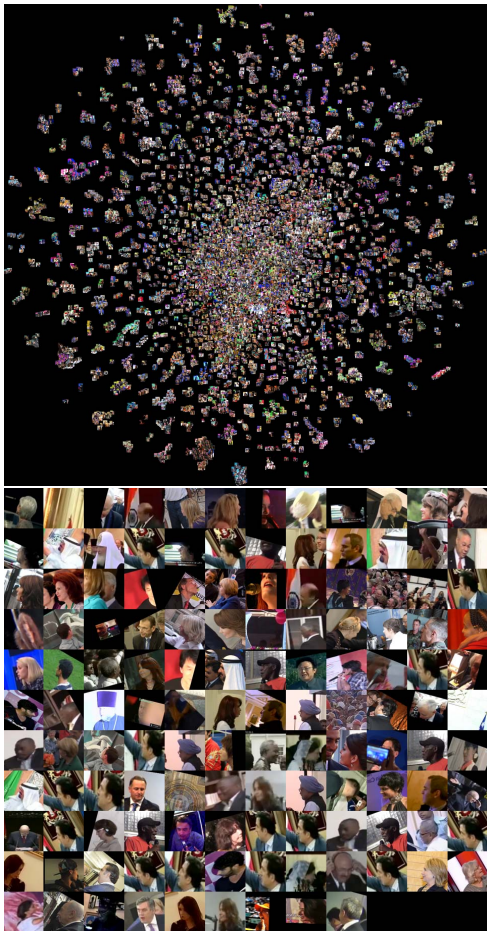

Fig. 4. Results of t-SNE applied to the DCNN top level features of Network A for all 24,502 images (top). An array of the 129 images closest to the center of the space (0.05%) in Network A. The upper-left image is the image closest to the center, and each image's distance from the center grows as you progress across the rows (bottom).

easy to detect across all views (e.g. Bono's oddly tinted sunglasses). Alternatively, when visualizing identities with the most *variant* features, many subjects show strongly separated clusters containing small ranges of similar views. This latter pattern resembles what we saw in Figure 1 for Vladimir Putin. The main point, though, is that *identity* is determining whether the features are organized in a view-dependent or view-invariant manner. Some identities are marked most strongly by characteristics which are static across shifts in pose, while others are marked by the way certain traits appear when seen from different viewing angles.

To determine the extent to which the nature of an identity code (view-variant or view-*in*variant) affects performance in a face recognition algorithm, we conducted the following experiment. We selected the 7 identities coded most invariantly over view-change. Next we compared the performance of Network A on template comparisons involving pairs of these

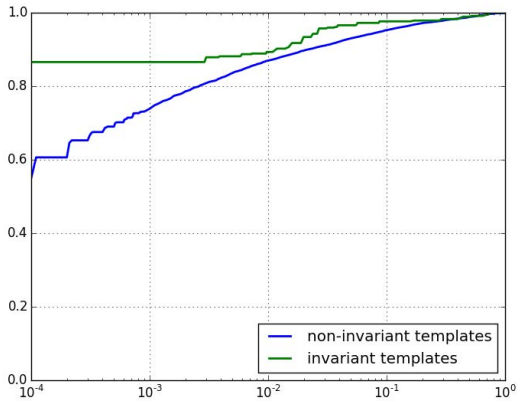

Fig. 5. Identity verification performance of Network A for template pairs where both identities are coded view-invariantly versus for all other template pairs. View-invariance of an identity is characterized by feature values across its images that do not dissociate for frontal and profile views.

7 identities against template comparisons involving pairs of all other identities. Note that a template is defined as a variably sized set of images and video frames of an individual identity, and that the contents of the templates were specified by the Janus protocol. The results of these comparisons appear in Figure 5 and show a strong advantage for recognizing identities that are coded invariantly, over those in which feature values dissociate for frontal and profile images.

B. Media Type (In)variance Coding

We repeated the same approach from the previous section to examine the way media type is coded across features and individuals, developing an index of feature robustness across still images and video frames. First, we sub-selected identities in the database (n = 34) for which there were at least 20 still images and 20 video frames. Second, within each of these 34 identities, for each of the 320 top-level DCNN features produced by Network A, we computed a t-test to determine whether a feature's values differed significantly for images of a particular individual pulled from either a still image or a video frame. We again set the alpha level for statistical significance at 0.000156. In this case, the p-values act as an index of the feature's invariance for coding media type. The results of this analysis are displayed in the bottom-panel heat map in Figure 2 and echo what is seen in the heat map distinguishing frontal and profile views. Individual identities tend to be coded in either a media-dependent or media-independent manner.

VI. When DCNN Features Fail They Leave a Trail

We returned to the use of t-SNE to help visualize the feature spaces of our two recognition networks. This time, rather than analyzing the feature space for a single individual, we applied t-SNE to the DCNN top level features for all 24,502 images in the JANUS CS2 dataset (see Figure 4, top). This was used as an exploratory analysis to help us visualize the DCNN feature space in more detail. The primary insight gained from this visualization is that the images located near the center appear to be of extremely poor "quality", where quality refers to a wide range of issues that would make the person in the image difficult to detect or identify. We therefore examined the images in order of closeness to the center of the raw feature space. We defined the center of the feature space to be equal to the space's origin, where all feature values are zero. Figure 4 (bottom) shows an array of the 129 images closest to the center of the space (0.05%) produced by Network A, arranged across the rows and starting from the image closest to the center. As seen in the array, the images closest to the center of the feature space are affected by a range of problems including extreme views, strong occlusion, blurring, distortion, and lack of an identifiable face.

Does distance from the center of the DCNN feature space index image quality? To examine this, we pulled images from different distances to the center of the space. We ranked the images according to their distance from the origin. Figure 6 shows 129 sampled images from each of the 20th, 50th, and 90th percentiles of these ranked distances. This figure illustrates that face quality seems to increase with distance from the center of the DCNN feature space.

VII. Conclusions

The analyses discussed in this paper point to the following conclusions. First, the top-level features produced by DCNNs trained for face recognition retain a surprising amount of information about the original input image. Yaw, pitch, and media type were readily available in the top-level DCNN codes of both networks we examined and could be classified with high accuracy.

Second, when characterizing the extent to which individual top-level DCNN features coded either view-dependent or view-invariant information about faces, we found that the tendency to develop a view-dependent code was a characteristic of the *identities* rather than the features. This suggests that some identities in this dataset present with appearance-based characteristics that are easy to detect and code across viewpoint, whereas other identities are marked by characteristics that tend to vary according to view. The data-dependent manner in which these face codes are produced is intriguing because it suggests that DCNNs and the human visual system alike might need to exploit both types of codes in order to operate efficiently and accurately in unconstrained viewing conditions. Notably, this general finding of data-dependency also held for media type. Some identities are consistently coded across different media types, while other identities have more disparate codes.

Finally, we made the unexpected finding that an image's distance from the origin of the DCNN's top-level feature space could be used to index the quality of an image. Further, the tendency for low-quality images to cluster at the origin of the feature space was notable since the degraded quality emanated from many distinct sources. This allows for the generic identification of images with limited or unusable information

Fig. 6. Images (n=129) sampled at the 20th (top), 50th (middle), and 90th (bottom) percentiles of ranked distances from the origin. Face image quality seems to increase with distance from the center of the DCNN feature space.

about a particular identity. Because low quality images cluster around the origin and quality increases with distance from the origin, we might speculate that feature descriptors with a large magnitude reflect robust identity information. This suggests a new method for screening out poor quality imagery in DCNNs.

In summary, a more in-depth look at the compact top-level feature codes learned by DCNNs trained for face recognition gave insight into the nature of the learned representation. These analyses point to data-dependent flexibility in the type of codes that emerge at the DCNNs top level, as well as the possibility to separate low- and high-quality imagery.

VIII. ACKNOWLEDGMENTS

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.

REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1701–1708.

[2] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Proceedings of the British Machine Vision*, 2015.

[3] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *BMVC*, vol. 60, no. 2, 2013.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] G. Hu, Y. Yang, D. Yi, J. Kittler, W. J. Christmas, S. Z. Li, and T. M. Hospedales, "When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition," *CoRR*, vol. abs/1504.02351, 2015. [Online]. Available: http://arxiv.org/abs/1504.02351

[6] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2518–2525.

[7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: a database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, 2007.

[8] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and Simile Classifiers for Face Verification," in *Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV)*, October 2009.

[9] J.-C. Chen, "Unconstrained face verification using deep cnn features," *arXiv preprint arXiv:1508.01722*, 2016.

[10] S. Sankaranarayanan, "Triplet probabilistic embedding for face verification and clustering," *arXiv preprint arXiv:1604.05417*, 2016.

[11] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Megaface: A million faces for recognition at scale," *arXiv preprint arXiv:1505.02108*, 2015.

[12] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. PAMI*, vol. 20, pp. 39–51, 1998.

[13] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings, SIGGRAPH'99*, 1999, pp. 187–194.
[14] 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 benchmark a," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1931–1939.
[15] L. van der Maaten, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, 2008.
[16] H. Hong, D. L. Yamins, N. J. Majaj, and J. J. DiCarlo, "Explicit information for category-orthogonal object properties increases along the ventral stream," *Nature neuroscience*, vol. 19, no. 4, pp. 613–622, 2016.
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[18] 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.