Discovering Interpretable Feature Directions in the Embedding Space of Face Recognition Models

Supplementary Material

Richard Plesh*[1] Janez Križaj*[2] Keivan Bahmani[1] Mahesh Banavar[1]
Vitomir Štruc[2] Stephanie Schuckers[1]
[1]Clarkson University, United States [2]University of Ljubljana, Slovenia

In this supplementary material we provide some larger images to better observe the properties of the discovered feature directions.

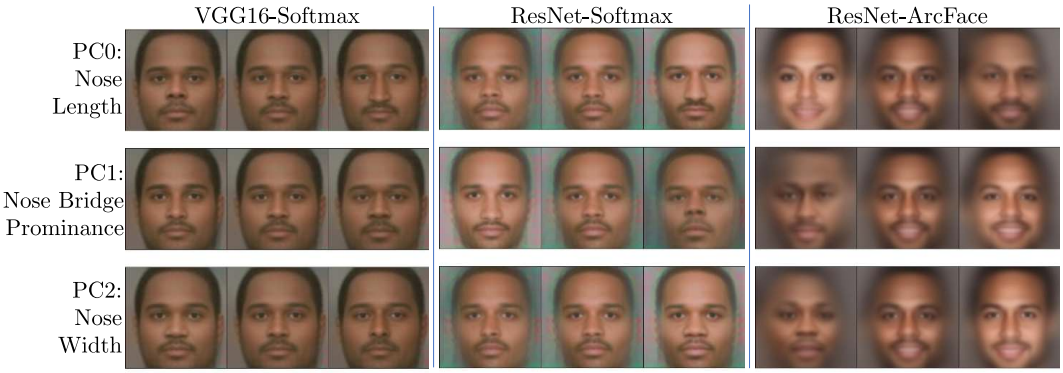

Figure 1. **Visualization of movements along the Nose directions.** Results are presented for 3 FR models (in columns) and 3 SSI deep features relating to the Nose for three different facial recognition models, VGG16-Softmax (Left), ResNet-Softmax (Middle), and ResNet-ArcFace (Right). Rows represent the different principal components of the PCA subspace, labeled with their apparent semantic meaning. PC0 is the most significant component of template variance, suggesting it is most important for face-matching decisions.

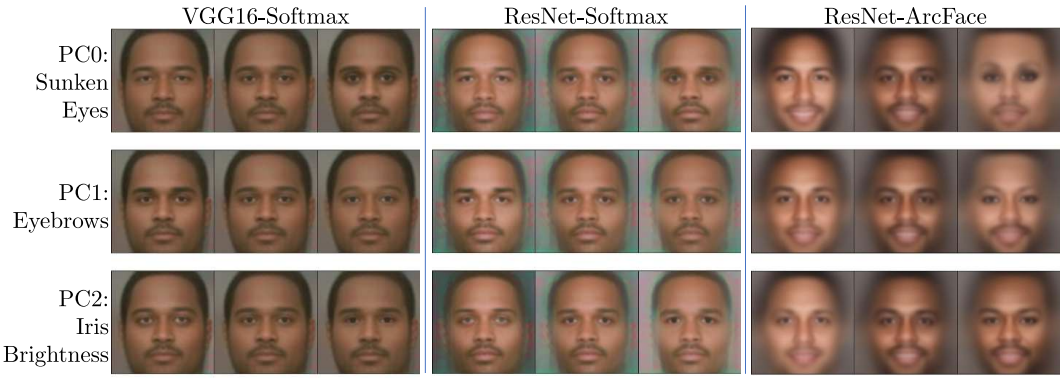

Figure 2. Eye Region Features: SSI deep features relating to the eye region for three different facial recognition models, VGG16-Softmax (Left), ResNet-Softmax (Middle), and ResNet-ArcFace (Right). Rows represent the different principal components of the PCA subspace, labeled with their apparent semantic meaning. PC0 is the most significant component of template variance, suggesting it is most important for face-matching decisions.

*Equal contribution.

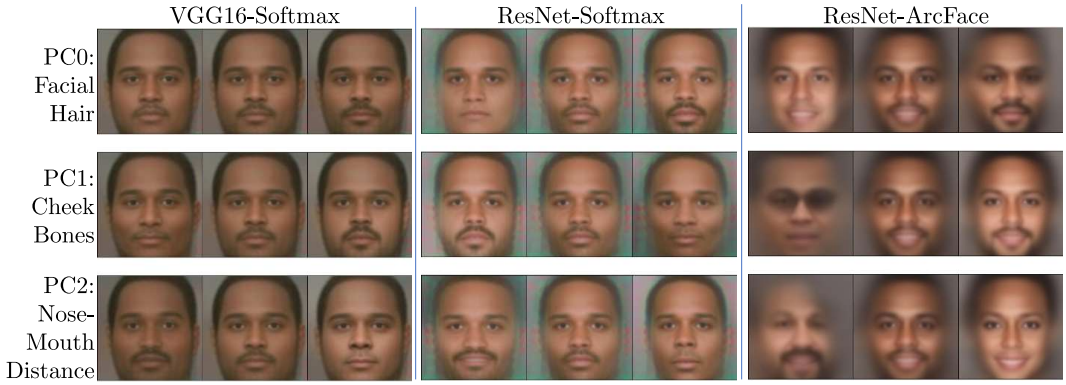

Figure 3. Lower Face Region Features: SSI deep features relating to the lower face region for three different facial recognition models, VGG16-Softmax (Left), ResNet-Softmax (Middle), and ResNet-ArcFace (Right). Rows represent the different principal components of the PCA subspace, labeled with their apparent semantic meaning. PC0 is the most significant component of template variance, suggesting it is most important for face-matching decisions.

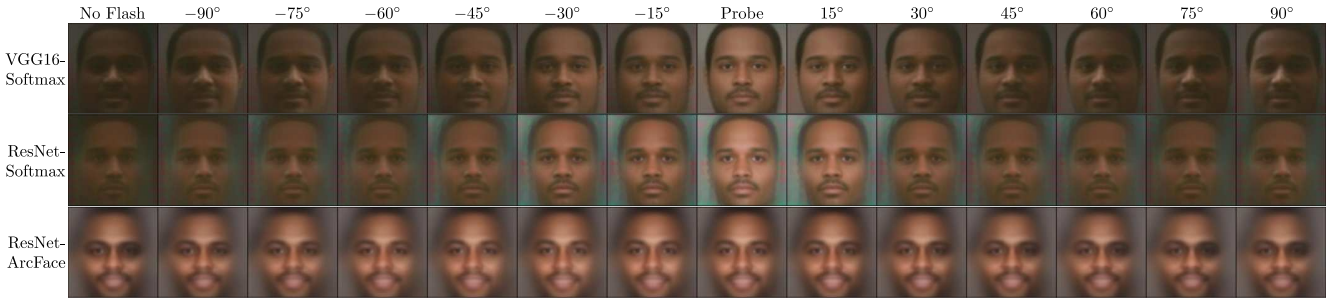

Figure 4. **Centroid-based Illumination Direction Features.** With the proposed approach, we are able to discover feature directions associated with the angle of illumination (in 15° increments) for three different face matchers. Best viewed zoomed-in.

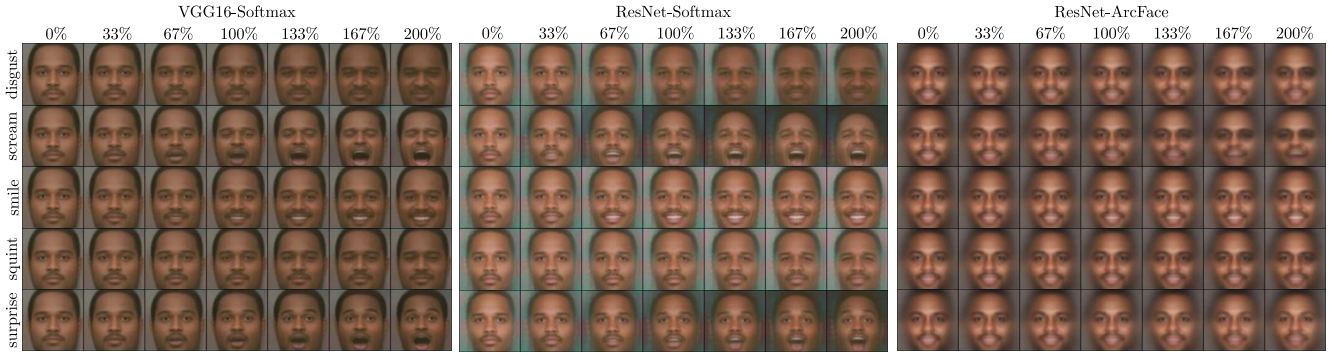

Figure 5. Centroid-based Expression Features: By using average template centroid differences we discovered the deep feature mapping associated with 5 different facial expressions (disgust, scream, smile, squint, and surprise), for three different face matchers. The leftmost column represents the probe image before modification. The template is adjusted in the learned-expression direction by up to 200% of the average centroid difference. As can be seen in the columns right of 100%, the expression strength relationship carries beyond the average centroid distance.

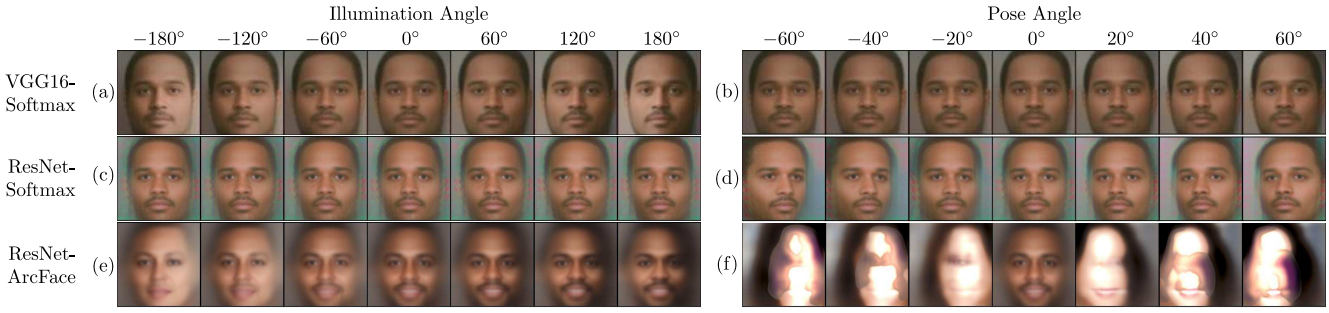

Figure 6. Learned Deep Features using Linear Regression: We learn features for particular deep features by fitting a linear model to predict dataset labels given deep features. The coefficients for the fitted linear regression are used for visually evaluating the ability of the regression to model the relationship in rows a-f. The ability of the linear model to learn the illumination and pose angle appears to differ significantly. This is likely due to the non-linear shape of the feature manifold in the template space, an intuition backed up by the arced structures observed in the MDS plots for Illumination and Pose angle.