Can Shadows Reveal Biometric Information?
– Supplementary Material –

Safa C. Medin[1], Amir Weiss[1], Frédo Durand[1], William T. Freeman[1], and Gregory W. Wornell[1]

[1]Massachusetts Institute of Technology

1. Implementation Details

In the following, we explain our data generation processes, the implementation of our maximum likelihood-based learning algorithm, and the training procedure of our neural network classifier in detail.

1.1. Maximum Likelihood Analysis

Dataset generation. According to the coordinate system definition shown in Figure 5(a), we have the following configurations and variations in the face dataset for each of the $M = 16$ identities. Hereafter, with a slight abuse of notation, we denote a point in the 3D-space by (x, y, z), where the unit of measure is meters.

- We fix the position of the faces at $(1.65, 0.0, 1.15)$.
- We vary the facial expressions by sampling the expression coefficients from $\mathcal{N}(\mathbf{0}_{k_{\text{exp}}}, c\,\mathbf{I}_{k_{\text{exp}}})$, which changes the face shape according to Equation (1). Here, c controls the amount of variations, which we set as $c = 0.5$.
- We rotate the faces around y– and z–axes, which we refer to as elevation and azimuth, respectively. We sample both elevation and azimuth uniformly from $[-15°, 15°]$.
- We simulate a white spotlight directed to the face. We sample its location uniformly along the line connecting $(0.15, -0.5, 1.50)$ and $(0.15, 0.5, 1.50)$.

We render face images of resolution 128×128 with Mitsuba2 [3] using 100 samples per pixel. Rendering one image takes ~ 100 miliseconds on NVIDIA GeForce RTX 2080 Ti GPU. We simulate a rectangular occluder with a 512×512 image that contains a square with a diagonal length of 400 pixels. After proper scaling in pixel values to ensure the conservation of energy [1, 4], we convolve this image with the rendered face images to obtain the shadow images, which we downsample to 128×128 resolution. Hence, we have $n = 128^2 = 16384$, i.e, each observation in the dataset is a 16384-dimensional vector.

Experiments. We analyze the ML algorithm for different SNR levels ranging from -35 dB to 80 dB with step size 5 dB. For each data point, running the algorithm for $M = 16$ identities takes ~ 150 minutes. For each data point shown in Figure 6, we compute the accuracies by averaging the results over 5 independent trials.

A particular setting that requires careful attention is when the noise variance σ^2 is sufficiently small, which makes the covariance matrices $\{\mathbf{Q}^m\}$ nearly singular. This makes the "tail" eigenvalues of the sample covariance matrix $\hat{\mathbf{Q}}^m$ close to 0, causing numerical instabilities during inversion. To resolve this, if the estimated noise variance $\widehat{\sigma}^2$ drops below some chosen threshold σ^2_{th}, we work with the *pseudoinverse* of the covariance matrix as $(\hat{\mathbf{Q}}^m)^{-1} \triangleq \mathbf{U}^m(\mathbf{\Lambda}^m)^+(\mathbf{U}^m)^T$ where $(\mathbf{\Lambda}^m)^+ \triangleq \operatorname{diag}(1/\lambda^m_1, \ldots, 1/\lambda^m_p, 0, \ldots, 0)$. Here, λ^m_p is the smallest eigenvalue larger than a refined threshold $\bar{\sigma}^2_{\text{th}} \triangleq \max(\sigma^2_{\text{th}}, k\widehat{\sigma}^2)$, which is heuristically chosen to ensure the continuity of the algorithm between numerically singular and nonsingular covariance matrix regimes. In our experiments, we set $\sigma^2_{\text{th}} = 10^{-6}$ and $k = 5$.

1.2. Neural Network Classifier

Dataset Generation. According to the same coordinate system definition, we have the following configurations and variations in the synthetic data, which we use to train our neural network classifier.

- As before, we vary the facial expressions by sampling the expression coefficients from $\mathcal{N}(\mathbf{0}_{k_{\text{exp}}}, c\,\mathbf{I}_{k_{\text{exp}}})$ with $c = 0.5$.
- We rotate the faces around y– and z–axes, and sample both elevation and azimuth uniformly from $[-30°, 30°]$. Here, positive angles indicate clockwise rotations with respect to the xz– and xy–planes.
- We sample the position of the face uniformly along the line connecting $(1.55, 0.0, 1.15)$ and $(1.75, 0.0, 1.15)$.

That is, face position varies along the x–axis as variations in other axes are accounted for in the data augmentation step, where the final images are randomly cropped.

- We simulate a white spotlight directed to the face. We sample its location uniformly along the line connecting $(0.15, -1.0, 1.50)$ and $(0.15, 1.0, 1.50)$.
- Occluders are located 0.7 meters from the wall and situated on the ground, where we measure the distance from the center of mass of the occluder.

Training details. We train our classification network with the synthetic data for 30 epochs, using cross entropy loss and Adam optimizer [2] with a learning rate of 0.0001. We augment the training data by flipping the images randomly, resizing them to 280×280 resolution and randomly cropping a 224×224 patch from these images. At test time, we resize the images to 280×280 resolution and center-crop the 224×224 patch from them. In our experiments, we pick the epoch with the highest test accuracy, and use the network at that epoch as our baseline, on which we apply domain adaptation by updating the batch normalization statistics.

2. Fundamental Limitations

The performance of identity classification systems from shadows (like the one that we present in this work) is fundamentally influenced by several factors such as the similarities between identities of interest; the amount of variations covered in head poses, facial expressions, lighting conditions, occluding objects; and other less trivial factors.

As an illustrative example, it is expected that the faces under extreme head poses and lighting conditions are more likely to be classified incorrectly. To support this, we investigate the effect of the face appearance on the results by analyzing the impact of the head pose and light source location on the predictions. We illustrate our findings in Figure 1, where we show elevation-azimuth and light source position-azimuth plots for correctly and incorrectly classified examples. In the left plot, we observe that the elevation has an evident impact on classification performance, where faces with higher elevation are more likely to be misclassified. This can be explained by our scene geometry shown in Figure 5(a), where a more direct view of the face is reflected on the wall when the elevation is low. Taking the averages over all samples shown in the plot, incorrectly classified examples have an average elevation of $+2.83$ degrees whereas correctly classified examples have an average of -6.73 degrees. In the right plot, we observe a positive correlation (a Pearson correlation of 0.22) between the light source position (measured along the y-axis) and the azimuth for correctly classified examples, for which the faces are illuminated with lower incidence angles. That is, faces with less cast shadows are more likely to predicted correctly.

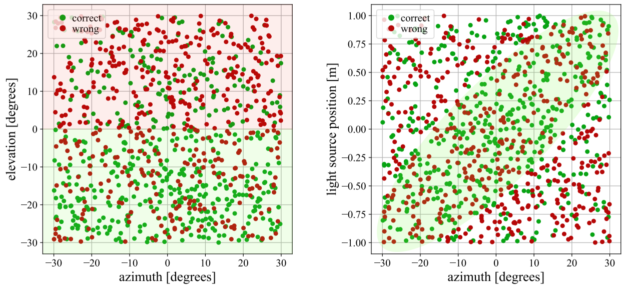

Figure 1: Correctly and incorrectly classified examples depending on azimuth, elevation and light source position. We observe that faces with lower elevations and less cast shadows are more likely to be classified correctly.

References

[1] Ganesh Ajjanagadde, Christos Thrampoulidis, Adam Yedidia, and Gregory Wornell. Near-optimal coded apertures for imaging via nazarov's theorem. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7690–7694. IEEE, 2019.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[3] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics*, 38(6):1–17, 2019.

[4] Adam Yedidia, Christos Thrampoulidis, and Gregory Wornell. Analysis and optimization of aperture design in computational imaging. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4029–4033. IEEE, 2018.