A model of familiar and unfamiliar 3D face recognition

Kelsey R. Allen[1], **Ilker Yildirim**[1], **Joshua B. Tenenbaum**[1]
[1]Department of Brain and Cognitive Sciences, MIT
krallen@mit.edu, ilkery@mit.edu, jbt@mit.edu

Abstract

We are very familiar with certain objects; we can recognize our laptops, cars, friends and collaborators instantaneously behind or among other objects, under heavy occlusion, when lit under an unusual light, or when viewed at an extreme angle, where we might be slow or even fail to recognize an arbitrary exemplar that we have little experience with. This difference between the recognition of familiar and unfamiliar exemplars is particularly important in the social domain – with a quick glance at their face, even under very poor viewing conditions, we are able to recognize not only the identities of family and friends, but also can catch their emotional or physiological state. However, we are also able to determine if two very different views of a stranger are indeed of the same person, rather than two separate people. How do we allow for exceptionally fast recognition of familiar exemplars and maintain the ability to do rich, but slower, recognition of unfamiliar objects? Here, in the domain of faces, we describe a method for implementing a memory module via an identity classification network, which interacts with a general recognition model and a 3D face model in order to make very fast, rich inferences about observed faces. We additionally show that an online clustering algorithm can allow for same day recognition of unfamiliar faces, and how memories can be consolidated into a new recognition network at the end of the day. This combination of deep learning, online Bayesian inference, nonparametric Bayesian clustering, and inverse graphics provides, to our knowledge, the first implemented account of familiar and unfamiliar face recognition.

1 Introduction

Recognizing objects under different viewing and lighting conditions is a task most of us are able to easily perform. With a quick glance at a person across the room, we can immediately determine whether they are someone we know or not, even if we catch a side or frontal view of their face, or if their face is mostly occluded by a drink, a hand or a window. If this person is a close friend, we are able to pick up on minute shifts in their facial expressions which cue us into their emotional and physiological state. For strangers, we may not be sensitive to these subtle changes, but we are still able to determine if the person we talk to at the end of the night was the same person we saw earlier in the evening. Both abilities require a rich model of a person's face which is invariant to lighting and pose conditions.

Some previous work has used recognition models (following the framework of a Helmholtz machine [1]) to efficiently invert a generative model for 3D faces by providing good initial guesses for the latent parameters [2]. However, there the recognition model is used to predict the latent parameters directly, and thus cannot learn parameters for individual, familiar people over many exposures. In this paper, we show how to combine this model for unfamiliar faces with a recognition model that operates on facial identity, including an online non-parametric clustering approach for identifying new faces. Identity is treated as the fundamental unit of operation, and each identity has an associated set of latent parameters that generate a face. Thus, the memory can be thought of as a mixture model in latent space, where each remembered person has their own distinct cluster with a tight variance, and new faces can likewise be added over the course of the day. The rest of the paper describes the model in more detail, and shows some quantitative results for how the model outperforms the previous approach. We additionally show how this model captures some results from human behavioral experiments in familiar/unfamiliar face processing tasks.

2 Model

2.1 Generative model

In the generative model, we consider a mixture model on top a 3D face model described in [3]. In the mixture model, we have clusters corresponding to identities. Each of these clusters has an associated 400 dimensional latent vector

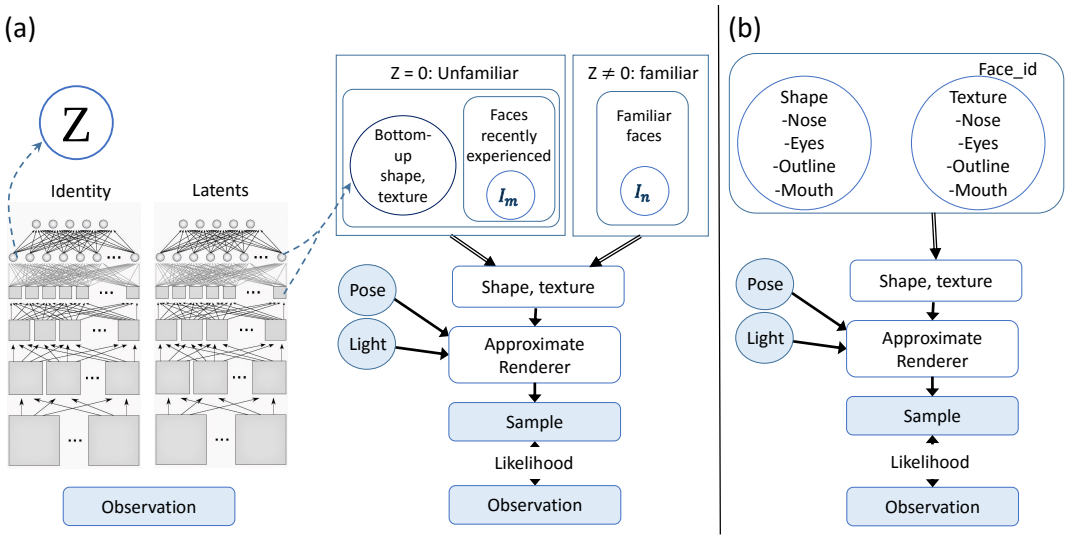


Figure 1: Pipeline used for recognizing an observed image. (a) shows our modified model which includes a mixture distribution over identities. The mixture distribution depends on both the indicator variable Z (as determined from the identity recognition model) and the recent history of experience. (b) shows the standard generative model operating only over the base distribution in latent space as in [4].

representing the shape S and texture T of the face, with a mean vector μ_i and isotropic variance Σ. Σ has been set to 0.01 to represent perceptually indistinguishable identities.

The prior over the space of shapes and textures is taken from the Morphable Face Model [3] as in [4]. The model is obtained from a set of 200 laser scanned heads of people, and provides a mean face in a part based manner (eyes, nose, mouth and outline), as well as a covariance matrix to determine new faces by eigendecomposition. Thus, the shape and texture are gaussian distributed, with $N(\mu_{\text{shape}}, \Sigma_{\text{shape}})$ and $N(\mu_{\text{texture}}, \Sigma_{\text{texture}})$.

2.2 Recognition models

In this work, we use two recognition networks to account for both identity and general face recognition capabilities. The general face recognition model has been previously described in [2], where Yildirim et al. trained a linear model with inputs from the top convolutional and first fully connected layers of an ImageNet trained CNN to predict the shape and texture variables of a set of generated faces. This recognition model will be referred to as the "latents recognition model".

The novel recognition model contributed by this work is a classification network for identity. Here, we use the first fully connected layer of the network described in [5] which was also trained on ImageNet, adding a linear layer from the 4096 activation units to the 30 identities. We use softmax to determine class probabilities. Two versions were created: an "older" network which knows 80 identities, and a "younger" network which only knows 30. Each network is trained using 400 different viewing conditions for each identity.

2.3 Pipeline

An observed image I_D is fed to both the identity recognition network and the latents recognition network. The first task is to determine whether this is a familiar person. In order to do this, the entropy across the n dimensional output vector from the identity recognition network is calculated, and if it falls below an empirically determined threshold, the face is said to be familiar and the indicator variable Z is set accordingly. In this case, our mixture distribution in latent space is composed of tight clusters for each of our familiar faces, and we draw a sample from the relevant cluster in order to initialize the forward inference.

If the entropy is above the threshold, the face is unfamiliar. There are two possible cases for unfamiliar faces: either the face is completely novel, or it is the same face as a person we saw today but who is not stored in long term memory (ie. as a class label in the identity recognition network). Thus, the mixture distribution we are operating over is composed of the clusters representing unfamiliar faces we have recently observed. In order to sample from this distribution, we

consider a sequential clustering algorithm with a CRP prior on cluster assignments for observation i:

$$P(k) = \begin{cases} \frac{n_k}{i+\alpha} & (n_k > 0, \text{old cluster}) \\ \frac{\alpha}{i+\alpha}, & (n_k = 0, \text{new cluster}) \end{cases}$$

where α is chosen to be 1 for the following experiments.

The likelihood of a specific cluster k for the current observation is computed in image space. Each image is rendered at the same pose and lighting as the observed image, and we assume a gaussian likelihood in pixel space. While the likelihood for already existing clusters is trivial to compute, determining the likelihood of a new cluster is more complex. We approximate it using an image rendered from the bottom-up guess on the latent parameters from the recognition network (I_{bu}).

$$P(z_k|k) \propto \begin{cases} e^{-\frac{(I_D - I_k)^2}{2\sigma^2}} & (n_k > 0, \text{old cluster}) \\ e^{-\frac{(I_D - I_{\text{bu}})^2}{2\sigma^2}} & (n_k = 0, \text{new cluster}) \end{cases}$$

We then choose as our estimate the local MAP, which gives us a good initialization for the latent parameters of the new face, even when we are unfamiliar with the observed individual. After forward inference, the cluster means in latent space are updated, reflecting the potential addition of a new cluster member. For these experiments, σ is set to 0.01.

2.4 Inference

Finally, the forward inference machinery is described in detail in [2] and [4]. After initializing the latent parameters as above, multi-site elliptical slice sampling [6], a form of MCMC, is performed on the vectors for shape and texture. At each MCMC sweep, we iterate a proposal-and-acceptance loop on four shape vectors and four texture vectors, where an image is rendered based on the latent parameters, pose and lighting, and compared to the observed image. The log-likelihood of this image is gaussian in pixel space.

3 Results

Here we show the performance of the model in several different scenarios. In order to run experiments, we generated a set of 100 identities, each rendered under 500 different pose and lighting conditions. We trained the identity recognition model on a subset of 30 identities (the young network) from 400 randomly selected viewpoints, leaving 100 viewpoints for testing. On the test set, the model achieves 99.82% accuracy. We additionally trained a second identity recognition model corresponding to knowing 80 faces (the "old" network) using the same number of viewpoints, which achieves 99.42% accuracy on the test set. Training was performed using stochastic gradient descent with a learning rate of 0.001 with a maximum number of iterations of 1500.

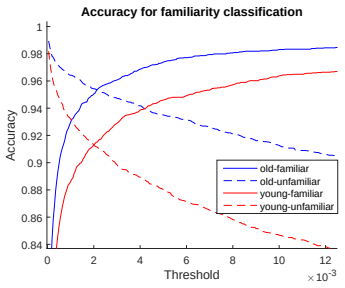


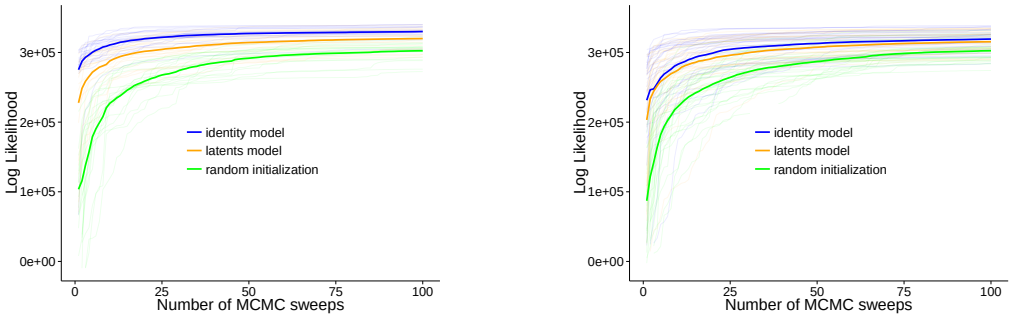
Figure 2: Familiarity classification based on a trained entropy threshold.

3.1 Familiarity classification

At the first stage of the pipeline, an incoming face is deemed to be familiar or unfamiliar based on the relative uncertainty of the network in the identity classification task. This is measured using the entropy over the classes. To determine a threshold for the pipeline, we use 400 views of 40 faces: 20 familiar, 20 unfamiliar. Training was performed for the young network on half of the faces, and then testing was done on both the young and old networks using the determined entropy threshold. This resulted in accuracies of 91.3% and 95.1% respectively. The older network seems to outperform the younger network in this task, which qualitatively matches the behavioral findings of Germine et al. [7]. Conveniently, the threshold trained for the young network works very well for familiarity classification in the older network as seen in Figure 2.

3.2 Initialization

We then check whether the addition of the identity model helps speed up our inference procedure compared to random initializations or initializations taken only from the latents recognition network. For this experiment, we randomly sampled 20 known and 20 unknown faces rendered at 3 different viewing/lighting conditions from our dataset. Each face was then presented to the identity model pipeline as described earlier, but without the added "online clustering" for unfamiliar faces. The resulting log likelihood trajectories are shown in Figure 3.



(a) Familiar faces

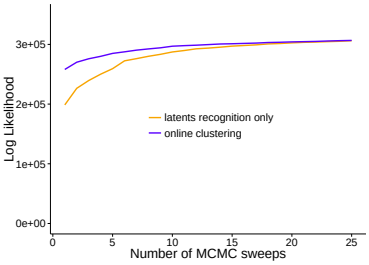
(b) Unfamiliar faces

Figure 3: Inference traces for (a) 20 familiar and (b) 20 unfamiliar faces.

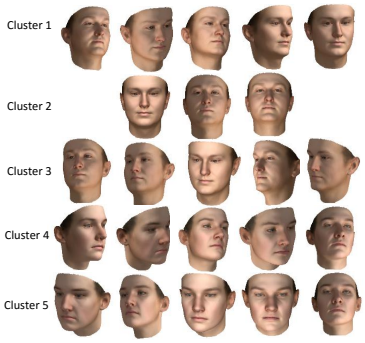
3.3 Online clustering

For this experiment, we chose 6 identities, each with two viewing conditions. The first viewing condition was chosen to be the frontal view under random lighting conditions, and the second view was chosen to be either a left or right side view under different lighting conditions from the frontal view. The 6 frontal view conditions were shown to the network first, followed by the 6 side views in scrambled order. The online clustering algorithm was able to successfully cluster $4/6$ of the second views of an identity (while it incorrectly made new clusters for the other two), and correctly made new clusters for all 6 of the initial views.

We additionally show that this method over-clusters unfamiliar faces in a qualitatively similar way to humans on the Jenkins face task [8]. In this task, 20 views of two Danish politicians were shown to participants, who were asked to state how many identities they perceived. Those who were not Danish consistently over clustered the space into 5-10 identities, while those who were familiar with the politicians correctly clustered the space into 2 identities. We perform a similar experiment with our model using the online clustering method we described. The model produces 5 distinct clusters of identities, with populations 7, 3, 9, 7 and 12. One view from each identity was incorrectly guessed as "familiar" due to the entropy threshold. The found clusters do not cross the original identity boundaries, and are reasonably consistent with differences in pose.



(a) Average inference traces for online clustering



(b) Sample faces from the found clusters in the online sequential clustering experiment.

Figure 4: Results from online clustering in two experiments.

3.4 Expanding the set of known identities

Finally, we tested how well the network could consolidate new faces at the end of the day. To do this, we made a training set by sampling 20 views from 3 novel identities, as well as 5 views from each of our previous 30 known identities. We initialized the weights of the linear layer to those from the previous day for the known identities, and randomly initialized the weights for the unknown identities. After training, we achieve 89.84% accuracy on the old faces, and 89.70% on the new faces.

4 Conclusions

We have shown an effective way to build a fast inference model for familiar and unfamiliar face recognition which combines approximate online bayesian sequential clustering with two deep networks for classifying identity for familiar faces and latent parameters for unfamiliar faces. By combining a network specializing in subclass identification (facial identities here) with a network that can give a rough approximation of latent parameters for a superclass, we are able to speed up inference for familiar subclasses, while still performing well on new exemplars. We hope that such an approach may be applicable to other types of visual objects, allowing for a general framework for the problem of subclass identification. In future work, we would like to investigate other clustering schemes, as well as novel ways for updating the recognition model during the consolidation phase, and finally using one-vs-all classifiers for identity.

Acknowledgements

We would like to deeply thank Tejas Kulkarni for sharing the Picture code which he developed. This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

[1] Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural Computation*, 7(5), pp. 889904.

[2] Yildirim, I., Kulkarni, T.D., Freiwald, W.A. & Tenenbaum, J.B. (2015) Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations, In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society*.

[3] Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 187194.

[4] Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: An imperative probabilistic programming language for scene perception. In *Computer Vision and Pattern Recognition*.

[5] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[6] Murray, I., Adams, R. P., & MacKay, D. J. (2009). Elliptical slice sampling. *arXiv preprint arXiv:1001.0175*

[7] Germine L.T., Duchaine B., Nakayama K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, 118, pp. 201210.

[8] Jenkins, R., White, D., Van Montfort, X., Burton, A.M. (2011). Variability in photos of the same face. *Cognition*, 3, pp. 313 - 323.