Integrating identification and perception: A case study of familiar and unfamiliar face processing

Kelsey R. Allen (krallen@mit.edu), **Ilker Yildirim** (ilkery@mit.edu), **Joshua B. Tenenbaum** (jbt@mit.edu)
Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

We are very familiar with certain objects; we can quickly recognize our cars, friends and collaborators despite heavy occlusion, unusual lighting, or extreme viewing angles. We can also determine if two very different views of a stranger are indeed of the same person. How can we recognize familiar objects quickly, while performing deliberate, perceptual inference on unfamiliar objects? We describe a model combining an identity classification network for familiar faces with an analysis by synthesis approach for unfamiliar faces to make rich inferences about any observed face. We additionally develop an online non-parametric clustering algorithm for recognition of repeatedly experienced unfamiliar faces, and show how new faces can become familiar by being consolidated into the identity recognition network. Finally, we show that this model predicts human behavior in viewpoint generalization and identity clustering tasks, and predicts processing time differences between familiar and unfamiliar faces. **Keywords:** face recognition; analysis-by-synthesis; neural networks; computational

Introduction

Walking to work in the morning, we may encounter familiar faces, buildings in which we regularly have meetings, or the passing car of a colleague on their way to the office. Glancing at these objects, we can effortlessly perceive details of their shape and appearance; we can also recall associated identity-specific content (like which colleague owns that car, or the name of that office building). These two abilities can be thought of as object perception and identification respectively. Thus, perception is noticing the shape, texture and expression of a face, even if the person is a stranger. Identification is recognizing a close friend even if she has had a dramatic hair cut and is wearing a new pair of large, dark sunglasses.

Recent work in machine vision has made significant progress on both of these problems, but very different techniques have been applied to each problem. Dramatic gains in object *identification* have come from deep neural networks (Simonyan & Zisserman, 2014). These methods learn to be invariant to certain object transformations and small differences in appearance. However, they require large amounts of training data, and do not generalize to novel objects without at least re-training the top classification layers. Rich object *perception* has become possible using an alternative approach to vision, known as "analysis by synthesis" or "inverse graphics". This approach posits that the perceptual system models the generative processes that form images from scenes, and works backwards from an observed image to infer the scene most likely to have generated it (Kulkarni, Kohli, Tenenbaum, & Mansinghka, 2015). Inverse graphics methods can often recover the fine-grained geometrical and physical properties of objects in an image, but are much slower than feed-forward neural networks and have not yielded practical object identification or recognition systems.

Figure 1: How many people are depicted here?

However, object identification and object perception are not separate. For example, extensive research on face perception has studied familiar face recognition, unfamiliar face perception, and the dynamics of how new faces are recognized differently as they become increasingly familiar (O'Toole, Edelman, & Bülthoff, 1998; Burton, Bruce, & Hancock, 1999). Our goal is to build models of these two aspects of vision and their interaction in the domain of faces, and more generally to integrate object perception and identification, learning to see objects differently as they become familiar to us.

There is a wealth of experimental data, including neurophysiological, fMRI, and behavioral studies, investigating the differences between familiar and unfamiliar face processing (Eifuku, De Souza, Nakata, Ono, & Tamura, 2011; Natu & O'Toole, 2011). Many behavioral studies have found dramatic differences in processing, including differences in viewpoint generalization, reaction times for recognition tasks, and a shift from external to internal facial feature processing as faces become more familiar (Johnston & Edmonds, 2009). As a quick example: looking at Figure 1, how many identities do you see? Do you recognize any of the individuals?

One of the early conceptual models seeking to capture some of these behavioral differences was proposed by Bruce and Young (1986). They suggested that face recognition begins with a structural encoding of the face, regardless of familiarity. The structural encoding is compared to stored representations for familiar faces, and a face recognition unit is activated if a similarity threshold is reached. The associated 'person identity nodes' can interface with other identity-specific semantic modules. Quantitative implementations of just the structural encoding aspect of this model include

O'Toole et al.'s RBF model (1998) and others (Leibo, Mutch, & Poggio, 2011).

Moving beyond structural encoding, we implement a modified version of the Bruce and Young system capturing notions of familiarity and identity. Our model suggests that 'person identity nodes' are recognized in a holistic way that depends on learned individual invariances rather than by a comparison of structural encodings. Recognition is accomplished by a 'long term memory' which is represented by a neural network trained to predict identities from face images. These identities are associated with latent parameters describing the 3D structure of that person's face (our structural encoding). With this representation of a person identity node, the set of familiar identities need not be fixed, and can be expanded over time. Thus we provide a computational account of how we become familiar with a new face, which also explains how the processing of familiar faces differs from those we have only seen a few times.

The rest of the paper describes the model in more detail, including its accuracy and inference curves, as well as an online clustering algorithm for unfamiliar faces. We validate the model on three different behavioral experiments, and suggest directions for future research.

Model

Our model represents one way of combining the richness of generative models with the speed of neural networks. Inspired by the Helmholtz machine (Dayan, Hinton, Neal, & Zemel, 1995), we describe an efficient analysis-by-synthesis approach by training a recognition model to approximate the latent parameters of a generative model in a fast, feed-forward way. The approximated parameters provide initializations for top-down inference in a generative model, allowing for some fine-tuning. The generative process, inference procedure, and learned recognition models are described below.

Generative Model

We consider the 3D Morphable Face Model as described in (Blanz & Vetter, 1999). This model is obtained from a set of 200 laser scanned heads, providing a mean shape and texture vector for the eyes, nose, mouth and outline of a face, as well as a covariance matrix to generate new faces by eigendecomposition. The shape and texture are Gaussian distributed, with $N(\mu_{\text{shape}}, \Sigma_{\text{shape}})$ and $N(\mu_{\text{texture}}, \Sigma_{\text{texture}})$.

Each of the shape and texture vectors are 200 dimensional, such that a given face lives in a 400 dimensional latent space. An identity can be thought of as a cluster in this latent representation, with a corresponding mean vector μ_i and isotropic variance Σ. Here Σ has been set to 0.01 to represent perceptually indistinguishable identities. An image can be created by sampling a latent vector for a given facial identity, and rendering it at a specific pose and lighting, as seen in Figure 2b. Figure 3 shows some example faces drawn from this model.

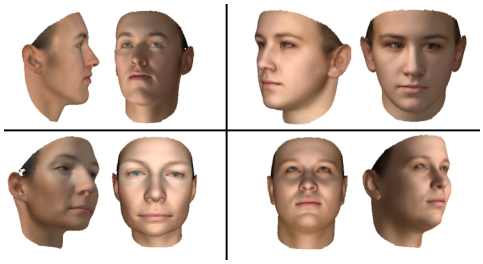


Figure 3: Pairs of images drawn from the generative model.

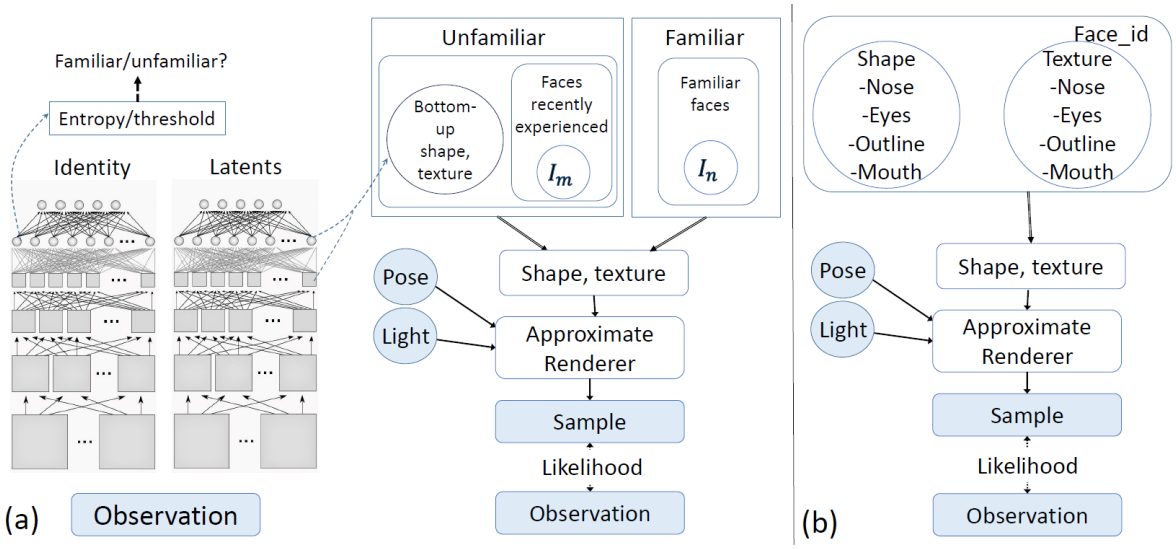


Figure 2: Pipeline used for recognizing an observed image. (a) shows our modified model, using the identity recognition network to determine familiarity, and then initializing with a draw from either the identified familiar cluster or an unfamiliar/new cluster as appropriate. (b) shows the standard generative model operating only over the base distribution in latent space.

Recognition models

Latents recognition network The first recognition network (Figure 2a, network labeled "Latents") is trained to predict the 400 dimensional latent vector that generated a given image of a face. This allows for an efficient, approximate, guess for the latent parameters of a face. The details of this network are described in (Yildirim, Kulkarni, Freiwald, & Tenenbaum, 2015). Yildirim et al. use the top convolutional and first fully connected layers of a convolutional neural network (CNN) (pre-trained on ImageNet) to train a linear model to predict the shape, texture, pose and lighting variables of a set of generated faces. Here we assume that pose and lighting are observed, so only shape and texture need to be predicted. Since the generative model created all training data for the network, it is self-supervised. This recognition model will be referred to as the 'latents recognition model'.

Familiar identity recognition network To mimic long-term memory, we include a classification network for familiar identities (Figure 2a, network labeled "Identity"). We use the first fully connected layer of the network (Simonyan & Zisserman, 2014) (also pre-trained on ImageNet) as input to a linear model, which outputs a probability for each familiar identity. Two versions of this network, an 'old' network which knows 80 identities (80 output class labels), and a 'young' network (30 familiar identities/class labels), were trained using 400 different viewing conditions for each familiar identity.

Processing Pipeline

An observed image I_D generated with the Morphable Face Model is fed to both the identity recognition network and the latents recognition network. The system first determines whether this is a familiar person by calculating the entropy across the familiar identities (Figure 2a, entropy/threshold).

Familiar faces If the entropy falls below a learned threshold, the face is classified as familiar and the identity is set to the most probable class. The latent parameters are then initialized by sampling from the stored representation associated with the determined familiar identity, rather than from the general purpose latents recognition network (Figure 2a, Familiar box).

Unfamiliar faces If the entropy falls above the threshold, the face is unfamiliar and we disregard the familiar identities. There are then two possible cases for unfamiliar faces: either the face is completely novel, or it is the same face as one which we have only seen a few times before (Figure 2a, Unfamiliar box). This can be viewed as a non-parametric clustering problem. The very first unfamiliar person we see will generate their own cluster. Each unfamiliar face we see afterwards will either be clustered with a previously encountered identity, or form its own cluster. We therefore model this process using a sequential clustering algorithm with a Chinese Restaurant Process (CRP) prior on cluster assignments for observation i, where n_k is the number of times you have seen person k before:

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

α is chosen to be 1 for the following experiments, but this choice has little effect on the results.

The likelihood of a specific cluster k for the current observation is computed in image space. We use the generative model to obtain an image from each cluster, rendered at the same pose and lighting as the observed image. While the likelihood for already existing clusters is trivial to compute (Gaussian in pixel space), determining the likelihood of a new cluster is more complex. We approximate it using an image rendered with the latent parameters from the latents recognition network (I_{lrn}). Thus the likelihood can be described by a Gaussian with mean I_k for old clusters, and mean I_{lrn} for the new cluster (and noise $\sigma = 0.01$).

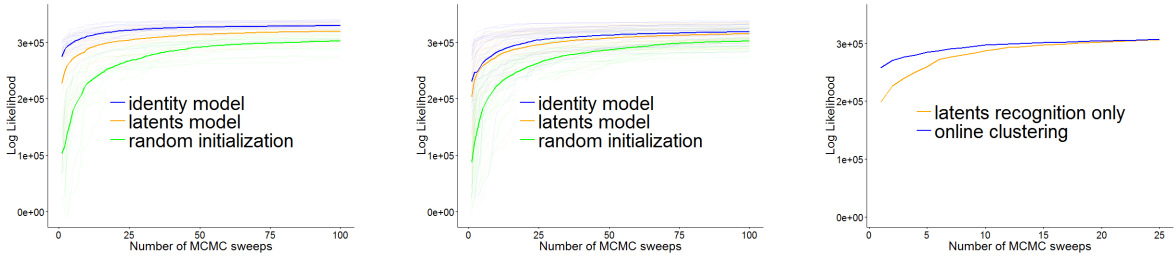
We choose as our estimate the local MAP, which gives us a good initialization for the latent parameters of the new face, even when we are unfamiliar with the observed individual. After forward inference, the cluster means in latent space are updated, reflecting the potential addition of a new cluster member. This learning procedure is the critical contribution of our approach: it presents an account of how we may become familiar with a previously unfamiliar face, even without any supervised training data.

Inference

In order to fine-tune the latent parameters for a given image, we iterate through a few sweeps of forward inference as described in (Yildirim et al., 2015) and (Kulkarni et al., 2015). After initializing the latent parameters for either a familiar face or an unfamiliar face as above, multi-site elliptical slice sampling (Murray, Adams, & MacKay, 2009), a form of MCMC, is performed on the vectors for shape and texture (Figure 2a, Approximate Renderer $\rightarrow$ Observation). At each MCMC sweep, we iterate a proposal-and-acceptance loop on the shape and texture vectors. Proposals are images that are rendered based on a set of latent parameters, a set pose, and a set lighting using a standard graphics engine. The log-likelihood with respect to the observed image is then computed (and assumed to be Gaussian in pixel space).

Simulation experiments

We analyze the performance of our model in several different scenarios. We first generated a set of 100 identities, each of which was rendered under 500 different pose and lighting conditions. We then trained an output layer on the last fully connected layer in the network from (Simonyan & Zisserman, 2014) to predict either a set of 30 identities (the young network) or 80 identities (the old network). For each identity, 400 views were used for training, leaving 100 viewpoints for testing. On the test set, the young network achieves 98.72% accuracy while the old network achieves 98.42% accuracy. Training was performed using stochastic gradient de-



(a) Familiar faces (b) Unfamiliar faces (c) Faces observed once before

Figure 4: Inference traces for (a) 20 familiar (b) 20 unfamiliar and (c) observed once before, faces. The addition of the identity recognition network improves performance for familiar identities, and doesn't hurt performance for unfamiliar faces.

scent with a learning rate of 0.001 and a maximum number of iterations of 1000.

Familiarity classification At the first stage of the pipeline, an incoming face is deemed to be familiar or unfamiliar based on the entropy over the network class labels. To determine an appropriate threshold, we maximize accuracy on a familiar/unfamiliar task using 400 views of 20 familiar faces and 400 views of 20 unfamiliar faces in the **young** network. This results in an accuracy of 91.3% for the young network. Using the same threshold for the old network yields an accuracy of 94.1%. The older network slightly outperforms the younger network, which qualitatively matches the behavioral findings of (Germine, Duchaine, & Nakayama, 2011), who showed that face recognition ability increases with age (up to a certain point).

Inference We check whether including the identity network yields a better initial estimate of the latent parameters for familiar faces compared to random initializations or initializations taken from the latents recognition network. For this experiment, we randomly sampled 20 known and 20 unknown faces rendered at 3 different viewing conditions. Each face was 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 Figures 4a and 4b.

Online clustering 6 identities were chosen, each with a frontal view under random lighting and a 1/4 side view under random lighting. The model was first presented with the 6 frontal views, and correctly made 6 new clusters for these faces. The 6 side views were then presented in scrambled order, and the clustering scheme was able to successfully cluster $4/6$ of the secondary views. The average likelihood traces are shown in Figure 4c.

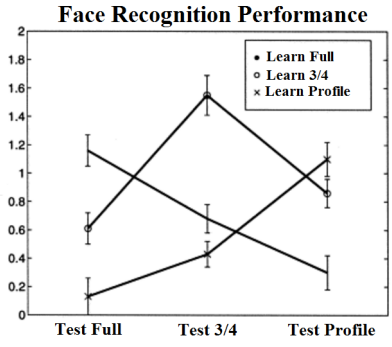
Expanding the set of familiar identities Finally, we tested how well the network consolidated new faces into long-term memory. We sampled 20 views from 3 novel identities, as well as 5 views from each of our previous 30 familiar identities as our training set (which might reflect dreaming of new faces, for example). We initialized the weights of the linear layer in the identity recognition network to those from the previous network for the familiar identities, and randomly initialized the weights for the unfamiliar identities. After training, we achieve 89.84% accuracy on the old faces, and 89.70% on the new faces, reflecting reasonable memory consolidation.

Comparisons with behavioral experiments

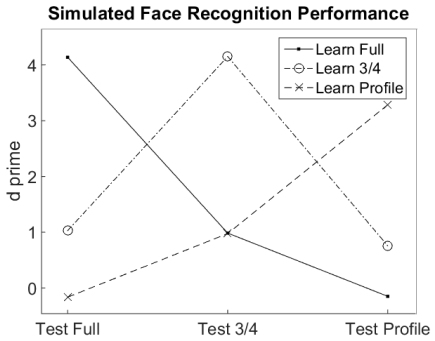
Experiment 1 In the first experiment, we show the power of the unfamiliar face processing component of the model by reproducing results from O'Toole et al. (1998). In this experiment, participants were trained on 36 unfamiliar faces from one of three views: frontal, 3/4 or profile. Participants were then shown 72 images from any of the three viewing conditions, and asked to classify each image as depicting an individual in the training set ('old'), or a new individual ('new'). D prime measures were then calculated for each individual in the task.

We simulated this task by collecting 36 random identities using the Morphable Face Model. Each individual was rendered under a profile, 3/4 and frontal view, with identical lighting conditions. We then used the latents recognition network to predict the latent parameters for each of the 36 unfamiliar faces, with all faces shown in the same viewing condition (either profile, 3/4 or frontal). This results in 36 distinct clusters for 36 individuals. In the test phase, the model observes a face at one of the three views. We then compute the likelihood for each of the 36 learned identities (in pixel space) by rendering its associated latent parameters in the same pose as the test image. These are compared to the likelihood computed with the latents predicted from the latents recognition network. If the likelihood of one of the learned clusters is higher than the likelihood from the latents recognition network, the face is classified as 'old'. Otherwise, it is classified as 'new'. The results from both the psychophysics experiment and from our simulations are shown in Figure 5.

Overall, our model is much more accurate within viewpoints than humans on this task, but follows the general trend for viewpoint-transfer generalization. The model provided by O'Toole in the paper (based on radial basis functions) also predicts this trend, although the old/new classification task would need to incorporate a learned threshold (with one free parameter), which we do not need. The most major discrep-



(a) Experimental results from (O'Toole et al., 1998)



(b) Model results

Figure 5: Results from experiment performed by O'Toole et al. showing viewpoint generalization compared with results from our model.

ancy lies in the relative inability of our model to generalize from a 3/4 view to a profile view. This may be mitigated by running a few sweeps of forward inference during the training phase (in the model) to more accurately determine latent parameters for faces viewed from the 3/4 and frontal views.

Experiment 2 We next show that our model can account for differences in processing speed for familiar and unfamiliar faces, even in very easy tasks with near ceiling performance. Balas, Cox, and Conwell (2007) performed a delayed match-to-sample task for identity, where participants were cued with a profile of either a familiar or unfamiliar person, and then asked to choose which of two individuals (shown at either a 3/4 or frontal view) matched the cue. They found that reaction times for personally familiar individuals was approximately 100 ms faster than for unfamiliar individuals, even though performance in both conditions was above 95%.

It is not obvious how models that rely on stored viewpoints for both familiar and unfamiliar faces could account for this difference in processing speed. Thus, the RBF models presented by O'Toole or those provided by Leibo et al. (2011) do not immediately explain the results of this experiment.

Our model can account for this difference regardless of whether a likelihood measure for the unfamiliar case is done in image or latent space. In latent space, the perceptual system may require a certain confidence in the latent parameters of a face before making a judgment. Therefore, in the test phase, if the images are detected as unfamiliar, they will need more MCMC inference steps to achieve the same likelihood as in the familiar case.

Alternatively, if likelihood is computed in image space, the pose and lighting of the test face need to be inferred (which requires at least one pass through the latents recognition network). The cue face must then be rendered in the appropriate viewpoint in order to compute likelihoods. This extra projection step could account for the longer reaction times.

To ensure that our model achieves comparable accuracy for this task, we trained it on 80 familiar faces, and then randomly chose 9 of these as well as 9 unfamiliar faces for the experiment. The model is first shown a cue face (in the profile view), and then shown two faces during the test phase. The test images are shown in either frontal or 3/4 view. As in the original experiment, each identity is used as a cue 4 times, giving 36 trials for the familiar cases and 36 for the unfamiliar. For 8/9 familiar identities, the model correctly identifies the individual and classifies both the cue image and one of the two test images as familiar. This requires only two passes through the feed-forward identity recognition network. In the last case, the cue was classified as unfamiliar, but the correct judgment was still made when choosing the image with the highest likelihood to the cued face.

For the unfamiliar faces, the model correctly matched the cued individual on each run. It also correctly classified every cued image as "unfamiliar" which meant that a projection back to image space was performed. In future work, we will run further controlled experiments looking at reaction times, and quantitatively compare the model's performance to human performance in recognition tasks for familiar and unfamiliar individuals.

Experiment 3 In this experiment, Jenkins et al. showed that there are massive differences between familiar and unfamiliar face recognition by asking participants to cluster images of two famous Dutch actresses (20 images for each individual) into identities (Jenkins, White, Van Montfort, & Burton, 2011). The experimenters did not specify how many identities were depicted in the collection. Strikingly, they found that participants who were unfamiliar with the actresses clustered the set of images into 6-10 identities (mode 9, range 3-16), while those who were familiar with the actresses correctly clustered the space into two individuals (mode 2, range 2-5). Interestingly, the rate of misidentification (ie. sorting the two different individuals into the same pile) was very rare for both groups.

We simulated this task in a sequential clustering experiment, with two individuals rendered under 20 different viewing conditions each. We used our 'old' network which was trained on 80 identities, but varied whether or not the two individuals were included in the training set (giving an unfamiliar condition as well as a familiar condition).

In the unfamiliar condition, the model created 5 distinct clusters for the two identities with memberships of 7, 3, 9, 7

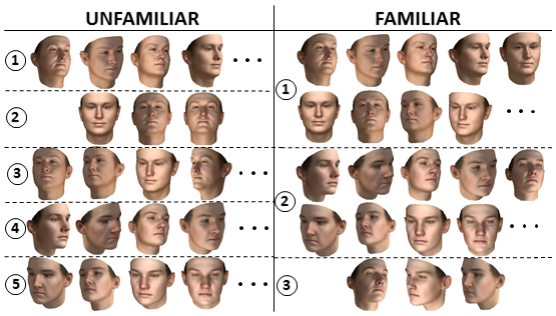


Figure 6: Clusters discovered by the model in the unfamiliar condition (left) and familiar condition (right).

and 12 images (selections from clusters are shown in Figure 6), while 2 faces were incorrectly classified as familiar (into two separate identities). In the familiar condition, the model correctly identified the first individual (with all 20 images being classified as the correct familiar person), while mostly correctly classifying the second individual (with 17/20 images). There were three minor misclassifications for the second individual, resulting in a third cluster being formed.

These results seem to reflect those found by Jenkins et al. (2011). Namely, the model also over-clusters the space when the identities are unfamiliar, but makes only three mistakes when the identities are familiar. In the latents recognition network, the inferred latents depend substantially on pose and lighting, and thus are not successfully ignored in clustering unfamiliar faces. In the familiar network, these invariances are successfully learned, allowing for accurate clustering. Additionally, matching humans, the model never forms clusters which have images from the two different identities in either the familiar or unfamiliar conditions.

Discussion

The model that we have described is both computationally powerful, and also qualitatively and quantitatively captures human behavior across a wide range of different experiments. To our knowledge, this is the first model that learns new identities in both an unsupervised and supervised way, and can account for both effects of familiar and unfamiliar face recognition.

First, we showed how the unfamiliar component of the model can predict the patterns of viewpoint generalization found by O'Toole et al. (1998), even with no explicit viewpoint dependence built in. Although the model is 3 dimensional, the reconstruction accuracy is constrained by the fact that there may be multiple sets of generative parameters that give rise to the same 2D view, which leads to this viewpoint dependent generalization. Second, we show how recognition of familiar faces can proceed significantly faster than for unfamiliar faces, predicting the experimental results from Balas et al. (2007). We discussed how this could result either from a comparison in the latent space (where a good estimate of the latents may be required, and the estimates get better faster for familiar faces) or by a projection back to image space, which is only necessary for unfamiliar faces. This cannot be immediately predicted by standard view-dependent models. Third, we show a major difference in the processing of familiar and unfamiliar identities by replicating the findings of Jenkins et al.'s clustering experiment (albeit in an online fashion). Our model over-clusters the space when faces are unfamiliar, but correctly clusters the space (with minor errors) when the faces are familiar.

We propose that this framework presents a general way of integrating identification and perception. One future line of work will investigate whether the same architecture might be applied to familiar and unfamiliar objects. We also plan to examine other methods of non-parametric clustering, run experiments using the face stimuli we generated, and do a more thorough model comparison. We will also investigate how much exposure you need with an individual in order for them to be consolidated in long-term memory, and whether or not this requires sleep.

References

Balas, B., Cox, D., & Conwell, E. (2007). The effect of real-world personal familiarity on the speed of face information processing. *PloS One*, *2*(11), e1223.

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. 187–194).

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305–327.

Burton, A. M., Bruce, V., & Hancock, P. J. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, *23*(1), 1–31.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural Computation*, *7*, 889–904.

Eifuku, S., De Souza, W. C., Nakata, R., Ono, T., & Tamura, R. (2011). Neural representations of personally familiar and unfamiliar faces in the anterior inferior temporal cortex of monkeys. *PLoS One*, *6*(4), e18913.

Germine, L., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118*, 201–210.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. (2011). Variability in photos of the same face. *Cognition*, *3*, 313–323.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577–596.

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

Leibo, J. Z., Mutch, J., & Poggio, T. (2011). Why the brain separates face recognition from object recognition. In *Advances in neural information processing systems* (pp. 711–719).

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

Natu, V., & O'Toole, A. J. (2011). The neural processing of familiar and unfamiliar faces: A review and synopsis. *British Journal of Psychology*, *102*(4), 726–747.

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision research*, *38*(15), 2351–2363.

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

Yildirim, I., Kulkarni, T., Freiwald, W., & Tenenbaum, J. (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.*