# An Evaluation of the Invariance Properties of a Biologically-Inspired System for Unconstrained Face Recognition

Nicolas Pinto[1] and David Cox[2]

[1] Massachusetts Institute of Technology, Cambridge, MA 02139, USA,
pinto@mit.edu
[2] Rowland Institute at Harvard, Cambridge, MA 02142, USA,
cox@rowland.harvard.edu

**Abstract.** A key challenge in building face recognition systems — biologically-inspired or otherwise — is evaluating performance. While much of face recognition research has traditionally used posed photographs for evaluation, recent efforts have emerged to build more naturalistic, unconstrained test sets by collecting large numbers of face images from the internet (e.g. the "Labeled Faces in the Wild"(LFW) test set [1]). While such efforts represent a large step forward in the direction of realism, the nature of posed photographs from the internet arguably represents an incomplete sampling of the range of variation in view, lighting, etc. found in the real world. Here, we evaluate a family of large-scale biologically-inspired vision algorithms that has previously proven to perform well on a variety of object and face recognition test sets [2], and show that members of this family perform at a level of performance that is comparable with current state-of-the-art approaches on the LFW challenge. As a counterpoint to internet-photo based approaches, we use synthetic (rendered) face images where the amount of view variation is controllable and known by design. We show that while there is gross agreement between the LFW benchmark and synthetic benchmarks, the synthetic benchmarks reveal a substantially greater degree of tolerance to view variation than is apparent from the LFW benchmark in models containing deeper hierarchies. Furthermore, such an approach yields important insights into which axes of variation are most challenging. These results suggest that parametric synthetic benchmarks can play an important role in guiding the progress of biologically-inspired vision systems.

**Key words:** biologically-inspired, computer vision, face recognition, performance evaluation

## 1 Introduction

The face recognition abilities of biological visual systems are currently unrivaled by artificial systems, particularly in unconstrained environments. A natural strategy that follows from this observation is to seek direct inspiration from biology, building artificial visual systems that attempt to capture aspects of the computational architecture of the brain, in the hope of eventually mimicking its abilities. Such efforts to model visual computations done by the brain have a long history, at least dating back to Fukushima's Neocognitron (1980; [3]). More recent experiments with biologically-inspired models have shown them to be highly competitive in a variety of different face and object recognition contexts [4, 5, 6, 7, 8].

Recently, interest in unconstrained face recognition has grown, driven largely by the creation of the *Labeled Faces in the Wild (LFW)* face recognition test set, which has provided a standardized benchmark against which to measure progress. While much work has been done on face recognition in relatively constrained environments (e.g. posed photographs, under controlled lighting conditions [9, 10, 11, 12, 13, 14]), until recently, relatively few available image sets have tackled face recognition in less controlled circumstances. More recently, thanks in large part to the rise of the internet, it has become possible to assemble large collections of face images "in the wild" in the sense that they come from a wide variety of sources and were not posed for the purpose of research.

As in other computer vision domains, biologically-inspired models have achieved highly-competitive performance on the LFW challenge since its inception [7, 15]. More recently, Pinto et al. [2] described a large-scale feature search approach in which thousands of candidate biologically-inspired feature sets are rapidly "screened" to find model architectures that are well suited to a given problem domain. Here, we apply this method to the LFW challenge, and find that it achieves high levels of performance, on par with state-of-the-art methods, even without using any particularly sophisticated machine-learning backend.

However, while these models achieve excellent performance on the LFW challenge set, this set provides little direct insight into why one model performs better than another, and the extent to which the LFW set — which is primarily composed of posed photographs of celebrities — is reflective of the "real" problem of unconstrained face recognition is not entirely clear. In particular, it is not clear that this set contains an accurate sampling of the range of view variation found in the real world [7, 15] since most images are frontal views, and some of the examples of a given individual are taken on the same day, at the same event (e.g. multiple photos of Halle Berry taken from the academy awards ceremony). Thus, while the LFW challenge is clearly useful, and an improvement over more controlled sets, it does not provide an obvious path to the full evaluation of a vision model, nor is it clear how performance on the LFW sets will transfer to other real-world scenarios.

As an complement to the LFW set, we here draw upon carefully-crafted synthetic image sets. While synthetic images have fallen out of favor in the computer vision community in recent years, advances in 3D rendering software have increasingly narrowed the gap between real and synthetic imagery, and rendered images offer several critical advantages over collected photographs. In particular, rendered images allow for complete knowledge and control over the view, position, scale, lighting, presence of other objects etc. in a scene. As a result, synthetic test sets that span whatever range of variation the experimenter desires can be easily generated, and tasks of parametrically variable difficulty can be constructed. Importantly, such data sets also allow one to specifically test the performance of a model as a function of variation in view, lighting, etc [6]. The ability of a model to tolerate such variation – referred to as "invariance" in the parlance of neuroscience — is a critical property of natural vision systems, and is a key stumbling block in the creation of artificial systems.

## 2 Methods

### 2.1 Biologically-inspired visual representations

In the experiments presented below, we studied a family of biologically-inspired visual representations designed to model various stages of visual cortex in the brain.

We used two basic sub-classes of models: 1) *V1-like*, a simple one-layer model with fixed parameters, designed to mimic cortical visual area V1 [6], and 2) multi-layer "High-Throughput" (HT) models, generated by way of a large scale screening approach [2].

Both models classes are characterized by a cascade of linear and nonlinear processing steps (see Figure 1), with *V1-like* having just one layer (and with filters kernels constrained to be Gabor wavelets), and the *HT* models having either two or three layers (referred to hereafter as *HT-L2* and *HT-L3*, respectively).

Our *V1-like* implementation was taken without modification from [6, 7]. Similarly, the *HT-L2* and *HT-L3* models were generated according the high-throughput screening approach described in [2] except with randomly generated filters instead of filters trained using an unsupervised learning approach. The details of the *HT-L2* and *HT-L3* models are described in greater detail below.

### 2.2 High-throughput-derived multilayer visual representations: *HT-L2* and *HT-L3*

In this study, we considered the best two- and three-layer models generated from a high-throughput screening model selection procedure. An important feature of the generation of these representations, according to the scheme set forth in [2], is the use of a massively parallel, high-throughput search over the parameter space of possible instances of a large class of biologically-inspired models. Details of this model class and the high-throughput screening (model selection) procedure have been described before [2] but are summarized below for convenience.

**Model architecture:** Candidate models were composed of a hierarchy of two (*HT-L2*) or three layers (*HT-L3*), with each layer including a cascade of linear and nonlinear operations that produce successively elaborated nonlinear feature-map representations of the original image. A diagram detailing the flow of operations is shown in Figure 1, and, for the purposes of notation, the cascade of operations is represented as follows:

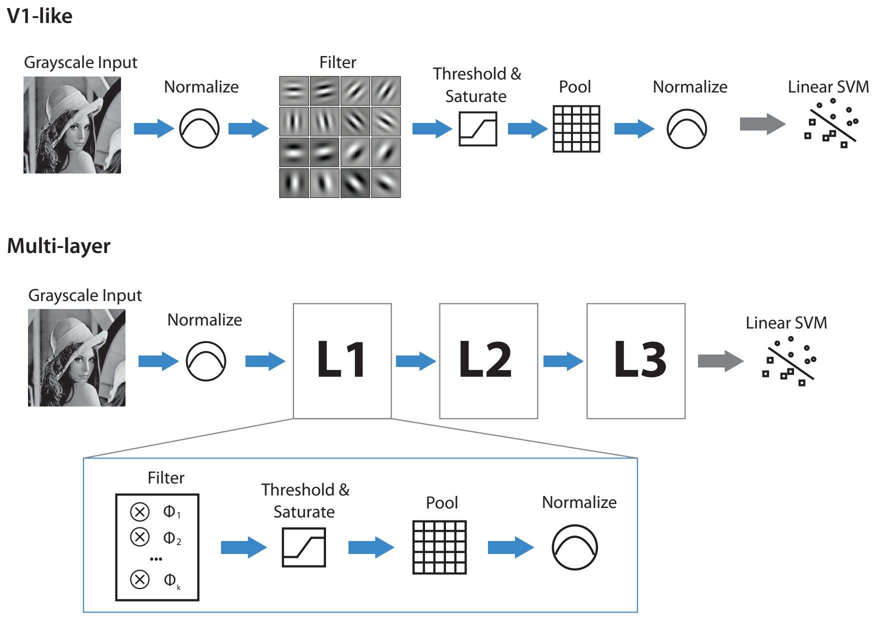


**Fig. 1. A schematic diagram of the system architecture of the family of models considered.** Each model consists of one to three feedforward filtering layers, with the filters in each layer being applied across the previous layer.

$$Layer^0:$$

$$\mathbf{Input} \xrightarrow{\mathbf{Grayscale}} \xrightarrow{\mathbf{Normalize}} \mathbf{N^0}$$

$$Layer^1:$$

$$\mathbf{N^0} \xrightarrow{\mathbf{Filter}} \mathbf{F^1} \xrightarrow{\mathbf{Activate}} \mathbf{A^1} \xrightarrow{\mathbf{Pool}} \mathbf{P^1} \xrightarrow{\mathbf{Normalize}} \mathbf{N^1}$$

and generally, for all $\ell \geq 1$:

$$Layer^\ell:$$

$$\mathbf{N}^{\ell-1} \xrightarrow{\mathbf{Filter}} \mathbf{F}^\ell \xrightarrow{\mathbf{Activate}} \mathbf{A}^\ell \xrightarrow{\mathbf{Pool}} \mathbf{P}^\ell \xrightarrow{\mathbf{Normalize}} \mathbf{N}^\ell$$

Details of these steps along with the range of parameter values included in the random search space are described next.

**Input and Pre-processing** The input of the *HT-L2* and *HT-L3* models were 100x100 and 200x200 pixel images, respectively. In the pre-processing stage, referred to as *Layer*$^0$, this input was converted to grayscale and locally normalized:

$$\mathbf{N^0} = \mathbf{Normalize}(\mathbf{Grayscale}(\mathbf{Input})) \tag{1}$$

where the **Normalize** operation is described in detail below. Because this normalization is the final operation of each layer, in the following sections, we refer to $N^{\ell-1}$ as the input of each $Layer^{\ell>0}$ and $N^{\ell}$ as the output.

**Linear Filtering** The input $N^{\ell-1}$ of each subsequent layer (i.e. $Layer^{\ell}, \ell \in \{1, 2, 3\}$) was first linearly filtered using a bank of $k^{\ell}$ filters to produce a stack of $k^{\ell}$ feature maps, denoted $F^{\ell}$. In a biologically-inspired context, this operation is analogous to the weighted integration of synaptic inputs, where each filter in the filterbank applied at a particular image location represents a different cell.

*Definitions:* The filtering operation for $Layer^{\ell}$ is denoted:

$$\mathbf{F}^{\ell} = \mathbf{Filter}(\mathbf{N}^{\boldsymbol{\ell-1}}, \boldsymbol{\Phi}^{\ell}) \tag{2}$$

and produces a stack, $F^{\ell}$, of $k^{\ell}$ feature maps, with each map, $F_i^{\ell}$, given by:

$$F_i^{\ell} = N^{\ell-1} \otimes \Phi_i^{\ell} \qquad \forall i \in \{1, 2, \ldots, k^{\ell}\} \tag{3}$$

where $\otimes$ denotes a correlation of the output of the previous layer, $N^{\ell-1}$ with the filter $\Phi_i^{\ell}$ (e.g. sliding along the first and second dimensions of $N^{\ell-1}$). Because each successive layer after $Layer^0$ is based on a stack of feature maps, $N^{\ell-1}$ is itself a stack of 2-dimensional feature maps. Thus, the filters contained within $\Phi^{\ell}$ are, in turn, 3-dimensional, with the their third dimension matching the number of filters (and therefore, the number of feature maps) from the previous layer (i.e. $k^{\ell-1}$).

*Parameters:*

- The filter shapes ${f_s}^{\ell} \times {f_s}^{\ell} \times {f_d}^{\ell}$ were chosen randomly with ${f_s}^{\ell} \in \{3, 5, 7, 9\}$ and ${f_d}^{\ell} = k^{\ell-1}$.
- Depending on the layer $\ell$ considered, the number of filters $k^{\ell}$ was chosen randomly from the following sets:
  - In $Layer^1$, $k^1 \in \{16, 32, 64\}$
  - In $Layer^2$, $k^2 \in \{16, 32, 64, 128\}$
  - In $Layer^3$, $k^3 \in \{16, 32, 64, 128, 256\}$

All filter kernels were fixed to random values drawn from a uniform distribution.

**Activation Function** Filter outputs were subjected to threshold and saturation activation function, wherein output values were clipped to be within a parametrically defined range. This operation is analogous to the spontaneous activity thresholds and firing saturation levels observed in biological neurons.

*Definitions:* We define the activation function:

$$\mathbf{A}^{\ell} = \mathbf{Activate}(\mathbf{F}^{\ell}) \tag{4}$$

that clips the outputs of the filtering step, such that:

$$\mathbf{Activate}(\mathbf{x}) = \begin{cases} {\gamma_{max}}^{\ell} & \text{if } x > {\gamma_{max}}^{\ell} \\ {\gamma_{min}}^{\ell} & \text{if } x < {\gamma_{min}}^{\ell} \\ x & \text{otherwise} \end{cases} \tag{5}$$

Where the two parameters ${\gamma_{min}}^{\ell}$ and ${\gamma_{max}}^{\ell}$ control the threshold and saturation, respectively. Note that if both minimum and maximum threshold values are $-\infty$ and $+\infty$, the activation is linear (no output is clipped).

*Parameters:*

- ${\gamma_{min}}^{\ell}$ was randomly chosen to be $-\infty$ or 0
- ${\gamma_{max}}^{\ell}$ was randomly chosen to be 1 or $+\infty$

**Pooling** The activations of each filter within some neighboring region were then pooled together and the resulting outputs were spatially downsampled.

*Definitions:* We define the pooling function:

$$\mathbf{P}^{\ell} = \mathbf{Pool}(\mathbf{A}^{\ell}) \tag{6}$$

such that:

$$\mathbf{P}^{\ell}_{\mathbf{i}} = \mathbf{Downsample}_{\alpha}\left( \sqrt[p^{\ell}]{(A^{\ell}_{i})^{p^{\ell}} \odot \mathbf{1}_{a^{\ell} \times a^{\ell}}} \right) \tag{7}$$

Where $\odot$ is the 2-dimensional correlation function with $\mathbf{1}_{a^{\ell} \times a^{\ell}}$ being an $a^{\ell} \times a^{\ell}$ matrix of ones ($a^{\ell}$ can be seen as the size of the pooling "neighborhood"). The variable $p^{\ell}$ controls the exponents in the pooling function.

*Parameters:*

- The stride parameter $\alpha$ was fixed to 2, resulting in a downsampling factor of 4.
- The size of the neighborhood $a^{\ell}$ was randomly chosen from $\{3, 5, 7, 9\}$.
- The exponent $p^{\ell}$ was randomly chosen from $\{1, 2, 10\}$.

Note that for $p^{\ell} = 1$, this is equivalent to blurring with a $a^{\ell} \times a^{\ell}$ boxcar filter. When $p^{\ell} = 2$ or $p^{\ell} = 10$ the output is the $L^{p^{\ell}}$-norm [1].

**Normalization** As a final stage of processing within each layer, the output of the Pooling step was normalized by the activity of their neighbors within some radius (across space and across feature maps). Specifically, each response was divided by the magnitude of the vector of neighboring values if above a given threshold. This operation draws biological inspiration from the competitive interactions observed in natural neuronal systems (e.g. contrast gain control mechanisms in cortical area V1, and elsewhere [16, 17])

*Definitions:* We define the normalization function:

$$\mathbf{N}^{\ell} = \mathbf{Normalize}(\mathbf{P}^{\ell}) \tag{8}$$

such that:

[1] The $L^{10}$-norm produces outputs similar to a *max* operation (i.e. *softmax*).

$$N^\ell = \begin{cases} \rho^\ell \cdot C^\ell & \text{if } \rho^\ell \cdot \left|\left|C^\ell \otimes \mathbf{1}_{b^\ell \times b^\ell \times k^\ell}\right|\right|_2 < \tau^\ell \\ \frac{C^\ell}{\left|\left|C^\ell \otimes \mathbf{1}_{b^\ell \times b^\ell \times k^\ell}\right|\right|_2} & \text{otherwise} \end{cases} \tag{9}$$

with

$$C^\ell = P^\ell - \delta^\ell \cdot \frac{P^\ell \otimes \mathbf{1}_{b^\ell \times b^\ell \times k^\ell}}{b^\ell \cdot b^\ell \cdot k^\ell} \tag{10}$$

Where $\delta^\ell \in \{0, 1\}$, $\otimes$ is a 3-dimensional correlation over the "valid" domain (i.e. sliding over the first two dimensions only), and $\mathbf{1}_{b^\ell \times b^\ell \times k^\ell}$ is a $b^\ell \times b^\ell \times k^\ell$ array full of ones. $b^\ell$ can be seen as the normalization "neighborhood" and $\delta^\ell$ controls if this neighborhood is centered (i.e. subtracting the mean of the vector of neighboring values) before divisive normalization. $\rho^\ell$ is a "magnitude gain" parameter and $\tau^\ell$ is a threshold parameter below which no divisive normalization occurs.

*Parameters:*

- The size $b^\ell$ of the neighborhood region was randomly chosen from $\{3, 5, 7, 9\}$.
- The $\delta^\ell$ parameter was chosen from $\{0, 1\}$.
- The vector of neighboring values could also be stretched by gain values $\rho^\ell \in \{10^{-1}, 10^0, 10^1\}$. Note that when $\rho^\ell = 10^0 = 1$, no gain is applied.
- The threshold value $\tau^\ell$ was randomly chosen from $\{10^{-1}, 10^0, 10^1\}$.

### 2.3 Final model output dimensionality

The output dimensionality of each candidate model was determined by the number of filters in the final layer, and the x-y "footprint" of the layer (which, in turn, depends on the subsampling at each previous layer). In the model space explored here, the possible output dimensionality ranged from 256 to 73,984.

### 2.4 Screening (model selection)

A total of 5,915 *HT-L2* and 6,917 *HT-L3* models were screened on the *LFW* View 1 "aligned" set [18]. We selected the best model from each "pool" for further analysis on the *LFW* View 2 set (Restricted Protocol). Note that *LFW* View 1 and View 2 do not contain the same individuals and are thus mutually exclusive sets. View 1 was designed as a model selection set while View 2 is used as an independent validation set for the purpose of comparing different methods.

Examples of the screening procedure for *HT-L2* and *HT-L3* models on the *LFW* View 1 task screening task are shown in Figure 2. Performance of randomly generated *HT-L3* models ranged from chance performance (50%) to better than 80% correct; the best five models were drawn from this set and are denoted *HT-L3-1st*, *HT-L3-2nd*, and so on. An analogous procedure was undertaken to generate five two-layer models, denoted *HT-L2-1st*, *HT-L2-2nd*, etc. For the purposes of the present paper, we only considered the best model from each group (i.e. *HT-L2-1st* and *HT-L3-1st*).

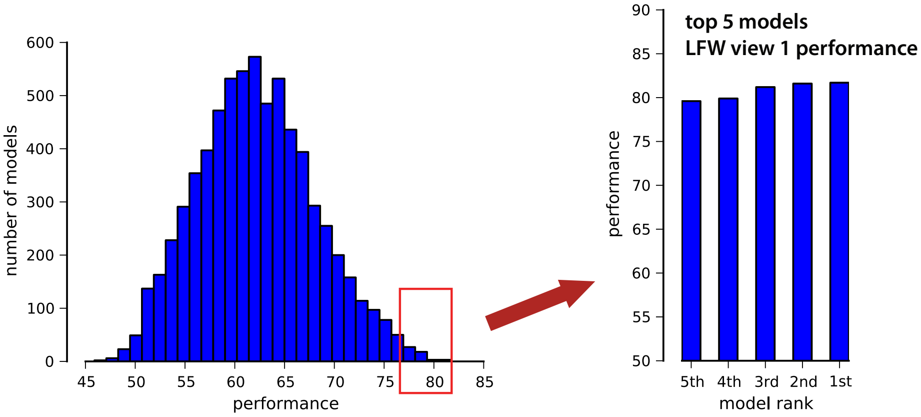


**Fig. 2. The high-throughput screening process used to find good representations.** Here, data is shown for the screening of *HT-L3* models. A distribution of the performance of approx. 7,000 randomly generated models is shown on the left, with the top five high-performing models replotted on the right. Following screening, the models were evaluated exclusively with sets that do not overlap with the screening set.

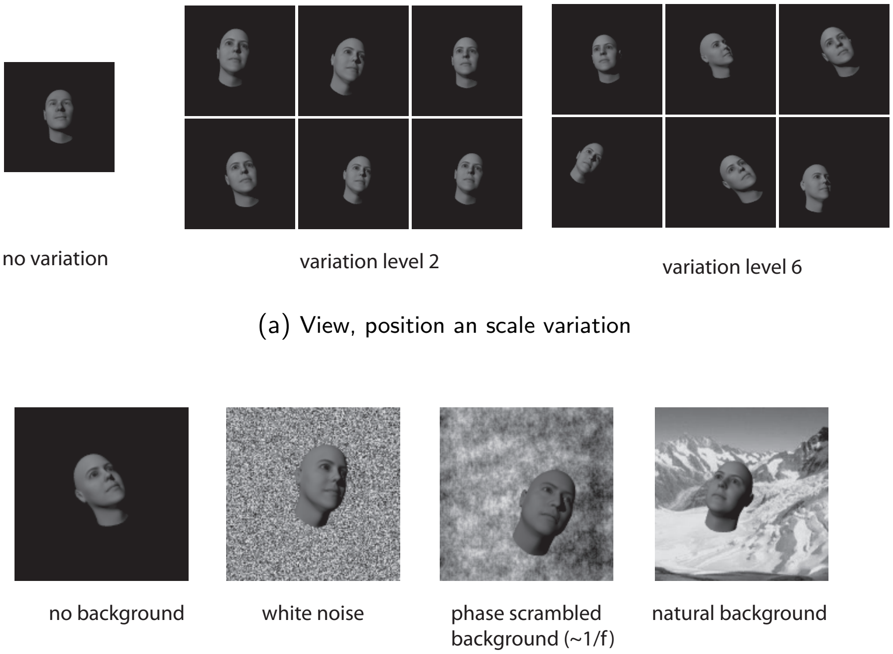
**Table 1.** Performance of the family of biologically-inspired models on the *LFW* challenge set (restricted view 2). For the *HT-L2* and *HT-L3* models, the cross-validated performance of the top 5 randomly-generated models is shown (e.g. 1st, 2nd, etc.). The performance of the simpler single layer *V1-like* model [7] is provided for comparison.

| | *5th* | *4th* | *3rd* | *2nd* | *1st* |
|---|---|---|---|---|---|
| *V1-like* | | | 77.0 ± 0.5 | | |
| *HT-L2* | 77.8 ± 0.4 | 81.3 ± 0.4 | 81.5 ± 0.6 | 80.8 ± 0.4 | 81.0 ± 0.3 |
| *HT-L3* | 82.8 ± 0.6 | 82.3 ± 0.4 | 83.3 ± 0.4 | 83.9 ± 0.3 | 84.1 ± 0.3 |

### 2.5 Synthetic Face Images

In order to assess model performance on an image set with a known amount of variation, we generated a set of 3D-rendered face images. 3D face meshes were randomly generated using the FaceGen [19] software package and were rendered using the free POV-Ray ray-tracer [20]. For each rendered image, a model rotation (azimuth and elevation), position (x and y), and scale were drawn from a uniform distribution and the models were rendered with a common light source (Figure 3(a)). For the experiments presented here, rotation, size, and position were combined into a single composite “variation level” wherein the variation in the pixel-level euclidean norm was equalized for each kind of variation (e.g. one “unit” of rotation variation produced an equivalent pixel-level change as one "unit" of position variation). Examples of several variation "levels" are shown in Figure 3(a).

The rendered face/head was next composited onto one of four kinds of backgrounds: no background, a white noise background, a phase-scrambled natural background (approximately equivalent to 1/f noise), and a randomly chosen natural background, chosen from a large pool of outdoor background images (Figure 3(b)). Care was taken to ensure that the same background image was never used in more than one final image.



(a) View, position an scale variation

(b) Background variation

**Fig. 3. Synthetic face stimuli**

### 2.6 Classification and Performance Evaluation

To evaluate the performance of a given model with a given stimulus set, we trained a multi-class support vector machine (SVM) classifier [21] using a one-vs-all configuration [22] for each target class. Training and test data were strictly segregated, and performance was evaluated using five 250 train / 50 test random folds of the data. Error bars in all plots show the standard deviation of performance across these five folds.

## 3 Results

### 3.1 LFW performance

Performance on the LFW data set for these models is presented in Table 1. Performance ranged as high as 84.1% percent correct for the best HT-L3 model, achieving performance within a few percent of state-of-the-art methods [23, 24]. While more sophisticated kernel blending techniques have previously been used to achieve better performance on the LFW challenge set by leveraging multiple feature representations (e.g. [15]), we here restrict ourselves here to unblended model performance for the sake of clarity. Further, for simplicity, we also here only consider the best-performing model from each group (i.e. HT-L2-1st and HT-L3-1st).

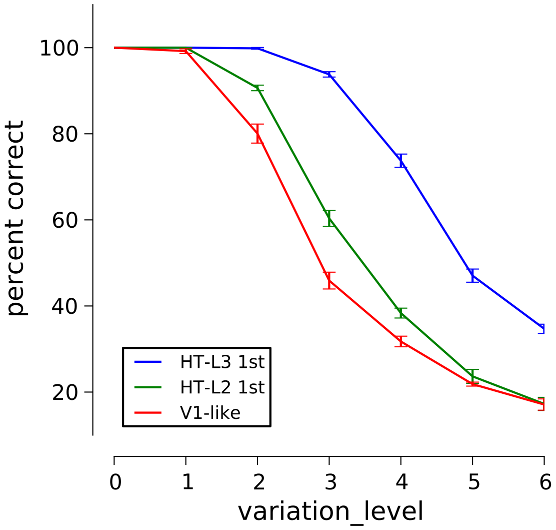


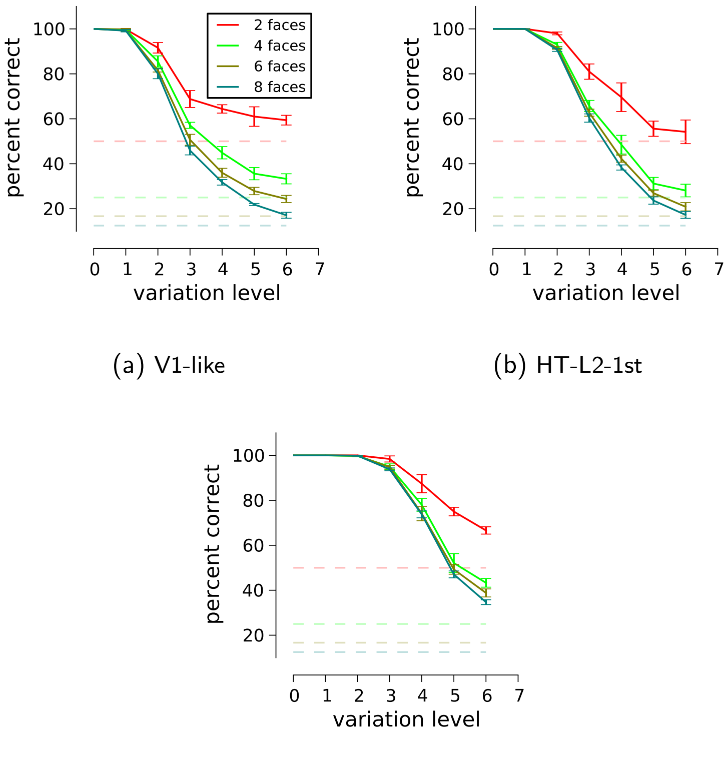
**Fig. 4. Model performance on synthetic faces as a function of level of variation**

### 3.2 Performance as a function of variation level

The synthetic face evaluation sets used here provide us with the ability to parametrically control the level of rotation, position and scale variation that our models are required to tolerate. Figure 4 shows the performance the best models from each model class (V1-like, HT-L2, HT-L3) as a function of (composite) variation level for an eight-way face classification task.

### 3.3 Effect of number of faces to be discriminated

To further explore the behavior of our models with a controlled stimulus, we examined model performance as a function of the number of faces to be discriminated. In particular, we considered cases with two, four, six, and eight faces. Performance, grouped by model is shown in Figure 5, and is shown grouped by variation level in Figure 6. Predictably, absolute performance level is depressed as a larger number of faces is considered, as is the chance performance level (dotted line). Interestingly, the rate at which performance falls off varies between models as a function of both number of faces to be discriminated, and as a function of variation level. The stability of the performance of the largest/deepest model — HT-L3-1st — is most pronounced when large number of faces and large amounts of variation are considered. Differences between models are far less pronounced with smaller numbers of faces and lesser degrees of variation.

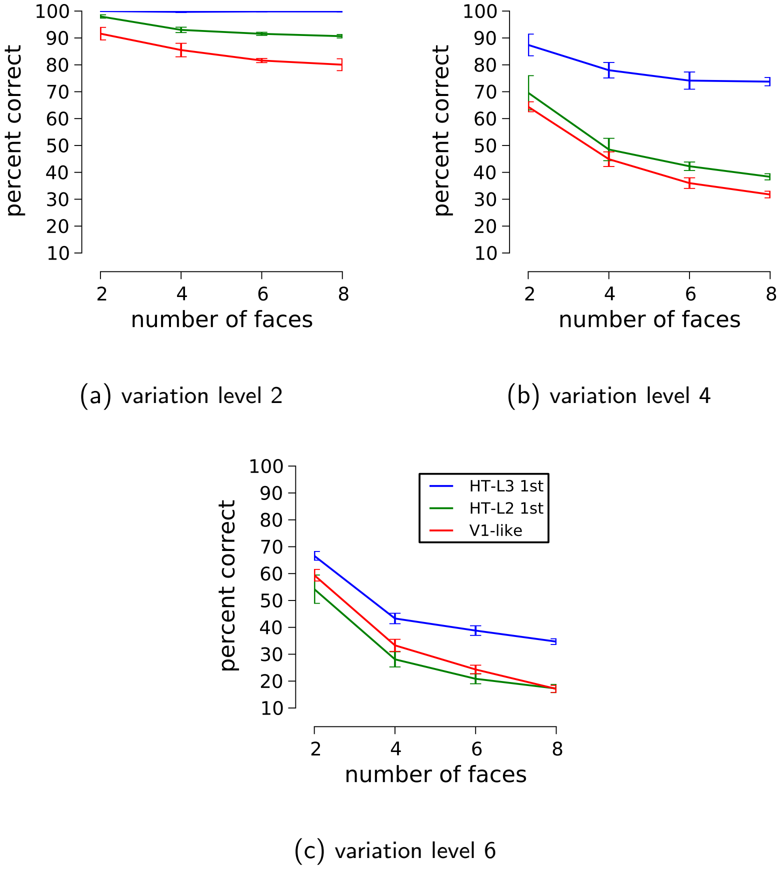


(a) V1-like

(b) HT-L2-1st

(c) HT-L3-1st

**Fig. 5. Effect of number of synthetic faces to be discriminated, sorted by model**



(a) variation level 2

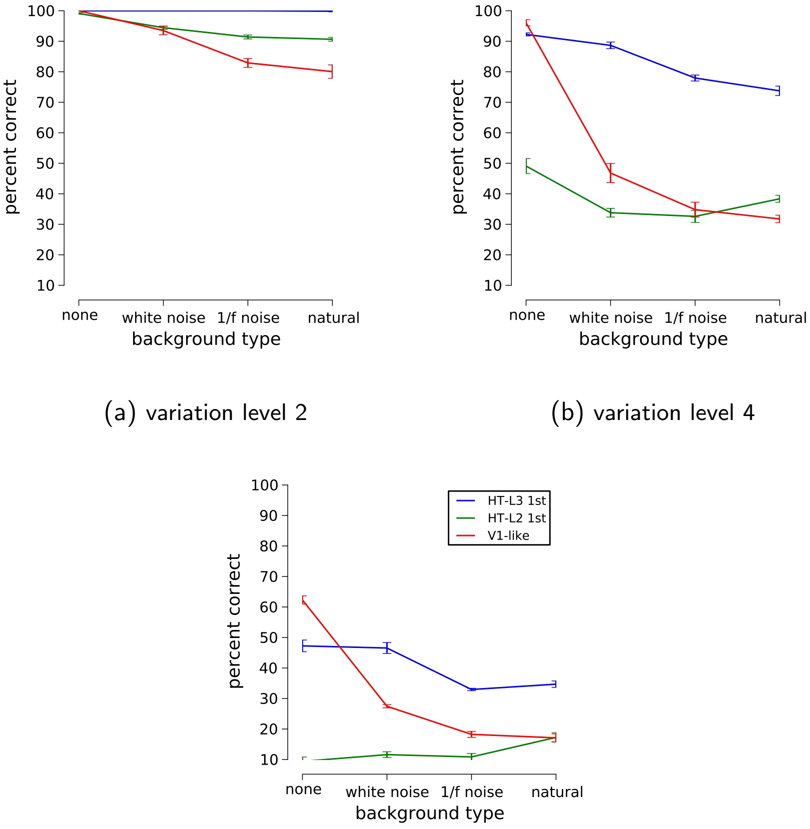
(b) variation level 4

(c) variation level 6

**Fig. 6. Effect of number of synthetic faces to be discriminated, sorted by variation level.** Note that the performance was 100% in all cases for the zero variation condition (data not shown).

### 3.4 Effect of background

To explore the role of background variation, we evaluated model performance with four different background conditions: no background, white-noise background, phase-scrambled natural backgrounds (i.e. approx. 1/f noise), and natural backgrounds. Performance as a function of background and variation level is shown Figure 7. Choice of background was found to have a profound effect on model performance. In the absence of a background, the performance for most models remained high, even at relatively high levels of variation in view, position, and scale (e.g. greater than 90% performance at variation level 4 for the HT-L3-1st and V1-like models). However, the inclusion of any background resulted in a precipitous drop-off in performance for all models, except for the HT-L3-1st model, whose performance degraded gradually. In general, progressively more realistic backgrounds proved increasingly difficult for all models.



(a) variation level 2

(b) variation level 4

(c) variation level 6

**Fig. 7. Effect of background type on performance with synthetic faces.** Note that the performance was 100% in all cases for the zero variation condition (data not shown).

## 4 Discussion

While it is standard practice to test computer vision algorithms with standardized "natural" image test sets such as the LFW set, the performance obtained on such a set provides a relatively narrow window onto behavior of a given system.

Here, we used synthetic test images, rendered with known amounts of variation, to provide a much richer multidimensional assessment of the invariance properties of a class of models that have achieved high levels of performance on the LFW set.

While the ordinal performance of the one-, two- and three-layer models considered here is roughly the same as is observed for the LFW set (i.e. V1-like < HT-L2 < HT-L3), tests with synthetic sets reveal that the model with the deepest hierarchy (HT-L3) is substantially better able to tolerate variation in view, position, scale and background as compared to the other models considered here. This dramatic difference was not at all apparent from the LFW performance, where the best HT-L3 model performed only a few percent higher than its nearest rivals. While there is no hard evidence one way or another, we speculate that the relatively compressed range of performance between the various models on the LFW set is reflective of the relatively limited range of view variation found in that set. Indeed, when we examine a relatively low level of variation with our synthetic faces, we see a similarly compressed range of performance variation across the models.

More broadly, our results suggest that the level of variation present in a set, both in terms of view and in terms of background can have a large effect on the "dynamic range" within which one has the ability to distinguish between models. Indeed, without any background, and at low levels of variation, the differences between models can become vanishing small, and in some cases can even reverse. These results underscore the importance of building sets, be they synthetic or natural, that contain more realistic ranges of variation.

## References

1. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild. TR UMass (2007)
2. Pinto, N., Doukhan, D., DiCarlo, J.J., Cox, D.D.: A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Comput Biol (2009)
3. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybernetics (1980)
4. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. PAMI (2007)
5. Mutch, J., Lowe, D.G.: Object class recognition and localization using sparse features with limited receptive fields. IJCV (2008)
6. Pinto, N., Cox, D.D., DiCarlo, J.J.: Why is Real-World Visual Object Recognition Hard. PLoS Comput Biol (2008)
7. Pinto, N., DiCarlo, J.J., Cox, D.D.: Establishing Good Benchmarks and Baselines for Face Recognition. ECCV (2008)
8. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition? ICCV (2009)
9. ORL Face Set: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html (Accessed 07/15/2010)

10. Yale Face Set: http://cvc.yale.edu (Accessed 07/15/2010)
11. CVL Face Set: http://www.lrv.fri.uni-lj.si/facedb.html (Accessed 07/15/2010)
12. AR Face Set: http://www2.ece.ohio-state.edu/ aleix/ardatabase.html (Accessed 07/15/2010)
13. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for face-recognition algorithms. PAMI (2000)
14. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image and Vision Computing (2009)
15. Pinto, N., DiCarlo, J.J., Cox, D.D.: How far can you get with a modern face recognition test set using only simple features? CVPR (2009)
16. Geisler, W.S., Albrecht, D.G.: Cortical neurons: isolation of contrast gain control. Vision Research (1992)
17. Rolls, E.T., Deco, G.: Computational neuroscience of vision. Oxford Univ Press (2002)
18. Taigman, Y., Wolf, L., Hassner, T., Tel-Aviv, I.: Multiple one-shots for utilizing class label information. BMVC (2009)
19. FaceGen: singularinversions.com (Accessed 07/15/2010)
20. POV-Ray Raytracer: www.povray.org (Accessed 07/15/2010)
21. Schölkopf, B., Smola, A.: Learning with kernels: Support vector machines, regularization, optimization, and beyond. the MIT Press (2002)
22. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. JMLR (2004)
23. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and Simile Classifiers for Face Verification. ICCV (2009)
24. Wolf, L., Hassner, T., Taigman, Y.: Similarity Scores based on Background Samples. ACCV (2009)