Scaling Up Biologically-Inspired Computer Vision: A Case Study in Unconstrained Face Recognition on Facebook

Nicolas Pinto[1,2], Zak Stone[3], Todd Zickler[3], and David Cox[1]

[1]The Rowland Insitute at Harvard, Harvard University, Cambridge, MA 02142
[2]McGovern Institute for Brain Research, MIT, Cambridge, MA 02139
[3]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138
{pinto,cox}@rowland.harvard.edu, zstone@post.harvard.edu, zickler@seas.harvard.edu

Abstract

Biological visual systems are currently unrivaled by artificial systems in their ability to recognize faces and objects in highly variable and cluttered real-world environments. Biologically-inspired computer vision systems seek to capture key aspects of the computational architecture of the brain, and such approaches have proven successful across a range of standard object and face recognition tasks (e.g. [23, 8, 9, 18]). Here, we explore the effectiveness of these algorithms on a large-scale unconstrained real-world face recognition problem based on images taken from the Facebook social networking website. In particular, we use a family of biologically-inspired models derived from a high-throughput feature search paradigm [19, 15] to tackle a face identification task with up to one hundred individuals (a number that approaches the reasonable size of real-world social networks). We show that these models yield high levels of face-identification performance even when large numbers of individuals are considered; this performance increases steadily as more examples are used, and the models outperform a state-of-the-art commercial face recognition system. Finally, we discuss current limitations and future opportunities associated with datasets such as these, and we argue that careful creation of large sets is an important future direction.

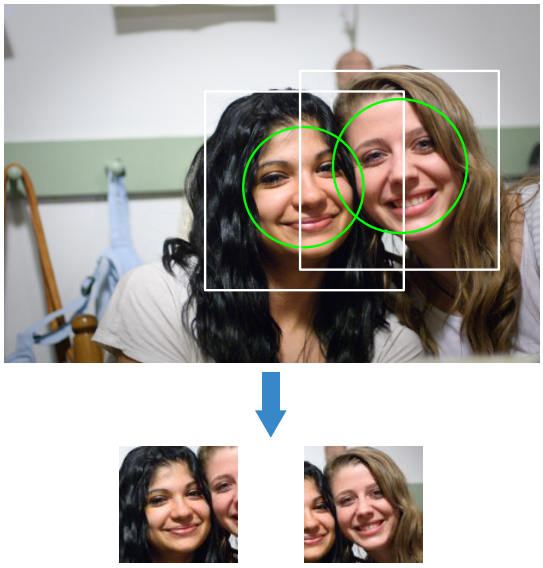


Figure 1. A diagram of the tagging procedure that produced the *Facebook100* set, using a representative publicly-available photo. Manually-applied "tags" are shown as white square outlines, and OpenCV face detections are superimposed as green circles. Detected face regions are matched with nearby tags to yield labeled face samples, as shown above. Because tags carry social meaning and can trigger notifications to hundreds of people when assigned, the identities that they specify for faces are typically extremely accurate. (Photo courtesy of Flickr user `wickenden` under a Creative Commons License [11].)

1. Introduction

In recent years, several serious efforts have emerged to move face recognition research towards less constrained, "real-world" settings. A major driver of this push has been the *Labeled Faces in the Wild (LFW)* data set, which brings together thousands of face images of public figures from the Internet. While some concerns have been raised about whether this set is an ideal surrogate for the "full" problem of real-world face recognition [17, 18], it nonetheless has focused the efforts of the community. More recently, a set in the same vein called *PubFig* [10] has been introduced to help facilitate larger-scale explorations in real-world face recognition.

One property that *LFW* and *PubFig* have in common (at least in their usage to date) is that they are designed primarily as tests of face *verification* — deciding whether two faces represent the same person — rather than face *identification*, which requires matching an unknown face or face set against a gallery of labeled face samples. Clearly, a continuum exists between verification and identification, and, within limits, a system built for one of these tasks can be reconfigured to perform the other. In practice, however, a verification system that tries to do identification may be very sensitive to verification errors, and it may not fully utilize the advantages of having a large, labeled training set per individual.

Verification is a natural paradigm in many contexts (e.g. biometric authentication), and it is obviously desirable to have face recognition systems that can function even without a large amount of training data. But experiments in a large-scale face identification regime have become increasingly practical and relevant. The explosion in usage of digital cameras has greatly increased the number of real-world photos that are captured and shared, and photo-sharing software and services (e.g. Facebook, Flickr, iPhoto, and Picasa) have aggregated and organized these photos. Today, it is not uncommon for individuals to have large personal databases of photos of familiar faces, with hundreds or even thousands of images per individual. The ubiquity of personal and shared photo databases presents opportunities to assemble novel, large-scale, realistic datasets to guide face recognition research and to explore potential use-cases for working face recognition systems. Already, several available software applications attempt to perform automatic face tagging with varying levels of success.

Unfortunately, existing verification datasets for artificial systems cannot always be converted into identification datasets. The *LFW* data set, for example, contains few face samples for most individuals and few individuals with large numbers of face samples.

To address this problem, we introduce two new datasets for identification research, both of which are derived from images taken "in the wild," and both of which include many face samples per individual. The first dataset we created *(Facebook100)* is a set of face samples drawn from photos shared online through the Facebook social network; images were collected in the manner of [25, 26]. Due to the vast size of the network, we were able to extract many labeled samples of many distinct individuals, and it will be straightforward to expand our benchmark set to increase the difficulty of the identification problem. As a public complement to this private set of Facebook photos, we assembled a subset of the *PubFig* dataset with an emphasis on removing near-duplicate images, which are commonly encountered when seeking images of celebrities online *(PubFig83)*.

To benchmark these sets, we used a family of biologically-inspired visual models. Because humans are currently unrivaled by artificial systems in their ability to recognize familiar faces, a biologically-inspired approach to the problem of face identification warrants study, particularly in the context of familiar face recognition. The models tested here seek to instantiate biologically-plausible neural-network-style computational elements organized either into a single- [17, 18] or multi-layer [19] architecture. These models have been shown to excel in a standard face verification task (*LFW*), previously achieving state-of-the-art performance on that set [15].

2. Datasets

2.1. The "Facebook100" Dataset

The *Facebook100* data set used in this study contains 100 distinct person categories, each of which is represented by 100 cropped face samples. These labeled face samples were extracted from a set of shared Facebook photos and their associated "tags", which identify the locations of particular people in specific photographs. Fig. 1 represents a typical Facebook photo with its manually-applied tag locations superimposed in white.

Facebook users tag themselves and their friends in photos for a variety of social purposes [14, 13, 1], and they typically manually assign a tag to a photograph by clicking somewhere on the photo and entering a name. The coordinates of the click are used to place the tag, and these coordinates are often conveniently centered on faces [25, 26]. At present, the Facebook interface does not allow users to specify the size of a tagged region, so the tags are assumed to label square regions of a standard size as shown in Fig. 1. Because the act of assigning a tag to a photo typically triggers a notification to the person tagged and all of the friends of that person and the photographer (at least), the identities assigned to faces with tags tend to be extremely accurate.

Given a photo and its associated tags, we ran the frontal OpenCV face detector to identify actual face locations (shown as green circles in Fig. 1), and we associated the detected face regions with nearby identity tags using a conservative distance threshold. In the mockup photo shown in Fig. 1, the detected foreground frontal faces are successfully matched with near-concentric manually-applied tags to yield two labeled face samples. The requirement that a face be both manually tagged and detected algorithmically helps to suppress the effect of "joke" tags, where non-face portions of an image are tagged. As a future direction, more intensive and sophisticated face detection techniques would allow us to harvest more challenging non-frontal tagged face images throughout the Facebook dataset.

The face samples used in the *Facebook100* dataset were drawn from the user-tagged photos of approximately 50 college-age volunteers and their friends on the Facebook

social network; each volunteer authorized a Facebook application to allow us to collect this data. While Facebook is currently experimenting with computer-vision assisted tagging, the data described here were collected prior to the introduction of these automated features, and thus the sampling of which faces were tagged was not influenced by factors external to the users and their preferences. For the purposes of these experiments, face samples were extracted and labeled as described above and then grouped by individual, and face samples with OpenCV detection diameters less than 80 pixels were discarded for the purposes of this study. The 500 individuals with the largest number of remaining face samples were selected to create a larger database of individuals, and 100 of those individuals were chosen at random to form the dataset used here. Each individual is represented by 100 face samples chosen at random from the set of their available samples. We reserve the full set of 500 individuals for ongoing work.

2.2. The "PubFig83" Data set

The main disadvantage of the Facebook data set is that the images it contains are currently private. Facebook has recently made it easier for users to share their photographs with "Everyone", and, as a consequence, we expect that many tagged photographs and videos of an extremely large number of individuals will eventually be available to the public from Facebook and other sources. In an effort to facilitate academic research on familiar face recognition in the wild at the current time, however, we have created a data set of public face images culled from the web that we call *PubFig83*. Our hope is that recognition performance on *PubFig83* will be broadly predictive of recognition performance on more realistic face images from personal photos such as those shared on Facebook, and we can then use the much larger repository of Facebook images to explore how various algorithms perform with increasingly difficult images and much larger databases of people.

To create the *PubFig83* dataset, we began with the recently released *PubFig* dataset [10], which consists of a set of nearly 60,000 image URLs that depict 200 people, most of whom are well-known celebrities. In a series of processing steps, we selected a subset of *PubFig* that we hope will provide a stable foundation for face recognition research. First, we downloaded all of the images that were still available from the original image lists for both the *development* and *evaluation* sets, and we obtained roughly 89% of the original images. We then ran the OpenCV face detector on all downloaded images and treated the provided face label locations as "tags"; we proceeded to match the face detections with identity labels just as we did for the Facebook data set. This OpenCV filtering step left us with 90.6% of the readable images (80.9% of the original *PubFig* set).

Upon examination of the remaining images, we noticed several sets of near-duplicate images in many individual identity categories. These near-duplicate copies of a single image varied from the original in many ways: some were scaled, cropped, and compressed differently, some had their color spaces altered, and some had been digitally edited more substantially, with whole backgrounds replaced or overlays added. With millions of within-class image pairs to consider, we could not evaluate each pair manually. To remove the vast majority of images that could be duplicates and produce the final *PubFig83* dataset, we applied a simple but coarse method: we globally ranked all within-class image pairs by the similarity of their labeled face samples, and we treated a portion of the most similar image pairs as duplicates. We compared images on the basis of their face samples to avoid the effects of extreme cropping, and we scored each pair by the maximum correlation of the central region of the face sample in one image to the central region of the face sample in the other. After browsing the globally-ranked list of image pairs manually, we determined that most obvious near-duplicates landed in the top 4% of the list, so we treated all pairs in that range as duplicates. Manual inspection of the remaining images suggested that this technique eliminated the majority of the near-duplicate image pairs, but it also eliminated pairs of images in which the same individual makes nearly identical expressions on different occasions. This artificial culling of similar facial expressions makes this dataset more challenging than it would have been if we could have removed only the true duplicate images.

For this study, we further selected all of the individuals in both the development and evaluation sets for whom 100 or more face samples remained. This yielded a final dataset of 83 individuals suitable for large-scale face identification testing.

3. Biologically-inspired visual representations

In these experiments, we relied on a family of biologically-inspired visual representations designed to model various stages of visual cortex in the brain. These models, inspired by Fukushima's Neocognitron [6], belong to the broader class of convolutional neural networks [12], and they have previously been used in a variety of machine vision contexts [23, 8, 9, 16, 17, 18, 19, 15].

Two basic model classes were tested (Figure 2). First, we used *V1-like-Plus* (*V1-like* for short), a simple one-layer model characterized by a cascade of linear and nonlinear processing steps and designed to encapsulate some of the known properties of the first cortical processing stage in the primate brain. Our *V1-like* implementation was taken without modification from [16, 17].

Second, we used two- and three-layer models following the basic multi-layer model scheme described in [19] and [15]. Briefly, these models consist of multiple stacked lay-

ers of linear-nonlinear processing stages, similar to those in *V1-like*. Importantly, in order to speed the processing of these models, we disabled the learning mechanisms described in [19] and instead used random filter kernels drawn from a uniform distribution. Prior experience of our group and others [8] has suggested that random filters can in many cases function surprisingly well for models belonging to this general class.

A more complete description of each model class follows.

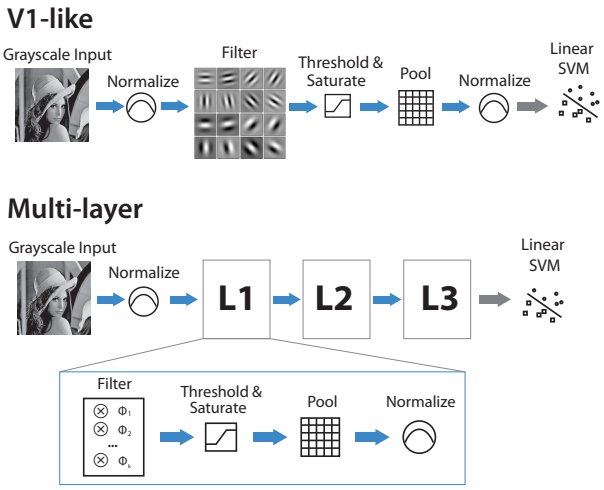


Figure 2. **A schematic diagram of the system architecture of the family of models considered.** Each model consists of one to three feedforward filtering layers, with the filters in each layer being applied across the previous layer (see Section 3).

3.1. *V1-like* visual representation

In the *V1-like* representation, features were taken without additional optimization from Pinto et al.'s V1S+ [16]. This visual representation is based on a first-order description of primary visual cortex V1 and consists of a collection of locally-normalized, thresholded Gabor wavelet functions spanning a range of orientations and spatial frequencies.

V1-like features have been proposed by neuroscientists as a "null" model (a baseline against which performance can be compared) for object and face recognition since they do not contain a particularly sophisticated representation of shape or appearance, nor do they possess any explicit mechanism designed to tolerate image variation (e.g. changes in view, lighting, position, etc. [4, 16]). Here, this model serves as a lower bound on the level of performance that can be achieved by relying only on relatively low-level regularities that exist in the test set.

Despite their simplicity, *V1-like* features have been shown to be among the best-performing non-blended feature sets on standard natural face and object recognition benchmarks (i.e. *Caltech-101*, *Caltech-256*, *ORL*, *Yale*, *CVL*, *AR*, *PIE*, *LFW* [16, 17, 18]), and they are a key component of the best blended solutions for some of these same benchmarks [7]. We used publicly available source code to generate these features and followed the same basic readout/classification procedure as detailed in [16], with two minor modifications. Specifically, no PCA dimensionality reduction was performed prior to classification (the full vector was used) and a different SVM regularization parameter was used ($C = 10^5$ instead of $C = 10$; see below).

3.2. High-throughput-derived multilayer visual representations: *HT-L2* and *HT-L3*

An important feature of the generation of these representations, according to the scheme set forth in [19], is the use of a massively parallel, high-throughput search over the parameter space of all possible instances of a large class of biologically-inspired models. Details of this model class and the high-throughput screening (model selection) procedure can be found in [15].

Candidate models were composed of a hierarchy of two (*HT-L2*) or three (*HT-L3*) layers, 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 Fig. 2, and, for the purposes of notation, the cascade of operations is represented as follows:

$Layer^0$:

$$\mathbf{Input} \overset{\mathbf{Grayscale}}{\longrightarrow} \overset{\mathbf{Normalize}}{\longrightarrow} \mathbf{N^0}$$

$Layer^1$:

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

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

$Layer^\ell$:

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

Details of these steps along with the range of parameter values included in the random search space are described in [15].

3.3. 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 [27]. Following [15], we selected the best five models from each "pool" for further analysis on the *Facebook100* , *PubFig83* and *LFW Restricted View 2* sets. Note that *LFW View 1* and *View 2* do not contain the same individuals and are thus fully mutually exclusive sets. *View 1* was designed as a model selection

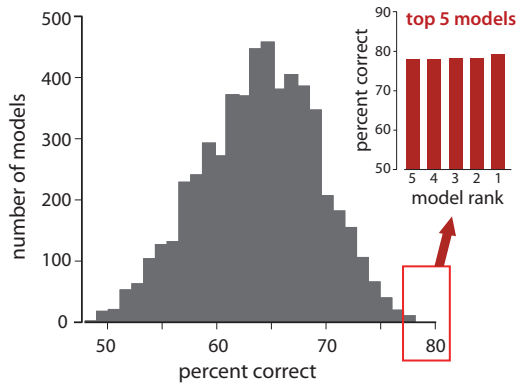


Figure 3. **An example of the high-throughput screening process used to find *HT-L2* and *HT-L3* representations.** Here, data is shown for the screening of *HT-L2* models. A distribution of the performance of 5,915 randomly generated models is shown on the left, with the top five high-performing models replotted on the right. Following screening, the models were evaluated exclusively with sets that do not overlap with the screening set.

set while *View 2* is used as an independent validation set for the purpose of comparing different methods. Importantly, no special optimization of these models was done for either *Facebook100* or *PubFig83*.

An example of the screening procedure for the *HT-L2* models on the *LFW View 1* task screening task is shown in Fig. 3. Performance of randomly generated *HT-L2* models ranged from chance performance (50%) to 80% correct; the best five models were drawn from this set and are denoted *HT-L2-1st*, *HT-L2-2nd*, and so on. An analogous procedure was undertaken to generate five three-layer models, denoted *HT-L3-1st*, *HT-L3-2nd*, etc.

3.4. Identification

To test in an *identification* mode for a given feature representation and data set, we first generated feature vectors for each image in the set. These feature vectors were then used to train a binary linear support vector machine (SVM) [22] per individual in a one-versus-all configuration [21] using the Shogun Toolbox [24] with the LIBSVM solver [3]. To avoid the computational cost of fitting the SVM's regularization hyperparameter C, we fixed C to a very high value (10^5), allowing no slack and thus resulting in a parameter-free hard-margin SVM.

Final performance values were computed as the average of ten random test/train splits of the data, with a variable number of training examples (see Figure 4) and ten testing examples per individual. In the case of the *Facebook100* set, all performance values presented here were the results of 100-way classification. For the *PubFig83* set, 83-way classifiers were used. For comparison experiments described in section 4.4, five random splits, instead of ten, were used.

3.5. Verification

To explore the relationship between identification and verification, we also used the *Facebook100* and *PubFig83* sets in a verification mode, following the structure of the *LFW* face verification protocol (*Restricted View 2*) as closely as possible. 6,000 different face image pairs (half "same", half "different") were drawn randomly from the sets and divided into 10-fold cross validation splits with 5,400 training and 600 testing examples each.

Because the biologically-inspired representations used here generate one feature vector per image, comparison functions were used to generate a new feature vector for each pair, and these "comparison" features were used to train binary ("same" / "different") hard-margin linear SVM classifiers. Following [18] and [15], we used four comparison functions: $|F_1 - F_2|$, $\sqrt{|F_1 - F_2|}$, $(F_1 - F_2)^2$, and $(F_1 \cdot F_2)$, where F_1 and F_2 are the feature vectors generated from the first and the second image of the pair, respectively.

As an additional point of reference, we also include verification performance on the *LFW* set. Verification performance was derived for the *Restricted View 2* portion of the set. Performance of the selected *V1-like*, *HT-L2*, and *HT-L3* models on *LFW* was also reported in [15]. While that work showed that relatively simple blended combinations of multiple models belonging to this class were able to significantly outperform the state-of-the-art on the *LFW* set ($>$ 88% performance), here we opted to use each model individually for the sake of simplicity (a total of 11 models were evaluated: one from *V1-like*, five from *HT-L2*, and five from *HT-L3*). Also, in contrast with [18, 15], we restricted ourselves to grayscale versions of the original image crops.

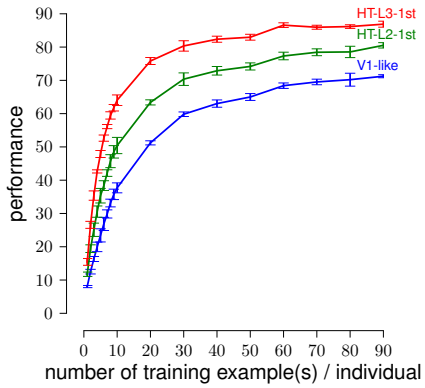
4. Results

4.1. Facebook100

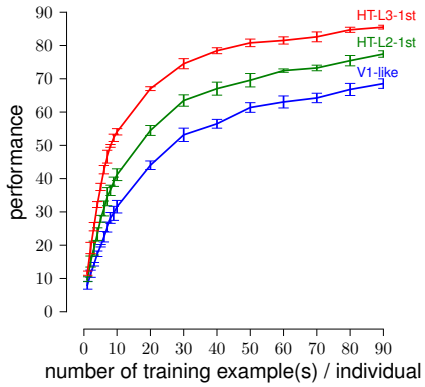
Performance using our biologically-inspired feature representations on the *Facebook100* followed the same basic pattern as had been previously observed for *Labeled Faces in the Wild* [15], with progressively more complex models (those with more layers) yielding progressively higher performance (i.e. *HT-L3* $>$ *HT-L2* $>$ *V1-like*). Figure 4(a) shows performance as a function of number of training examples per individual for the *V1-like*, *HT-L2-1st* (i.e. the best-ranked two-layer model, as ranked by its performance on the *Labeled Faces in the Wild View 1* set), and the *HT-L3-1st* models. Interestingly, we find that relatively high levels of performance (close to 90%) are possible on this 100-way identification task, especially as the number of training examples increases to 90.

4.2. Pubfig83

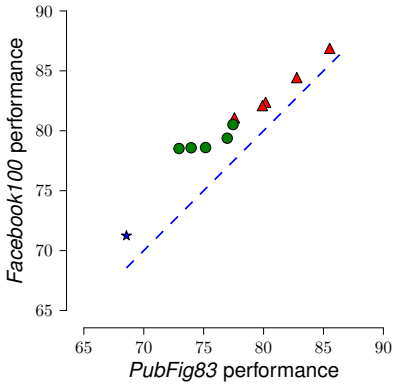
Performance on the *PubFig83* set followed appreciably the same trend as for the *Facebook100 set*. Figure 4(b)



(a) Facebook100



(b) *PubFig83*



(c) Comparison of identification performance on the *PubFig83* and *Facebook100* data sets.

Figure 4. Performance of three models as a function of the number of training examples per individual (top two plots); performance comparisons across models and data sets (bottom plot). Red triangles indicate *HT-L3* models, green circles indicate *HT-L2* models, and the blue star indiates *V1-like*.

shows performance of the *V1-like*, *HT-L2-1st*, and *HT-L3-1st* models as a function of the number of training examples per individual.

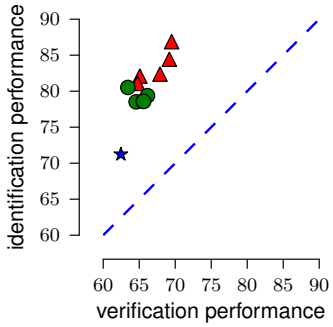
Asymptotic performance on the *PubFig83* set was lower for all feature representations as compared to performance on the *Facebook100* set. This is consistent with the fact that the creation of *PubFig83* involved an aggressive screening process designed to remove duplicates, which also removed many legitimate faces of each individual that were too similar to other distinct face samples of that individual. We hypothesize that these "typical" faces would be easier to classify, because their presence increases the odds that, for each test face, one or more similar faces would normally exist in the training set. Figure 4(c) shows a scatter plot of the relative performance on these two sets for each of the 11 models considered here (*V1-like*, five *HT-L2* models, and five *HT-L3* models). While the performance on the *PubFig83* set is displaced downward for all models, the relationship between performance on the *PubFig83* and *Facebook100* sets is remarkably linear.

4.3. Comparing Verification and Identification Paradigms

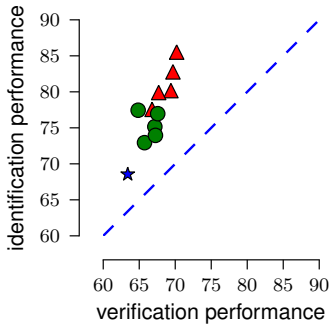
To explore the relationship between face verification and identification paradigms, we ran verification-mode experiments (in the style of *Labeled Faces in the Wild*) using the *Facebook100* and *PubFig83* sets. Verification performance on the *Facebook100* set ranged from 62.45%, with the *V1-like* model, to 69.5% for the best *HT-L3* model. Verification performance on the *PubFig83* set followed a similar range, with the *V1-like* model achieving 63.4% and the best *HT-L3* achieving 70.2%. Figure 5 shows the verification-mode performance of each of the 11 models considered here, plotted against their identification-mode performance. Interestingly, the rough rank order of models (from *V1-like* to *HT-L2* to *HT-L3*) is preserved in both verification and identification modes, and the approximately linear relationship between verification and identification in the *Facebook100* and *PubFig83* is quite similar despite these sets' substantially different provenance.

4.4. Comparison with a state-of-the-art commercial system

In order to provide some comparative grounding for these results, we also performed experiments using the the face recognition machinery provided online by Face.com. Figure 1 shows the performance of the top peforming *HT-L3* system, *V1-like*, and Face.com. Because Face.com additionally employs an alignment preprocessing step, and because the company makes this alignment functionality available through its web API, we ran the *HT-L3* and *V1-like* models on both raw and pre-aligned images. It should be noted that none of the systems evaluated here (*HT-L3*,



(a) *FB100* (verif.) vs *FB100* (ident.)



(b) *PubFig83* (verif.) vs *PubFig83* (ident.)

Figure 5. **Comparison of face verification and identification for 11 biologically-inspired models.** Symbols and colors follow the same conventions as in Figure 4.

V1-like , or Face.com) were specifically optimized in any way for these particular data sets.

Both of the "state-of-the-art" models (Face.com and *HT-L3-1st*) perform surprisingly well with the *Facebook100* and *PubFig83* set (Table 1), achieving 80%-90% accuracy in spite of the large number of individuals to be discriminated (i.e. 83 in the case of *PubFig83*, and 100 in the case of *Facebook100*). The *HT-L3* produced slightly higher performance, indicating that models of this class are highly competitive with other face recognition approaches. Predictably, the *V1-like-Plus* "baseline" model performs at a substantially lower level of performance, achieving ∼70% correct. Pre-alignment of the face images yielded little boost for the *HT-L3* model, while *V1-like* saw a comparatively larger boost from pre-aligning.

Table 1. Comparison of performance (accuracy ± std. err.) for *V1-like*, *HT-L3*, and the commercial face recognition system, *Face.com* (performance measured 10/2010)

		PubFig83	*Facebook100*
V1-like-Plus	unaligned	68.69 ± 0.65	74.08 ± 0.80
	aligned	75.64 ± 0.25	80.06 ± 0.68
Face.com	(aligned)	82.09 ± 0.47	84.50 ± 0.69
HT-L3-1st	unaligned	85.22 ± 0.45	87.02 ± 0.57
	aligned	87.11 ± 0.56	89.30 ± 0.33

5. Discussion

Here we have presented experiments in biologically-inspired face identification and verification in real-world settings. We introduced two new large-scale face identification sets: *Facebook100*, a naturalistic set of face images from users of the Facebook social networking website, and *PubFig83*, a filtered subset of the original *PubFig* data set with many near-duplicate images removed. While the *Facebook100* cannot be shared due to privacy concerns, our results indicate that, at least for the set of representations considered here, performance on *PubFig83* is highly predictive of performance on the *Facebook100* set, and we have made the *PubFig83* dataset available online[1]. As privacy norms continue to evolve on Facebook, we anticipate that much larger face sets (more individuals, more examples per individual) will eventually become available for research purposes.

The methods used to collect our datasets are samples from a larger space of possibilities. The original *PubFig* dataset leveraged text-image co-occurrence on the web to harvest facial images of famous individuals, and similar results can be obtained by exploiting captions in news feeds and videos [2, 5] or by combining image and video data [28]. In fact, because clothing and hair features allow faces in videos to be tracked through partial occlusion and drastic pose changes, face datasets harvested from video can more easily be built to include these large-scale effects [20]. In contrast, the faces in our datasets are currently filtered by a frontal face detector and therefore include only limited variations in pose. Overcoming this limitation in the future is an important line of research.

Another important finding from this study is that high levels of performance (85+%) are achievable with multi-layer biologically-inspired systems when reasonable quantities of training data are available. We note that we did not attempt to optimize any of the representations used

[1] http://www.eecs.harvard.edu/~zak/pubfig83

here for face identification, nor did we pursue any blending strategies to combine together multiple representation (such strategies have been demonstrated to yield even better performance [18, 15]). Consequently, the performance numbers presented here likely serve as a lower bound on performance that might be possible. Similarly, as even larger numbers of examples per individual are included (Facebook users are routinely tagged in hundreds if not thousands of photos), we anticipate higher performance still.

6. Acknowledgements

This study was funded in part by the Rowland Institute of Harvard, the NVIDIA Graduate Fellowship, the Singleton Fellowship, and the National Science Foundation (IIS0963668, CNS0708895, and IIS0905387).

References

[1] M. Ames and M. Naaman. Why We Tag: Motivations for Annotation in Mobile and Online Media. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 971–980, 2007. 26

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2004. 31

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. 2001. Software available at http://bit.ly/2rwxl. 29

[4] J. J. DiCarlo and D. D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007. 28

[5] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy – Automatic naming of characters in TV video. In *British Machine Vision Conference (BMVC)*, 2006. 31

[6] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980. 27

[7] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. *International Conference on Computer Vision (ICCV)*, 2009. 28

[8] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? *International Conference on Computer Vision (ICCV)*, 2009. 25, 27, 28

[9] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. *Computer Vision and Pattern Recognition Conference (CVPR)*, 2009. 25, 27

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. *International Conference on Computer Vision (ICCV)*, 2009. 25, 27

[11] D. LaVange. "Best Friends Forever" http://www.flickr.com/photos/wickenden/3259826856/. 25

[12] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, pages 255–258, 1995. 27

[13] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read. In *Conference on Hypertext and Hypermedia*, 2006. 26

[14] O. Nov, M. Naaman, and C. Ye. What drives content tagging: the case of photos on flickr. *SIGCHI Conference on Human Factors in Computing Systems*, 2008. 26

[15] N. Pinto and D. D. Cox. Beyond Simple Features: A Large-Scale Feature Search Approach to Unconstrained Face Recognition. *IEEE Automated Face and Gesture Recognition (FG)*, 2011. 25, 26, 27, 28, 29, 32

[16] N. Pinto, D. D. Cox, and J. J. DiCarlo. Why is Real-World Visual Object Recognition Hard? *PLoS Computational Biology*, 4(1):e27, 2008. 27, 28

[17] N. Pinto, J. J. DiCarlo, and D. D. Cox. Establishing Good Benchmarks and Baselines for Face Recognition. *European Conference on Computer Vision (ECCV)*, 2008. 25, 26, 27, 28

[18] N. Pinto, J. J. DiCarlo, and D. D. Cox. How far can you get with a modern face recognition test set using only simple features? *Computer Vision and Pattern Recognition Conference (CVPR)*, 2009. 25, 26, 27, 28, 29, 32

[19] N. Pinto, D. Doukhan, J. J. DiCarlo, and D. D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput Biol*, 2009. 25, 26, 27, 28

[20] D. Ramanan, S. Baker, and S. Kakade. Leveraging archival video for building face datasets. In *International Conference on Computer Vision (ICCV)*, 2007. 31

[21] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004. 29

[22] B. Scholkopf and A. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. the MIT Press, 2002. 29

[23] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411, 2007. 25, 27

[24] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large Scale Multiple Kernel Learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006. 29

[25] Z. Stone, T. Zickler, and T. Darrell. Autotagging Facebook: Social Network Context Improves Photo Annotation. In *IEEE Workshop on Internet Vision*, 2008. 26

[26] Z. Stone, T. Zickler, and T. Darrell. Toward Large-Scale Face Recognition Using Social Network Context. In *IEEE Special Edition on Internet Vision (To Appear)*, 2010. 26

[27] Y. Taigman, L. Wolf, T. Hassner, and I. Tel-Aviv. Multiple one-shots for utilizing class label information. In *British Machine Vision Conference (BMVC)*, 2009. 28

[28] M. Zhao, J. Yagnik, H. Adam, and D. Bau. Large scale learning and recognition of faces in web videos. In *Automatic Face and Gesture Recognition Conference (FG)*, 2008. 31