Establishing Good Benchmarks and Baselines for Face Recognition

Nicolas Pinto, James J. Dicarlo, David D. Cox

Establishing Good Benchmarks and Baselines for Face Recognition

Nicolas Pinto[1], James J. DiCarlo[1], and David D. Cox[2]

[1] Massachusetts Institute of Technology, Cambridge, MA 02139, USA,
{pinto, dicarlo}@mit.edu
[2] The Rowland Institute at Harvard, Harvard University, 100 Edwin Land Blvd., Cambridge, MA 02142, USA,
cox@rowland.harvard.edu

Abstract. Progress in face recognition relies critically on the creation of test sets against which the performance of various approaches can be evaluated. A good set must capture the essential elements of what makes the problem hard, while conforming to practical scale limitations. However, these goals are often deceptively difficult to achieve. In the related area of object recognition, Pinto et al. [2] demonstrated the potential dangers of using a large, uncontrolled natural image set, showing that an extremely rudimentary vision system (inspired by the early stages of visual processing in the brain) was able to perform on par with many state-of-the-art vision systems on the popular Caltech101 object set [3]. At the same time, this same rudimentary system was easily defeated by an ostensibly "simpler" synthetic recognition test designed to better span the range of real world variation in object pose, position, scale, etc. These results suggested that image sets that look "natural" to human observers may nonetheless fail to properly embody the problem of interest, and that care must be taken to establish baselines against which performance can be judged. Here, we repeat this approach for the "Labeled Faces in the Wild" (LFW) dataset [1], and for a collection of standard face recognition tests. The goal of the present work is not to compete in the LFW challenge, per se, but to provide a baseline against which the performance of other systems can be judged. In particular, we found that our rudimentary "baseline" vision system was able to achieve ∼68% correct performance on the LFW challenge, substantially higher than a pure "chance" baseline. We argue that this value might serve as a more useful baseline against which to evaluate absolute performance and argue that the LFW set, while perhaps not perfect, represents an improvement over other standard face sets.

1 Introduction

Highly accurate, "in-the-wild" face recognition is one of the holy grail applications in the field of artificial vision. While substantial progress has been made in the last several decades, the problem of face recognition in real-world images remains a largely unsolved challenge. At the heart of this challenge is the

considerable amount of image variation (e.g. position, size, orientation, lighting, clutter, occlusion, etc.) that a successful recognition system must tolerate, while maintaining its specificity for individual faces.

As with any engineering effort, it is essential to lay out a specification of what the problem is and what would constitute its solution. In the context of face recognition in real-world environments, this operationally amounts to constructing image test sets and "challenges" that capture the problem of interest. In practice, this is a daunting task, both because of the substantial effort associated with building the set (e.g. collecting and labeling images), and because it is difficult to construct a test set that is fully representative of the staggering image variation that is present in the real world.

At a deeper level, a fundamental problem is that no test set can practically be large enough to span the full range of variation observed in the "wild" (cf. the work of Torralba and colleagues [5, 6]). Compounding this problem, it is difficult to escape bias in the selection of images — most photographs are implicitly or explicitly centered and framed; frontal views of faces are typically over-represented, either by accident or by design (see the work . Sometimes, individual identity is correlated with background features (indeed, Shamir recently showed that relatively high performance was possible on a variety of standard face recognition sets using image patches taken from the background of images [7]). Taken together, these factors make it difficult to know whether or not "cheats" (i.e. trivial regularities, which exist in the test set, but not in the real world) exist for a given test set. Likewise, it is difficult to know what fraction of the performance achieved by a particular approach arises from exploitation of these low-level regularities, as opposed to from real progress towards a robust, general solution. This problem is compounded by the increasing complexity of artificial vision systems and the power of machine learning approaches [8], both of which make it difficult to determine which aspects of an image set a given system is actually utilizing to achieve its performance.

In recent years, it has become increasingly popular to evaluate artificial vision systems using large collections of "natural" images, such as can be harvested from the internet. Such image collections are appealing because they are relatively easy to assemble, and they typically include a wide range of sources, settings, etc. However, there is no guarantee that sets that "look" like they span the range of situations that would be encountered in the real world actually do so in reality. Due the nature of these tests, it is practically impossible to control for low-level statistical regularities that may significantly bias the results and potentially lead to wrong interpretations and conclusions.

In the related domain of object recognition, Pinto et al. [2] previously demonstrated some of the potential dangers associated with large uncontrolled image sets, showing that an extremely rudimentary vision system (inspired by the early stages of visual processing in the brain) was able to perform on par with many state-of-the-art vision systems on the popular Caltech101 object recognition set [3]. At the same time, this same rudimentary system was easily defeated by an ostensibly "simpler" synthetic recognition test designed to better span the

range of real world variation in object pose, position, scale, etc. These results suggest that a substantial fraction of the Caltech101 problem can be solved using low-level features, without solving the core problem of image variation. It is important to note that this does not necessarily mean that the Caltech101 set is not useful or that systems that perform well on the Caltech101 do not contain good ideas; rather, it suggests that performance reports might better be judged relative to a baseline that takes these low-level regularities into account.

In the present work, we undertake a similar approach to investigate what might constitute a reasonable baseline benchmark for various face recognition image sets. In particular, we focus on a collection of old "standard" publicly-available face image sets in wide use: ORL [10], Yale [11], AR [12], CVL [13] and on the relatively new "Labeled Faces in the Wild" challenge set [1].

2 Simple baseline models: Pixel and V1-like representations

In the following experiments, we considered three basic representations to serve as potential baselines: 1) a raw grayscale pixels representation, 2) a "V1-like" representation, inspired by the known properties of cortical area V1, and 3) a "V1-like+" representation, which includes all of the V1 features, plus a grab-bag of easily-computed additional features (e.g. histograms).

The "V1-like" and "V1-like+" representations were constructed as described in [2], without any optimization or modification for the task at hand. Briefly, the model was composed of a population of locally-normalized, thresholded Gabor functions spanning a range of orientations and spatial frequencies. From a neuroscientist's perspective, these models are "null" models, because they include only a first-order description of the earliest stage of visual processing in the brain. Importantly, these models do not contain any particularly sophisticated representation of shape, nor do they possess any explicit mechanism designed to tolerate image variation (e.g. changes in view, lighting, position, etc.).

For the purposes of the analyzes that follow, the processing of images was divided into two phases: a *representation* phase, and a *classification* stage. For the pixel representation, the representation phase consisted of resizing each image by bicubic interpolation, then converting them to grayscale and finally unwrapping the pixels into a n-dimensional vector. For the V1-like models, each element in the representation corresponded to the "activity" of a simulated V1-simple-cell-like unit. Each response was computed by first locally normalizing the image (dividing each pixel's intensity value by the norm of the pixels in the 3x3 neighboring region), then applying a set of 96 spatially local (43x43 pixel) Gabor wavelet filters to the image (with a 1 pixel stride), and normalizing the output values (dividing by the norm of the output values of all 3x3 spatial region across all Gabor filter types). Output values were finally thresholded (values below 0 were clipped to zero) and clipped (values above 1 were clipped). The 96 Gabor filters were chosen such that they spanned an exhaustive cross of 16 orientations

(evenly spaced "around the clock") and 6 spatial frequencies ($\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{11}$, $\frac{1}{18}$ cycles/pixel). See [2] for a detailed description of these methods.

After each image was converted to an n-dimensional vector, for each of the "standard" face datasets, these vectors were then used as inputs to a linear SVM after dimensionality reduction by PCA. Where required, multi-class classification was implemented using a one-against-rest approach. For the LFW challenge set, a slightly different procedure was used (see below).

3 Commonly-used face datasets

In this section we briefly present the performance of these baseline models with previous face image sets (ORL, YALE, AR and CVL).To facilitate comparison with previous results, we followed established testing protocols in the literature for each image set (described below).

3.1 Olivetti Research Lab (ORL) dataset

The ORL face dataset [10] consists of images of 40 subjects, with 10 grayscale images (92x112) per subject, with random variations in facial expression, pose, and lighting. The standard task for this set is to identify which individual is present in a given image, based on some number of training examples. Because there are 40 individuals, theoretical chance (i.e. from guessing) is $\frac{1}{40}$, or 2.5%.

Following previously published protocols, classifiers were trained using 4 or 8 training examples per individual (reserving the remaining 6 or 2, respectively, for testing), with a 10-trial random subsampling cross-validation scheme. Figure 1 shows the performance using our three baseline representations, along with several performance reports from the literature. In general, in spite of their simplicity the "baseline" models perform very well on the ORL set, with the pixel representation yielding better than 98% correct (with 8 training examples), and the V1-like model achieving *perfect* performance. Given the triviality of these baseline models, these results call into question whether the ORL database provides any real leverage in evaluating face recognition models.

3.2 Yale dataset

The Yale face set [11] consists of images of 15 subjects, with 11 gray scale images (320x243) per subject with fixed variations in lighting (e.g. center, right, or left lighting) and expression (e.g. neutral, sad, sleepy, happy). As with the ORL set, the standard task is to identify the individual on the basis of some number of training examples. Theoretical chance is $\frac{1}{15}$, or 6.67%.

Performance was assessed in a manner comparable to that described above. Classifiers were trained with 4 or 8 training examples per individual (reserving the remaining 7 or 3 images, respectively, for testing), with 10-trial random subsampling cross-validation. Results are shown in Figure 2. Again, the "baseline" models perform extremely well — the V1-like model achieves near perfect performance ($> 99\%$) with 8 training examples.

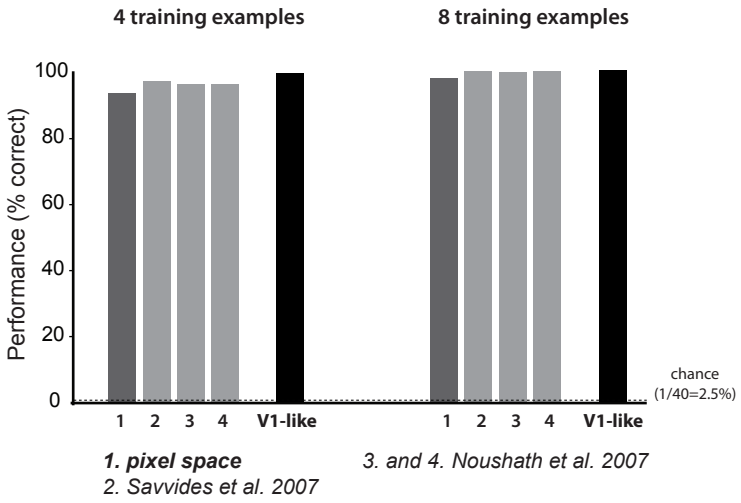

Fig. 1. Performance of Baseline Models on the ORL Set

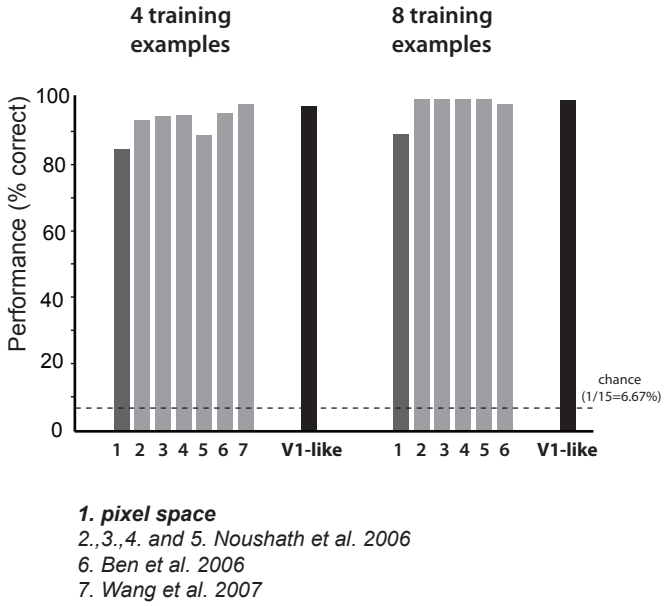

Fig. 2. Performance of Baseline Models on the Yale Set

3.3 Aleix and Robert (AR) dataset

The AR face set [12] consists of over 4,000 color face images (768x576) of 126 subjects. The majority of subjects (65 men and 55 women) participated in two session, separated by two weeks, where 26 images per subject (13 per session) were

taken with fixed variations in facial expressions (e.g. neutral, smile or anger), illumination conditions (e.g. left or right light), and occlusions (e.g sun glasses and scarf). Theoretical chance performance on this set is $\frac{1}{120}$, or 0.83%. Again, a 10-trial random subsampling cross-validation procedure was used. To facilitate comparison with existing literature, we trained with 5 and 8 training examples per individual (reserving the remaining images in each cross-validation split for testing). For the pixel representation, each color image was converted to grayscale and resized to 192x144. Results are shown in Figure 3. Once again, performance of the baseline models is comparable to or better than previous performance report from the literature, with the V1-like model achieving greater than 98% performance with 8 training examples.

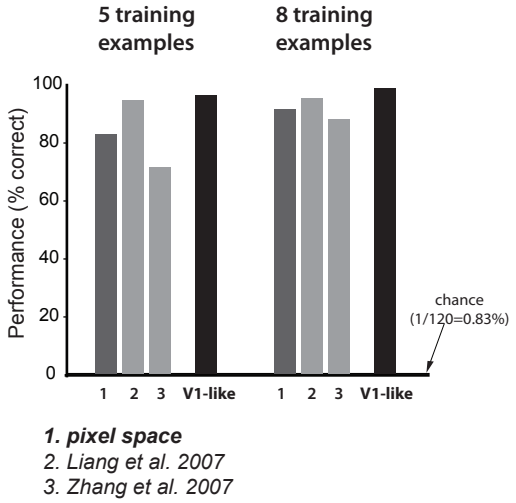

Fig. 3. Performance of Baseline Models on the AR Set

3.4 Computer Vision Laboratory (CVL) dataset

The CVL image set [13] contains 114 subjects, 7 color images (640x480) per subject with fixed variations in facial expression (i.e. smile or laugh) and pose (i.e. right, midright, frontal, midleft and left). Theoretical chance performance on this set is $\frac{1}{114}$, or 0.88%.

Following the existing literature using the CVL data set, we report performance using two distinct protocols. First, we trained and tested using only frontal views (2 training examples, 1 test example). In the second protocol, we used three examples drawn randomly from the 7 available images, and tested performance using the remaining 4. In both train/test protocols, 10-trial random subsampling was used for cross-validation. For the pixel representation,

each color image was converted to grayscale and resized to 160x120. Results are shown in Figure 4. Performance of the baseline models was again quite high, with the V1-like model outperforming reported performances from the literature in both training/testing protocols. In the case of protocol based on frontal views only, the V1-like model achieved better than 90% correct, indicating that the task under this protocol can be be largely solved using trivial regularities in the test set images. In the case of the protocol using all views, performance of the V1-like model was substantially lower (∼50% correct), though still substantially higher than the "chance" baseline (0.88%)

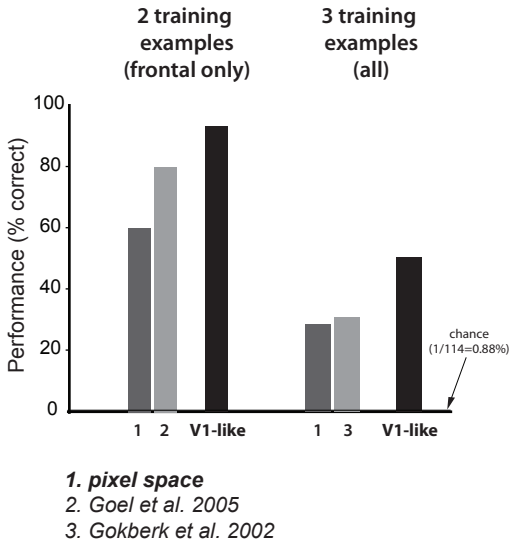

Fig. 4. Performance of Baseline Models on the CVL Set

4 Labeled Faces in the Wild (LFW) dataset

The recent Labeled Faces in the Wild (LFW) face set [1] contains 13,233 images of 5,749 individuals. This database is described by the creators as "unconstrained", meaning that face images are subject to a large range of "natural" variation (pose, lighting, focus, facial expression, background, age, etc).

The operational goal of this new image set is different from those presented above; it is aimed at studying the problem of face pair matching (i.e. given two face images, decide if they are from the same person or not). To accommodate this alternate goal, we took the vectors produced as the output of each representation (150x150 grayscale pixels, V1-like, V1-like+), and for each pair, we computed the element-wise squared difference. For each training pair, these

squared-difference vectors were labeled as "same" and "different," and the task of labeling new (test) examples was thus treated as a two-class classification problem (theoretical chance is 50%).

Prior to training a two-class linear SVM, training data were sphered (zero-mean and unit-variance feature wise), and dimensionality was reduced using principal components analysis (PCA), keeping as many dimensions as there were data points in the training set. Test data were transformed in an identical manner, using parameters (mean, standard deviation and principal components) computed exclusively from the training set.

Table 1. Performance of Baseline Models on the LFW Challenge Set

	Pixels	V1-like	V1-like+
mean (%)	59.95	64.21	68.08
std. error	0.64	0.69	0.45

Table 1 summarizes the mean classification accuracy on the "View 2" portion of the LFW set, using three baseline models: Pixels, V1-like and V1-like+. It is important to note that the V1-like models we used were taken verbatim from [2] — no attempt was made to optimize model parameters for the LFW challenge (i.e. using "View 1" images). Portable code (written in Python) and a minimal virtual machine environment (available in VMware or Amazon EC2 AMI format) are available upon request to facilitate reproducing these results.

While not quite as good, these baseline results are nonetheless reasonably close to early reported results on the LFW set (e.g. ∼72-73% [4]).

5 Counterpoint: A "simple" synthetic face dataset

To rule out the logical possibility that our baseline models (particularly the V1-like models) are actually effective face recognition systems, we constructed a synthetic image set that spans a range of view variation by design. In particular, the set consisted of two textured 3D face meshes (one male, one female; created with FaceGen, Singular Inversions [22]) that were ray-traced using POV-Ray [23] and randomly overlaid on a variety of backgrounds (as described in [2]). Critically, because the faces were rendered, known amount of variation in view, lighting, etc. could be introduced into the set, and this variation can be parametrically controlled. By the logic of most face recognition challenges, the set should be easy as only two faces must be discriminated, and ample training examples are available for each face.

Figure 5 shows the performance of the various baseline models with this synthetic set, as a function of the amount of view variation parametrically applied.

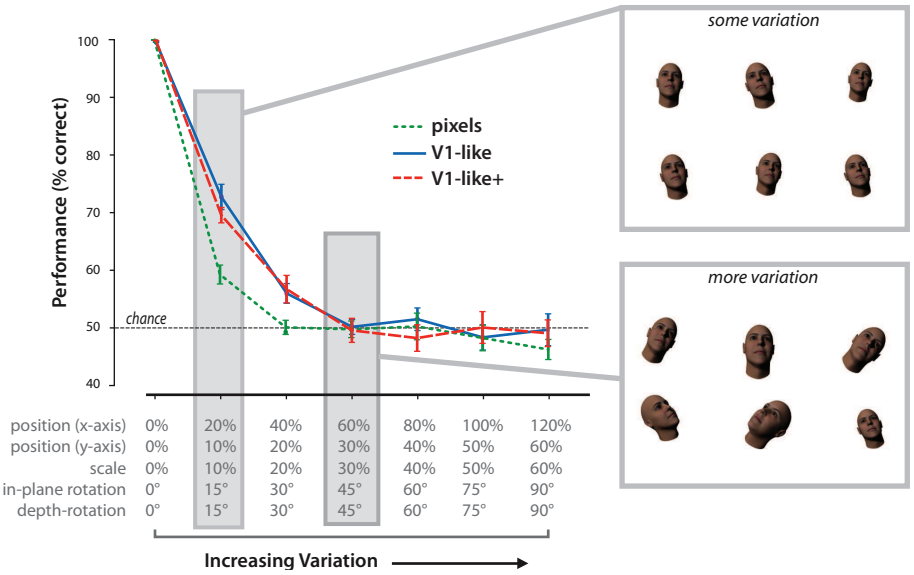

Fig. 5. Performance of Baseline Models on a Synthetic Face Set

As even modest amounts of view variation are included, performance rapidly declines. This rapid decline verifies that the baseline representations are not able to tolerate the sorts of image variation observed in the real world.

6 Discussion

Our results show that a simple V1-like vision system can perform extremely well on a variety of standard face recognition tests, and that it can perform moderately well on the LFW challenge test set. At the same time, we have shown that this same simple model performs at or near chance in a "simpler" face recognition task comprised of just two synthetic faces undergoing a wider range of view transformations. Taken together, these results suggest that while the V1-like model is demonstrably not a good general face recognition system, sufficient low-level regularities exist in each test set such that it can nonetheless perform surprisingly well. In the case of the "standard" face recognition sets that we tested (ORL, Yale, AR and CVL), the V1-like model can perform at or near 100%.

Interestingly, the V1-like model performs at ~68% correct on the new Labeled Faces in the Wild challenge, indicating that some, but not all, of the problem can be solved using a simple system relying on low-level cues. Clearly, there remains a substantial gap between this performance level and 100%, indicating that the LFW set has potential for guiding face recognition progress. We would argue, however, that performance reports on this set should be considered with

this number in mind. Given that a trivial algorithm can perform at close to 70% correct, models should ideally target substantially higher performance.

7 Acknowledgments

This work was funded in part by The National Institutes of Health (NIH-R01-EY014970), The McKnight Endowment Fund for Neuroscience, and The Rowland Institute at Harvard. We would like to thank David Doukhan and Youssef Barhomi for help in assembling the face databases for this work.

References

1. Huang G.B., Ramesh M., Berg T. and Learned-Miller E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49, 2007. http://vis-www.cs.umass.edu/lfw
2. Pinto N., Cox D.D. and DiCarlo J.J.: Why is Real-World Visual Object Recognition Hard? PLoS Computational Biology, 2008.
3. Fei-Fei L., Fergus R. and Perona P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE CVPR, 2004.
4. Nowak E. and Jurie F.: Learning visual similarity measures for comparing never seen objects. IEEE CVPR, 2007. http://vis-www.cs.umass.edu/lfw/results.html
5. Antonio Torralba's webpage: http://web.mit.edu/torralba/www
6. Ponce J., Berg T.L., Everingham M., Forsyth D.A., Hebert M., Lazebnik S., Marszalek M., Schmid C., Russell B.C., Torralba A., Williams C.K.I., Zhang J. and Zisserman A.: Dataset issues in object recognition. Lect Notes Comput Sci., Toward category-level object recognition, 2006.
7. Shamir L.: Evaluation of Face Datasets as Tools for Assessing the Performance of Face Recognition Methods. IJCV, 2008.
8. Hand D.J.: Classifier Technology and the Illusion of Progress. Statistical Science, 2006.
9. Gross R.: Face Databases. Handbook of Face Recognition, 2005.
10. ORL Dataset: http://www.cl.cam.ac.uk/Research/DTG/attarchive/facedatabase.html
11. YALE Dataset: http://cvc.yale.edu/projects/yalefaces/yalefaces.html
12. Martinez A.M. and Benavente R.: The AR Face Database. CVC Technical Report 24, June 1998. http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html
13. CVL Dataset: http://www.lrv.fri.uni-lj.si/facedb.html
14. Savvides M., Bhagavatula R., Li Y. and Abiantun R.: Face Recognition, Chapter 26: Frequency Domain Face Recognition, 2007.
15. Noushath S., Hemantha Kumar G., and Shivakumara P.: Diagonal Fisher linear discriminant analysis for efficient face recognition. Neurocomputing, 2006.
16. Ben N., Shiu S.C.K. and Pal S.K.: Two Dimensional Laplacianfaces Method for Face Recognition. LNCS RSCTC, 2006.
17. Wang L., Li Y., Wang C. and Zhang H.: Face Recognition using Gaborface-based 2DPCA and (2D) 2PCA Classification with Ensemble and Multichannel Model. IEEE CISDA, 2007.

18. Zhang L., Gao Q. and Zhang D.: Block Independent Component Analysis for Face Recognition. ICIAP, 2007.
19. Liang Y., Li C., Gong W. and Pan Y.: Uncorrelated linear discriminant analysis based on weighted pairwise Fisher criterion. Pattern Recognition, 2007.
20. Goel N., Bebis G. and Nefian A.: Face recognition experiments with random projection. SPIE, 2005.
21. Gokberk B., Akarun L. and Alpaydin E.: Feature selection for pose invariant face recognition. ICPR, 2002.
22. FaceGen: http://www.singularinversions.com
23. Persistence of Vision Raytracer (POV-Ray): http://www.povray.org