Subtasks of Unconstrained Face Recognition

Joel Z Leibo*,[1], Qianli Liao*,[1] and Tomaso Poggio[1]
*Authors contributed equally
[1]Center for Brains, Minds and Machines, McGovern Institute for Brain Research, MIT
Cambridge MA 02139, USA
jzleibo@mit.edu, lql@mit.edu, and tp@ai.mit.edu

Keywords: Invariance, Face identification, Same-different matching, Labeled Faces in the Wild, Synthetic data.

Abstract: Unconstrained face recognition remains a challenging computer vision problem despite recent exceptionally high results ($\sim 95\%$ accuracy) on the current gold standard evaluation dataset: Labeled Faces in the Wild (LFW) (Huang et al., 2008; Chen et al., 2013). We offer a decomposition of the unconstrained problem into subtasks based on the idea that invariance to identity-preserving transformations is the crux of recognition. Each of the subtasks in the *Subtasks of Unconstrained Face Recognition* (SUFR) challenge consists of a same-different face-matching problem on a set of 400 individual synthetic faces rendered so as to isolate a specific transformation or set of transformations. We characterized the performance of 9 different models (8 previously published) on each of the subtasks. One notable finding was that the HMAX-C2 feature was not nearly as clutter-resistant as had been suggested by previous publications (Leibo et al., 2010; Pinto et al., 2011). Next we considered LFW and argued that it is too easy of a task to continue to be regarded as a measure of progress on unconstrained face recognition. In particular, strong performance on LFW requires almost no invariance, yet it cannot be considered a fair approximation of the outcome of a detection→alignment pipeline since it does not contain the kinds of variability that realistic alignment systems produce when working on non-frontal faces. We offer a new, more difficult, natural image dataset: SUFR-in-the-Wild (SUFR-W), which we created using a protocol that was similar to LFW, but with a few differences designed to produce more need for transformation invariance. We present baseline results for eight different face recognition systems on the new dataset and argue that it is time to retire LFW and move on to more difficult evaluations for unconstrained face recognition.

1 INTRODUCTION

Current approaches to face recognition perform best on well-posed photographs taken for identification purposes, e.g., passport photos. However, in the real world, images of faces undergo many transformations—including aging, pose, illumination, expression, and many more. Not only do transformations degrade the performance of current algorithms, but in many cases they are known to lead to their catastrophic failure (Pinto et al., 2008a; Grother et al., 2010).

The computer vision and biometrics communities have responded to this challenge by shifting their focus to unconstrained benchmark datasets, of which Labeled Faces in the Wild (LFW) is generally considered to be the gold standard (Huang et al., 2008). LFW and similar datasets (e.g., PubFig83) consist of publicly available images of celebrities gathered from the internet and thus contain considerable variability.

The state-of-the-art on LFW has steadily improved in recent years to the point that it now arguably rivals human performance (on same-different matching of unfamiliar faces). At the time of writing, the best LFW performance is above 95% (Chen et al., 2013). However, we argue in this paper, there are several reasons that a declaration of victory over unconstrained face recognition remains premature.

1. The strong performance achieved on Labeled Faces in the Wild does not transfer to another, ostensibly quite similar, dataset we gathered.
2. The failure modes of state-of-the-art algorithms remain unclear. Moreover, when an algorithm does not perform well on an unconstrained test like LFW, it is not clear what aspect of the task is responsible.
3. Another goal is to understand the brain's solution to the unconstrained face recognition problem. In the Visual Psychophysics and Cognitive Neuroscience literature there is a wealth of available in-

formation concerning the robustness of human vision with respect to specific transformations, e.g, (Troje and Bülthoff, 1996; Braje et al., 1998). This data is typically gathered in highly controlled laboratory settings with one transformation varied at a time. Unless artificial systems are tested in comparable settings then there is no way to connect to this large body of previous work.

In this paper, we argue that in order to make further progress, it is necessary to simultaneously consider unconstrained face recognition along with its component subtasks. To that end, we contribute a collection of synthetic datasets (produced using 3D graphics) which, taken together, constitute a (partial) decomposition of unconstrained face recognition into its component subtasks. Our parsing of the full problem into subtasks is based on the premise that transformation invariance is the crux of recognition (Poggio et al., 2012; DiCarlo et al., 2012). We also gathered a new unconstrained dataset, similar to LFW (publicly available images on the Internet), but apparently more difficult. The entire collection of new datasets is available to researchers[1].

2 SUBTASKS

Our decomposition of unconstrained face recognition into subtasks is based on the idea that invariance to transformations is the main computational problem of recognition. The subtasks can be used to test face recognition systems. Unlike LFW, and similar datasets for which only a single accuracy score is measured, testing on all the subtasks gives a detailed analysis in terms of which transformations a system handles well and which cause it to fail.

The Subtasks of Unconstrained Face Recognition (SUFR) challenge is a collection of datasets which we call subtasks. Each subtask was designed to test specific aspects of the unconstrained face pair-matching (same-different) task. There are 400 individuals in each subtask. The total numbers of images range from $2,000$ for some of the smaller subtasks, to $10,000$ for some of the larger interaction tasks (tests with two transformations applied simultaneously). Each image is 512×512 pixels and in color. Since our goal in creating these datasets was precise control of transformation parameters, we employed 3D graphics software to synthesize the images. In section 3.1 we also describe a separate component of the challenge which uses natural images: SUFR-W.

[1]It can be downloaded from `http://cbmm.mit.edu/`.

The 400 textured head models were randomly generated using FaceGen (Singular_Inversions, 2003) and rendered onto a transparent background with Blender (Blender.org, 2013) using the CYCLES ray tracing engine. Most of the transformations required 3D information, e.g., rotation in depth and simulated movement of the illumination source. These transformations were applied with Blender. In other cases, images were transformed by explicitly specifying an affine matrix and using Matlab's image processing toolbox.

The SUFR challenge can be divided up in different ways. The "core" of the challenge is a set of six datasets which test transformation invariance directly. They consist of images of faces on a uniform black background. Another set of subtasks are concerned with transformation invariance in the presence of background clutter. Each image has a different randomly chosen natural scene or semi-structured random noise image in the background. Several subtasks are suitable for studying robustness to occlusion. Strong performance on these tasks requires invariance to whether or not a face is wearing sunglasses. Finally, there are also interaction tests. It is possible for a face recognition system to employ methods that successfully ensure invariance to any single transformation, but fail in combination. The interaction tests could quickly diagnose such issues. The full list of subtask datasets and benchmark results (without the random noise background sets for space reasons) is in table 1.

Testing face recognition algorithms on all the SUFR datasets yields a lot of information. However, it should be noted that SUFR is still only a partial decomposition of the unconstrained face recognition problem. In general, it would have to include transformations that are quite difficult to parametrize, e.g., facial expressions and aging. Thus our parsing of the full task remains somewhat incomplete since it only contains the transformations which we were able to simulate using 3D graphics software. Nevertheless, the SUFR benchmark contains many tests which are quite difficult for recent face recognition systems.

2.1 Performance of benchmark face recognition models

The intended use of the SUFR datasets is same-different matching of unfamiliar individuals (never seen during the training phase). This problem is sometimes called face-verification. It is identical to the standard procedure used with LFW. Unless mentioned otherwise, each test was performed by training a Support Vector Machine (SVM) using the difference

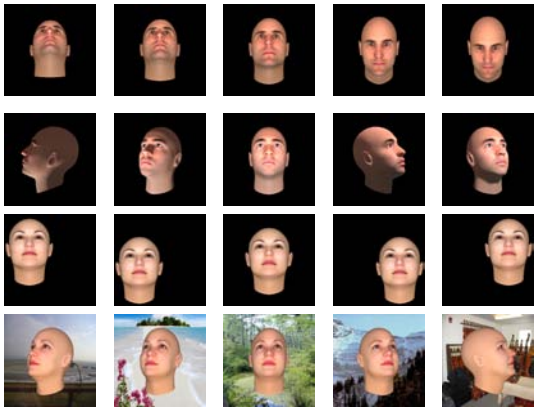

Figure 1: Example images

between the feature representations of the two images to be compared. 4000 image pairs were used for training and 4000 independent pairs for testing.

The SUFR benchmark results in table 1 include all nine models we tested. However, some care in interpretation is needed since they are not all directly comparable with one another. For example, some entries in table 1 correspond to testing concatenated vectors of local descriptors for their translation invariance. Obviously, they are not translation invariant—they were never intended to be.

2.1.1 Local descriptors

Many computer vision features (e.g. Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBP)) are extracted independently on a block-by-block basis. That is, first each image is subdivided into many relatively small blocks, then each is fed into a "feature extraction blackbox" which returns a vector of feature values. Vectors from all the blocks are concatenated to represent the entire image. Within this category, we tested Histograms of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Local Phase Quantization (LPQ). *Note:* Many of these features could be used as components of Global methods using bag of words or spatial pyramid approaches. We list them as "local" since their particular variant tested here was local.

Histograms of Oriented Gradients (HOG)

Originally proposed by Dalal and Triggs (2005), our experiments are based on the variant proposed by Felzenszwalb et al. (2010). The image was divided into blocks. For each one, a histogram of gradient orientations for each pixel is accumulated. The histogram of each block is then normalized with respect to neighboring blocks. We used an open source implementation from the VLFeat library (Vedaldi and Fulkerson, 2008).

Local Binary Patterns (LBP)

LBP (Ojala et al., 2002) and its generalizations to three-patch-LBP, four-patch-LBP and Local Quantized Patterns have been shown to be powerful representations for face recognition with LFW (Guillaumin et al., 2009; Wolf et al., 2011; Hussain et al., 2012). These methods work by thresholding the pixel intensities in a small region surrounding a central pixel and treating the resulting pattern as a binary number. As in HOG, histograms of local descriptors are accumulated in non-overlapping blocks. We used the implementation from VLFeat (Vedaldi and Fulkerson, 2008).

Local Phase Quantization (LPQ)

LPQ (Ojansivu and Heikkilä, 2008) is a blur-insensitive feature computed by quantizing the Fourier transform phase in local neighborhoods. Variants of LPQ were previously shown to outperform LBP on several datasets including LFW (Chan et al., 2013). We used an implementation provided by the author.

2.1.2 Features inspired by primary visual cortex

Hierarchical Model and X — C1 (HMAX-C1)

HMAX is a (partial) model of the primate ventral stream (Riesenhuber and Poggio, 1999), the part of cortex that is believed to be involved in object recognition. The elements of its C1 layer model complex cells in primary visual cortex (V1). We used the open source "CVPR06" implementation of HMAX which is distributed with the CNS simulation system (Mutch et al., 2010).

V1-like model (V1-like)

V1-like features are another family of low-level features intended to model the output of primary visual cortex (Pinto et al., 2008a). Variants of V1-like features were shown to be effective in various object and face recognition tasks (Pinto et al., 2008a, 2009). In all of our experiments, we used V1-like(A)—the best performing variant according to Pinto et al. (2009). We used an implementation provided by the author. Following their testing procedure, we reduced the dimensionality of the V1-like features by PCA[2] (Pinto et al., 2008b).

2.1.3 Global features

Hierarchical Model and X — C2 (HMAX-C2)

Another layer of HMAX. It was developed as a model for regions involved in later stages of ventral stream visual processing beyond primary

[2]Due to the large size of the features (86,400 per image) we only used 1,000 random training samples (out of 4,000) to compute principal components.

visual cortex. We used the open source "PNAS" implementation of HMAX from CNS (Mutch et al., 2010). This version corresponds to the "C2b" layer of Serre et al. (2007).

Scale-Invariant Feature Transform + Bag of Words or Spatial Pyramid (SIFT-BoW and SIFT-Pyr)

The Scale-invariant feature transform (or SIFT) (Lowe, 1999) is performed on a point-by-point basis. Canonically, 128 dimensional features can be extracted from a keypoint, but one cannot directly use it for classification. A common practice is to use a Bag-of-words (BoW) or spatial pyramid representation (Pyr), which treats each keypoint as a visual word and ignore its spacial location in the whole image (BoW) or each block (Pyr). A histogram of all visual words is computed as the final features. We used k-means clustering to quantize these visual words into 1024 clusters producing a final feature size of 1024 (BoW) or N*1024 (Pyr), where N is the number of blocks in the spatial pyramid. The open source implementation is provided by van de Sande et al. (2011).

2.1.4 An alignment-based system

SIFT-RANSAC→Warping→HOG features

We developed and tested the following pipeline—SIFT-RANSAC→Warping→HOG features. The idea is: given a pair of test images, warp the first image, A, to match the other image, B. If the warping is successful then A could be aligned with B and substantial affine transformations discounted. Since many other transformations are approximately affine (e.g. small yaw rotations) it is possible that this approach may also be successful in those cases. We implemented the common SIFT-RANSAC algorithm that is usually used for panoramic photo stitching. Then we extracted HOG features from image B and the warped image A. After that, we followed the same testing process as with the HOG features.

2.1.5 The SUFR benchmark clusters models by type

We used multidimensional scaling (MDS) to visualize the similarities between the pattern of results obtained with each feature set (fig. 2). Distance between points in the scatter plot corresponds to the Euclidean distance between each model's vector of accuracy values on the "core SUFR" subset: all single transformation subtasks with a uniform background. It shows that the feature types can be distinguished from one another by their pattern of SUFR results. Unsurprisingly, one MDS dimension appears to represent "globalness", HMAX-C2, the two SIFT-based models, and the RANSAC-HOG system are located at its extremes. The more local models inspired by primary

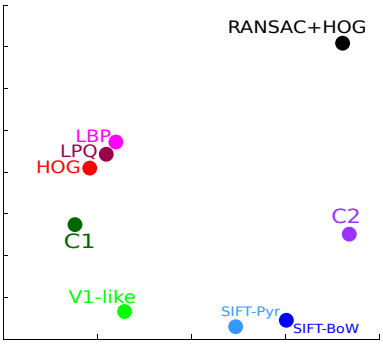

Figure 2: Multidimensional scaling based on the vector of performances on cure SUFR. Distance in the scatter point corresponds to the Euclidean distance between each model's vector of accuracies on the core SUFR tests.

visual cortex: HMAX-C1 and Pinto's V1-like model also cluster closer to one another than to other models, though interestingly, they are farther apart than we expected. A more surprising finding was that HOG, LPQ, and LBP all had quite similar patterns of results on the SUFR benchmark and all were relatively different from the local biologically-inspired features. As expected, the RANSAC-HOG system is isolated and far from other features. It works very well for all the affine transformations (even with background). But for non-affine transformations, it is fairly unstable and largely compromised, the same reason it is not applicable to real-world data.

2.1.6 Disrupting translation invariance with cluttered backgrounds

HMAX-C2 and SIFT-Bag-of-Words performed nearly perfectly on the tests of translation invariance without background clutter. However, both failed the same test in the presence of natural image clutter. This result was surprising since there are at least two previous reports in the literature that HMAX-C2 was translation-invariant on tasks with cluttered backgrounds (Leibo et al., 2010; Pinto et al., 2011).

Leibo et al. (2010) tested translation-invariant face pair-matching with and without background clutter. They reported that there was very little loss of accuracy due to clutter. However, it is likely that the clutter they used was too noise-like and not similar enough to the target class (natural faces). We observed that random semi-structured noise backgrounds do not have much effect on translation invariance for either HMAX-C2 or SIFT-BoW (fig. 3).

Pinto et al. (2011) followed a similar approach to ours. They also generated datasets of transforming objects using 3D graphics. However, they studied a basic level categorization task: cars vs. airplanes. They found that HMAX C2's performance was unaffected by translation over natural clutter. It is pos-

sible that this result was due to a difference between subordinate level face matching and their basic level task. But there were many other differences between the two studies that may also have been responsible. We also investigated a pure background-invariance task which was trivially easy for the local features and found that C2 and the SIFT-BoW method were quite disrupted by very small amounts of clutter—even when no translation invariance is necessary (fig. 4).

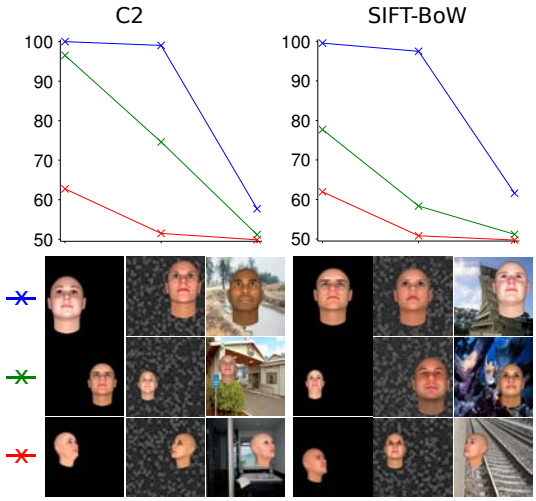

Figure 3: Top diagrams: Accuracy curves of C2 and SIFT-BoW over different transformations and background types (Blue: translation, Green: translation + scaling, Red:translation + yaw rotation). Y axis is verification accuracy in percentage. X axis is background type. 1 = no background. 2 = noise. 3 = natural images. Bottom row shows the example images used for the three curves, respectively.

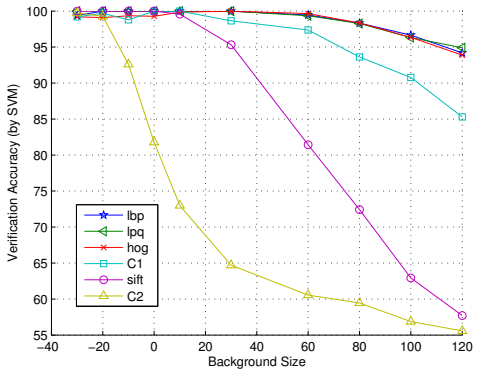

Figure 4: Performances of different models over different background sizes. It shows that global features (e.g., C2, SIFT) are much less tolerant of clutters, adding even a small amount of background lower their performances significantly.

3 FACE RECOGNITION IN THE WILD

If you accept the premise that transformation invariance is the crux of visual recognition then performance on the subtasks ought to be a good predictor of performance on the unconstrained task. However, if the meaning of "the unconstrained task" is "Labeled Faces in the Wild", this turns out not to be true. Figure 6 shows that many of the models we tested actually perform better on LFW than they do on most of the subtasks. How can this be?

It turns out that LFW doesn't really require substantial invariance to many of the transformations that the SUFR datasets were designed to test. The creators of LFW filtered its set of candidate images by the Viola-Jones face detector (Viola and Jones, 2004) which, for the most part, only detects nearly frontal faces. Thus LFW contains hardly any rotation in depth. Also, the faces are all centered and roughly the same size so translation and scale invariance are also unnecessary.

3.1 SUFR in the Wild (SUFR-W)

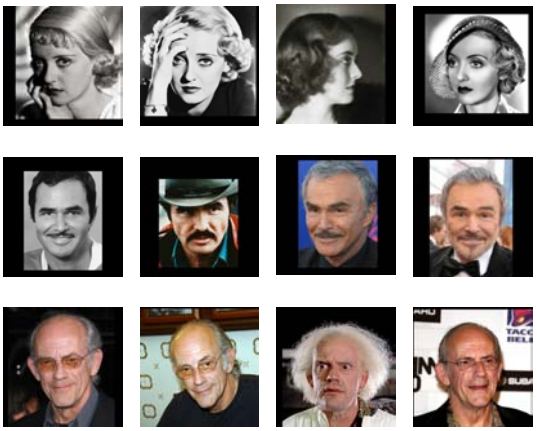

Figure 5: Example images in the SUFR-in-the-Wild dataset (SUFR-W). Top row: Bette Davis. Middle row: Burt Reynolds. Bottom row: Christopher Lloyd. The degree of alignment shown here is typical for the dataset. Profile faces as in the top row are rare.

In order to address these shortcomings of LFW, we created a new "unconstrained" natural image dataset using a very similar protocol to the one used by the creators of LFW. The new dataset, which we call SUFR-in-the-Wild (SUFR-W), is similar in size to LFW. It contains 13,661 images, slightly more than LFW's 13,233. While LFW contains a small number of images per person and a large number of people (5749 individuals), SUFR-W contains a much larger number of images of exactly 400 people (picked for uniformity with the synthetic SUFR datasets). See figure 5 for example SUFR-W images.

Table 1: Subtasks of Unconstrained Face Recognition benchmark results (% correct).

Core	C1	C2	HOG	LBP	LPQ	SIFT-BoW	SIFT-Pyr	V1-like	RANSAC+HOG
translation	52.8	**99.6**	53.0	55.0	55.9	**98.0**	**89.6**	69.6	**93.7**
scaling	61.7	**87.5**	61.7	61.0	62.7	64.7	63.7	55.3	**80.5**
in-plane rotation	61.4	**85.9**	71.3	79.3	71.2	77.9	71.5	63.1	**99.4**
pitch rotation	79.5	**90.0**	79.8	**84.1**	76.5	79.7	75.9	70.5	76.2
yaw rotation	57.1	**70.8**	58.6	64.8	60.3	67.1	63.1	59.8	55.1
illumination	**96.0**	**94.6**	**93.2**	**92.5**	**87.2**	**93.1**	**95.5**	**96.3**	71.7
Core + clutter	C1	C2	HOG	LBP	LPQ	SIFT-BoW	SIFT-Pyr	V1-like	RANSAC+HOG
translation	55.5	57.7	57.1	57.6	57.3	61.6	55.5	49.6	**97.1**
scaling	49.6	48.4	53.3	53.5	52.6	51.0	52.2	49.4	**89.5**
in-plane rotation	54.6	50.7	54.5	60.2	55.7	51.3	51.0	53.2	**96.6**
pitch rotation	54.1	52.5	54.5	60.1	55.9	51.0	52.7	55.4	**68.2**
yaw rotation	49.6	48.5	50.7	52.2	51.4	49.7	49.8	50.5	52.7
illumination	56.0	49.6	**67.0**	62.9	60.6	50.1	50.6	58.2	54.7
Interactions	C1	C2	HOG	LBP	LPQ	SIFT-BoW	SIFT-Pyr	V1-like	RANSAC+HOG
translation + scaling	53.5	**96.5**	53.0	53.2	53.3	77.7	67.6	51.5	**84.5**
translation + in-plane rotation	53.4	**87.1**	53.3	53.3	52.5	79.2	57.6	51.5	**91.8**
translation + yaw rotation	50.5	**62.7**	51.3	51.2	51.3	**62.0**	52.1	51.3	51.7
yaw rotation + illumination	56.5	58.5	52.6	54.2	54.9	59.3	57.1	57.4	52.7
Occlusion	C1	C2	HOG	LBP	LPQ	SIFT-BoW	SIFT-Pyr	V1-like	RANSAC+HOG
sunglasses + pitch rotation	76.6	69.5	79.7	**84.5**	77.6	75.8	73.5	64.2	63.6
sunglasses + yaw rotation	57.0	50.0	59.8	**69.3**	61.3	**67.9**	63.6	59.5	54.8

We gathered the images for SUFR-W using Google images. In order to avoid the same Viola-Jones filtering issue that prevented LFW from containing non-frontal faces, we did the following: First we manually eliminated all the images for each name that did not have a single isolated face, were not the correct person, or were too low resolution. Next, to prevent the dataset from being too difficult, we ran the Zhu and Ramanan (2012) face detection and landmark localization method. This method works particularly well with rotations in depth. It managed to detect all but ~ 30 of the candidate faces (which we then removed). To introduce some additional difficulty, but not too much, we allowed the Zhu and Ramanan (2012) system to attempt to align the images based on the landmarks it localized. However, it frequently failed to achieve a good alignment. Many of the faces (but not too many) remain clearly misaligned. Since we performed no further alignment, all these images are still misaligned in the final dataset.

SUFR-W contains none of the same individuals as LFW so it is straightforward to conduct experiments that train on one dataset and test on the other. As an unintended consequence of this, since so many celebrities are already in LFW, we had to look farther afield to find the individuals for SUFR-W. Many of them are actors and politicians who were active in the first half of the 20th century. Since these individuals are older today, we found that SUFR-W has considerably more variation in age than LFW. Of course, one unfortunate bias is that age is very clearly correlated with photography style (e.g. 'younger' implies 'probably black and white'). This is not a problem for the same-different matching task; though it does mean that successful algorithms will need to be reasonably tolerant of "the aging transformation".

While the systems we tested are not quite at the state-of-the-art, it is clear from the difference in performance between LFW and SUFR-W that the latter is a considerably more difficult dataset (fig. 6). At the same time, it is also clear that it is not so difficult that it cannot be used productively to guide future research.

3.2 On the use of LFW-a

Upon seeing figure 7, a colleague of ours remarked that the images in the *bottom* row are the ones for the task of face *recognition*. Depending on what part of the community you come from, that statement will either be obviously true or completely absurd.

Of the 123 papers indexed by Google Scholar that report results on LFW, at least 95 of them actually used a different, even more tightly aligned version[3]. Most of these paper (at least 58 of them) used LFW-a, a version of LFW which was *very* finely aligned with

[3]There were 9 papers that reported results on both and 23 papers for which we were unable to determine which dataset was used.

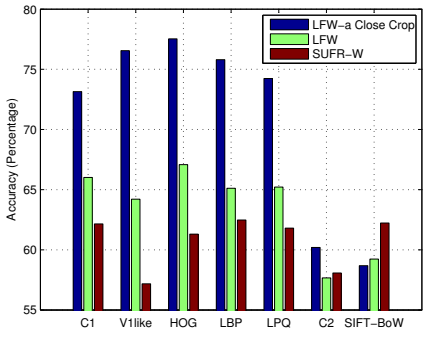

Figure 6: Results on natural image tasks (LFW-a closely cropped, LFW original and SUFR-W). The x axis is type of features. All the results are from our experiments, except that the LFW V1like is from Pinto et al. (2008b) and LFW-a close crop V1like is reported in Pinto et al. (2009). Our attempts to replicate these were stymied by a lack of computational resources.

Figure 7: Top: Three typical images from LFW. Bottom: The same three images in LFW-a.

a commercial software package (Wolf et al., 2011). The vast majority of papers using LFW-a crop all the images to an extremely tight, fixed, bounding box like the one shown in 7.

Even the relatively simple features we tested here are improved by up to 10% by using (cropped) LFW-a (fig. 6). Similar results have been reported before (e.g. Wolf et al. (2011)).

The argument in favor of taking results on the tightly cropped LFW-a test as a proxy for performance on unconstrained face recognition appeals to the detection $\rightarrow$ alignment $\rightarrow$ recognition (DAR) pipeline. In that framework, recognition is only the last step in a process through which transformations have already been discounted. It is acceptable to focus on a dataset containing hardly any transformations since normalizing those was already supposed to have been accomplished at earlier stages. However, there are several reasons not to take this argument at face value.

1. At best, the DAR framework guarantees that recognition systems will receive data that is as well-normalized as detection and alignment systems can deliver within application demands (e.g. processing time or cost). The creators of LFW referred to this issue when they wrote

"every face image in our database is the output of the Viola-Jones face detection algorithm. The motivation for this is as follows. If one can develop a face alignment algorithm (and subsequent recognition algorithm) that works directly on LFW, then it is likely to also work well in an end-to-end system that uses the Viola-Jones detector as a first step." (Huang et al., 2008).

This view of LFW is very conservative with respect to its implications for the full unconstrained face recognition problem. In this vein, the honest interpretation of the fact that the state-of-the-art on LFW-a is now 95% is: Consider the algorithm that first runs Viola-Jones (missing all the non-frontal faces), then has humans manually remove false positives, then passes the remaining images to the commercial system used to create LFW-a, and finally, then runs the best performing system on LFW-a. 5% of this algorithm's error rate would be attributed to the last step.

2. Within the DAR framework, a more fair natural image recognition test along the lines of LFW would, at least, have to include the kinds of images obtained by the errors of the previous stages. At least, these images should be included if the results are to be understood as measuring progress on unconstrained face recognition. Even if one expects to have relatively strong detection and alignment in the pipeline, it is still desirable for the last step to tolerate transformations. This allows the recognition system to "rescue" some alignment errors. It introduces redundancy into the system and prevents alignment from being a single point of failure.

3. It is interesting to consider to what extent, if any, the DAR framework is applicable to the brain's method of recognizing faces. Eye movements serve to approximately align images across time. However, numerous studies have shown that the brain's visual system is surprisingly tolerant of transformations, even when the images are flashed more quickly than the eyes can move (Hung et al., 2005). One interpretation is that the brain's visual system has two operating modes. One mode is faster and more automatic; it does not involve eye movements. The other mode operates more slowly, engages specific task-related information, and employs eye movements for alignment.

4 CONCLUSION

It has long been appreciated that the development of appropriate recognition tests to isolate subprob-

lems is essential to advancing computer vision. Notable datasets in this tradition include the Face Recognition Grand Challenge (FRGC) (Phillips et al., 2005) and Multi-PIE (Gross et al., 2010) datasets. Approaches based on synthetic data have fallen out of favor in recent years. While synthetic tests clearly have limitations: the variability within the class of synthetic faces does not approach that of natural faces. Tests with synthetic data also have numerous advantages. In particular, appearance transformations can be specified with a level of detail that could never be obtained in a dataset of natural photographs. Very large synthetic datasets can be created with no extra cost, in the case of the SUFR challenge, it was simple to include tests that address interaction effects between transformations. This could not have been done in a set of natural photographs without a costly investment.

We advocate an approach that combines tests on unconstrained natural image datasets like Labeled Faces in the Wild with detailed testing of particular subtasks. However, the results presented here, and (much more so) the work of Chen et al. (2013)—the creators of the current (95%) state-of-the-art system for LFW—argue that LFW may simply be too easy of a dataset to guide future progress. We suggested that the next generation of datasets ought to focus more on the problem of transformations. To that end, we are making the new SUFR-W dataset, as well as the complete set of synthetic datasets, available to interested researchers.

ACKNOWLEDGMENTS

This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF STC award CCF-1231216.

REFERENCES

Blender.org (2013). Blender 2.6.

Braje, W., Kersten, D., Tarr, M., and Troje, N. (1998). Illumination effects in face recognition. *Psychobiology*, 26(4):371–380.

Chan, C., Tahir, M., Kittler, J., and Pietikainen, M. (2013). Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1164–1177.

Chen, D., Cao, X., Wen, F., and Sun, J. (2013). Blessing of Dimensionality: High-dimensional Feature and Its Efficient Compression for Face Verification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

DiCarlo, J., Zoccolan, D., and Rust, N. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image and Vision Computing*, 28(5):807–813.

Grother, P., Quinn, G., and Phillips, P. (2010). Report on the evaluation of 2d still-image face recognition algorithms. *NIST Interagency Report*, 7709.

Guillaumin, M., Verbeek, J., and Chmid, C. (2009). Is that you? Metric learning approaches for face identification. In *IEEE International Conference on Computer Vision*, pages 498–505, Kyoto, Japan.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in real-life images: Detection, alignment and recognition (ECCV)*, Marseille, Fr.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749):863–866.

Hussain, S., Napoléon, T., and Jurie, F. (2012). Face recognition using local quantized patterns. In *Proc. British Machine Vision Conference (BMCV)*, volume 1, pages 52–61, Guildford, UK.

Leibo, J. Z., Mutch, J., Rosasco, L., Ullman, S., and Poggio, T. (2010). Learning Generic Invariances in Object Recognition: Translation and Scale. *MIT-CSAIL-TR-2010-061, CBCL-294*.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee.

Mutch, J., Knoblich, U., and Poggio, T. (2010). CNS: a GPU-based framework for simulating cortically-organized networks. *MIT-CSAIL-TR*, 2010-013(286).

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987.

Ojansivu, V. and Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. In *Image and Signal Processing*, pages 236–243. Springer.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W. (2005). Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 947–954. IEEE.

Pinto, N., Barhomi, Y., Cox, D., and DiCarlo, J. J. (2011). Comparing state-of-the-art visual features on invariant object recognition tasks. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 463–470. IEEE.

Pinto, N., Cox, D., and DiCarlo, J. J. (2008a). Why is real-world visual object recognition hard? *PLoS computational biology*, 4(1):e27.

Pinto, N., DiCarlo, J. J., and Cox, D. (2009). How far can you get with a modern face recognition test set using only simple features? In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2591–2598. IEEE.

Pinto, N., DiCarlo, J. J., Cox, D. D., et al. (2008b). Establishing good benchmarks and baselines for face recognition. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*.

Poggio, T., Mutch, J., Anselmi, F., Leibo, J. Z., Rosasco, L., and Tacchetti, A. (2012). The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work). *MIT-CSAIL-TR-2012-035*.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–6429.

Singular_Inversions (2003). FaceGen Modeller 3.

Troje, N. and Bülthoff, H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36(12):1761–1771.

van de Sande, K. E. A., Gevers, T., and Snoek, C. G. M. (2011). Empowering visual categorization with the gpu. *IEEE Transactions on Multimedia*, 13(1):60–70.

Vedaldi, A. and Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

Wolf, L., Hassner, T., and Taigman, Y. (2011). Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1978–1990.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, Providence, RI.