Attribute and Simile Classifiers for Face Verification

Neeraj Kumar Alexander C. Berg Peter N. Belhumeur Shree K. Nayar
Columbia University*

Abstract

We present two novel methods for face verification. Our first method – "attribute" classifiers – uses binary classifiers trained to recognize the presence or absence of describable aspects of visual appearance (e.g., gender, race, and age). Our second method – "simile" classifiers – removes the manual labeling required for attribute classification and instead learns the similarity of faces, or regions of faces, to specific reference people. Neither method requires costly, often brittle, alignment between image pairs; yet, both methods produce compact visual descriptions, and work on real-world images. Furthermore, both the attribute and simile classifiers improve on the current state-of-the-art for the LFW data set, reducing the error rates compared to the current best by 23.92% and 26.34%, respectively, and 31.68% when combined. For further testing across pose, illumination, and expression, we introduce a new data set – termed PubFig – of real-world images of public figures (celebrities and politicians) acquired from the internet. This data set is both larger (60,000 images) and deeper (300 images per individual) than existing data sets of its kind. Finally, we present an evaluation of human performance.

1. Introduction

There is enormous variability in the manner in which the same face presents itself to a camera: not only might the pose differ, but so might the expression and hairstyle. Making matters worse – at least for researchers in computer vision – is that the illumination direction, camera type, focus, resolution, and image compression are all almost certain to differ as well. These manifold differences in images of the same person have confounded methods for automatic face recognition and verification, often limiting the reliability of automatic algorithms to the domain of more controlled settings with cooperative subjects [33, 3, 29, 16, 30, 31, 14].

Recently, there has been significant work on the "Labeled Faces in the Wild" (LFW) data set [19]. This data set is remarkable in its variability, exhibiting all of the differences mentioned above. Not surprisingly, LFW has proven difficult for automatic face verification methods [25, 34, 17, 18, 19]. When one analyzes the failure cases for some of the existing algorithms, many mistakes are found that would seem to be avoidable: men being confused for women, young people for old, asians for caucasians, *etc*. On the other hand, small changes in pose, expression, or lighting can cause two otherwise similar images of the same person to be mis-classified by an algorithm as different. However, we show that humans do very well on the same data – given image pairs, verification of identity can be performed almost without error.

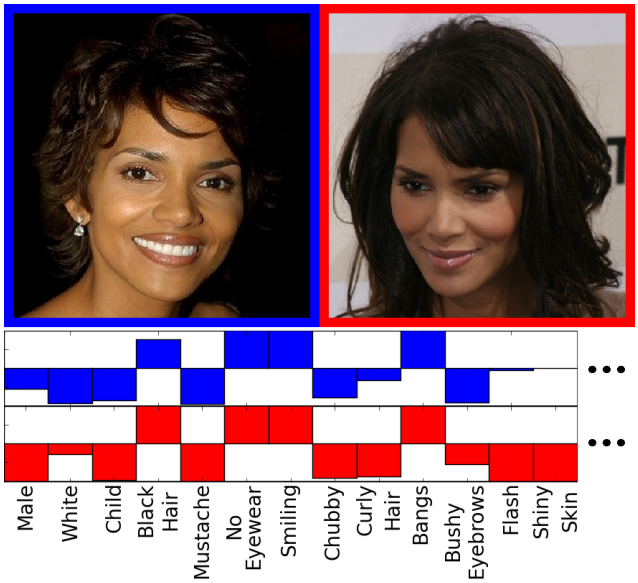

Figure 1: **Attribute Classifiers:** An attribute classifier can be trained to recognize the presence or absence of a *describable aspect* of visual appearance. The responses for several such attribute classifiers are shown for a pair of images of Halle Berry. Note that the "flash" and "shiny skin" attributes produce very different responses, while the responses for the remaining attributes are in strong agreement despite the changes in pose, illumination, expression, and image quality. We use these attributes for face verification, achieving a 23.92% drop in error rates on the LFW benchmark compared to the existing state-of-the-art.

In this paper, we attempt to advance the state-of-the-art for face verification in uncontrolled settings with non-cooperative subjects. To this end, we present two novel and complementary methods for face verification. Common to both methods is the idea of extracting and comparing "high-level" visual features, or traits, of a face image that

*{neeraj,aberg,belhumeur,nayar}@cs.columbia.edu

are insensitive to pose, illumination, expression, and other imaging conditions. These methods also have the advantage that the training data they require is easier to acquire than collecting a large gallery of images per enrolled individual (as is needed by traditional face recognition systems).

Our first method – based on attribute classifiers – uses binary classifiers trained to recognize the presence or absence of *describable aspects of visual appearance* (gender, race, age, hair color, *etc.*). We call these visual traits "attributes," following the name and method of [21]. For example, Figure 1 shows the values of various attributes for two images of Halle Berry. Note that the "flash" and "shiny skin" attributes produce very different responses, while the responses for the remaining attributes are in strong agreement despite the changes in pose, illumination, and expression. To date, we have built sixty-five attribute classifiers, although one could train many more.

Our second method – based on simile classifiers – removes the manual labeling required to train attribute classifiers. The simile classifiers are binary classifiers trained to recognize the *similarity of faces, or regions of faces, to specific reference people*. We call these visual traits "similes." The idea is to automatically learn similes that distinguish a person from the general population. An unseen face might be described as having a mouth that *looks like* Barack Obama's and a nose that *looks like* Owen Wilson's. Figure 2 shows the responses for several such simile classifiers for a pair of images of Harrison Ford. Rj denotes reference person j, so the first bar on the left displays similarity to the eyes of reference person 1. Note that the responses are, for the most part, in agreement despite the changes in pose, illumination, and expression. To date, we have used sixty reference people to build our simile classifiers, although many more could be added with little effort.

Our approach for face verification does not use expensive computation to align *pairs* of faces. The relatively short (65–3000 dimensional) vector of outputs from the trait classifiers (attribute and simile) are computed on each face independently. Comparing two faces is simply a matter of comparing these trait vectors. Remarkably, both the attribute and simile classifiers give state-of-the-art results, reducing the previous best error rates [34] on LFW [19] by 23.92% and 26.34%, respectively. To our knowledge this is the first time visual traits have been used for face verification.

As the attribute and simile classifiers offer complementary information, one would expect that combining these would further lower the error rates. For instance, it is possible for two people of different genders to have eyes like Salma Hayek's and noses like Meryl Streep's. So, while the simile classifier might confuse these, the attribute classifier would not. Our experiments seem to support this, as combining the attributes and similes together reduce the previous best error rates by 31.68%.

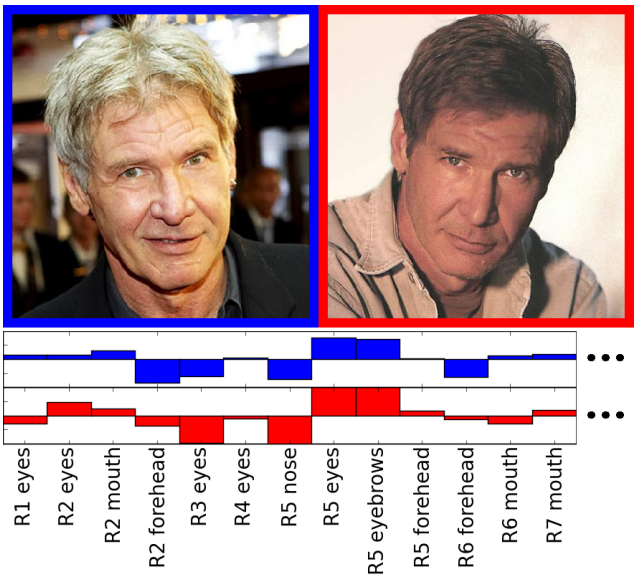

Figure 2: **Simile Classifiers:** We use a large number of "simile" classifiers trained to recognize the *similarities of parts of faces to specific reference people*. The responses for several such simile classifiers are shown for a pair of images of Harrison Ford. Rj denotes reference person j, so the first bar on the left displays the similarity to the eyes of reference person 1. Note that the responses are, for the most part, in agreement despite the changes in pose, illumination, and expression. We use these similes for face verification, achieving a 26.34% drop in error rates on the LFW benchmark compared to the existing state-of-the-art.

For testing beyond the LFW data set, we introduce PubFig – a new data set of real-world images of public figures (celebrities and politicians) acquired from the internet. The PubFig data set is both larger (60,000 images) and deeper (on average 300 images per individual) than existing data sets, and allows us to present verification results broken out by pose, illumination, and expression.

We summarize the contributions of the paper below:

1. **Attribute Classifiers:** We introduce classifiers for face verification, using 65 describable visual traits such as gender, age, race, hair color, *etc.*; the classifiers improve on the state-of-the-art, reducing overall error rates by 23.92% on LFW.
2. **Simile Classifiers:** We introduce classifiers for face verification, using similarities to a set of 60 reference faces; the classifiers improve on the state-of-the-art, reducing overall error rates by 26.34% on LFW. The simile classifiers do not require the manual labeling of training sets.
3. **PubFig Data set:** We introduce PubFig, the largest data set of real-world images (60,000) for face verification (and recognition), publicly available at http://www.cs.columbia.edu/CAVE/databases/pubfig/.
4. **Human Performance Evaluation:** We present an evaluation of human performance on the LFW data set.

2. Related Work

It is well understood that variation in pose and expression and, to a lesser extent, lighting cause significant difficulties for recognizing the identity of a person [35]. The Pose, Illumination, and Expression (PIE) data set and follow-on results [33] showed that sometimes alignment, especially in 3D, can overcome these difficulties [3, 4, 16, 33, 7].

Unfortunately, in the setting of real-world images such as those in Huang *et al*.'s "Labeled Faces in the Wild" (LFW) benchmark data set [19] and similar data sets [2, 10], 3D alignment is difficult and has not (yet) been demonstrated. Various 2D alignment strategies have been applied to LFW – aligning all faces [17] to each other, or aligning each pair of images to be considered for verification [25, 11]. Approaches that require alignment between each image pair are computationally expensive. Our work does not require pairwise alignment. Neither does that of the previously most successful approach on LFW from Wolf *et al*. [34], which uses a large set of carefully designed binary patch features. However, in contrast to Wolf *et al*. [34], our features are designed to provide information about the identity of an individual in two ways: by recognizing describable attributes (attribute classifiers), and by recognizing similarity to a set of reference people (simile classifiers).

Our low-level features are designed following a great deal of work in face recognition (and the larger recognition community) which has identified gradient direction and local descriptors around fiducial features as effective first steps toward dealing with illumination [6, 28, 22, 23, 10].

Automatically determining the gender of a face has been an active area of research since at least 1990 [15, 9], and includes more recent work [24] using Support Vector Machines (SVMs) [8]. This was later extended to the recognition of ethnicity [32], pose [20], expression [1], *etc*. More recently, a method for automatically training classifiers for these and many other types of attributes was proposed, for the purpose of searching databases of face images [21]. We follow their method for training our attribute classifiers, but improve on their feature selection process and the number of attributes considered. Gallagher and Chen [13] use estimates of age and gender to compute the likelihood of first names being associated with a particular face, but to our knowledge, no previous work has used attributes as features for face verification.

3. Our Approach

The first step of our approach is to extract "low-level" features from different regions of the face, *e.g.,* normalized pixel values, image gradient directions, or histograms of edge magnitudes. But as our aim is to design a face verification method that is tolerant of image changes, our second step is to use these low-level features to compute "high-level" visual features, or traits, which are insensitive to changes in pose, illumination, and expression. These visual traits are simply scores of our trait classifiers (attribute or simile). To perform face verification on a pair of images, we compare the scores in both images. Our steps are formalized below:

1. **Extract Low-level Features:** For each face image I, we extract the output of k low-level features $f_{i=1\ldots k}$ and concatenate these vectors to form a large feature vector $F(I) = \langle f_1(I), \cdots, f_k(I) \rangle$.

2. **Compute Visual Traits:** For each extracted feature vector $F(I)$, we compute the output of n trait classifiers $C_{i=1\ldots n}$ in order to produce a "trait vector" $\mathbf{C}(I)$ for the face, $\mathbf{C}(I) = \langle C_1(F(I)), \cdots, C_n(F(I)) \rangle$.

3. **Perform Verification:** To decide if two face images I_1 and I_2 are of the same person, we compare their trait vectors using a final classifier D which defines our verification function v:

$$v(I_1, I_2) = D\left(\mathbf{C}(I_1), \mathbf{C}(I_2)\right) \tag{1}$$

   $v(I_1, I_2)$ should be positive when the face images I_1 and I_2 show the same person and negative otherwise.

Section 3.1 describes the low-level features $\{f_i\}$. Our trait classifiers $\{C_i\}$ are discussed in Section 3.2 (attribute classifiers) and Section 3.3 (simile classifiers). The final classifier D is described in Section 3.4.

3.1. Low-level Features

To extract low-level features, we follow the procedure described in [21], summarized here. We first detect faces and fiducial point locations using a commercial face detector [26]. The faces are then rectified to a common coordinate system using an affine warp based on the fiducials. The low-level features are constructed by choosing a face region, a feature type to extract from this region, a normalization to apply to the extracted values, and an aggregation of these values.

The regions were constructed by hand-labeling different parts of the rectified face images, such as the eyes, nose, mouth, *etc*. (To handle the larger variation of pose in our data, we slightly enlarged the regions shown in [21].) Feature types include image intensities in RGB and HSV color spaces, edge magnitudes, and gradient directions. Normalization can be done by subtracting the mean and dividing by the standard deviation, or by just dividing by the mean, or not at all. Finally, the normalized values can be aggregated by concatenating them, collapsing them into histograms, or representing them only by their mean and variance.

This produces a large number of possible low-level features, $\{f_i\}$, a subset of which is automatically chosen and used for each trait classifier C_i, as described next.

Figure 3: **Attributes for Training:** Each row shows training examples of face images that match the given attribute label (positive examples) and those that don't (negative examples). Accuracies for all of our 65 attributes are shown in Table 1.

3.2. Attribute Classifiers

We build classifiers C_i to detect the describable attributes of faces, *e.g.*, as shown in Figure 3. While coarse attributes such as gender, age, and race will of course provide strong cues about a person's identity, these alone are not sufficient for successful face verification – we will need the outputs of as many different attributes as we can get.

We thus train several attribute classifiers, using an approach much like [21]. Their work treats attribute classification as a supervised learning problem. Training requires a set of positive and negative example images for each attribute, and uses a simplified version of adaboost [12] to choose from the set of low-level features described in the previous section. However, one downside to their simplification of ababoost is that the weak learners are only trained once during feature selection. To get around this drawback, we use forward feature selection, where we consider appending each remaining feature to the current feature set and choose the one that drops error rates the most. We do this greedily to pick up to a maximum of 6 features.

Each attribute classifier is an SVM with an RBF kernel, trained using libsvm [5]. The accuracies on held out data of 65 attribute classifiers trained using our system are shown in Table 1. We note that although a few are lower than [21], the images used in our system are not limited to only frontal poses (as in theirs). Examples of some of the training data used for a few of our attributes are shown in Figure 3.

Obtaining Training Data: Our attribute training procedure is fully automatic given the initial labeling of positive and negative examples. At 1,000+ examples per attribute (at least 500 positive and 500 negative), this quickly becomes the main bottleneck in our attribute training process – for our set of 65 attributes, we had to obtain at least 65,000 labels for training, and more for validation.

To collect this large number of labels, we used Amazon Mechanical Turk.[1] This service matches online workers to online jobs. "Requesters" can submit jobs to be completed by workers, optionally setting various quality controls such as confirmation of results by multiple workers, filters on minimum worker experience, *etc*. The jobs we created asked workers to mark face images which exhibited a specified attribute. (A few manually-labeled images were shown as examples.) Each job was submitted to 3 different workers and only labels where all 3 people agreed were used. In this way, we collected over 125,000 confirmed labels over the course of a month, for less than $5,000.[2]

3.3. Simile Classifiers

The attribute classifiers described in the previous section, while offering state-of-the-art performance on LFW, require each attribute to be describable in words. However, one can imagine that there are many visual cues to people's identities that cannot be described – at least not concisely.

In order to use this complementary information, we introduce the concept of a "simile" classifier. The basic idea is that we can describe a person's appearance in terms of the similarity of different parts of their face to a limited set of "reference" people. For example, someone's mouth might

[1] http://mturk.com

[2] We submitted 73,000 jobs showing 30 images to each of the 3 workers per job, gathering a total of 6.5 million user inputs.

Attribute	Accuracy	Attribute	Accuracy
Asian	92.32%	Mouth Wide Open	89.63%
Attractive Woman	81.13%	Mustache	91.88%
Baby	90.45%	No Beard	89.53%
Bags Under Eyes	86.23%	No Eyewear	93.55%
Bald	83.22%	Nose Shape	86.87%
Bangs	88.70%	Nose Size	87.50%
Black	88.65%	Nose-Mouth Lines	93.10%
Black Hair	80.32%	Obstructed Forehead	79.11%
Blond Hair	78.05%	Oval Face	70.26%
Blurry	92.12%	Pale Skin	89.44%
Brown Hair	72.42%	Posed Photo	69.72%
Child	83.58%	Receding Hairline	84.15%
Chubby	77.24%	Rosy Cheeks	85.82%
Color Photo	95.50%	Round Face	74.33%
Curly Hair	68.88%	Round Jaw	66.99%
Double Chin	77.68%	Semi-Obscured Forehead	77.02%
Environment	84.80%	Senior	88.74%
Eye Width	90.02%	Shiny Skin	84.73%
Eyebrow Shape	80.90%	Sideburns	71.07%
Eyebrow Thickness	93.40%	Smiling	95.33%
Eyeglasses	91.56%	Soft Lighting	67.81%
Eyes Open	92.52%	Square Face	81.19%
Flash Lighting	72.33%	Straight Hair	76.81%
Frowning	95.47%	Sunglasses	94.91%
Goatee	80.35%	Teeth Not Visible	91.64%
Gray Hair	87.18%	Teeth Visible	91.64%
Harsh Lighting	78.74%	Visible Forehead	89.43%
High Cheekbones	84.70%	Wavy Hair	64.49%
Indian	86.47%	Wearing Hat	85.97%
Male	81.22%	Wearing Lipstick	86.78%
Middle-Aged	78.39%	White	91.48%
Mouth Closed	89.27%	Youth	85.79%
Mouth Partially Open	85.13%		

Table 1: **Attribute Classification Results:** We present accuracies of the 65 attribute classifiers trained using the procedure described in Sec. 3.2. Example training images for the attributes in **bold** are shown in Figure 3.

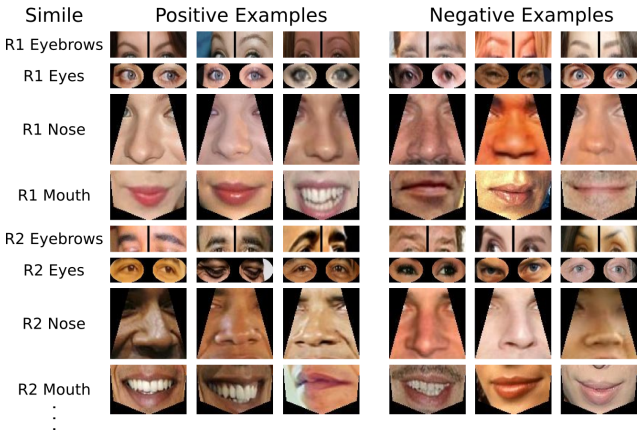

Figure 4: **Similes for Training:** Each simile classifier is trained using several images of a specific reference person, limited to a small face region such as the eyes, nose, or mouth. We show here three positive and three negative examples for four regions on two of the reference people used to train these classifiers.

be described as similar to Angelina Jolie's, or their nose as similar to Brad Pitt's. Dissimilarities also provide useful information – *e.g.*, her eyes are *not* like Jennifer Aniston's.

Figure 4 shows examples of regions selected from subjects "R1" and "R2" in the training data. For each reference person in the training set, several simile classifiers are trained for each face region (one per feature type), yielding a large set of total classifiers.

We emphasize two points. First, the individuals chosen as reference people *do not appear* in LFW or other benchmarks on which we produce results. Second, we train simile classifiers to recognize similarity to *part* of a reference person's face in *many* images, not similarity to a single image.

For each reference person, we train classifiers to distinguish a region (*e.g.*, eyebrows, eyes, nose, mouth) on their face from the same region on other faces. We choose eight regions and six feature types from the set of possible features described in Sec. 3.1 and train classifiers for each simile using the training procedure described in the previous section. Each simile classifier is trained using at most 600 positive example face images of the reference person, and at most 10 times as many negative examples, randomly sampled from images of other people in the training set.

Obtaining Training Data: The data required for training simile classifiers is simpler than for attribute classification: for positive examples, images of a particular person; for negative examples, images of other people. This training data is part of the PubFig data set, described in Sec. 4.3.

3.4. Verification Classifier

In order to make a decision about whether two face images I_1 and I_2 show the same person, we use the final classifier D to compare the trait vectors $\mathbf{C}(I_1)$ and $\mathbf{C}(I_2)$ obtained by one or both of the methods above.

We build our final classifier D based on some observations about our approach: (1) corresponding values $C_i(I_1)$ and $C_i(I_2)$ from the ith trait classifier should be similar if the images are of the same individual, (2) trait values are raw outputs of binary classifiers (in the range $[-1, 1]$), and so the signs of values should be important, and (3) our particular choice of classifier, SVMs, optimize for separating data at the separation boundary, and so differences in values close to 0 are more important than differences between those with greater absolute values.

Let $a_i = C_i(I_1)$ and $b_i = C_i(I_2)$ be the outputs of the ith trait classifier for each face. For each of the n trait classifiers, we compute a pair $p_i = (|a_i - b_i|, (a_i \cdot b_i)) \cdot g\left(\frac{1}{2}(a_i + b_i)\right)$, where the first term is the absolute value of the difference between the two trait vectors and second term is their product, and both are weighted by a gaussian g with mean 0 and variance 1. These pairs are concatenated to form the $2n$ dimensional vector that we actually classify:

$$v(I_1, I_2) = D(\langle p_1, \ldots, p_n \rangle) \quad (2)$$

We found that changing the exact nature of D (*e.g.*, using either the difference or the product, or not applying the gaussian weighting) did not affect accuracy by more than 1%. Training D requires pairs of positive examples (both images of the same person) and negative examples (images of different people). In our experiments, we use an SVM with an RBF kernel for D.

4. Experiments

All of our experiments evaluate performance on a face verification task: given two images of faces, determine if they show the same individual. For each computational experiment, a set of pairs of face images is presented for training, and a second set of pairs is presented for testing. Not only are the images in the training and test sets disjoint, but there is also no overlap in the individuals used in the two sets. High-level model selection choices (*e.g.*, representation for the final classifier D) were made using a separate training/test set (*e.g.*, View 1 of the LFW set, as described in the next section). Also, both our trait classifiers – attribute and simile – were trained on data disjoint (by image and identity) from the train and test sets in the experiments.

We explore performance on the LFW benchmark (Sec. 4.1) and on our PubFig benchmark (Sec. 4.3), varying the set of traits used. In both cases, we use the detected yaw angles to first flip images so that they always face left. We also use six additional features for verification: the three pose angles, the pose confidence, and two quality measures based on image and file sizes. These boost performance slightly, especially for off-frontal faces. Because the PubFig data set has more images per individual, we also evaluate performance as a function of pose, lighting, and expression on that data set. Finally, we present results showing

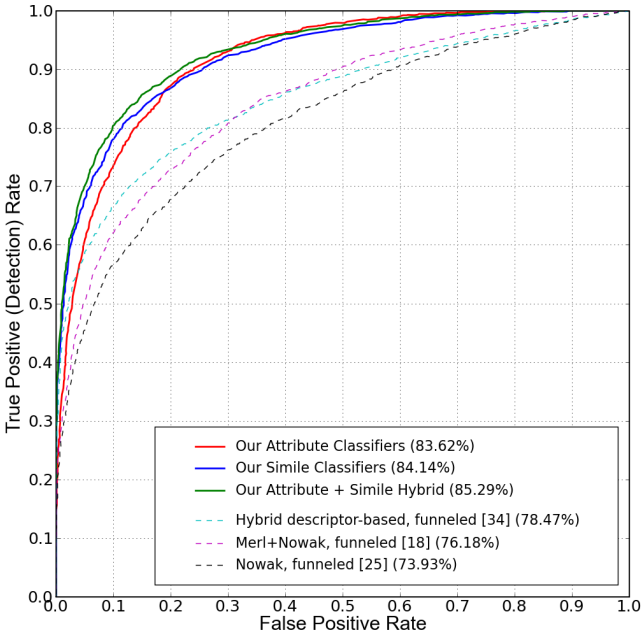

Figure 5: **Face Verification Results on LFW:** Performance of our attribute classifiers, simile classifiers, and a hybrid of the two are shown in solid red, blue, and green, respectively. All 3 of our methods outperform all previous methods (dashed lines). Our highest accuracy is 85.29%, which corresponds to a 31.68% lower error rate than the previous state-of-the-art.

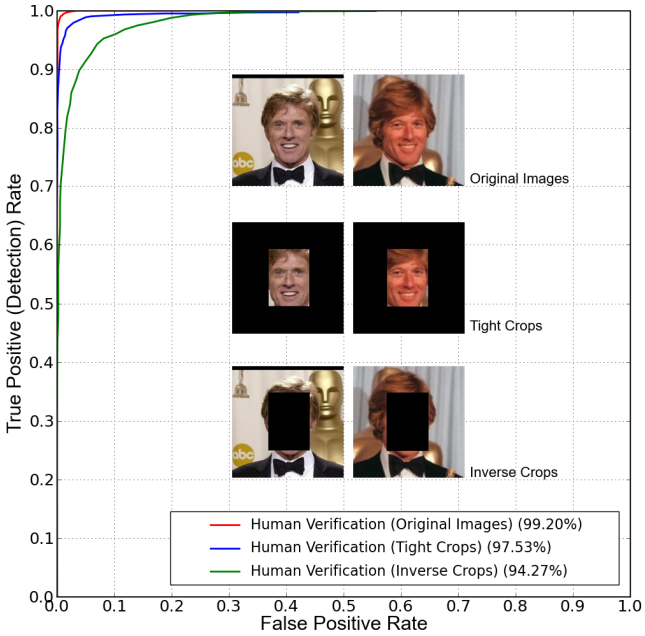

Figure 6: **Human Face Verification Results on LFW:** Human performance on LFW is almost perfect (99.20%) when people are shown the original images (red line). Showing a tighter cropped version of the images (blue line) drops their accuracy to 97.53%, due to the lack of context available. The green line shows that even with an inverse crop, *i.e.*, when *only* the context is shown, humans still perform amazingly well, at 94.27%. This highlights the strong context cues available on the LFW data set. All of our methods mask out the background to avoid using this information.

human performance on the LFW set (Sec. 4.2). Unlike the algorithms, humans were not shown any training data.

4.1. Labeled Faces in the Wild

The Labeled Faces in the Wild (LFW) [19] data set consists of 13,233 images of 5,749 people, which are organized into 2 views – a development set of 2,200 pairs for training and 1,000 pairs for testing, on which to build models and choose features; and a 10-fold cross-validation set of 6,000 pairs, on which to evaluate final performance. We use View 1 for high-level model selection and evaluate our performance on each of the folds in View 2 using the "image restricted configuration," as follows.

For each split, we train a final classifier D on the training data and evaluate on the test data. Receiver Operating Characteristic (ROC) curves are obtained by saving the classifier outputs for each test pair and then sliding a threshold over all values to obtain different false positive/detection rates. An overall accuracy is obtained by using only the signs of the outputs and counting the number of errors in classification. The standard deviation for the accuracy is obtained by looking at the accuracies for each fold individually.

Figure 5 shows results on LFW for our attribute classifiers (red line), simile classifiers (blue line), and a hybrid of the two (green line), along with several previous methods (dotted lines). The accuracies for each method are $83.62\% \pm 1.58\%$, $84.14\% \pm 1.31\%$, and $85.29\% \pm 1.23\%$, respectively.[3] Each of our methods out-performs all previous methods. Our highest performance is with the hybrid method, which achieves a 31.68% drop in error rates from the previous state-of-the-art.

4.2. Human Performance on LFW

While many algorithms for automatic face verification have been designed and evaluated on LFW, there are no published results about how well people perform on this benchmark. Furthermore, it is unknown what characteristics of the data set might make it easier or harder to perform the verification task. To this end, we conducted several experiments on human verification. To obtain this data, we followed the procedure of [27], but on Amazon Mechanical Turk, averaging the replies of 10 different users per pair to get smoothed estimates of average human performance. Thus, for the 6,000 image pairs in LFW, we gathered 60,000 data points from users for each of the three tests described below (for a total of 240,000 user decisions). To create an ROC curve for the results, the users were asked to rate their confidence in labeling each pair of images as belonging to the same person or not.

[3] Our face detector [26] was unable to detect one or more faces in 53 of the 6,000 total pairs. For these, we assumed average performance.

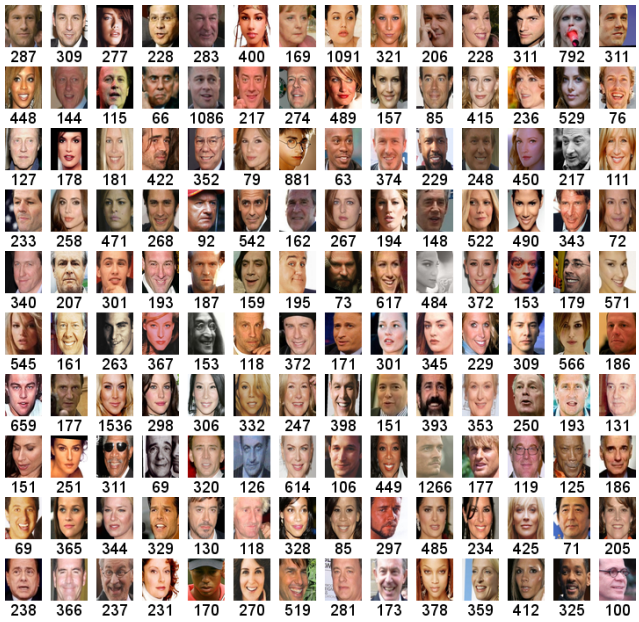

Figure 7: **The PubFig Data Set:** We show example images for the 140 people used for verification tests on the PubFig benchmark. Below each image is the total number of face images for that person in the entire data set.

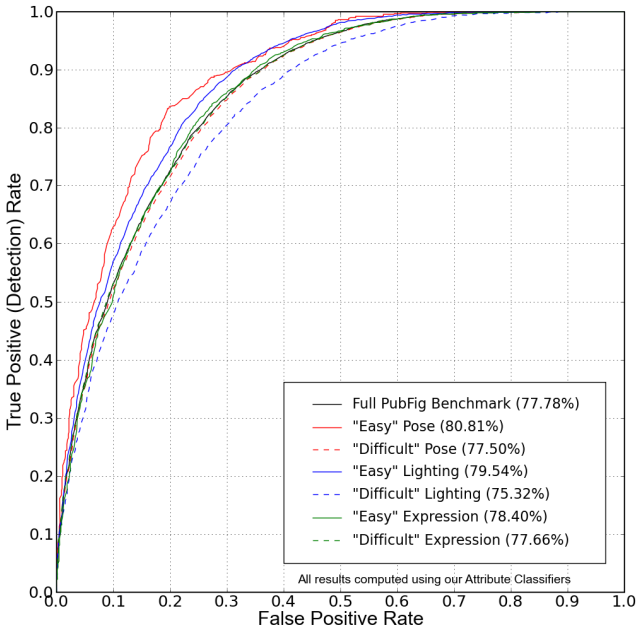

Figure 8: **Face Verification Results on PubFig:** Our performance on the entire benchmark set of 20,000 pairs using attribute classifiers is shown in black. Performance on the pose, illumination, and expression subsets of the benchmark are shown in red, blue, and green, respectively. For each subset, the solid lines show results for the "easy" case (frontal pose/lighting or neutral expression), and dashed lines show results for the "difficult" case (non-frontal pose/lighting, non-neutral expression).

We first performed a test using the original LFW images. The results are shown in red in Figure 6. At 99.20% accuracy, people are essentially perfect on this task. We now look at tougher variants of this test.

The first variant is to crop the images tightly around the face. We do this by blacking out most of the image, leaving only the face visible (including at least the eyes, nose and mouth, and possibly parts of the hair, ears, and neck). This test measures how much people are helped by the context (sports shot, interview, press conference, *etc*.), background (some images of individuals were taken with the same background), and hair (although sometimes it is partially visible). The results are shown in blue in Figure 6. Performance drops quite a bit to 97.53% – a tripling of the error rate.

To confirm that the region outside of the face is indeed helping people with identification, we ran a second test where the mask was inverted – *i.e*., we blacked out the face but showed the remaining part of the image. Astonishingly, people still obtain 94.27% accuracy, as shown by the green line in Figure 6. These results suggest that automatic face verification algorithms should not use regions outside of the face, as they could artificially boost accuracy in a manner not applicable on real data. (In all experiments involving the attribute and simile classifiers, we only used features from the face region, masking out the rest of the image.)

4.3. PubFig Data Set

As a complement to the LFW data set, we have created a data set of images of public figures, named PubFig. This data set consists of 60,000 images of 200 people. The larger number of images per person (as compared to LFW) allows us to construct subsets of the data across different poses, lighting conditions, and expressions, while still maintaining a sufficiently large number of images within each set. Further, this data set is well-suited for recognition experiments, an avenue we wish to pursue in future work.

Images in the data set were downloaded from the internet using the person's name as the search query on a variety of image search engines, such as Google Images and flickr. We ran face and fiducial point detection on the downloaded images to obtain cropped face images [26]. Finally, we rectified these images using an affine transform.

The first evaluation benchmark in PubFig is much like the LFW one: face verification is performed on 20,000 pairs of images of 140 people, divided into 10 cross-validation folds with mutually disjoint sets of 14 people each.[4] The larger size and more varied image sources used to gather the PubFig data set make this a tougher benchmark than LFW, as shown by our performance on this test, displayed in black in Figure 8.

The second benchmark in PubFig consists of subsets of the full 20,000 pairs, divided by pose, lighting, and expression. The objective is to measure the sensitivity of algo-

[4]These people are disjoint from the 60 used for our simile classifiers.

rithms to these confounding factors. Training is performed using the same data as in the first evaluation, but the testing pairs are split into "easy" and "difficult" subsets for each type of variation. Figure 8 shows results for pose (red), lighting (blue), and expression (green), with "easy" results plotted using solid lines and "difficult" using dashed lines. For pose, "easy" is defined as pairs in which both images have frontal pose (less than 10 degrees of pitch and yaw), while the remaining pairs are considered "difficult." Similarly for lighting, pairs of frontally-lit images are "easy" and remaining pairs are "difficult." For expression, "easy" means both images have a neutral expression, while "difficult" pairs have at least one image with a non-neutral expression, *e.g.*, smiling, talking, frowning, *etc*.

All the data and evaluation benchmarks in PubFig (including fiducials and pose angles) are publicly available at http://www.cs.columbia.edu/CAVE/databases/pubfig/.

5. Discussion

We have presented and evaluated two approaches for face verification using traits computed on face images – based on describable attributes and our novel simile classifiers. This is the first time such attributes have been applied to face verification. Both approaches result in error rates significantly lower (23.92% to 31.68%) than the state-of-the-art for face verification on the LFW data set. Furthermore, this is achieved using only the face region of images (without the background or context). This is important because our experiments measuring human performance show that people perform suprisingly well (94.27%) at this task even if the central portion of the face is artificially occluded. However, humans perform quite well (97.53%) when shown only a tight crop of the face, leaving a great deal of room for improvement in the performance of algorithms for face verification in the unconstrained setting.

Finally, in order to further encourage research on face verification and recognition, we introduce the new PubFig data set, which is both larger and deeper than previous data sets, allowing for exploration of subsets focusing on pose, illumination, and expression changes.

Acknowledgments: This research was funded in part by NSF award IIS-03-25867 and ONR award N00014-08-1-0638. We are grateful to Omron Technologies for providing us the OKAO face detection system. We thank flickr users UltimateGraphics, brava_67, and igorjan for their creative-commons licensed images.

References

[1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. *CVPRW*, 05, 2003.

[2] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. *CVPR 2004*.

[3] V. Blanz, S. Romdhani, and T. Vetter. Face identification across different poses and illuminations with a 3d morphable model. *FGR*, 2002.

[4] C. D. Castillo and D. W. Jacobs. Using stereo matching for 2-d face recognition across pose. *CVPR, 2007*.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[6] H. F. Chen, P. N. Belhumeur, and D. W. Jacobs. In search of illumination invariants. *CVPR, 2000*.

[7] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. *FGR 2000*.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3), 1995.

[9] G. W. Cottrell and J. Metcalfe. Empath: face, emotion, and gender recognition using holons. In *NIPS*, pages 564–571, 1990.

[10] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... Buffy – automatic naming of characters in TV video. *BMVC 2006*.

[11] A. Ferencz, E. Learned-Miller, and J. Malik. Learning to locate informative features for visual identification. *IJCV Spec. Issue on Learning and Vision*, 2007.

[12] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. *ICML*, 1996.

[13] A. Gallagher and T. Chen. Estimating age, gender, and identity using first name priors. *CVPR, 2008*.

[14] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *PAMI*, 23(6):643–660, 2001.

[15] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. SexNet: A neural network identifies sex from human faces. In *NIPS*, pages 572–577, 1990.

[16] R. Gross, J. Shi, and J. Cohn. Quo vadis face recognition? In *Workshop on Empirical Evaluation Methods in Computer Vision*, December 2001.

[17] G. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. *ICCV, 2007*.

[18] G. Huang, M. Jones, and E. Learned-Miller. LFW results using a combined Nowak plus MERL recognizer. 2008.

[19] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. *UMass Amherst Technical Report 07-49*, October 2007.

[20] J. Huang, X. Shao, and H. Wechsler. Face pose discrimination using support vector machines (SVM). *ICPR*, pages 154–156, 1998.

[21] N. Kumar, P. N. Belhumeur, and S. K. Nayar. FaceTracer: A search engine for large collections of images with faces. *ECCV*, 2008.

[22] H. Ling, S. Soatto, N. Ramanathan, and D. Jacobs. A study of face recognition as people age. *ICCV, 2007*.

[23] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2003.

[24] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *PAMI*, 24(5):707–711, May 2002.

[25] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. *CVPR, 2007*.

[26] Omron. OKAO vision. http://www.omron.com/r_d/coretech/vision/okao.html, 2009.

[27] A. O'Toole, P. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi. Face recognition algorithms surpass humans matching faces over changes in illumination. *PAMI*, 29(9):1642–1646, Sept. 2007.

[28] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. *CVPR, 1994*.

[29] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, and W. Worek. Preliminary face recognition grand challenge results. *FGR 2006*, pages 15–24.

[30] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *PAMI*, 22(10):1090–1104, 2000.

[31] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. *Workshop on Applications of Computer Vision, 1994*.

[32] G. Shakhnarovich, P. Viola, and B. Moghaddam. A unified learning framework for real time face detection and classification. *FGR, 2002*.

[33] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression (PIE) database. *FGR*, 2002.

[34] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Real-Life Images workshop at ECCV*, 2008.

[35] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.