W.J. Scheirer, N. Kumar, V.N. Iyer, P.N. Belhumeur, and T.E. Boult, "How Reliable are Your Visual Attributes?" Biometric and Surveillance Technology for Human and Activity Identification X, I. Kakadiaris, W.J. Scheirer and L.G. Hassebrook, Editors, Proc. SPIE 8712, 87120Q (May 31, 2013).

http://dx.doi.org/10.1117/12.2018974

# How Reliable are Your Visual Attributes?

W. J. Scheirer[a] and N. Kumar[b] and V.N. Iyer[c] and P.N. Belhumeur[d] and T.E. Boult[c]

[a]Harvard University, Cambridge, MA, USA;
[b]University of Washington, Seattle, WA, USA;
[c]Securics, Inc., Colorado Springs, CO, USA;
[d]Columbia University, New York, NY, USA

## ABSTRACT

Describable visual attributes are a powerful way to label aspects of an image, and taken together, build a detailed representation of a scene's appearance. Attributes enable highly accurate approaches to a variety of tasks, including object recognition, face recognition and image retrieval. An important consideration not previously addressed in the literature is the reliability of attribute classifiers as the quality of an image degrades. In this paper, we introduce a general framework for conducting reliability studies that assesses attribute classifier accuracy as a function of image degradation. This framework allows us to bound, in a probabilistic manner, the input imagery that is deemed acceptable for consideration by the attribute system – without requiring ground truth attribute labels. We introduce a novel differential probabilistic model for accuracy assessment that leverages a strong normalization procedure based on the statistical extreme value theory. To demonstrate the utility of our framework, we present an extensive case study using 64 unique facial attributes, computed on data derived from the Labeled Faces in the Wild (LFW) data set. We also show that such reliability studies can result in significant compression benefits for mobile applications.

**Keywords:** Computer Vision, Biometrics, Visual Attributes, Extreme Value Theory, Support Vector Machines

## 1. INTRODUCTION

When considering an object in an image, we often want to determine its category by examining its features. For an application such as recognition, this most typically involves detecting relevant feature points, computing a descriptor, and building a model that can be compared to others via some classification technique. The most common texture-based recognition techniques (for example, SIFT, LBP, and subspace methods) operate in this fashion. However, as Ferrari and Zisserman[1] point out, "an object also has many other qualities apart from its categories." For instance, a face corresponds to an identity (i.e., a category), but also may have a hat, a beard, African ethnicity and a round nose.

These *describable visual attributes*[1–3] are a powerful way to label aspects of an image, and taken together, build a detailed representation of a scene's appearance. Attributes correspond to semantically meaningful labels (as opposed to abstract or very primitive feature descriptors) for characteristic image properties. Interest in attributes has grown substantially in the computer vision community, thanks in part to their ease of implementation and flexibility of representation.

Attributes enable highly accurate approaches to a variety of tasks, including object recognition,[3] face recognition,[4] and image retrieval.[5] Farhadi et al.[3] assigned attributes to the categories in the PASCAL VOC 2008 object set with a mean classification accuracy of 83.4%. Kumar et al.[4] showed that attribute and simile (relative attribute similarity) classifiers are able to achieve a mean verification accuracy of 85.29% on the challenging Labeled Faces in the Wild (LFW)[6] data set. For a face search application, Kumar et al. [2,5] also demonstrated the ability to construct textual queries over 64 unique attributes, many of which possess classification error rates less than 10%. With results like these, we expect to see more effort placed into attribute based applications serving real users in the near future.

Send correspondence to W.J. Scheirer, E-mail: wscheirer@fas.harvard.edu


Figure 1. In this work, we introduce a method of evaluation for conducting reliability studies for describable visual attributes. Attributes are descriptive labels for an image such as "pointy nose," "brown hair," "pale skin," etc., that can be used for a variety of recognition tasks. Starting with an initial set of unlabeled images, we show how to generate controlled data sets reflecting different levels of degradation along various axes (e.g., scale, blur, and JPEG quality). By applying a statistical normalization technique to the attribute outputs on these data sets, and analyzing the resulting values, we can determine the conditions at which classifiers are no longer useful. This is shown above visually, where the images highlighted in red represent conditions that may produce unreliable results from the attribute classifier for "pointy nose."

An important consideration not previously addressed in the literature is the reliability of attribute classifiers as the relative quality of an image degrades. Attribute reliability is different from the reliability of the underlying features, since it must consider multiple features often computed over specific regions of varying color, size and shape. For instance (Fig. 1), what happens to the attribute classification decision scores as the scale of an image decreases, when the image is blurry, or when an image is highly compressed? Existing work has hinted at these effects using cross validation accuracy assessment[2] for attribute classifiers. But with that type of test, an emphasis is placed on generalization over the training data, as opposed to understanding the limits of classifiers.

What benefit do reliability studies provide? For one, they can help us improve the performance of real-world applications with resource constraints – such as in the rapidly-growing mobile computing space. In this regime, processing power, memory, and bandwidth are all quite limited. Yet, many applications, such as recognition, fine-grained visual categorization, or automatic text summarization, often only need access to computed attribute values – the original images are not required. In this kind of scenario, knowing how severely one can compress or downsample images and maintain high classification accuracy would be quite valuable for reducing the usage of the limited resources available on the device.

Beyond mobile applications, measuring attribute classifier reliability enables numerous improvements in attribute systems. Special versions of classifiers can be trained for harsh conditions expected in a given operating regime. If second stage classification is built on top of attributes, e.g.,[4, 7] reliability information can be used as an additional input to improve results. Going even further, relative reliabilities and correlations between attributes can be used to improve attribute reliability. For example, the failure of a particular attribute due to a given type of image degradation can be mitigated by using a weighted average of related attributes that are less affected by that condition. One could even train multiple versions of an attribute classifier, each tuned for a given transformation type, and intelligently switch between (or weight) them – similar to how separate face detectors are often trained for separate poses, and combined at the end.

In this paper, we introduce a general framework for conducting reliability studies that assesses attribute classifier accuracy as a function of image degradation. Our contributions include:

1. **Criteria for Data Generation and Evaluation:** We define a methodology for generating images with various degradation effects in a controlled manner. These images can be used to evaluate any set of attribute classifiers.

2. **Novel Differential Probabilistic Model for Accuracy Assessment:** We propose a probabilistic model based on the statistical extreme value theory to provide an indication of attribute reliability. This

model is far more powerful than a naive treatment of classifier decision scores as binary decisions, or any analysis of raw or weakly normalized classifier outputs. Moreover, the evaluation technique does not require any ground-truth data.

3. **Case Study on Labeled Faces in the Wild:** We present an extensive case study applying our methodology to the face attribute approach of,[2] evaluated over data derived from the popular LFW data set,[6] and demonstrate the value of our approach with several new results.

## 2. PRIOR WORK IN DESCRIPTIVE VISUAL ATTRIBUTES

Visual attributes were first described in the vision community by Ferrari and Zisserman[1] as a model for understanding object appearance and for generating human understandable descriptions. In that work, a probabilistic generative feature model is coupled with an optimized likelihood ratio approach to learning, enabling classification for simple color and pattern attributes. Kumar et al.[5] demonstrated that the attribute approach could be extended to a large set of specific facial attributes. Using a variety of simple feature descriptors and large scale learning, Kumar et al. achieve high levels of attribute classifier accuracy. Farhadi et al.[3] also make the case for describing objects by their attributes, and highlight the challenge presented by the need to generalize attributes across object categories. They choose to address this at the feature selection level.

An important aspect of most attribute approaches is machine learning, and several novel techniques have been explored for different unconstrained source data scenarios. Berg et al.[8] look at boosting for automatically discovering attributes from images and associated text found on the web. Russakovsk and Fei-Fei[9] build SVM attribute classifiers from the relationships they find in the ImageNet data set. In some cases, no training examples for a class of interest are available. To handle such instances, Lampert et al.[10] apply a set of low-level attributes (shape, color, geographic information) across different classes in a "zero-shot" transfer learning approach. The specific interest in images from the web brings with it the problem of poor quality images, which can lead to bad results because of attribute reliability issues.

Even more specialized learning approaches have been developed to understand fine-grained attribute relationships. Parikh and Grauman[11] introduced a semi-automatic learning process that puts humans in the loop to improve detailed attribute category annotation. In follow-up work[12] to facilitate relative attributes, whereby similarity between faces or objects can be assessed, Parikh and Grauman introduce a novel form of zero-shot learning.[10] That approach produces results that are superior to typical binary attribute classification applied to the same data. Again, with the possibility of poor quality images, all of these learning methods are candidates for the reliability study framework we introduce in this paper.

For faces, there are two primary areas that attributes have been applied to: face image retrieval and face recognition. Kumar et al. created a search engine[2,5] for very accurate face image retrieval over 64 different attributes. The search engine supports multi-attribute queries such as "smiling Indian men with glasses", "attractive women wearing lipstick", etc. Siddiquie et al.[13] introduced a multi-attribute query approach that supports attribute inference to refine the pool of images considered relevant for a query. While not specific to faces, Douze et al.[7] looked at a method for fusing attribute scores and Fisher vectors to achieve more accurate retrieval results.

Face recognition is a particularly interesting area for attribute research. Kumar et al.[2,4] showed that by comparing the decision scores from their attribute traits from their face search work[5] via a product-and-sum distance computation using SVM learning, very accurate face verification results can be achieved. Depending on the application domain, however, there could be systematic quality issues that impact attribute-based recognition performance – for instance, the resolution of a security camera, or blur from a long-range surveillance sensor.

The wide range of possible application domains for recognition and search (security, human computer interaction, or marketing, to name a few) will, of course, have quite different performance requirements. However, even though an understanding of the limits of classification is critical for building useful attribute applications, no prior work has examined this issue. And while attribute classifiers are typically trained for generalization, there could be systematic biases[14] on particular input data due to blur, scale, or any other factor. Conducting an attribute reliability study is a principled way of characterizing classification accuracy with respect to these factors.

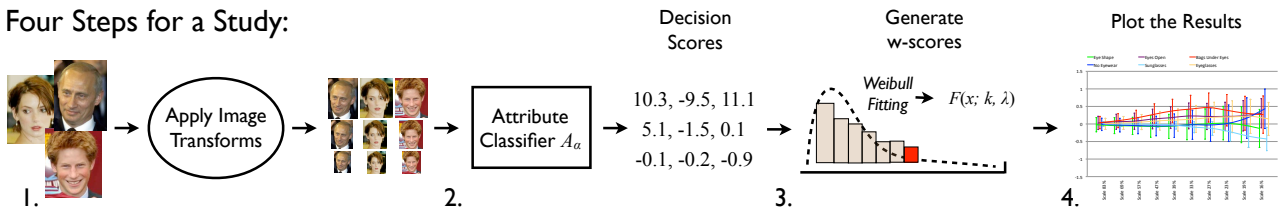



Figure 2. The goal of an attribute reliability study is to assess the conditions under which an attribute classifier can produce accurate results. To accomplish this, we propose a four-step process: 1. Generate data for many different image transformations using a well-known data set as a basis; 2. Process the generated data using a set of attribute classifiers; 3. Apply a probabilistic w-score model to assess classifier accuracy based on decision scores; 4. Produce results that can be analyzed both visually and analytically.

## 3. HOW TO CONDUCT A RELIABILITY STUDY

The goal of an attribute reliability study is to assess the conditions under which an attribute classifier can produce accurate results. Ideally, we'd like to be able to vary a series of image transformation parameters in a controlled manner to manipulate the original images from a well-known data set. The transformed images should represent a spectrum of conditions that an attribute classifier might encounter in the real world. By identifying the boundary between "good" and "bad" imagery for a particular attribute classifier, we can ensure that only appropriate images are considered during operational classification, thus improving the reliability of the results. To accomplish this, we propose a four step process.

**Step 1: Generate Data.** The first step involves selecting a data set $D$ that will provide a significant number of images to be used as a basis for the evaluation. If the chosen set is well-known, it can be used as a good comparison point across attribute methods evaluated in prior work. Ground-truth attribute data for each image is *not* necessary; accuracies will be assessed using our probabilistic model described in Step 3. A series of transformation functions $T_1, \ldots, T_m$ must also be selected, reflecting the types of conditions that might be encountered during attribute classification. In this paper, we examine scale, blur and JPEG quality, though any other transformation of interest (pose, noise, etc.) could be applied as well. Each transformation function can be parameterized as $T_i(x, y)$, in order to vary the magnitude/severity ($x$) of the condition being assessed for an image ($y$). Thus, the process will consider the set of transformations and a range of parameters $j_1, \ldots, j_n$ to generate the test images $J_{i,j_p}$ from the base set $I \subseteq D$:

$$\begin{aligned}(J_{1,j_1} = T_1(j_1, I), \ldots, J_{1,j_n} = T_1(j_n, I)), \ldots, \\ (J_{m,j_1} = T_m(j_1, I), \ldots, J_{m,j_n} = T_m(j_n, I))\end{aligned} \tag{1}$$

**Step 2: Process Data.** The image sets generated in Step 1 are used as the input to a collection of attribute classifiers. Most often, attribute classifiers have been binary in nature (examples: male / female; Asian / not Asian), with a decision score indicating positive or negative classification based on its sign, or some threshold over the raw classifier output. By collecting the decision scores (either raw or weakly normalized), we can go beyond binary classification to a probabilistic analysis of the actual values in Step 3. For each set of test images $J_{i,j_p}$, a set of attribute classifiers $A_1(x), \ldots, A_\eta(x)$ are applied:

$$\begin{aligned}(S_{1,1,j_1} = A_1(J_{1,j_1}), \ldots, S_{1,1,j_n} = A_1(J_{1,j_n})), \ldots, \\ (S_{\eta,1,j_1} = A_\eta(J_{1,j_1}), \ldots, S_{\eta,1,j_n} = A_\eta(J_{1,j_n}))\end{aligned} \tag{2}$$

The above equations show the decision score generation process for a single transformation with its complete set of parameters and a set of attribute classifiers. With the application of each classifier $A_\alpha(x)$, a set of scores $S_{\alpha,i,j_p}$ is generated for a transformation at a certain parameter.

**Step 3: Estimate Probabilities.** The goal of our probabilistic analysis is to map a decision score to a probability that a given image matches its attribute label, as determined by the decision score. A collection of probabilities for an individual attribute classifier and a particular image condition, taken as a whole, give

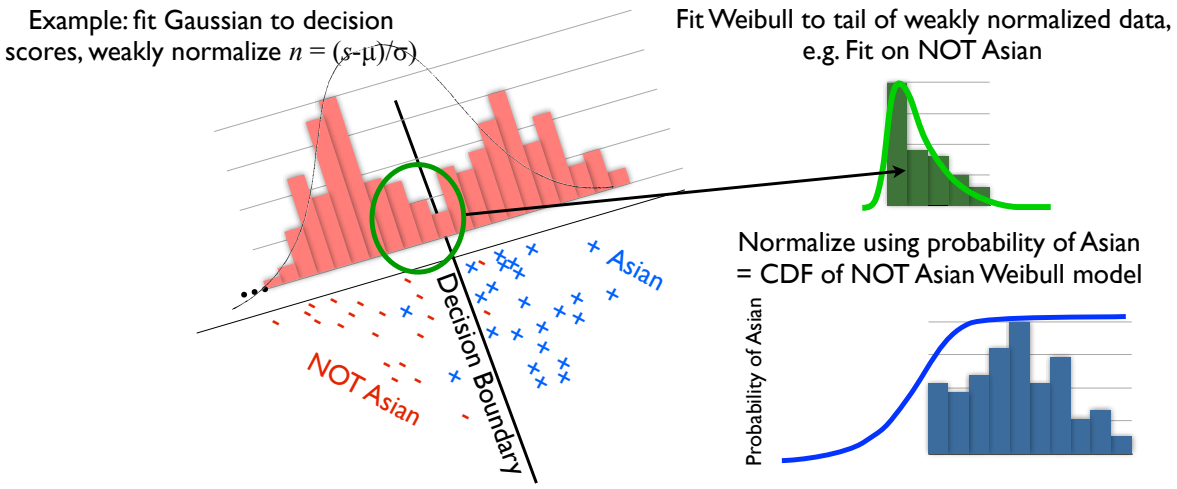



Figure 3. Typical attribute approaches such as[2] perform a Gaussian normalization of decision scores from an attribute classifier (left). However, this is not accurate for most classifiers; instead, extreme value theory shows that fitting a Weibull to the *tail* of this weakly normalized data results in a much better fit (top-right), which can then be used to normalize the data into significantly more reliable *w-scores*[15,16] (bottom-right).

us an indication of the reliability of the classifier. To compute the probabilities, we use the recently proposed w-score method of Scheirer et al.[16] The w-score is a general normalization technique that leverages the statistical extreme value theory (EVT), which has been shown[15] to be an appropriate model for computer vision problems where the tails of the data are what influence classification – regardless of the distribution of the rest of the data. (Weaker normalizations such as the Gaussian utilize non-EVT distributions.) While different techniques to convert decision scores to probabilities could be used – sigmoid, min-max, or other statistical modeling – the w-score is particularly attractive because it gives us an indication of the correctness of an attribute label. This is accomplished by assessing the label's formal probability of being an outlier (match) in the extreme value "non-match" model.

Algorithmically, the shape parameter $k > 0$ and scale parameter $\lambda > 0$ for a Weibull distribution must first be determined. This is done by fitting a Weibull distribution to the scores closest to the decision boundary from the non-match distribution of $S_{\alpha,I} = A_\alpha(I)$.[15] For binary attribute classifiers (Fig. 3), the non-match distribution is the side of the attribute classifier not of interest. For example, if you are interested in the "Asian" attribute, you want to fit the model on the scores labeled "NOT Asian." (The data may need to be transformed to ensure that the extreme values are the largest positive values, if the scores are distances where "smaller is better.") This statistical fitting can be considered a "training" phase, where the attribute values computed on the original images $I$ are used as a basis for the model of each attribute. (Ground-truth is not needed.)

To calculate the w-scores, we use the CDF of the Weibull distribution defined by the parameters $k$ and $\lambda$:

$$F(x;k,\lambda) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k} \tag{3}$$

for $x > 0$. For each individual score $s_c \in S_{\alpha,i,j_p}$, we can apply the CDF function defined by the Weibull distribution $\mathcal{W}_\alpha$ for a particular attribute:

$$w_c = F(s_c;\mathcal{W}_\alpha); k,\lambda \in \mathcal{W}_\alpha \tag{4}$$

where $w_c$ is the resulting w-score. We can fit two Weibull distributions for each binary classifier (fitting on female to test for male, and vice versa). Labels are needed to distinguish between them. For a fitting from the first class, the w-scores are between 0 and 1. For a fitting from the second class, we include a negative prefix, putting the w-scores in the range of -1 to 0. Since we want to perform a large scale study over many scores, we create

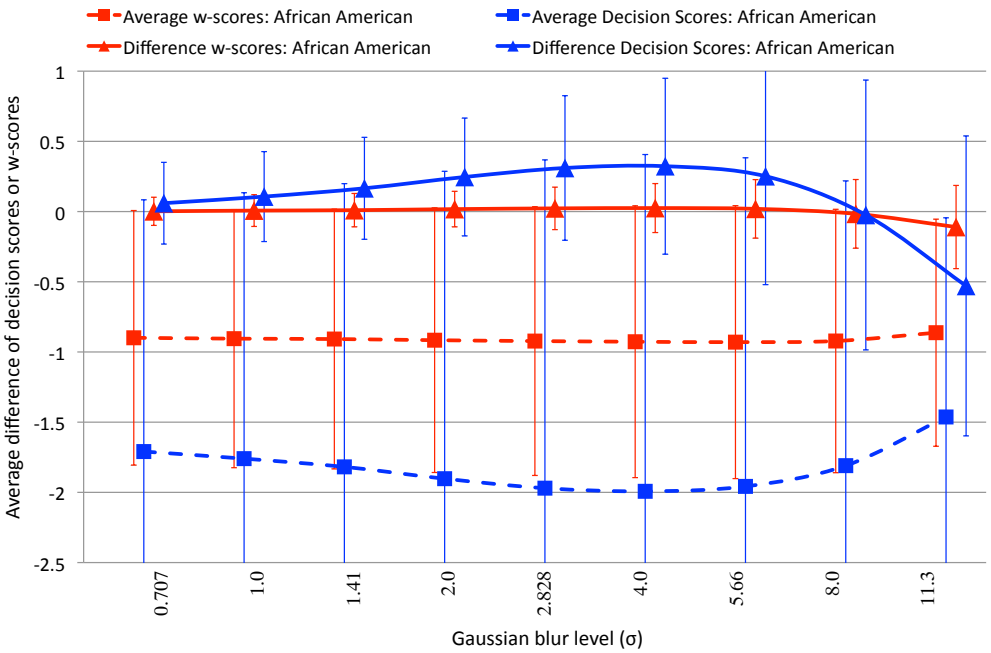



Figure 4. For increasing levels of Gaussian blur along the x-axis, we see that the averages of raw decision scores (blue squares) or w-scores (red squares) are not useful indicators of reliability, due to the high standard deviations (shown as error-bars). Nor are differences of raw scores (blue triangles), which are not even monotonic. However, the differences of w-scores (red triangles) *do* provide a useful cue, as they consistently deviate away from 0 at severe blur levels, with much lower variance.

w-scores for every score available, yielding a series for each parameter of each transformation of interest:

$$(W_{1,1,j_1},\ldots,W_{1,1,j_n}),\ldots,(W_{\eta,1,j_1},\ldots,W_{\eta,1,j_n}) \tag{5}$$

**Step 4: Analyze Results.** Step 3 generates a large amount of data that can be summarized in an intuitive visual manner supporting a useful analysis. To accomplish this, we have designed a plotting procedure that expresses attribute classifier reliability at different intervals of image degradation. Our first step is to take the average of each w-score set $W_{\alpha,i,j_p}$, yielding $\mu_{\alpha,i,j_p}$. A first inclination might be to plot averages directly, however, this is not a good approach.

Fig. 4 shows multiple alternative approaches to plotting the attribute "African American," for a series of Gaussian blur levels applied to LFW. For the plotted w-score averages, the x-axis denotes different levels of Gaussian blur, while the y-axis denotes an average w-score. Averages alone do not tell us very much about the varying impact of image degradations – the dashed red curve marked with squares falls in a region that is attribute specific, making comparisons across attributes impossible. Further, if we take the standard deviation at each point, we also see a high degree of variability. What is being shown in this case is a normalized summary of the decision scores from the attribute classifiers – not a representation of the results with low variability that yields better conclusions.

To produce a more useful reliability representation, we compute the difference between the average w-score for the original images for a particular attribute $\mu_{\alpha,I}$ and the average w-scores across the transformation intervals:

$$\Delta_{\alpha,i,j_p} = \mu_{\alpha,I} - \mu_{\alpha,i,j_p} \tag{6}$$

This "normalizes" the data in a way that puts the curves near 0 on the y-axis (which, when plotting differences, represents $\Delta_{\alpha,i,j_p}$) if the degradation is not significantly impacting the attribute classifier. This can be seen in Fig. 4 (red curve marked with triangles), where blur starts to have an impact on reliability at $\sigma = 8.0$. With a consistent representation, it becomes easy to visually examine different attributes and compare their

reliability on a common basis. We note that the use of the original decision scores (also shown in Fig. 4 as the blue curves) with or without the difference calculation also inflates variability. w-scores work well because they flatten regions of decision scores that represent insignificant changes as a function of probability. Without the probabilistic representation, change is inflated within and between parameter bins.

## 4. CASE STUDY: THE LFW DATASET

To highlight the utility of our proposed methodology for conducting attribute reliability studies, we present a large set of experiments using the complete Labeled Faces in the Wild (LFW) data set[6] as our source data. We evaluate reliability over three types of image transformations – Gaussian blur, image scale, and JPEG quality – at a variety of parameters. Note that ground truth attribute values for LFW are *not* required, as our methodology can quantify reliability by comparing differences of w-scores between the original and transformed images.

We generated all transformed images using the popular image processing package ImageMagick*. To create blurry data, we took the entire set of images from LFW, and convolved each with a Gaussian distribution parameterized by standard deviation $\sigma$. We used nine different standard deviations, reflecting moderate to severe blur, as is found in web imagery. For the second transformation, we wanted to study attribute reliability at face resolutions commonly found on the web, but from which faces are still automatically detectable (to some degree). So we repeatedly scaled the original $250 \times 250$ images by 83% to generate 10 different scales, with the smallest being $0.83^{10} = 16\%$ the size of the original. ($1/1.2 = 83\%$ is the factor between different octaves in scale space.) Finally, with the prevalence of low-quality (often mobile phone-captured) face imagery on the web, we wanted to evaluate the effect of recompression artifacts. We generated image sets at 17 different JPEG quality levels by recompressing the images from their original setting of 90[6] down to 5, in increments of 5.

Next, we compute visual attributes on all original and transformed images using the approach of Kumar et al.,[2,4,5] a top performer on the LFW face verification benchmark. This method uses hundreds of labeled training examples for each attribute to automatically learn SVM classifiers. First, a large set of low-level features are computed from an affine-aligned face image. These features consist of various feature types extracted from different parts of the face (shown in Fig. 5), such as "mean-normalized RGB pixel values from the eyes" or "histograms of oriented gradients from the mouth." A greedy, iterative feature selection process chooses the most appropriate set of features for a given attribute from this collection, as measured by cross-validation accuracy.

We use the authors' own implementation of these classifiers, publicly available as a webservice.† This service computes 64 unique visual attributes, from which we generate w-scores as described in Step 3 of our methodology. Note that although the cross-validation accuracies for these classifiers have been published,[2] these numbers do not give us a complete reflection of their usefulness in operation for the following reasons:

1. Cross-validation takes place using the original training data, which may not reflect problem conditions encountered in the real world;
2. The conditions of the imagery used for training are not controlled to a specific transformation or parameter, since the goal is generalization; and
3. Cross-validation doesn't actually give us an indication of how well the classifiers did generalize.

In contrast, our methodology will compute reliability information with respect to blur, scale, and quality that can prove crucial in many real-world applications, such as recognition from low-quality surveillance cameras, or at a distance using low-resolution imagery.

*http://www.imagemagick.org/

†http://afs.automaticfacesystems.com/

Figure 5. The face regions used for automatic feature selection in the approach proposed by Kumar et al.[2] Left: a region covering the entire face is considered for attributes such as gender and ethnicity. Right: nine regions correspond to the dominant features of the face for finer-grained attributes corresponding to facial parts and facial regions (eyes, nose, mouth, etc.).

### 4.1 Experimental Results

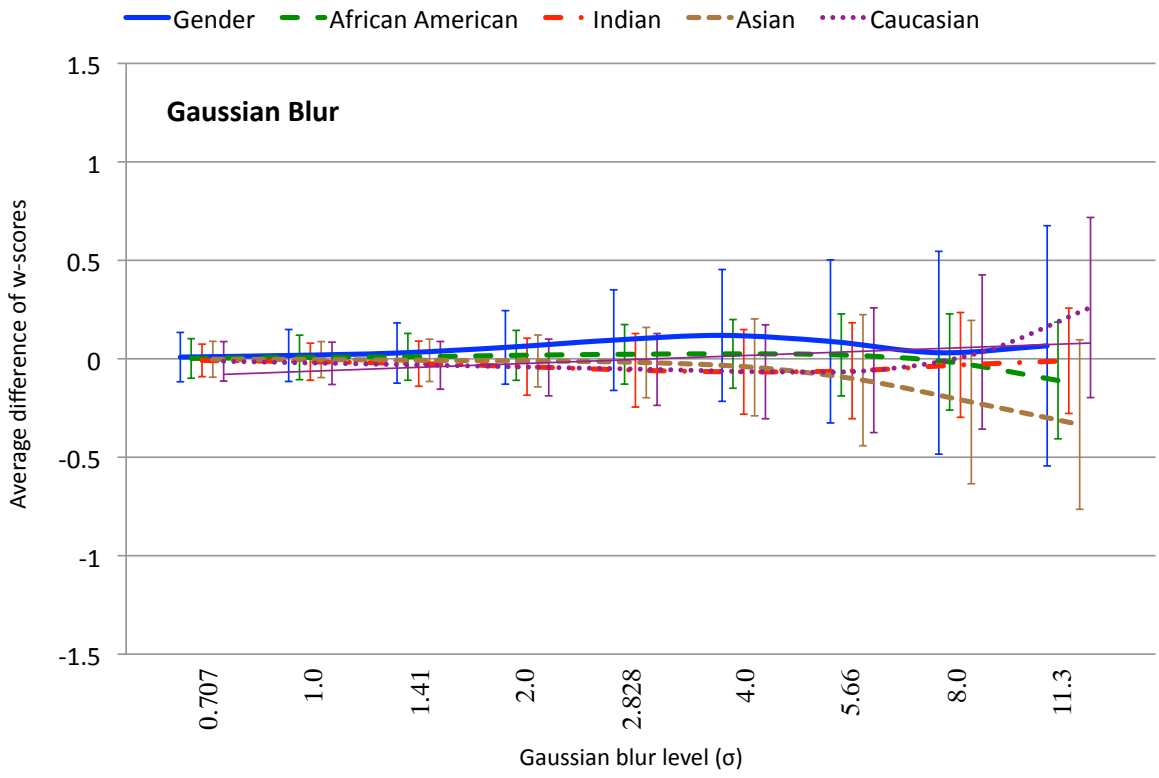
From the available w-scores, we generated reliability curves for each attribute over each transformation and associated parameter intervals. Figs. 6–8 highlight some of the more interesting results that were gathered during the course of our study. Looking at the Gender- and Ethnicity-related attributes in Figs. 6(a) & 7(a), we see a great deal of reliability (curves close to 0) until we reach the extremes of the parameters ($\sigma = 8.0$ in the case of blur, and 27% scaling in the case of scale). This is because these attributes use the entire face region shown in Fig. 5, which provides a large amount of feature information despite degradation. For the Forehead and Brow attributes shown in Figs. 6(b) & 7(b), we see some more variation, particularly with respect to eyebrow shape and eyebrow thickness – and it's easy to see why. Looking at Fig. 5, the underlying features for these attributes are small – especially for the eyebrows – with less content than we had with the attributes computed over the entire face. Thus, we now know the limitations with respect to small feature regions when using these classifiers. Finally, compared to blur and scale, JPEG quality (Fig. 8) does not significantly impact reliability.

A few other conclusions can be drawn on the overall impact of significant image transformations and processing for attribute classification. In Fig. 9, we see that the parameters which induce severe image degradation in the cases of blur and scale drastically reduce the percentage of faces detected (blue curves), before we even begin attribute classification. This means that some transformation parameters are not inherently useful for operational evaluation. An example of this is blur parameter 9 ($\sigma = 11.3$), where almost no faces are detected, and hence, no images would be considered for classification.

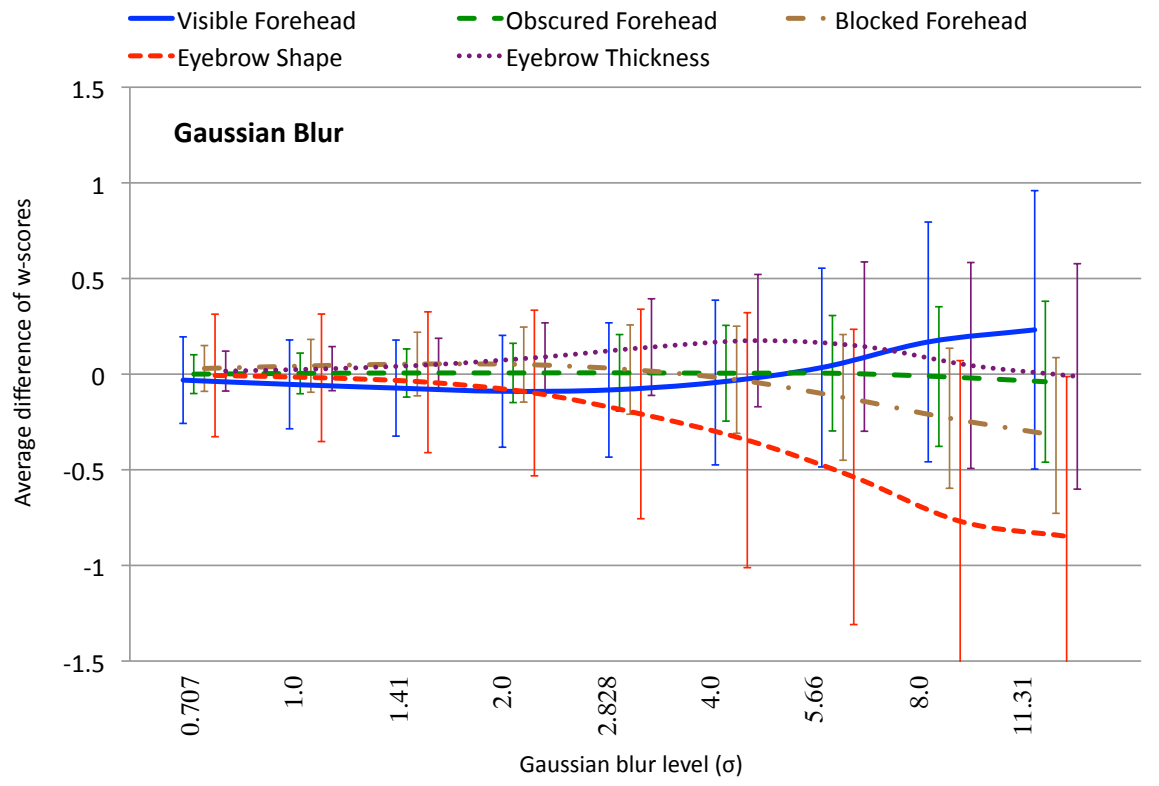
As described in the introduction, the ability to reduce file size and still achieve good classification accuracy is highly desirable for mobile attribute applications. The red curves in Fig. 9 depict decreasing file size ratios (compared to the original images) and standard deviations as a function of each transformation. Comparing these file size curves with the attribute curves in Figs. 6–8, we can find points of good reliability that minimize file size. For instance, a JPEG Quality of 15 requires less than 20% space, and yet is within the reliable region for all of the attributes shown in Fig. 8. Thus, we could use these highly compressed images to compute these (and other) attributes, saving an enormous amount of bandwidth.

## 5. DISCUSSION

An understanding of classifier behavior is critical for building useful applications around attributes. Even classifiers that have been trained for generalization can quickly become unreliable when presented with unconstrained data from the web. This is often due to prevalent systematic biases in the training data, which may not be discovered until a classifier is applied to data from outside the laboratory. All future visual attribute work should consider an evaluation framework such as that proposed herein to determine the conditions under which a particular approach can succeed. For instance, our study on face attributes revealed that blur and scale tend to affect classifier reliability much more than JPEG quality. Further, we learned that reliability is dependent to some degree on the face regions used to compute particular attributes, and that detector constraints should

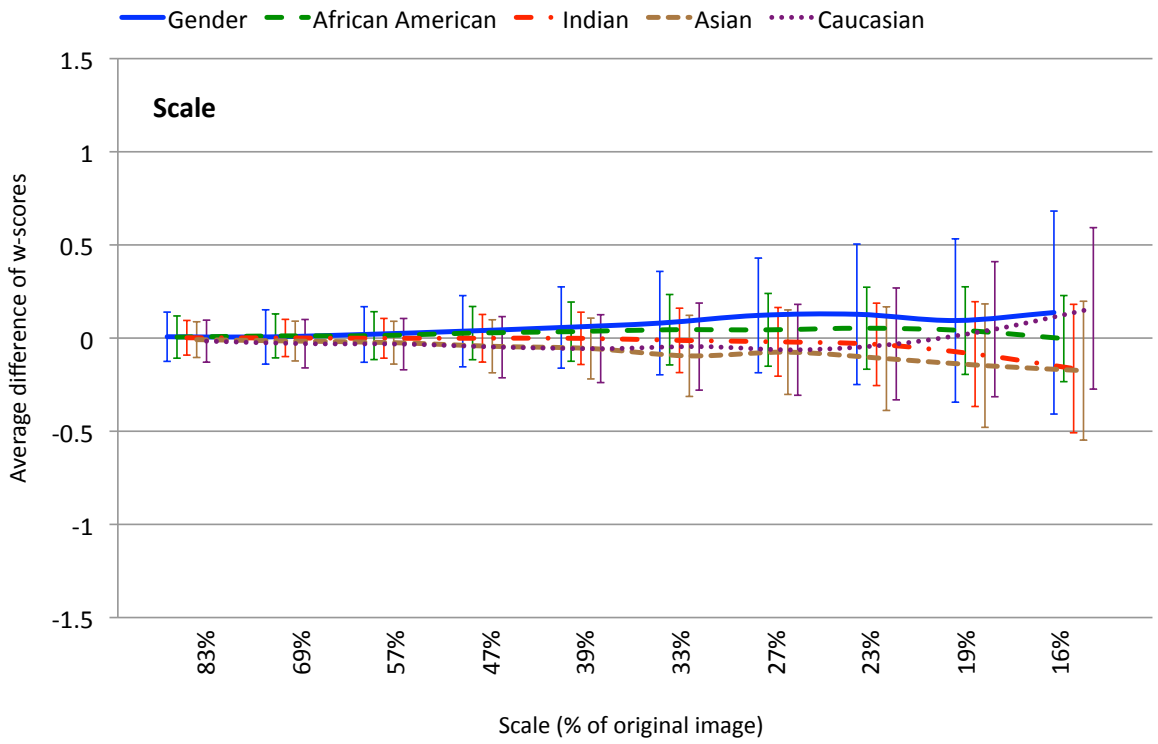



(a) Gender and Ethnicity

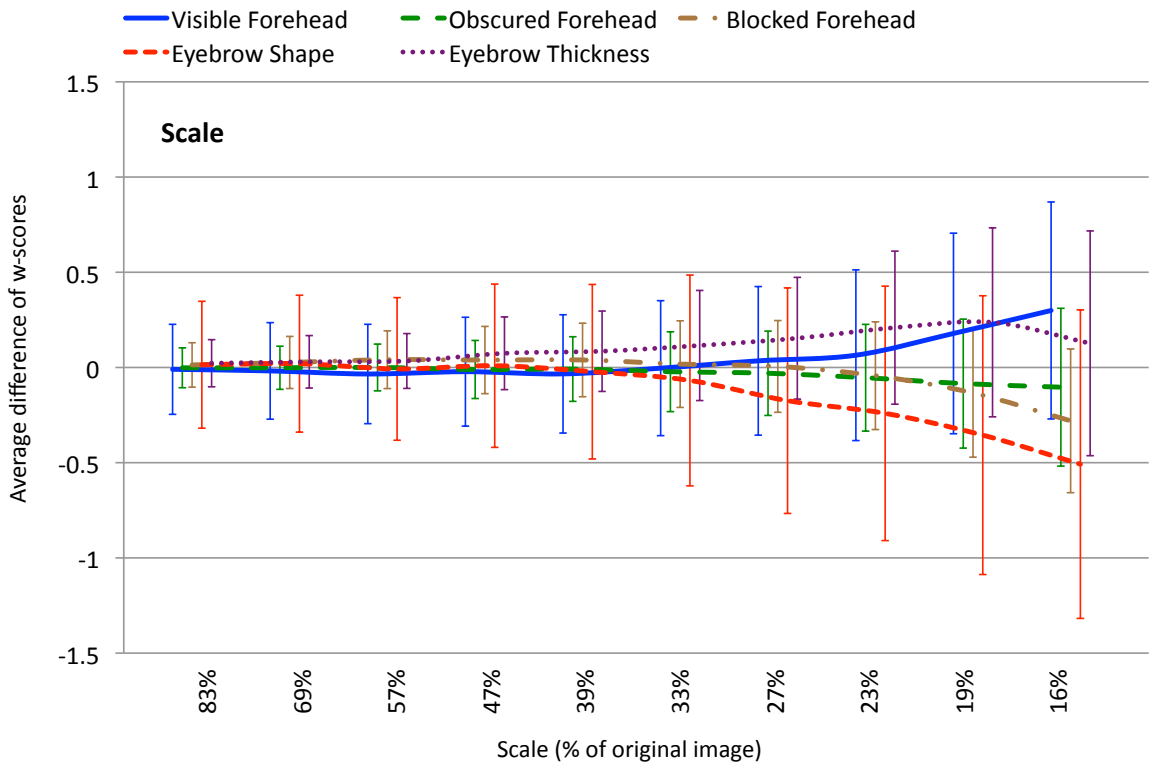



(b) Forehead and Brow

Figure 6. Classifier reliabilities with increasing Gaussian blur for a selection of attributes. Points close to 0 on the y-axis indicate higher levels of reliability. Classifiers that consider features from larger image regions are more reliable (e.g., ethnicity, which uses the entire face), whereas those based on smaller regions are much less reliable as the test images degrade (e.g., eyebrow shape, which uses the area around the eyes).




(a) Gender and Ethnicity




(b) Forehead and Brow

Figure 7. Classifier reliabilities with decreasing image scale for a selection of attributes. Points close to 0 on the y-axis indicate higher levels of reliability. These results are similar to those from Fig. 6, with classifiers considering features from larger image regions exhibiting more reliability than those derived from smaller regions.

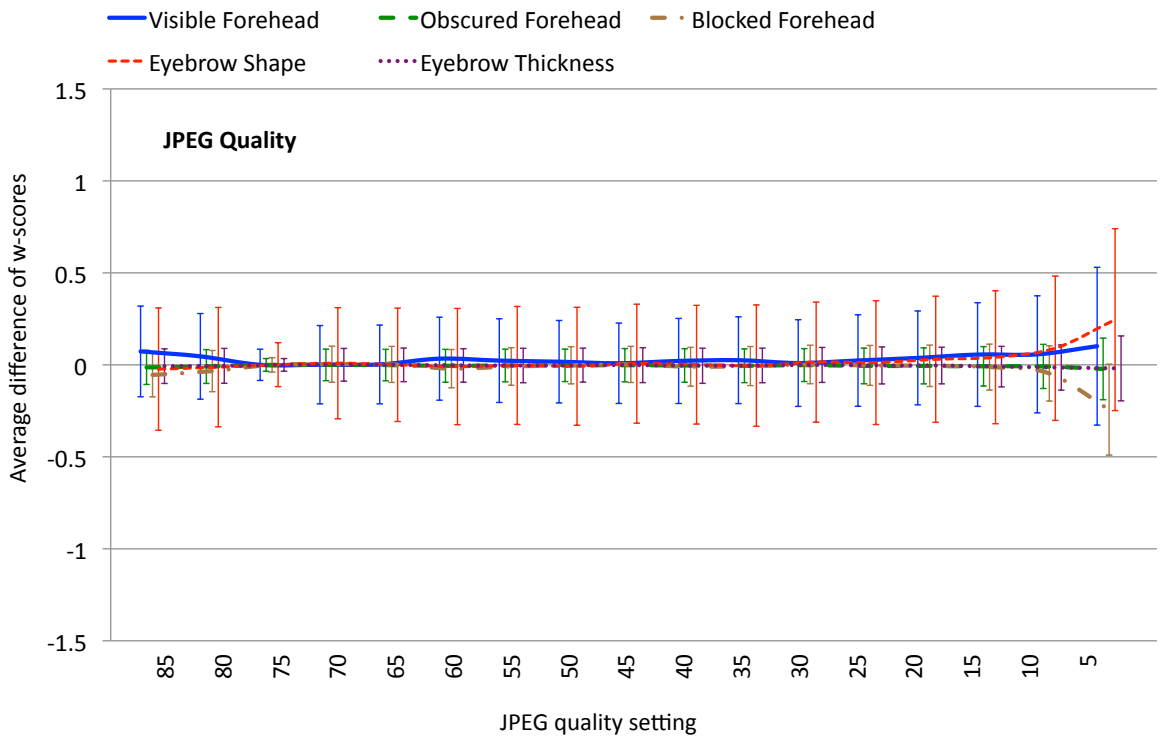



Figure 8. Attribute classifier reliabilities with decreasing JPEG quality for forehead and brow attributes. Points close to 0 on the y-axis indicate higher levels of reliability. Unlike blur and scale, JPEG quality does not significantly impact the reliability of these classifiers. (Gender and ethnicity are exceptionally stable and thus not shown here.) The slight bulge around settings 85 and 80 is due to resampling artifacts caused by the input LFW images being saved at a Quality setting of 90, regardless of source camera settings, which often default to 75.

influence the choice of parameters for a study. Finally, we saw the potential for greatly reducing file size without comprising accuracy for some attributes, which is a promising direction for mobile attribute applications.

## ACKNOWLEDGMENTS

This work was supported by ONR SBIR Award N00014-11-C-0243 and ONR MURI Award N00014-08-1-0638.

## REFERENCES

[1] Ferrari, V. and Zisserman, A., "Learning Visual Attributes," in [*NIPS*], (December 2007).

[2] Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K., "Describable Visual Attributes for Face Verification and Image Search," *IEEE TPAMI* **33**, 1962–1977 (October 2011).

[3] Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D., "Describing Objects by their Attributes," in [*IEEE CVPR*], (June 2009).

[4] Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K., "Attribute and Simile Classifiers for Face Verification," in [*IEEE ICCV*], (October 2009).

[5] Kumar, N., Belhumeur, P. N., and Nayar, S. K., "FaceTracer: A Search Engine for Large Collections of Images with Faces," in [*ECCV*], (October 2008).

[6] 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," Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007).

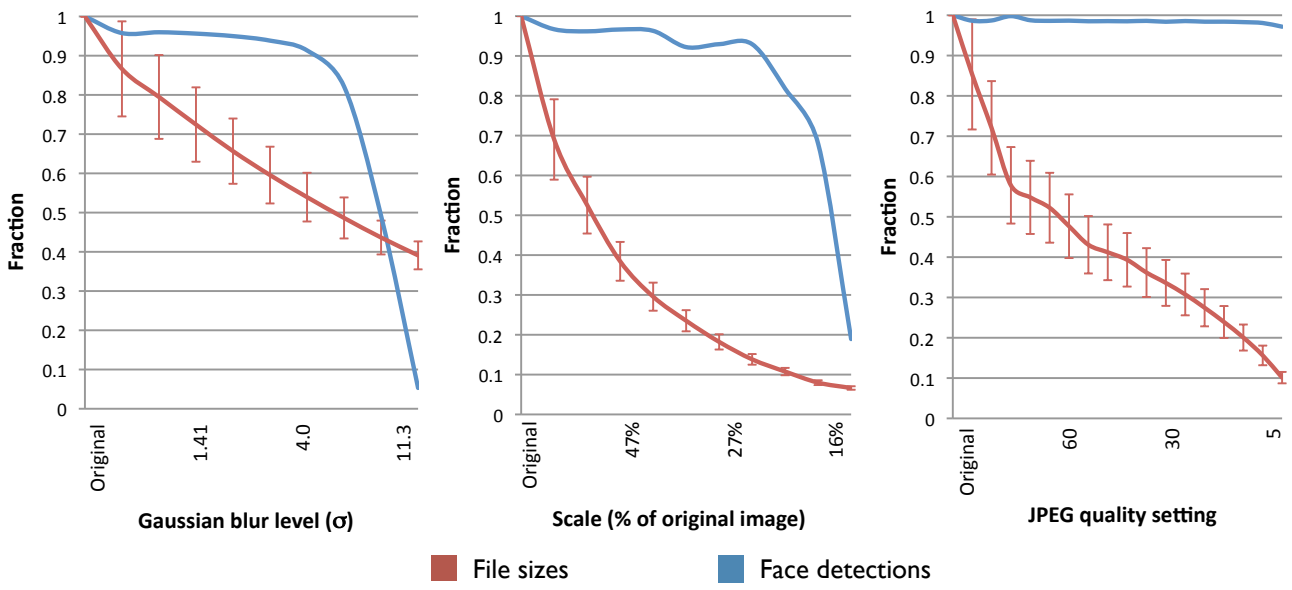



Figure 9. The impact of our image transformations on face detection and file size for our LFW-derived data. As conditions get more severe (x-axis), the fraction of faces detected (blue curves) start dropping quite drastically after a certain point for blur and scale, but remain fairly constant for JPEG quality. In contrast, the ratio of input file sizes to the original images (red curves) drops with all transformations, and especially so with scale and JPEG quality (error bars indicate standard deviation). By exploiting the difference between these curves, and jointly considering reliability results (as shown in previous figures), attribute systems can operate on greatly reduced file sizes without significantly hurting classification reliability. For example, a JPEG quality of 15 uses less than 20% of the original space, and yet is still reliable for most attributes (Fig. 8).


[7] Douze, M., Ramisa, A., and Schmid, C., "Combining Attributes and Fisher Vectors for Efficient Image Retrieval," in [*IEEE CVPR*], (June 2011).

[8] Berg, T. L., Berg, A. C., and Shih, J., "Automatic Attribute Discovery and Characterization from Noisy Web Data," in [*ECCV*], (October 2010).

[9] Russakovsky, O. and Fei-Fei, L., "Attribute Learning in Large-scale Datasets," in [*ECCV Workshop on Parts and Attributes*], (September 2010).

[10] Lampert, C., Nickisch, H., and Harmeling, S., "Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer," in [*IEEE CVPR*], (June 2009).

[11] Parikh, D. and Grauman, K., "Interactively Building a Discriminative Vocabulary of Nameable Attributes," in [*IEEE CVPR*], (June 2011).

[12] Parikh, D. and Grauman, K., "Relative Attributes," in [*IEEE ICCV*], (November 2011).

[13] Siddiquie, B., Feris, R., and Davis, L., "Image Ranking and Retrieval Based on Multi-Attribute Queries," in [*IEEE CVPR*], (June 2011).

[14] Torralba, A. and Efros, A., "Unbiased Look at Dataset Bias," in [*IEEE CVPR*], (June 2011).

[15] Scheirer, W. J., Rocha, A., Michaels, R., and Boult, T. E., "Meta-Recognition: The Theory and Practice of Recognition Score Analysis," *IEEE TPAMI* **33**, 1689–1695 (August 2011).

[16] Scheirer, W. J., Rocha, A., Micheals, R., and Boult, T. E., "Robust Fusion: Extreme Value Theory for Recognition Score Normalization," in [*ECCV*], (September 2010).