Pre-print of article that will appear at CVPR 2012.

Multi-Attribute Spaces: Calibration for Attribute Fusion and Similarity Search

*Walter J. Scheirer[1,2] Neeraj Kumar[3] Peter N. Belhumeur[4] Terrance E. Boult[1,2]
[1]University of Colorado at Colorado Springs [2]Securics, Inc.
[3]University of Washington [4]Columbia University

Abstract

Recent work has shown that visual attributes are a powerful approach for applications such as recognition, image description and retrieval. However, fusing multiple attribute scores – as required during multi-attribute queries or similarity searches – presents a significant challenge. Scores from different attribute classifiers cannot be combined in a simple way; the same score for different attributes can mean different things. In this work, we show how to construct normalized "multi-attribute spaces" from raw classifier outputs, using techniques based on the statistical Extreme Value Theory. Our method calibrates each raw score to a probability that the given attribute is present in the image. We describe how these probabilities can be fused in a simple way to perform more accurate multi-attribute searches, as well as enable attribute-based similarity searches. A significant advantage of our approach is that the normalization is done after-the-fact, requiring neither modification to the attribute classification system nor ground truth attribute annotations. We demonstrate results on a large data set of nearly 2 million face images and show significant improvements over prior work. We also show that perceptual similarity of search results increases by using contextual attributes.

1 Introduction

Visual attributes are a powerful representation for a variety of vision tasks including recognition, classification, image description and retrieval. First proposed in the computer vision community by Ferrari and Zisserman [4], visual attributes are text labels that can be automatically assigned to scenes, categories, or objects using standard machine learning techniques. Kumar *et al*. [9] demonstrated the first system to automatically train several attribute classifiers for faces, such as "brown hair," "pointy nose," "thin eyebrows," or "wearing lipstick." Later work has looked at attributes in the context of recognition [11, 3, 10], zero-shot learning [11], discovery [1, 13], automatic image description [8], scene parsing [21], and attribute refinement [7, 10, 14, 19].

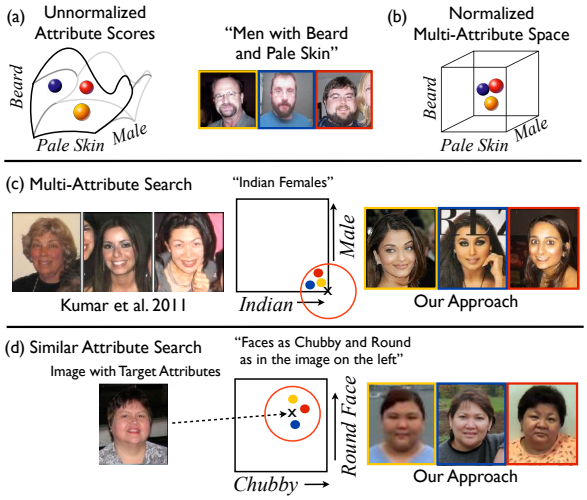

Figure 1. In this work, we show how to calibrate (a) raw attribute scores into (b) a "multi-attribute space" where each normalized value approximates the probability of that attribute appearing in the given image. This allows different attributes to be fused in a unified way – unlike the raw attribute scores, which have no uniform interpretation between different attributes. Distances between points in this space correspond to perceptual similarity between images, allowing for (c) better multi-attribute search results than prior work and (d) similarity searches based on target attributes, defined using a given query image.

Attribute classifiers take an image as input and return a real-valued score representing the presence of the given attribute in the image. Using these values directly suffices for some applications, but if multiple attributes are involved, issues of score calibration become paramount because (1) the distribution of scores for each attribute is usually not Gaussian, and (2) these distributions are often radically different for each attribute. Thus, looking at the distances between raw attribute values for different images does *not* correspond to similarity between the images (Fig. 1a). If, however, the scores were calibrated to the range $[0, 1]$ *and* these scores could be interpreted uniformly across different attributes, one could then use distances in this "multi-attribute space" to measure similarity as expected (Fig. 1b).

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

Unfortunately, current approaches do not calibrate attributes in this way. Some methods [18] do no normalization at all, while others [2, 10] normalize output scores assuming a Gaussian distribution, which meets the first condition but not the second – identical scores for two attributes don't correspond to any perceptual quality (*i.e.*, the degree to which those attributes are exhibited, or the probability that the given attributes are present). This has drastic implications for applications built on these scores. For multi-attribute searches, individual attributes can greatly bias search results, over- or under-emphasizing that attribute. For applications needing to find all instances "close" to a set of target attribute values, there is no principled way to define closeness that matches perceptual similarity. Indeed, in many systems the attribute scores are simply fed into another classifier with the hope that the second-stage classifier can figure out how to combine attribute values as needed.

In this paper, we show how to calibrate each attribute score to the probability that approximates how humans would label the image with the given attribute. Using a principled technique based on the statistical Extreme Value Theory (EVT) [17, 16], we fit a distribution to attribute scores close to *but on the opposite side of* the decision boundary for the attribute in question, *e.g.*, the scores for images classified only slightly negatively for the "female" attribute are used to estimate the probability of being "male." Counter-intuitively, the statistical fit from these "extreme values" is much more robust than one based on the strongly positive scores of a classifier. In fact, under mild assumptions, this distribution *must* be a Weibull. This allows for a normalization of raw classifier scores into a multi-attribute space, where comparisons and combinations of different attributes become "apples-to-apples." A significant advantage of our method is that it is done after-the-fact, requiring neither changes to the underlying attribute classifier nor ground truth attribute annotations.

One can visualize this normalization as a mapping of N attributes to the unit hypercube in $\mathbb{R}^N$ (see Fig. 1b). An image maps to a point in this hypercube based on its calibrated attribute scores. Simple metrics in this space measure distance between images in an intuitive way, and projecting the hypercube down to $\mathbb{R}^K$ allows for measuring similarities using only a subset of K attributes. Keeping this metaphor in mind, we can now define different search applications in a straightforward way.

Multi-attribute searches, such as "Indian females," map to a corner of a hypercube in $\mathbb{R}^K$ (see Fig. 1c), in this example with *Indian=1* and *male=0*. These are the kinds of queries possible in prior work [9, 18], but as shown in Fig. 1c and Sec. 5, our approach results in superior matches.

Target attribute similarity searches, such as "images similar to a given specified face with respect to face shape and weight," map the specified face to a point inside the hypercube defined by the given attributes, and return images close to that point, sorted by distance (see Fig. 1d). Since our scores are normalized, this results in perceptually similar matches. This is more powerful than prior methods that allow for similar functionality, as described next.

Simile classifiers [10] only measure similarity to a given part of the face, and have to be trained individually for each person, requiring additional data; in contrast, we can perform similarity searches on-the-fly using just a single image, and they can be based on any combination of target attributes. The work on relative attributes [14] uses pairs of images labeled with relative strengths of attribute values to learn a better ranking of attribute values (similar to methods using relational phrasing [5, 19] in the context of objects), but does not address the issue of combining multiple attributes. Other work [6, 7, 12, 20] has looked at ways to build more complete representations of faces for performing similarity searches; our approach is more general because it works within the attribute framework, which is not specific to faces, and can include appearance as well as geometry.

We also explore the effect of including **contextual attributes** in similarity searches for better perceptual similarity. This is done by raising the dimensionality of the similarity search hypercube to include other attributes (*e.g.* gender, age, hair) that form a context for understanding the image, and measuring distances in this space. Intuitively, if one were looking for similarity based on "curly hair," the gender of the person in an image would influence the perceptual similarity to query results, because hair style is usually evaluated in the context of gender.

Through extensive experiments, we show that search results for the above types of queries are far better when performed using our multi-attribute spaces, than those using Gaussian normalization. Since the quality of search results is subjective, we measure improvements quantitatively by asking hundreds of humans to compare the relevance of search results returned by different methods, and then evaluating the statistical significance of their preferences over hundreds of thousands of trials. We also show comparisons against the work of Kumar et al. [9] on a large data set of almost 2 million faces, highlighting the greater relevance of our matches. With the exception of that work itself, no other previous method has looked at data sets of this size.

2 Multi-Attribute Spaces

To analyze or combine multiple attributes in a meaningful way, their scores need to be properly normalized and, ideally, tied to how people would label an image.

Let $P(L(j)|I),\ j = 1 \ldots N$, be the probability that humans would assign label $L(j)$ to a given image I; $A_j(I)$ be attribute classifiers that map images to real-valued scores; and $E(A_j) \equiv |A_j(I) - P(L(j)|I)|$ be the expected labeling error in A_j approximating the labeling probability.

Def. 1 A continuous function $A_j : I \mapsto [0, 1]$ is called a *well normalized attribute function* when $E(A_j(I)) \leq \epsilon$ with a probability of at least 1 - δ.

Def. 2 A *multi-attribute space* $M : I \mapsto [0, 1]^N$ is a product space formed from well normalized attribute functions, $M(I) = A_1(I) \times A_2(I) \times \ldots \times A_N(I)$

With this idealized definition of multi-attribute spaces, attribute classifiers A_j for similar images produce similar scores, and scores in each dimension are an ϵ approximation of probabilities associated with an image being assigned a label. A multi-attribute space will, at least locally, support meaningful similarity measures, and since it approximates probabilities, its dimensions can be compared or fused.

The framework presented in Defs. 1 & 2 is a general one. Many different normalization schemes that produce a mapping into $[0, 1]$ will also satisfy Def. 1, given a large enough ϵ and δ. The key contribution in this paper is a calibration that not only conforms to Def. 1, but is based in a strong statistical theory that is appropriate for the raw attribute scores obtained via Support Vector Machine (SVM) classification. Note that quantifying ϵ and δ depends on the accuracy of the SVM, which is beyond the scope of this paper, and thus left for future work.

2.1 Calibration of SVM decision scores

The goal of our calibration is to map raw decision scores from a binary SVM to a probabilistic decision $A_j(I)$ that a given image I matches some attribute label $L(j)$. If we had enough ground truth data, we could estimate this function directly, but this is rarely the case. Instead, let us assume that we are only given the outputs of an existing classifier (without ground truth labels), which we would like to normalize. In general, the distribution of scores is *not* Gaussian, and thus difficult to estimate robustly. In particular, the "head" of the distribution (values around and greater than 1) can be quite fat, and yet these values are the least informative – they should all simply map to a probability of 1. However, the distribution of scores around 0 is much more informative and also more constrained – in fact, according to the statistical Extreme Value Theory (EVT) [16], if scores are bounded from above and below, it must be the Weibull distribution, which has shape parameter $k > 0$ and scale parameter $\lambda > 0$.

This idea was first exploited in the *w-score* technique [17] for normalizing scores from recognition systems with tolerance to "failure" cases. In a multi-class recognition scenario, suppose an algorithm is given a single input and outputs a set of scores, one for each class. If the algorithm succeeds, *i.e.*, assigns the highest score to the correct class, then the top score should be an outlier with respect to the distribution of all other scores (the "non-match distribution"). The EVT shows that the probability of a score being an outlier – a correct match – can be robustly estimated from

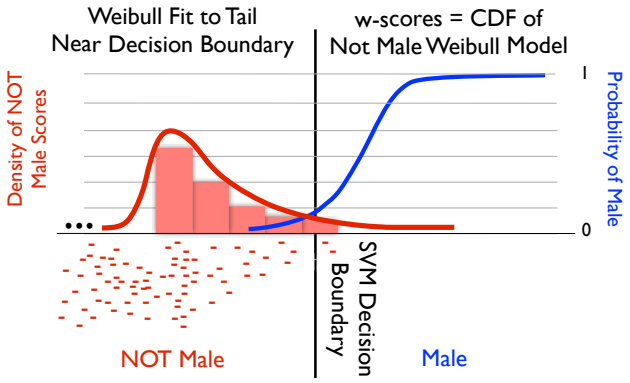

Figure 2. An overview of the score calibration algorithm introduced in this work. SVM decision scores are normalized by fitting a Weibull distribution (the red curve) to the tail of the *opposite* side of the classifier, where instances of the attribute of interest ("male" in this case) are outliers. The CDF of this distribution (the blue curve) is used to produce the normalized attribute w-scores. Note that no assumptions are made about the entire score distribution (which can vary greatly); our model is applied to only the tail of the distribution, which is much better behaved.

only the few highest values (excluding the correct match), not exceeding 50% of the total scores. This is the tail of the non-match distribution, its "extreme values." The following CDF is then used directly for score normalization once the parameters k and λ are found from the Weibull fitting:

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

In this formulation, x must be positive, and the extreme values must be the largest scores in the set of scores.

Unfortunately, this w-score formulation cannot be directly used with SVM-based attribute classifiers, as it assumes that the score set S comes from matching a single input image against a large gallery of classes, and also that there is a single hypothesized outlier with respect to the extrema of a non-match distribution – the true match. With binary SVMs, neither assumption is true: there are only two classes, and there will be many true match scores. We therefore reformulate the w-score method in a way that is consistent with both the EVT and binary SVM classification.

This can be done by an EVT fitting on the classification non-match distribution, followed by a hypothesis test that estimates the probability of a score being drawn from this distribution. In an SVM attribute context, the non-match distribution is the set of negative values obtained from the classifier, and so we must look at the scores from a classifier for the *opposite* attribute. As an illustration, consider the gender classifier in Fig. 2. If we are interested in estimating the probability of a face being "male," we look at the distribution of scores that are negative with respect to the *non*-male side (*i.e.*, the "female" side). The CDF of the Weibull fit to these scores directly gives us the probability that a face is male. For attributes which are multinary or not strictly binary, *e.g.*, "black hair," the negative class is composed of a mixture of "opposite" attributes – blonde hair, brown hair, grey hair, and red hair.

Algorithm 1 EVT Norm. of binary SVM decision scores

Input: A vector of decision scores $S = \{s_i\}$, of length m, from a single binary SVM attribute classifier
Input: n, the number of elements used for fitting
Let $V \subset S$ be scores from the opposite decision space
if the negative decision space is chosen **then** $\phi = 1$
else $\phi = -1$
end if
$\hat{V} = \phi * V$ ▷ If needed, flip so extrema are largest
Sort $\hat{V}$ retaining n largest scores: $D = d_n > \ldots > d_1$
Let $\hat{D} = D - d_1 + 1$ ▷ Shift to be positive
Fit a Weibull distribution W to $\hat{D}$
Let $T(x) = \phi * x - d_1 + 1$;
for $i = 1 \rightarrow m$ **do**
    $s'_i = F(T(s_i); W)$
end for
return normalized decision scores $\{s'_i\}$

The calibration process is detailed in Alg. 1. Given an input feature vector, an SVM outputs a score s. Normally, the sign of s determines its class – positive or negative. In our work, however, we are interested in calibrating the score itself by fitting a Weibull $W(k, \lambda)$ to the extreme values of the non-match distribution. From an input set of scores $s_i \in S$, these are the negative values closest to 0, *i.e.*, the highest scores that *don't* correspond to the positive class. We first apply a transform T that flips and shifts these scores as necessary to satisfy the two conditions needed by Eq. 1 (the data must always be positive, regardless of the side of the decision boundary we're considering), then fit a Weibull to the transformed scores, and finally normalize each score using its CDF:

$$F(T(s_i); W) \tag{2}$$

Note that ground truth data is not necessary for this process (assuming the attribute classifier SVMs have already been trained using appropriate data). As a formal probability estimate for the attribute label, this normalization meets the definition of a well-normalized attribute function and creates multi-attribute spaces consistent with Def. 2.

3 Fusion for Multi-Attribute Search

A multi-attribute search, such as "Indian females" (as shown in Fig. 1c), requires fusing the scores for each attribute in query q into a combined score s^q. Since our calibration procedure described in the previous section converts attribute values into a multi-attribute space, where scores A_j represent probabilities, one might assume that performing a search would simply be a matter of multiplying the appropriate attribute values. Unfortunately, there are a few issues with this scheme. First, many attributes are correlated, *e.g.*, "male", "beard", and "mustache." Teasing out these correlations can be very difficult, as some are due to inherent correlations in real life and some are due to classifier biases. Second, if the search involves many attributes, there might not be any images in the database that exhibit all of those attributes. Therefore, we formulate the search problem slightly differently:

$$\begin{aligned} \text{maximize over } I \quad & s^q = \| A_j(I) \|_1 \\ \text{subject to} \quad & A_j(I) = F(T(s_j(I)); W_j); \\ \text{for } \forall\, j \in J \text{ satisfying} \quad & 0 \leq \alpha_j \leq A_j(I) \leq \beta_j \leq 1; \end{aligned} \tag{3}$$

The goal here is to find the images that maximize the L_1 norm of estimated probabilities for each attribute j in the query set of attributes J that also satisfy constraints given by parameters α_j & β_j, which define a range of scores of interest for each attribute. For example, setting $\alpha_j = 0.95$ and $\beta_j = 1$ restricts the problem to use images that have at least 95% confidence in label j. Since calibrated attribute scores A_j can be precomputed for a images in a database, this optimization reduces to selecting the images satisfying the search constraints and then computing the maximum of the sum of the attribute scores.

The choice of the L_1 norm here is important, since it provides for robustness in the case that no images matching all attributes can be found in the database. If this happens, it is probably preferable to return the images that have high probabilities for $n - 1$ or $n - 2$ attributes, and perhaps not as high for the remaining 1 or 2. In contrast, multiplying attribute values would tend to favor images where all attributes are somewhat likely, but none is particularly low. As the bound α_j is lowered, this would start returning images where none of the attributes match, whereas our formulation would still return images with some relevance.

4 Multi-Attribute Space Similarity

Measuring the similarity of two images is an important and well-studied problem. However, most approaches do so on simple transformations of the original image space and are thus strongly biased by the "configuration" of the objects in question: pose, illumination, expression, *etc*. For instance, in face recognition, images of different individuals taken under the same pose and lighting are often more similar than those of the same individual under different conditions. Since attributes are designed to capture aspects of appearance independent of imaging configuration, we should be able to measure similarity better in a multi-attribute space.

What makes this problem particularly challenging – even with the use of attributes – is that perceptual similarity is not uniform across all values for a given attribute, or between attributes. For example, the perceived difference between images with calibrated attribute values A_j in the range of $0.81 - 0.84$ might be similar to that perceived between those in the much larger range of $0.2 - 0.4$. In this case, our similarity function should weight small distances in the first range more strongly than in the latter range. Or we might

be much more sensitive to small differences in one attribute as compared to another, and should thus weight distances accordingly. Since it is infeasible to directly measure these perceptual distances over an entire multi-attribute space, we must rely on a method that can estimate these directly from individual classifier outputs, without ground truth labels.

Let us consider the following scenario: a user selects a target image and wants to find images that are similar to it, with respect to a given set of k attributes. The search function should compute distances in the k-dimensional multi-attribute subspace corresponding to the given attributes, but in such a way as to respect the distribution of attribute values in that "neighborhood." The size of the neighborhood considered should be changeable to allow for different levels of similarity, *i.e.*, small neighborhoods result in searching only for "very similar" images, while larger neighborhoods correspond to images which might be only "somewhat similar." Finally, the relative neighborhood size of different attributes should be changeable as well, for emphasizing one attribute compared to another.

Our method for solving this problem applies the EVT normalization to *distances* for each of the target attributes individually, and then sums these for the final similarity score. We first gather images with calibrated attribute values $\alpha_j \leq A_j(I) \leq \beta_j$, where the neighborhood is defined by the range $\alpha_j - \beta_j$ for each attribute (allowing for both different neighbor sizes, and relative weighting of different attributes). For each attribute, we compute the set of L_1 distances between the target attribute value and each of the gathered images. The largest of these distances is assumed to be just outside the "similar" range – the outlier with respect to similarity – and thus the distances immediately smaller (the tail of the extreme values) can be used to fit a Weibull. Intuitively, we are measuring the local distribution of distances for this attribute, close to the target attribute value. As before, we can then use the CDF of the Weibull to estimate the probability that a particular image is "similar" with respect to this attribute and given search range. Finally, we maximize over the L_1 sum of these calibrated probabilities for each attribute (analogously to Eq. 3) and return the corresponding images to the user. The full code for this algorithm is available on the paper's companion website[1]. Since this normalization depends on the particular attribute values of the target image as well as the search ranges, we must perform this calibration at run-time for each query; however, this process is quite fast, typically requiring only a few seconds at most.

Examples of similarity searches are shown in Fig. 3. On the left, we show the results for faces with nose most like that of the target image of Jackie Chan; on the right, we show the results for faces with smiles most like that of the image of Angelina Jolie. The first query maps to the attribute "big nose", while the latter maps to a 3-dimensional subspace consisting of the attributes "smiling, "lipstick", and "high cheekbones." The most similar results are shown around the target images. Note that we have zoomed in on the relevant parts of the face for ease of comparison.

[1] http://www.metarecognition.com/

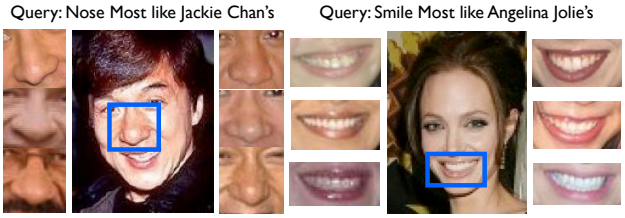

Figure 3. Results for similarity searches using a set of target attributes and a query image. By calibrating attribute distances in a local neighborhood around the normalized target attribute values, we can compute similarity in a consistent manner for any set of attributes and query images, despite the fact that perceptual similarity changes quite drastically with the attribute values in question.

One final issue is whether people can truly consider only a given set of query attributes in isolation. For example, face shape is often viewed in context of gender and age of the person, and it might be difficult to separate those components when looking at similarity results. To examine this effect, we can add "contextual attributes" to a query set, in effect forcing results to be similar with respect to not only the query attributes, but also these contextual attributes. In our experiments, these consist of gender (male, female), age (baby, child, youth, middle-aged, senior), and hair color (blonde, black, brown). Fig. 5c provides an illustration of the resulting reordering achieved with these contextual attributes. Note how the reordered results are perceptually more similar to the target.

5 Quantitative Experiments

Having presented the conceptual models for multi-attribute spaces and algorithms for computing them, we now turn to evaluation. While these concepts apply to many types of attributes and a variety of applications, our evaluation here focuses on face attributes for search. Face search based on visual similarity is not a new topic, with the notable *Photobook* work of Pentland et al. [15] appearing in 1996. Recently, more powerful approaches to the problem [10, 18] have shown excellent promise for general use. The work of Kumar et al. [10] is especially relevant, as it represents the largest scale face attribute approach in terms of data and number of attributes available in the literature. Where applicable, we compare directly to that work.

While Figs. 1, 3, 4, 5 show some qualitative results illustrating the merits of our method, we also perform rigorous and extensive quantitative evaluation. Since search results are subjective (by definition), we obtain quantitative results by asking workers on Amazon's Mechanical Turk service to rate search results. We gathered hundreds of thousands of responses from hundreds of workers, and then used sta-

Figure 4. Comparisons between the weighted SVM decision score fusion approach of Kumar *et al*. [10] (left) and our multi-attribute space fusion approach (right) for the top five results on a selection of queries, made over nearly 2 million face images from the web. Without proper normalization (left), certain attributes can dominate a query, *e.g*., gender in the first query, and ethnicity and age in the second.

tistical analysis to determine the validity (significance) of their responses. In particular, we ran 4 sets of experiments, corresponding to testing these 4 hypotheses:

H_1**:** For attribute-based search queries, multi-attribute space fusion with EVT normalization is more consistent with human rankings than fusion with state-of-the art Gaussian normalization.

H_2**:** For target-based similarity queries, using L_1 distance in the query-only multi-attribute space provides an ordering consistent with human similarity ranking.

H_3**:** For target-based similarity queries, using L_1 distance in a **contextual** multi-attribute space provides an ordering consistent with human similarity ranking.

H_4**:** For target-based similarity queries, using L_1 distance in a contextual multi-attribute space provides better ordering than just distance in query-space.

In each case, our null hypothesis H_0 is that there is no difference between what is being compared. We then perform a statistical test to reject H_0.

We evaluated these hypotheses using $1,932,987$ face images downloaded from the web, including a large portion of the Columbia Face Database [10], as well as many other images. Raw attribute scores for each image are computed using the method of Kumar et al. [10], via their publicly available implementation[2]. Given all scores for an attribute, we normalize them using Alg. 1. This provides the well-normalized attribute functions from which we obtain our multi-attribute spaces used for experiments.

5.1 Hypothesis H_1: Multi-Attribute Search

For the first hypothesis, the queries to our system come from a text string naming one or more attributes. For multi-attribute queries, the scores for each image are evaluated using Eq. 3, with $\alpha = 0$ and $\beta = 1$, and the images are shown to the user sorted by scores, highest first. We compare our multi-attribute space approach directly with the face retrieval approach of Kumar et al. [10], which converts weighted SVM decision scores into probabilities by fitting Gaussian distributions on a separate set of positive and negative attribute examples, and then uses the product of probabilities to rank the results. From the qualitative comparison shown in Fig. 4, one can see that despite the weighting and Gaussian normalization, certain attributes dominate the queries in the approach of Kumar et al. The Gaussian normalization is likely overemphasizing variations within the tails – *e.g*., weighting minor "baby" variations more heavily than scores for hats. Our multi-attribute space normalization and optimization emphasizes more meaningful differences in terms of label probabilities.

For quantitative evaluation, we generated 30 different textual queries, sampling from combinations of 42 different attributes, and submitted side-by-side top 10 face retrieval results to 30 different Mechanical Turk workers (order and side were randomized per test to remove presentation bias). For all 30 queries, each worker was asked to identify the set of results that was more relevant, yielding 900 individual comparisons. H_0 in this case indicates equal votes for each algorithm for each query, which was evaluated with a one-sided paired t-test. The results from our approach were chosen as "more relevant" by the workers for the given queries 86.9% of the time, yielding rejection of the null hypothesis with a p-value $< 10^{-16}$. **Thus, we accept our hypothesis H_1 and conclude that the EVT-based multi-attribute space provides significantly better fusion for search.**

5.2 Target-based Similarity Search

The remaining experiments test the three hypotheses related to target-based similarity queries (*e.g*., "find people with a round nose and black hair similar to the given image"). There is no prior work for direct comparison to this novel and useful capability, and so we focus on the quantitative results. These hypotheses are related to the ordering produced for a target-based attribute query, and are used to determine if a particular automatic ordering of the results is consistent with human orderings of similarity. Our statistical tests are designed to not only evaluate the overall viability of the

[2] http://afs.automaticfacesystems.com/

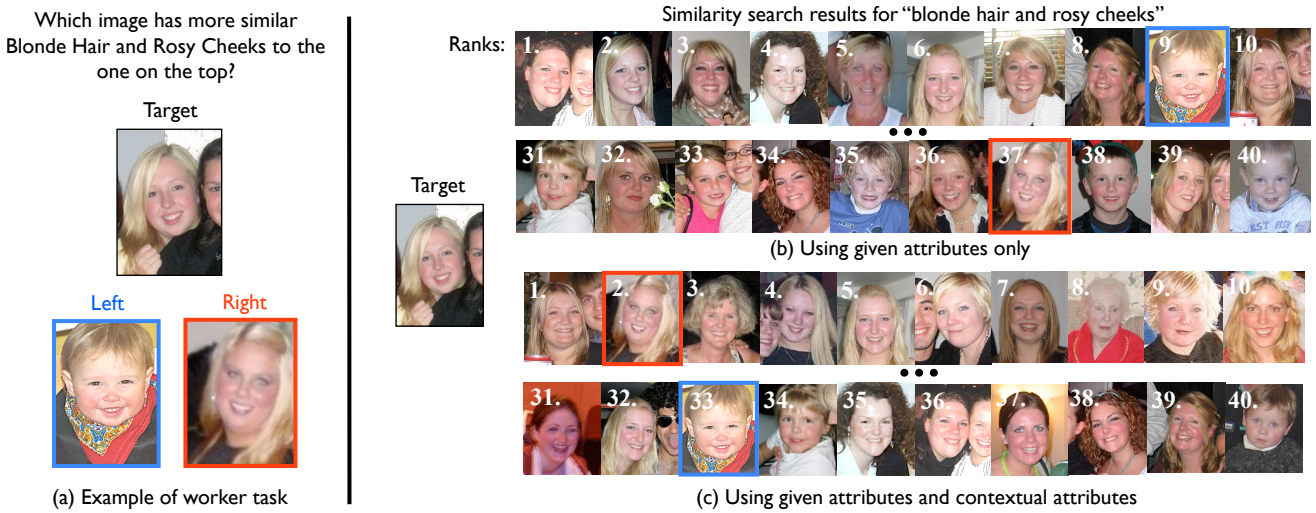

Figure 5. Using the task template shown in (a), workers were asked to rate the similarity search results for the target image w.r.t. attributes "blonde hair" and "rosy cheeks." The results of our algorithm are shown in (b) using only the query attributes, and in (c) using additional contextual attributes as well. The latter looks better (*e.g.*, see the relative order of the highlighted faces). Our quantitative experiments (summarized in Fig. 6) used hundreds of thousands of pairwise comparisons to assess hypotheses H_2 : (b) is similar to human rankings, H_3 : (c) is similar to human rankings, and H_4 : (c) is better than (b).

technique, but also the utility of these orderings for search tasks. We present a summary of our results in this paper. A more detailed analysis, including the full statistics, is also available on our companion website.

Experimental design is important here; asking human subjects to rate attribute strength directly is prone to strong biases (as became apparent during our experimentation). Thus, our measurements are based on a relative comparison approach, shown in Fig. 5a. Each worker was asked to select the image that most resembled the target image *with respect to the target attributes* from two candidates. The candidates were chosen from a set of the 40 closest images in query-only distance (described in Sec. 4) and 10 additional randomly selected images as "negative" examples, to reduce bias and provide meaningful comparisons for the bottom images in the top 40. These 50 images were fixed for each query, and were used to generate 100 unique pairs, with balanced left-right presentation and different random selections for each worker.

For our experiments, we generated 12 different attribute combinations (four 1-attribute, four 2-attribute and four 3-attribute queries) and chose score targets for each attribute by sampling a face from the available images and selecting the relevant normalized attribute scores as the targets for a query. With 100 workers evaluating 100 pairs for each of 12 queries, we collected a total of $120,000$ comparisons. Our null hypothesis H_0 is that our similarity search ordering is not consistent with human responses. A test is "consistent" if the automatic distance rank and worker response both indicate that a particular image from a pair is closer to the target image than the other image in the pair. We test each set of results by comparing to random chance, with a χ^2 test for significance, rejecting H_0 if $p < 0.01$. To measure how finely we can distinguish similarities, we test statistical significance separately for different intervals of results, looking first at the top 5 matches only, then the top 10 matches, and so on down to the top 40. This will show us if, for instance, our algorithm finds very similar images correctly, but not those which are only somewhat similar.

Fig. 6 shows the results for our hypothesis tests H_2, H_3, and H_4 for each query and each top n results set. A particular ring is marked with ** for a given query if the results were statistically significant to the $p = 0.01$ level, * for the $p = 0.05$ level, and - for not significant. The results indicate that H_2 and H_3 are valid and statistically significant for most queries, *i.e.*, that results from both the original query and from the additional contextual attribute results are consistent with human rankings.

Finally, we consider the hypothesis H_4, which seeks to identify if measuring distance in a contextual multi-attribute space provides better ordering than just distance in query-space. Here H_0 indicates that the orderings are the same. Using the collected results for each ordering approach, we applied a pooled sample test for differences in the means between the results for the two different distance measures, considering the null hypothesis rejected when $p < 0.01$ (see Fig. 6, right). Interestingly, the contextual-attribute ordering was significant only for some queries, indicating that adding these contextual attributes is not always helpful.

6 Discussion

In this paper, we have formalized the notion of multi-attribute spaces and shown how to calibrate attribute values into probabilities that an image exhibits a given at-

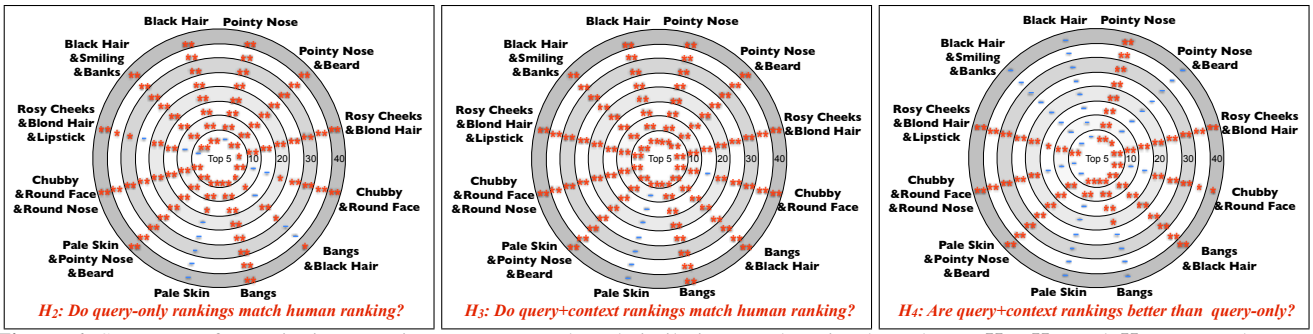

Figure 6. Summary of quantitative experiments on target-based similarity search, using hypotheses H_2, H_3, and H_4. For each set of query attributes (on the outside of the circles) the hypothesis was evaluated for each cumulative set of 5 ranks (moving from the center circle outwards: top-5, top-10, *etc.*). A ** means the results were statistically significant at $p = 0.01$, * means $p = 0.05$, and - means not statistically significant. Both algorithmic rankings (query-only and query+context) match human rankings well. Adding contextual attributes improves results for some queries but not others. See text for details.

tribute. Through extensive experiments on a large data set of almost 2 million faces, we have shown that our principled probabilistic approach to score normalization greatly improves the accuracy and utility of face retrieval using multi-attribute searches, and allows for the new capability of performing similarity searches based on target attributes in query images. We have publicly released our calibration code on our companion website[3].

As the use of attributes continues to expand to other application areas, we expect that techniques based on multi-attribute spaces will be crucial for effectively using the outputs of multiple attribute classifiers. For example, face verification using attributes [10] is currently done by training a second stage verification classifier on raw attribute scores. It would be exciting to obtain similar or better results using simple metrics in a multi-attribute space – an approach which would be more intuitive and easier to reason about.

References

[1] T. L. Berg, A. C. Berg, and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *ECCV*, October 2010.

[2] M. Douze, A. Ramisa, and C. Schmid. Combining Attributes and Fisher Vectors for Efficient Image Retrieval. In *IEEE CVPR*, June 2011.

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

[4] V. Ferrari and A. Zisserman. Learning Visual Attributes. In *NIPS*, December 2007.

[5] A. Gupta and L. S. Davis. Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers. In *ECCV*, October 2008.

[6] A. Holub, Y. Liu, and P. Perona. On Constructing Facial Similarity Maps. In *IEEE CVPR*, June 2007.

[7] A. Holub, P. Moreels, and P. Perona. Unsupervised Clustering for Google Searches of Celebrity Images. In *IEEE AFGR*, September 2008.

[8] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby Talk: Understanding and Generating Simple Image Descriptions. In *IEEE CVPR*, November 2011.

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

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable Visual Attributes for Face Verification and Image Search. *IEEE TPAMI*, 33(10):1962–1977, October 2011.

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

[12] A. K. Moorthy, A. Mittal, S. Jahanbin, K. Grauman, and A. C. Bovik. 3D Facial Similarity: Automatic Assessment versus Perceptual Judgements. In *IEEE BTAS*, 2010.

[13] D. Parikh and K. Grauman. Interactively Building a Discriminative Vocabulary of Nameable Attributes. In *IEEE CVPR*, November 2011.

[14] D. Parikh and K. Grauman. Relative Attributes. In *IEEE ICCV*, November 2011.

[15] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-Based Manipulation of Image Databases. *IJCV*, 18(3):233–254, June 1996.

[16] W. J. Scheirer, A. Rocha, R. Michaels, and T. E. Boult. Meta-Recognition: The Theory and Practice of Recognition Score Analysis. *IEEE TPAMI*, 33(8):1689–1695, August 2011.

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

[18] B. Siddiquie, R. Feris, and L. Davis. Image Ranking and Retrieval Based on Multi-Attribute Queries. In *IEEE CVPR*, June 2011.

[19] B. Siddiquie and A. Gupta. Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Class Active Learning. In *IEEE CVPR*, June 2010.

[20] B. Smith, S. Zhu, and L. Zhang. Face Image Retrieval by Shape Manipulation. In *IEEE CVPR*, June 2011.

[21] J. Tighe and S. Lazebnik. Understanding Scenes on Many Levels. In *IEEE ICCV*, November 2011.

[3] http://www.metarecognition.com