Exploring Locally Rigid Discriminative Patches for Learning Relative Attributes

Yashaswi Verma
http://researchweb.iiit.ac.in/~yashaswi.verma/
C. V. Jawahar
http://www.iiit.ac.in/~jawahar/

CVIT
IIIT-Hyderabad, India
http://cvit.iiit.ac.in

Abstract

Relative attributes help in comparing two images based on their visual properties. These are of great interest as they have been shown to be useful in several vision related problems such as recognition, retrieval, and understanding image collections in general. In the recent past, quite a few techniques have been proposed for the relative attribute learning task that give reasonable performance. However, these have focused either on the algorithmic aspect or the representational aspect. In this work, we revisit these approaches and integrate their broader ideas to develop simple baselines. These not only take care of the algorithmic aspects, but also take a step towards analyzing a simple yet domain independent patch-based representation for this task. This representation can capture local shape in an image, as well as spatially rigid correspondences across regions in an image pair. The baselines are extensively evaluated on three challenging relative attribute datasets (OSR, LFW-10 and UT-Zap50K). Experiments demonstrate that they achieve promising results on the OSR and LFW-10 datasets, and perform better than the current state-of-the-art on the UT-Zap50K dataset. Moreover, they also provide some interesting insights about the problem, that could be helpful in developing the future techniques in this domain.

1 Introduction

The goal of relative attribute learning is to learn attribute-specific models such that given a pair of images, we can predict which image exhibits a particular attribute more [20]. Relative attributes have been successfully employed in a variety of problems such as interactive image search [13, 14], scene classification [25], active learning [1, 21], etc., thus making the problem of learning accurate relative attribute models of immense practical interest.

Relative attribute learning is challenging primarily for two reasons: **(a)** First, in relative attributes, each data-point denotes a pair of images with an associated ordering information. For a given image-pair, this ordering indicates the relative strength of a particular attribute (more/less/similar). As visual differences within image pairs become more and more subtle, there may arise intransitive orderings (*e.g.*, there may exist pairs with orderings $A \succ B$, $B \succ C$ and $C \succ A$). This is because of human-inconsistencies in ground-truth annotation [2]. Due to this, learning a single (global) model for all the pairs may not be sufficient [27]. **(b)** The second difficulty arises due to weak-labeling; i.e., lack of correspondence between attribute

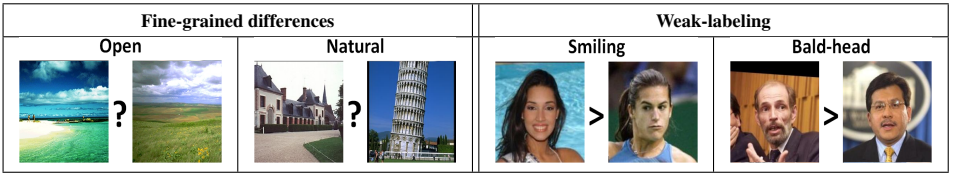

Figure 1: Two primary challenges in the visual comparison task are **(a)** fine-grained within-pair differences (left) that make a global model insufficient, and **(b)** weak-labeling (right) as training data has no information about which region(s) correspond to a particular attribute.

label and image regions in the training data. Since the attribute-specific ordering is assigned to an image-pair as a whole, it remains unknown which regions in the images correspond to that attribute. This makes it difficult to address the correspondence problem under a discriminative framework. Figure 1 illustrates these two challenges.

To our best knowledge, four approaches are proposed that *particularly* focus on "relative attribute learning" problem: **(a)** Parikh and Grauman [20], who were the first to introduce the idea of relative attributes, use a single attribute-specific ranking model learned using global features (*e.g.*, GIST or colour histogram); **(b)** Li *et al.* [18] learn a hierarchy of ranking functions, successively trained using smaller subsets of data; **(c)** Yu and Grauman [27] do local learning of ranking model for each test pair using a subset of the few most analogous (training) pairs; and **(d)** Sandeep *et al.* [23] learn a single ranking model per attribute similar to [20], however use a part-based representation rather than global features. Among these, Yu and Grauman [27] target fine-grained differences and focus on the algorithmic aspect. Whereas, Sandeep *et al.* [23] target weak-labeling and focus on the representational aspect.

Inspired from these works, we present a family of baseline techniques for the visual comparison task that take into consideration both representational as well as algorithmic aspects. These are based on adopting principles from state-of-the-art visual comparison methods [20, 23, 27], and integrating them with a locally rigid yet discriminative patch-based representation [4]. As we will illustrate and discuss in the later parts of the paper, this representation turns-out to be computationally more efficient and easily adaptive for diverse domains than the one proposed in [23], thus offering a wider applicability.

We perform extensive experiments on three challenging and diverse datasets, that include UT-Zap50K (shoes) [27], LFW-10 (faces) [23], and OSR (outdoor scenes) [19] datasets. Results indicate that even with the simplest novelty, the proposed baselines achieve promising results on the LFW-10 and OSR datasets, and state-of-the-art results on the (most challenging) UT-Zap50K dataset. Moreover, they also provide interesting insights about the visual comparison task, and thus may be informative for future methods addressing this.

2 Related Work

Here we briefly review some of the works related to attributes, relative attributes, and their applications. In a broad sense, attributes are mid-level properties that are generally comprehensible by humans. These may carry semantic information, and have been found to be useful in several problems such as object recognition [15, 26], describing unseen objects [6], recognizing unseen categories [7, 16], image search [24], etc. The focus of most of these works was on identifying the presence or absence of some attribute. However, usually the

Figure 2: Diversity in domains such as shoes and outdoor scenes makes it difficult to use some off-the-shelf part-detection technique.

spectrum of an attribute's visibility across images is very wide, and it makes more sense to deal with its perceived strength rather than absolute presence/absence. This led to the idea of relative attributes [20]. The fundamental concept behind relative attributes is to perform pairwise comparison rather than absolute prediction. In the recent years, relative attributes have been adopted in several applications. In interactive image search [12, 13, 14], a human can incrementally refine the search results based on relative attribute based feedback. In active learning [1, 21], similar feedback is used by a teacher (human) for correcting the learner (machine), and communicating the reasons for error. Other applications include zero-shot learning and image description generation [20], relative human descriptions that help in comparative biometrics (such as relative height) [22], etc.

Inspired by the success of [20], a few approaches have been proposed that focus on improving the efficacy of the visual comparison task. Rather than learning a single ranking function, a hierarchy of functions was learned in [18]. Recently, in [23], rather than using a global GIST descriptor or colour histogram (as in [20, 27]), image representation was formed by first detecting facial regions using [29], and then concatenating their features analogous to [8, 17]. This was then used for learning a global ranking function [20]. Due to (available) correspondence across parts, it was able to identify attribute-specific semantically meaningful regions. In [27], the ideas of nearest-neighbour based local learning [9, 28] and metric learning [5] were integrated to learn local (per test pair) ranking functions that were similar to [20]. This particularly improved the accuracy for fine-grained attribute comparisons. All these approaches were experimentally shown to outperform [20]. It is important to note that among the existing relative attribute learning approaches, [18, 20, 27] are domain-independent. However, that of [23] requires a domain-specific pre-trained part-detection model in order to form part-based representation, and thus was shown to work only on near-frontal faces where this is an almost solved problem [29]. For the same reason, [23] can not be directly applied to domains that are highly deformable by nature (*e.g.* outdoor scenes and shoes as shown in Figure 2), where learning part-detection models as accurate as those for faces is still an open problem.

3 Approach

Our approach for relative attribute learning involves three main steps: forming a patch-based representation, identifying analogous pairs, and learning a local ranking function. Initially, a patch-based representation is formed for all the training images. Given a test image pair, first

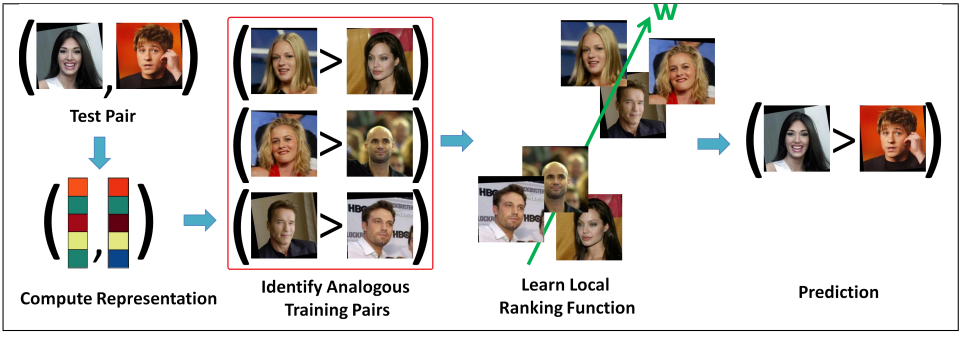

Figure 3: Approach overview: Given a test pair, first its patch-based representation is computed. Then using this representation, its analogous training pairs are identified. These pairs are used to learn a (local) ranking function, which is finally used for relative attribute prediction ("smiling" in the above illustration).

its images are represented using similar representation. Then, in this representation space, its few most analogous training pairs are identified using a pairwise distance function [27]. These pairs can be thought of as the K nearest-neighbours of the test pair. Finally, a ranking function is learned [20] using only these training pairs, and is used for comparing the two images in the test pair. Figure 3 gives an overview of this pipeline. Below we describe each of these steps in detail.

3.1 Forming Patch-based Representation

Rather than comparing two images as a whole using a global descriptor such as GIST or colour histogram, a patch-based representation helps in comparing regions between two images, and thus reduces the impact of the weak-labeling issue. With this motivation, a patch-based representation for visual comparison was proposed in the Relative Parts approach [23]. However, this approach has the following limitations: **(a)** It relies on a pre-trained domain-specific model for detecting local regions, which restricts its applicability to domains for which such models are easily available (*e.g.*, faces [29]). For other domains such as shoes or outdoor scenes which exhibit large variations in their images, identifying and learning discriminative patches remains a challenge (Figure 2). **(b)** It represents each of the detected regions using a bag-of-words histogram over SIFT descriptors with 100 visual words. Since the number of regions detected in an image is usually large (83 for faces), the high-dimensional concatenated representation becomes computationally restrictive, particularly for local learning of ranking models (Section 3.2).

At this point, it is worth re-iterating that our final goal is pairwise ranking, and this is conceptually different from the conventional object/scene categorization task. Given a test pair, first the learned ranking function evaluates each of its two images (i.e., their descriptors) individually, and then their scores are compared for final prediction. This means that though data is in the form of pairs of images, data representation is still at the level of *individual* image rather than image pair. With this insight, we adopt ideas for image representation that have been successfully applied in the categorization domain [9], and discuss a representation based on patches. In Section 4.4, we will illustrate how this representation can be used to implicitly learn the region that are discriminative with respect to a given attribute.

As a baseline patch-based representation, given an image, we compute HOG [4] descriptors from non-overlapping square patches (details are provided in Section 4.2) and concatenate them. This basic representation efficiently captures local shape in an image, as well as spatially rigid correspondences across regions in an image pair. Moreover, this is generic (i.e., equally applicable to all the images irrespective of their domain or visual content), and involves no explicit part-learning as in [23]. In the past, this has been successfully adopted in several visual recognition tasks (*e.g.*, object detection [8]). The motivation behind using this for the relative attribute learning task is the observation that images in several domain-specific datasets (such as shoes and faces) are largely aligned, and spatial variations in the regions of interest are globally minimal (Figure 4). As we will demonstrate in the experiments, this can provide significant boost in performance compared to using traditional representations (such as GIST and colour histogram).

3.2 Identifying Analogous Pairs

As visual differences within an image-pair become more and more subtle, a single prediction model trained using the whole dataset may become inaccurate. This is because it captures only the coarse details, and smoothens the fine-grained properties. The Local Learning approach [27] is motivated by this fact, and proposed to consider only the few training pairs for each test pair that are most analogous to it. The selected analogous pairs should be such that not only their individual images match with images in the test-pair, but their within-pair differences should also resemble that of the test-pair, thus allowing to capture local fine-grained differences.

Let $\mathcal{I} = \{I_i\}$ be a collection of images, with each image I_i being represented by a descriptor $\mathbf{x}_i \in \Re^d$. For every attribute $\mathcal{A}$ in a dataset, there is a set $\mathcal{O}_\mathcal{A} = \{X_p\}$ where each $X_p = (I_p, I_{p'})$ denotes an ordered pair of images, such that image I_p exhibits the attribute $\mathcal{A}$ more than image $I_{p'}$. Similarly, there is another set $\mathcal{S}_\mathcal{A} = \{X_q\}$ where each $X_q = (I_q, I_{q'})$ denotes a pair of images such that the visibility of $\mathcal{A}$ is similar in both I_q and $I_{q'}$.

In order to identify analogous pairs, a paired distance function is adopted [27]. Given a test pair $X_t = (I_t, I_{t'})$ and a training pair $X_p = (I_p, I_{p'}) \in \mathcal{O}_\mathcal{A}$, the distance between them is:

$$D_\mathcal{A}(X_t, X_p) = min\Big(D'_\mathcal{A}((\mathbf{x}_t, \mathbf{x}_{t'}), (\mathbf{x}_p, \mathbf{x}_{p'})), D'_\mathcal{A}((\mathbf{x}_t, \mathbf{x}_{t'}), (\mathbf{x}_{p'}, \mathbf{x}_p))\Big), \tag{1}$$

where $D'_\mathcal{A}$ is the product of two distances:

$$D'_\mathcal{A}((\mathbf{x}_i, \mathbf{x}_j), (\mathbf{x}_k, \mathbf{x}_l)) = d_\mathcal{A}(\mathbf{x}_i, \mathbf{x}_k) \times d_\mathcal{A}(\mathbf{x}_j, \mathbf{x}_l). \tag{2}$$

This product indicates that for a query pair and training pair, if both the query-training couplings are dissimilar, the distance will be high, and vice-versa. Also, the "minimum" of two terms in Eq. 1 signifies that while the true orderings for all the training pairs are known, it is unknown for the test pair. Hence, the distance accounts for both the possible orderings (more or less) for the test pair [1]. In Eq. 2, the distance $d_\mathcal{A}$ is defined as a Mahalanobis distance:

$$d_\mathcal{A}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M}_\mathcal{A} (\mathbf{x}_i - \mathbf{x}_j), \tag{3}$$

Rather than using a simple Euclidean distance, this helps in giving more importance to those feature dimensions that are more representative of a particular attribute, and less to others.

[1] While testing, pairs with similar attribute strength are not considered similar to previous methods.

The metric $\mathbf{M}_{\mathcal{A}}$ is a square positive semi-definite matrix. It is learned using the information-theoretic metric learning (ITML) algorithm [5]. The pairs of images in the sets $S_{\mathcal{A}}$ and ordered $O_{\mathcal{A}}$ constitute the pairwise constraints for ITML, such that images within similar pairs should be pulled closer, and those within ordered pairs should be pushed farther. The second set of constraints particularly helps in tuning $\mathbf{M}_{\mathcal{A}}$ towards attribute-specific fine-grained differences within image-pairs.

3.3 Learning A Local Ranking Function

Given a test pair $X_t = (I_t, I_{t'})$ and an attribute $\mathcal{A}$, our final step is to learn a function that can predict the relative strength of $\mathcal{A}$ for the two images in X_t. Similar to [20], we treat this task as a ranking problem. In order to learn X_t-specific local ranking function, we compute its distance from all the pairs in $\mathcal{O}_{\mathcal{A}}$ using Eq. 1, and pick the K pairs with least distance to form $\mathcal{O}'_{\mathcal{A}}$. The pairs in $\mathcal{O}'_{\mathcal{A}}$ are then used to learn a linear ranking function $R_{\mathcal{A}}$ defined as:

$$R_{\mathcal{A}}(\mathbf{x}) = \mathbf{w}_{\mathcal{A}}^T \mathbf{x}. \tag{4}$$

Here $\mathbf{w}_{\mathcal{A}} \in \Re^d$ is the parameter vector. The goal is to learn $\mathbf{w}_{\mathcal{A}}$ such that the orderings of the pairs in $\mathcal{O}'_{\mathcal{A}}$ are satisfied as much as possible. That is, $\forall X_p \in \mathcal{O}'_{\mathcal{A}} : \mathbf{w}_{\mathcal{A}}^T \mathbf{x}_p > \mathbf{w}_{\mathcal{A}}^T \mathbf{x}_{p'}$. Since this problem is NP-hard, its approximate version is solved by introducing a large-margin regularizer for $\mathbf{w}_{\mathcal{A}}$ and slack variables for ordering constraints [11]. This leads to the following optimization problem:

$$\min_{\mathbf{w}_{\mathcal{A}}} \frac{1}{2}||\mathbf{w}_{\mathcal{A}}||_2^2 + C\sum \gamma_p^2 \quad s.t. \;\; \mathbf{w}_{\mathcal{A}}^T(\mathbf{x}_p - \mathbf{x}_{p'}) \geq 1 - \gamma_p, \gamma_p \geq 0, \;\; \forall X_p \in \mathcal{O}_{\mathcal{A}} \tag{5}$$

Here, $||\cdot||_2^2$ denotes squared l_2-norm, and the positive real constant C balances the trade-off between the two competing terms. The above optimization problem is solved in the primal form itself using Newton's method [3]. Once the ranking model $\mathbf{w}_{\mathcal{A}}$ is learned, prediction on a test pair X_t is done by evaluating $R_{\mathcal{A}}(\cdot)$ on both of its images. If $R_{\mathcal{A}}(\mathbf{x}_t) > R_{\mathcal{A}}(\mathbf{x}_{t'})$, then I_t is predicted to exhibit $\mathcal{A}$ more than $I_{t'}$, and less otherwise. It should be noted that the objective function adopted in [20] also considers "similar" image-pairs from $\mathcal{S}_{\mathcal{A}}$. However, as noted in [27], such pairs do not impact the performance much, and hence these are omitted.

4 Experiments

To examine the performance and behaviour of the proposed baselines, we compare with state-of-the-art relative attribute learning methods on challenging relative attribute datasets.

4.1 Datasets

We evaluate on three datasets: **UT-Zap50K** [27] (shoes), **LFW-10** [23] (faces) and **OSR** [20] (outdoor scene). UT-Zap50K dataset has two collections: UT-Zap50K-1 and UT-Zap50K-2, which are collections of *coarser* and *fine-grained* pairs respectively. For OSR dataset, we pick 1000 random pairs per attribute for training/testing from the train/test splits of [20]. For the other two datasets, we use the same train/test splits as in [27] and [23] respectively. Among these datasets, UT-Zap50K and LFW-10 are comparatively more challenging than OSR, as they contain a wide variety of images in terms of image categories (by category, we

	Smile	Young	Teeth	Mascu.	Eyes	Mouth	GdLook	FrHead	DkHair	Bald	Avg.
RelativeParts [23]	81.3	82.4	76.2	67.0	90.5	77.6	77.6	80.2	80.5	71.8	78.5
Global [20]	54.6	65.8	56.0	71.3	52.6	55.0	68.4	64.5	75.7	70.4	63.4
LocalPair [27]	59.7	66.2	53.5	70.1	49.6	53.4	64.7	65.6	73.6	67.9	62.4
LocalPair+ML [27]	57.9	68.4	52.9	73.0	51.9	52.7	65.1	69.1	72.7	72.6	63.6
Global+Hog	67.4	68.4	71.7	84.5	70.7	67.8	67.6	79.3	72.4	78.8	72.9
LocalPair+Hog	65.9	63.7	71.1	89.7	69.9	63.8	70.2	78.1	74.2	76.3	72.3
LocalPair+ML+Hog	67.0	64.9	69.0	90.9	65.2	65.2	73.5	80.9	74.7	77.0	72.8

Table 1: Accuracy comparison for the LFW-10 dataset. The best results (this work and those of the previous approaches) are underlined.

mean specific person in LFW-10 dataset, shoe-type in UT-Zap50K dataset, and scene category in OSR dataset). While the number of categories in OSR dataset is 8, it is few hundreds in the other two. Also, UT-Zap50K-2 is the most challenging dataset among these since it targets fine-grained comparisons, and thus emphasizes the importance of local learning [27] to capture subtle differences within image pairs.

4.2 Experimental Set-up

While the images in OSR dataset are of size 256×256 pixels, those in LFW-10 are 250×250 pixels. In UT-Zap50K dataset, the images are of varying size, and roughly around 100×150 pixels. We resize the images in LFW-10 dataset to 256×256 pixels, and those in UT-Zap50K dataset to 144×96 pixels. While computing the HOG descriptor, the patch size is empirically set to be 16×16 pixels for OSR and LFW-10 datasets, and 8×8 pixels for UT-Zap50K dataset. After computing the HOG descriptor for all the patches, these are concatenated and projected into a 1200-dimensional space using principal component analysis (PCA). In our preliminary experiments, we found this to improve computational efficiency by a significant amount, and efficacy by a small margin.

In all the experiments, we keep $C = 0.1$ (Eq. 5) similar to [20], initialize $\mathbf{M}_{\mathcal{A}} = \mathbf{I}$ (identity matrix), and the number of analogous pairs to be $K = 100$ as in [27]. For other parameters, default settings are used. Similar to previous approaches, the results are reported in terms of percentage of test pairs that are assigned correct orderings. It should be noted that since we use a fixed value of 'C' throughout, we get slightly reduced performance than what is reported in the respective papers [23, 27] (which tune it separately for every attribute in every dataset). However, since we are interested in relative comparison of performances of different methods, the absolute performance should not be a major concern. Moreover, this also makes the results easily reproducible.

4.3 Comparison

We compare with the following approaches: Global [20], RelativeParts [23], LocalPair [27], and LocalPair+ML [27] (local pair with metric learning). We compare with RelativeParts [23] only on LFW-10 dataset since this approach requires identifying specific facial regions (such as those around eyes, mouth, etc.) using [29] , and was shown to be applicable only for face images. Also, we do not compare with [18] since code is not available. Nevertheless, since [27] was shown to consistently outperform [18], we believe comparing

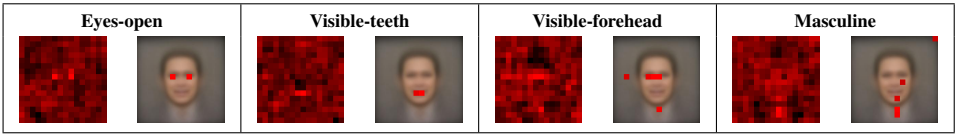

Figure 4: Learned HOG weights using Global+Hog baseline for four attributes from the LFW-10 datasets. Left: Normalized distribution of weights. Right: Top few largest weights overlaid on the average image of this dataset (best viewed in colour).

Dataset →	UT-Zap50K-1 (coarser) dataset					UT-Zap50K-2 (fine-grained) dataset				
Method ↓	**Open**	**Pointy**	**Sporty**	**Comf.**	**Avg.**	**Open**	**Pointy**	**Sporty**	**Comf.**	**Avg.**
Global [20]	87.3	87.5	89.5	88.1	88.1	60.8	59.1	62.8	64.0	61.7
LocalPair [27]	86.8	86.4	88.9	88.2	87.6	71.6	59.6	61.1	60.2	63.1
LocalPair+ML [27]	87.4	87.8	89.9	89.5	88.7	72.3	62.6	62.5	61.5	64.7
Global+Hog	90.1	89.9	90.3	90.4	90.2	70.9	63.8	62.4	66.2	65.8
LocalPair+Hog	88.8	88.1	88.8	89.6	88.8	76.4	64.8	63.0	63.9	67.0
LocalPair+ML+Hog	89.9	88.3	91.5	90.5	90.0	76.2	65.3	64.8	63.6	67.5

Table 2: Accuracy comparison for the UT-Zap50K-1 (coarser) dataset (left) and UT-Zap50K-2 (fine-grained) dataset (right). The best results (this work and those of the previous approaches) are underlined.

with [27] throughout will fill the comparison gap.

While computing performance of the above approaches, we use the same features as theirs. For OSR dataset, we use GIST [19] descriptor. For UT-Zap50K dataset, we use GIST descriptor concatenated with colour histogram. For LFW-10 dataset, we use the (8300-dimensional) part-based representation proposed in [23] to evaluate their method, and GIST descriptor to evaluate [20, 27]. This is because it is computationally prohibitive to work with the representation of [23] while learning local models [27], and using this representation with [20] is actually the approach of [23] (though part-weights were also learned in [23], we omit them as they had little impact on the performance).

4.4 Results and Discussion

Table 1 compares the accuracy on LFW-10 dataset. Compared to [20, 27], the approach of [23] performs significantly better. Recall that the same procedure is used for learning the ranking models in both [20, 23]. Hence, these performance gains should be attributed to the part-based representation of [23]. The results obtained using the baseline methods further validate this claim. This is because the concatenated HOG descriptor can be thought of as a naïve version of the part-based representation of [23]. The simplest Global+HOG baseline achieves an improvement of 9.5% in the average accuracy compared to Global [20], with maximum ($13-18\%$) gains on the attributes "visible-teeth", "eyes-open", "masculine" and "visible-forehead". To further validate the performance gains achieved by Global+HOG, we try to visualize what a global ranking model learns using the HOG descriptor (without dimensionality reduction). Since all the bins in the HOG descriptor have non-negative values, the aggregate weight of the ranking model in the interval corresponding to each cell can be thought of as a measure of confidence for identifying the relative importance of cells as

Method	Natural	Open	Perspective	LargeSize	Diagonal	Depth	Avg.
Global [20]	94.3	89.7	83.8	86.7	84.8	88.7	88.0
LocalPair [24]	92.3	88.4	82.5	85.0	77.3	88.7	85.7
LocalPair+ML [24]	95.1	91.3	86.2	88.8	89.0	90.8	90.2
Global+Hog	95.1	88.4	82.7	83.4	84.2	87.1	86.8
LocalPair+Hog	96.4	89.8	81.4	86.2	85.4	88.3	87.9
LocalPair+ML+Hog	96.2	90.0	82.3	86.9	87.8	88.5	88.6

Table 3: Accuracy comparison for the OSR dataset. The best results (this work and those of the previous approaches) are underlined.

	RelativeParts [13]	Global [20]	LocalPair [13]	LocalPair+ML [13]	Global+Hog	LocalPair+Hog	LocalPair+ML+Hog
LFW-10	78.5	63.4	62.4	63.6	72.9	72.3	72.8
OSR	–	88.0	85.7	90.2	86.8	87.9	88.6
UTZ-1	–	88.1	87.6	88.7	90.2	88.8	90.0
UTZ-2	–	61.7	63.1	64.7	65.8	67.0	67.5

Table 4: Average accuracy comparison for all the three datasets. The best results (this work and those of the previous approaches) are underlined.

learned by the model. In Figure 4, we show these weights for the four attributes mentioned above. Collectively visualizing all the weights does not provide much insight. However, when the top few cells with maximum aggregate weights are overlaid on the average image of this dataset, the performance gains seem justifiable. For "eyes-open" and "visible-teeth", the top two cells surprisingly fall at almost the right place. For "visible-forehead", these cells mostly focus on the upper-half of the image. The model for "masculine" learns that regions near cheeks and neck are indicative of this attribute. A possible reason can be that people with prominent cheek-bones and facial hair might have been considered more masculine than others by the annotators of LFW-10 dataset.

In Table 2, we compare the accuracy on the two collections of UT-Zap50K dataset. Interestingly, the simplest baseline Global+Hog itself outperforms the current state-of-the-art approach of [27] on both coarser as well as fine-grained collections. For UT-Zap50K-2 dataset, further improvement of around 2% is achieved by using local learning approach. However, for UT-Zap50K-1 dataset, the performance of both global and local learning approaches are comparable. This attributes to the inherent characteristic of this dataset, in which differences within pairs are coarse (or more spread-out). Table 3 compares the results on OSR dataset. Unlike the other two datasets, here the baselines perform slightly inferior than their counterparts that use GIST descriptor. This is because the images in OSR dataset are not as aligned as those in the other two datasets.

Table 4 summarizes the performance of all the methods on all the datasets. These results suggest that along with the learning algorithm, choosing the right representation also plays a crucial role in the visual comparison task, and it is possible to achieve significant performance gains even by employing a simple but more appropriate representation. Domain knowledge can also prove to be useful in designing/selecting the representation and learning algorithm, as observed in the case of LFW-10 and UT-Zap50K datasets.

Along with accuracy, an important concern while learning visual attributes is to avoid learning the wrong things; i.e., the properties that are (positively/negatively) re-

lated with a particular attribute but are not its characteristics as such. E.g., while learning the attribute "open" using more of the coastal images, the model may learn "blue" or some other correlated property instead. Hence the learned models are expected to be decorrelated with each other while also being discriminative. This is necessary because several applications are based on using attribute models' response as the mid-level representation. To analyze this, we compare the average absolute correlation among the ranking models for all the datasets using Global and Global+Hog in Figure 5. We can observe that the correlation among the models of Global+Hog is consistently less than that of Global, and thus the former are more decorrelated than the latter. This indicates that accounting for local structure in the representation plays a key role in learning decorrelated ranking models. Note that a recent work [10] also arrived at similar conclusion, though for the attribute classification task and not for the comparison task.

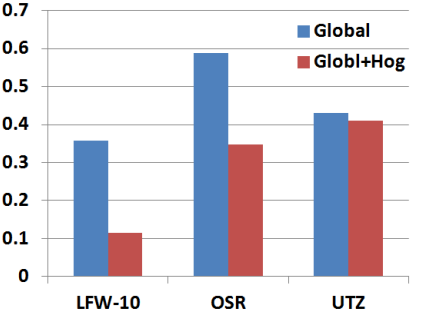

Figure 5: Average absolute correlation among the ranking models learned using Global and Global+Hog approaches.

5 Conclusion

The problem of relative attribute learning involves several other sub-problems such as efficient learning from paired samples, learning with weak supervision and domain understanding. However, all these topics are of intense research, and one would require to solve these first to be able to solve the attribute comparison problem at the human level. In this work, our goal was to develop intuitively simple baselines rather than to create a new method for learning relative attributes. Our baselines combine the existing methods with a simple yet locally discriminative patch-based representation, that tries to encode local shape as well as spatial information. Experiments demonstrated that comparing existing techniques for learning relative attributes with the proposed baselines helps us better appreciate the importance of the modeling steps involved in the existing techniques. As evident from the general performance level of the proposed baselines as well as existing methods, there is a lot of scope for improvement, especially on the challenging UT-Zap50K-2 dataset.

Acknowledgement

Yashaswi Verma is partially supported by Microsoft Research India PhD fellowship 2013.

References

[1] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*, 2013.

[2] C Boutilier. Preference elicitation and preference learning in social choice. In *CPAIOR*, 2011.

[3] Olivier Chapelle. Training a support vector machine in the primal. *Neural Comput.*, 2007.

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[6] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[7] Ali Farhadi, Ian Endres, and Derek Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[9] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *PAMI*, 1996.

[10] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014.

[11] Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2002.

[12] A. Kovashka and K. Grauman. Attribute adaptation for personalized image search. In *ICCV*, 2013.

[13] A. Kovashka and K. Grauman. Attribute pivots for guiding relevance feedback in image search. In *ICCV*, 2013.

[14] A. Kovashka, D. Parikh, and K. Grauman. WhittleSearch: Image search with relative attribute feedback. In *CVPR*, 2012.

[15] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attributes and simile classifiers for face verification. In *ICCV*, 2009.

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[17] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[18] S. Li, S. Shan, and X. Chen. Relative forest for attribute prediction. In *ACCV*, 2012.

[19] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[20] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[21] A. Parkash and D. Parikh. Attributes for classifier feedback. In *ECCV*, 2012.

[22] D. Reid and M. Nixon. Using comparative human descriptions for soft biometrics. In *IJCB*, 2011.

[23] Ramachandruni N. Sandeep, Yashaswi Verma, and C. V. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *CVPR*, 2014.

[24] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011. doi: 10.1109/CVPR.2011.5995329.

[25] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[26] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, pages 155–168, 2010. doi: 10.1007/978-3-642-15555-0_12.

[27] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014.

[28] H. Zhang, A.C. Berg, M. Maire, and J. Malik. SVM-KNN:Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.

[29] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.