Relative Parts: Distinctive Parts for Learning Relative Attributes

Ramachandruni N. Sandeep Yashaswi Verma C. V. Jawahar

Center for Visual Information Technology, IIIT Hyderabad, India - 500032

Abstract

The notion of relative attributes as introduced by Parikh and Grauman (ICCV, 2011) provides an appealing way of comparing two images based on their visual properties (or attributes) such as "smiling" for face images, "naturalness" for outdoor images, etc. For learning such attributes, a Ranking SVM based formulation was proposed that uses globally represented pairs of annotated images. In this paper, we extend this idea towards learning relative attributes using local parts that are shared across categories. First, instead of using a global representation, we introduce a part-based representation combining a pair of images that specifically compares corresponding parts. Then, with each part we associate a locally adaptive "significance-coefficient" that represents its discriminative ability with respect to a particular attribute. For each attribute, the significance-coefficients are learned simultaneously with a max-margin ranking model in an iterative manner. Compared to the baseline method, the new method is shown to achieve significant improvement in relative attribute prediction accuracy. Additionally, it is also shown to improve relative feedback based interactive image search.

1. Introduction

Visual attributes (or simply attributes) are perceptual properties that can be used to describe an entity ("*pointed* nose"), an object ("*furry* sheep"), or a scene ("*natural* outdoor"). These act as mid-level representations that are comprehensible for both human as well as machine, thus providing a strong means of filling-up the so-called semantic-gap.

Attributes have recently been used as a source of semantic cues in diverse tasks such as object recognition [17, 18], image description [24], learning unseen object categories (or zero-shot learning) [18], etc. While most of these works have focused on binary attributes (indicating presence or absence of some visual property), Parikh and Grauman [24] proposed that it is more natural to consider the *strength* of an attribute rather than its absolute presence/absence. This led to the notion of *"relative attributes"*, where the strength of an attribute in a given image can be described with respect to some other image/category; *e.g.* "given face is less chubby than person A and more chubby than person B". In [24], given a set of pairs of images depicting similar and/or different strengths of some particular attribute, the problem of learning a relative attribute classifier is posed as one of learning a ranking model for that attribute similar to Ranking SVM [12].

In this work, we build upon this idea by learning relative attribute models using local parts that are shared across categories. First, we propose a part-based representation that jointly represents a pair of images. A part corresponds to a block around a landmark point detected using a domain-specific method. This representation explicitly encodes correspondences among parts, thus better capturing minute differences in parts that make an attribute more prominent in one image than another, as compared to a global representation as in [24]. Next, we update this part-based representation by additionally learning weights corresponding to each part that denote their contribution towards predicting the strength of a given attribute. We call these weights as *"significance-coefficients"* of parts. For each attribute, the significance-coefficients are learned in a discriminative manner simultaneously with a max-margin ranking model. Thus, the best parts for predicting the relative attribute "more smiling" will be different from those for predicting "more eyes-open". The steps of the proposed method are illustrated in Figure 1. While the notion of parts is not new, we believe that ours is the first attempt that explores the applicability of parts in a ranking scenario, and for learning relative attribute ranking models in particular.

We compare the baseline method of [24] with the proposed method under various settings. For this, we have collected a new dataset of 10000 pairwise attribute-level annotations using images from the "Labeled Faces in the Wild" (LFW) dataset [11], particularly focusing on (i) large variety among samples in terms of poses, lighting condition, etc., and (ii) completely ignoring the category information while collecting attribute annotations. Extensive experiments demonstrate that the new method significantly improves the prediction accuracy as compared to the baseline

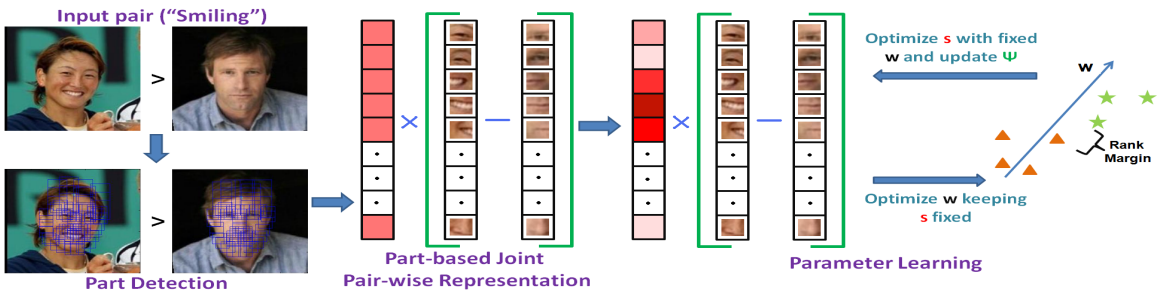

Figure 1. Given ordered pair of images, first we detect parts corresponding to different (facial) landmarks. Using these, a joint pairwise part-based representation is formed that encodes (i) correspondence among different parts, & (ii) relative importance of each part for a given attribute. Using this, a max-margin ranking model $\mathbf{w}$ is learned simultaneously with part weights $\mathbf{s}$ (red blocks) in an iterative manner.

method. Moreover, the learned parts also compare favorably with human selected parts, thus indicating the intrinsic capacity of the proposed framework for learning attribute-specific semantic parts.

The paper is organized as follows. In Sec. 2, we give an overview of some of the recent works based on attributes and relative attributes. In Sec. 3, we discuss the method of [24] for learning relative attribute ranking models. Then we present the new part-based representations in Sec. 4, followed by an algorithm for learning model variables in Sec. 5. Experiments and results are discussed in Sec. 6, and finally we conclude in Sec. 7.

2. Related Works

As discussed earlier, attributes are properties that are understandable by both human as well as machine. Because of this, attributes have recently gained significant popularity among several vision applications, where attribute identification is not the final goal but just an intermediate step. In [8], objects are described using their attributes; *e.g.* instead of classifying an image as that of a "sheep", it is described based on its properties such as "has horn", "has wool", etc. This helps in describing even those objects which have few or no examples during training phase. Similar idea is used in [7, 18] where attribute-based feedback is used for unseen category recognition. Attribute-based feedback has been shown to be useful for anomaly detection [28] within an object category, and adding unlabeled samples for category classfier learning [4]. Attributes have also been used for multiple-query image search [30], where input attributes along with other related attributes are used in a structured-prediction based model. Along with various applications, attributes have been used in several mid-level tasks. These include identification of color/texture [10], specific objects such as faces [17], and general object categories [18, 33]. In some cases, since it might not be possible to learn discriminative attributes from individual images, in [21], pairs of images are used to learn such attributes based on human feedback.

While most of the above methods have focused on presence/absence of some attribute, in [24] the notion of relative attributes was introduced. In this, two images are compared based on the relative strength of some given attribute, thus providing a semantically richer way of describing the visual world than using binary attributes. Since then, relative attributes have been used in several applications, such as customized image search [15, 16], where a user can interactively describe and refine visual properties while searching for some specific object. This has been further extended in recent works [14, 25]. In [14], generic attribute models are learned that can adapt to different users' preferences. In [25], novel features are introduced based on user's implied feedback, which subsequently help in improving search performance. In [26], an active learning framework based on relative attribute feedback is proposed. Here, the teacher (human) not only corrects an incorrect prediction made by learner (machine), but also tells why the prediction is incorrect using attribute based feedback. This helps the learner in propagating this understanding among other examples, which subsequently improves the learning process. This idea is extended in [2] where the learner learns attribute classifiers along with category classifiers. In [29], a semi-supervised constrained bootstrapping approach is proposed that tries to benefit from inter-class attribute-based relationships to avoid semantic drift during the learning process. In [32], a novel framework for predicting relative dominance among attributes within an image is proposed. In [27], rather than using either binary or relative attributes, their interactions are modeled to better describe images.

Our work closely relates with recent works [1, 5, 6, 13] that use distinctive part/region-based representations for scene classification [13] or fine-grained classification [1, 5, 6]. However, rather than identifying category-specific distinctive parts, our aim is to compare similar parts that are shared across categories. This makes our problem somewhat more challenging, since our representation is expected to capture small relative differences in the appearance of se-

mantically similar parts, which contribute in making some attribute prominent in one image than another.

3. Preliminaries

In [24], a Ranking SVM based method was used for learning relative attribute classifiers. Ranking SVM [12] is a max-margin ranking framework that learns linear models to perform pairwise comparisons. This is conceptually different from the conventional one-vs-rest SVM that learns a model using individual samples rather than pairs. Though SVM scores can also be used to perform pairwise comparisons, usually Ranking SVM has been known to perform better than SVM for such tasks. In [24] also, Ranking SVM was shown to perform better than SVM on the task of relative attribute prediction. We now briefly discuss the method used in [24] for learning relative attribute classifiers.

3.1. The Ranking SVM Model

Let $\mathcal{I} = \{I_1, \ldots, I_n\}$ be a collection of n images. Each image I_i is represented by a global feature vector $\mathbf{x}_i \in \mathcal{R}^N$. Suppose we have a fixed set of attributes $A = \{a_m\}$. For each attribute $a_m \in A$, we are given a set $\mathcal{D}_m = \mathcal{O}_m \cup \mathcal{S}_m$ consisting of ordered pairs of images. Here, $\mathcal{O}_m = \{(I_i, I_j)\}$ is such that image I_i has more strength of attribute a_m than image I_j. And, $\mathcal{S}_m = \{(I_i, I_j)\}$ is such that both I_i and I_j have nearly the same strength of attribute a_m. Using $\mathcal{D}_m$, the goal is to learn a ranking function f_m that, given a new pair of images I_p and I_q represented by $\mathbf{x}_p$ and $\mathbf{x}_q$ respectively, predicts which image has greater strength of attribute a_m. Under the assumption that f_m is a linear function of $\mathbf{x}_p$ and $\mathbf{x}_q$, it is defined as:

$$f_m(\mathbf{x}_p, \mathbf{x}_q; \mathbf{w}_m) = \mathbf{w}_m \cdot \Psi(\mathbf{x}_p, \mathbf{x}_q), \quad (1)$$

$$\Psi(\mathbf{x}_p, \mathbf{x}_q) = \mathbf{x}_p - \mathbf{x}_q \quad (2)$$

Here, $\mathbf{w}_m$ is the parameter vector for attribute a_m, and $\Psi(\mathbf{x}_p, \mathbf{x}_q)$ is a joint representation formed using $\mathbf{x}_p$ and $\mathbf{x}_q$. Using f_m, we determine which image has higher strength for attribute a_m based on $y^m_{pq} = sign(f_m(\mathbf{x}_p, \mathbf{x}_q; \mathbf{w}_m))$. $y^m_{pq} = 1$ means I_p has higher strength of a_m than I_q, and $y^m_{pq} = -1$ means otherwise. In order to learn $\mathbf{w}_m$, following constraints need to be satisfied:

$$\mathbf{w}_m \cdot \Psi(\mathbf{x}_i, \mathbf{x}_j) > 0 \;\; \forall (I_i, I_j) \in \mathcal{O}_m \quad (3)$$

$$\mathbf{w}_m \cdot \Psi(\mathbf{x}_i, \mathbf{x}_j) = 0 \;\; \forall (I_i, I_j) \in \mathcal{S}_m \quad (4)$$

Since this is an NP-hard problem, its relaxed version is solved by introducing slack variables. This leads to the following optimization problem ($OP1$):

$$OP1: \min_{\mathbf{w}_m} \; \frac{1}{2}||\mathbf{w}_m||_2^2 + C_m(\sum \xi_{ij}^2 + \sum \alpha_{ij}^2) \quad (5)$$

$$s.t. \; \mathbf{w}_m \cdot \Psi(\mathbf{x}_i, \mathbf{x}_j) \geq 1 - \xi_{ij}, \; \forall (I_i, I_j) \in \mathcal{O}_m \quad (6)$$

$$||\mathbf{w}_m \cdot \Psi(\mathbf{x}_i, \mathbf{x}_j)||_1 \leq \alpha_{ij}, \; \forall (I_i, I_j) \in \mathcal{S}_m \quad (7)$$

$$\xi_{ij} \geq 0; \quad \alpha_{ij} \geq 0. \quad (8)$$

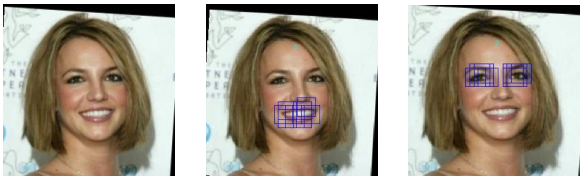

Figure 2. Given an input image (left), the parts that correspond to "visible-teeth" (middle) and "eyes-open" (right).

Here, $||\cdot||_2^2$ denotes squared L_2 norm, $||\cdot||_1$ denotes L_1 norm, and $C_m > 0$ is a constant that takes care of the trade-off between regularization term and loss term. Note that along with pairwise constraints as in [12], the optimization problem now also includes similarity constraints. This is solved in the primal form itself using Newton's method [3].

4. Proposed Representations

The Ranking SVM method discussed above uses a joint representation based on globally computed features (Eq. 2) while determining the strength of some given attribute. However, several attributes such as "visible-teeth", "eyes-open", etc. are not representative of whole image, and correspond to only some specific regions/parts. This means there exists a weak association between an image and its attribute label. E.g., Figure 2 shows the parts corresponding to attributes "visible-teeth" and "eyes-open". This inspires us to build a representation that (i) encodes part/region-specific features, without confusing across parts; and (ii) explicitly encodes the relative significance of each part with respect to a given attribute. With this motivation, next we propose two part-based joint-representations for the task of learning relative attribute classifiers.

4.1. Part-based Joint Representation

Given an image I, let $\mathcal{P} = \{p^1, \ldots, p^K\}$ be the set of its K parts. These parts can be obtained using a domain-specific method; *e.g.*, the method discussed in [35] can be used for determining a set of localized parts in face images. Each part $p^k, \forall k \in \{1, \ldots, K\}$ is represented using an N_1-dimensional feature vector $\tilde{\mathbf{x}}^k \in \mathcal{R}^{N_1}$. Here, $N_1 = K \times d_1$ such that each $\tilde{\mathbf{x}}^k$ is a sparse vector with only d_1 non-zero entries in the k^{th} interval representing part p^k. Based on this, given a pair of images I_p and I_q, we define a joint part-based feature representation as below:

$$\tilde{\Psi}(\tilde{\mathbf{x}}_p, \tilde{\mathbf{x}}_q) = \sum_{k=1}^{K} (\tilde{\mathbf{x}}_p^k - \tilde{\mathbf{x}}_q^k), \quad (9)$$

where $\tilde{\mathbf{x}}_p = \{\tilde{\mathbf{x}}_p^k \mid \forall k \in \{1, \ldots, K\}\}$. The advantage of this representation is that it specifically encodes correspondence among parts; i.e., now the k^{th} part of I_p is compared with just the k^{th} part of I_q. The assumption here is that such

a direct comparison between localized pairs of parts would provide stronger cues for learning relative attribute models than using a single global representation as in Eq. 2. (This assumption is also validated by improvements in prediction accuracy as discussed in Sec. 6.)

4.2. Weighted Part-based Joint Representation

Though the joint representation proposed in the previous section allows direct part-based comparison between a pair of images, it does not provide information about which parts actually symbolize some given attribute. This is particularly desirable in case of local attributes, where only a few parts are important in predicting attribute strength. With this motivation, we update the joint representation of Eq. 9 to precisely encode relative importance of parts.

As discussed in Sec. 4.1, let each image I be represented by a set of K parts. Additionally, let $s_m^k \in [0, 1]$ be a weight associated with the k^{th} part. This weight denotes the relative importance of the k^{th} part compared to other parts for predicting the strength of attribute a_m; i.e., larger the weight, more important is that part, and vice-versa. Using this, given a pair of images I_p and I_q, the new weighted part-based joint feature representation is defined as:

$$\tilde{\Psi}_s(\tilde{\mathbf{x}}_p, \tilde{\mathbf{x}}_q, \mathbf{s}_m) = \sum_{k=1}^{K} s_m^k(\tilde{\mathbf{x}}_p^k - \tilde{\mathbf{x}}_q^k), \tag{10}$$

where $\mathbf{s}_m = [s_m^1, \ldots, s_m^K]^T$. Since s_m^k expresses the relative significance of the k^{th} part with respect to a_m, we call it as the significance-coefficient of the k^{th} part. These help in explicitly encoding the relative importance of individual parts in the joint representation.

5. Parameter Learning

Now we discuss how to learn the parameters for each attribute using the two joint representations discussed above. Note that we still need to satisfy the constraints as in Eq. 3 and Eq. 4 depending upon the representation followed.

5.1. For Part-based Joint Representation

In order to learn a ranking model based on the part-based representation in Eq. 9, we optimize the following problem:

$$OP2: \min_{\mathbf{w}_m} \ \frac{1}{2}||\mathbf{w}_m||_2^2 + C_m(\sum \xi_{ij}^2 + \sum \alpha_{ij}^2) \tag{11}$$

$$s.t.\ \mathbf{w}_m \cdot \tilde{\Psi}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \geq 1 - \xi_{ij},\ \forall (I_i, I_j) \in \mathcal{O}_m \tag{12}$$

$$||\mathbf{w}_m \cdot \tilde{\Psi}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)||_1 \leq \alpha_{ij},\ \forall (I_i, I_j) \in \mathcal{S}_m \tag{13}$$

$$\xi_{ij} \geq 0; \quad \alpha_{ij} \geq 0. \tag{14}$$

This is similar to $OP1$, except that now we use part-based representation instead of global representation. This allows us to use the same Newton's method [3] for solving $OP2$.

5.2. For Weighted Part-based Joint Representation

For the weighted part-based joint representation in Eq. 10, we need to learn two sets of parameters corresponding to every attribute: ranking model $\mathbf{w}_m$, and significance-coefficients $\mathbf{s}_m$. To do this, we solve the following optimization problem ($OP3$):

$$OP3: \min_{\mathbf{w}_m, \mathbf{s}_m} \ \frac{1}{2}||\mathbf{w}_m||_2^2 + C_m(\sum \xi_{ij}^2 + \sum \alpha_{ij}^2) \tag{15}$$

$$s.t.\ \mathbf{w}_m \cdot \tilde{\Psi}_s(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j, \mathbf{s}_m) \geq 1 - \xi_{ij},\ \forall (I_i, I_j) \in \mathcal{O}_m \tag{16}$$

$$||\mathbf{w}_m \cdot \tilde{\Psi}_s(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j, \mathbf{s}_m)||_1 \leq \alpha_{ij},\ \forall (I_i, I_j) \in \mathcal{S}_m \tag{17}$$

$$\xi_{ij} \geq 0; \quad \alpha_{ij} \geq 0; \tag{18}$$

$$s_m^k \geq 0, \ \forall 1 \leq k \leq K; \quad \mathbf{e} \cdot \mathbf{s}_m = 1. \tag{19}$$

where $\mathbf{e} = [1, \ldots, 1]^T$ is a constant vector with all entries equal to 1. Note that the overall weight of all the parts is constrained to be unit; i.e., $s_m^k \geq 0$, $\mathbf{e} \cdot \mathbf{s}_m = 1$, which ensures that all parts are fairly used. This is equivalent to constraining the L_1-norm of $\mathbf{s}_m$ to be 1 (i.e., L_1-regularization), thus implicitly imposing sparsity on $\mathbf{s}_m$ [22, 31]. This is desirable since usually only a few parts contribute towards determining the strength of a given attribute.

5.2.1 Solving the optimization problem

We solve $OP3$ in the primal form itself using a block coordinate descent algorithm. We consider each set of parameters $\mathbf{w}_m$ and $\mathbf{s}_m$ as two blocks, and optimize them in an alternate manner. In the beginning, we initialize all entries of $\mathbf{w}_m$ to be zero, and all entries of $\mathbf{s}_m$ to be equal to $1/K$.

First we fix $\mathbf{s}_m$ to optimize $\mathbf{w}_m$. For a fixed $\mathbf{s}_m$, the problem becomes equivalent to $OP2$ (Eq. 11 to 14), and hence can be solved in the same manner using [3].

Then we fix $\mathbf{w}_m$ to optimize $\mathbf{s}_m$. Let $\tilde{\mathbf{X}}_i = [\tilde{\mathbf{x}}_i^1 \ldots \tilde{\mathbf{x}}_i^K] \in \mathcal{R}^{N_1 \times K}$ be a matrix formed by appending features corresponding to all parts of image I_i. Using this, we compute $\tilde{\mathbf{z}}_{im} = \tilde{\mathbf{X}}_i^T \mathbf{w}_m \in \mathcal{R}^K$. This gives

$$\mathbf{w}_m \cdot \tilde{\Psi}_s(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j, \mathbf{s}_m) = \mathbf{s}_m \cdot \tilde{\mathbf{z}}_{ijm}, \tag{20}$$

$$\tilde{\mathbf{z}}_{ijm} = \tilde{\mathbf{z}}_{im} - \tilde{\mathbf{z}}_{jm}. \tag{21}$$

Substituting this in $OP3$ leads to the following optimization problem for learning $\mathbf{s}_m$ (for fixed $\mathbf{w}_m$):

$$OP4: \min_{\mathbf{s}_m} \ C\ \Big(\sum_{(I_i, I_j) \in \mathcal{Q}_m} (1 - \mathbf{s}_m \cdot \tilde{\mathbf{z}}_{ijm})^2 \ + \sum_{(I_i, I_j) \in \mathcal{S}_m} ||\mathbf{s}_m \cdot \tilde{\mathbf{z}}_{ijm}||_1^2 \ \Big) \tag{22}$$

$$s.t. \quad s_m^k \geq 0, \ \forall 1 \leq k \leq K; \quad \mathbf{e} \cdot \mathbf{s}_m = 1. \tag{23}$$

where $\mathcal{Q}_m \subseteq \mathcal{O}_m$ is the set of pairs that violate the margin constraint. Note that $\mathcal{Q}_m$ is not fixed, and may change at every iteration. We solve $OP4$ using an iterative gradient descent and projection method similar to [34].

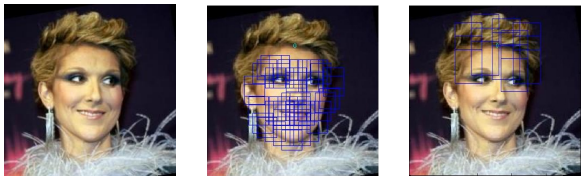

Figure 3. Input image (left), parts detected using [35] (middle), and additional parts detected by us (right).

5.3. Computing Parts

The two joint representations as proposed in Sec. 4 are based on an ordered set of corresponding parts computed from a given pair of images. Given a method for computing such parts, our framework is applicable irrespective of the domain. This makes our framework domain adaptable.

In this work, we consider the domain of face images. To compute parts from a given face image, we use the method proposed in [35]. It is based on a mixture-of-tress model to learn a shared pool of facial parts. Given a face image, it computes a set of 68 parts covering facial landmarks such as eyes, eyebrows, nose, mouth and jawline. Figure 3 shows a face image (left) and its parts (middle) computed using this method. Though these parts can be used to represent several attributes such as "smiling", "eyes-open", etc., there are few other attributes which are not covered by these parts such as "bald-head", "visible-forehead" and "dark-hair". In order to cover these attributes as well, we compute additional parts using image-level statistics such as image-size and distance from the earlier 68 parts. This gives an extended set of 83 parts for a given face image. Figure 3 (right) shows this extended set of parts computed for the given image (left).

5.4. Relation with Latent Models

In the last few years, latent models have become popular for several tasks, particularly for object detection [9]. These models usually look for characteristics (e.g., parts) that are shared within a category but distinctive across categories. (As discussed in Sec. 2, recent works such as [1, 5, 13] also have similar motivation, though they do not explicitly investigate the latent aspect.) Our work is similar to theirs in the sense that we also seek attribute-specific distinctive parts by incorporating significance-coefficients. However, in contrary to them, we require these parts to be shared across categories. This is because our ranking method uses these parts to learn attribute-specific models which are independent of categories being depicted in training pairs.

6. Experiments

We compare the proposed method with that of [24] under different settings on two datasets. First is the PubFig-29 dataset as used in [26]. It consists of 60 face categories and 29 attributes, with attribute annotations being collected

Figure 4. Example pairs and their ground-truth annotations from Pubfig-29 dataset. Due to category-level annotations, there exist inconsistencies in (true) instance-level attribute visibility.

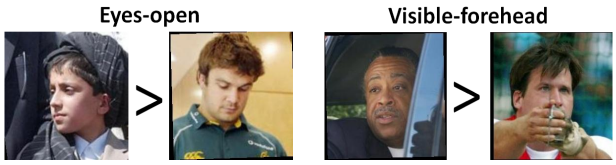

Figure 5. Example pairs from LFW-10 dataset. The images exhibit high diversity in terms of age, pose, lighting, occulusion, etc.

at category-level; i.e., using pairs of categories rather than pairs of images. Due to this, the annotations in this dataset are not consistent for several attributes (see Figure 4) ; e.g., Scarlett Johansson may not be smiling more than Hugh Laurie in all their images. To address this limitation, we have collected a new dataset using a subset of LFW [11] images. The new dataset has attribute-level annotations for 10000 image pairs and 10 attributes, and we call this as LFW-10 dataset. While collecting the annotations, we particularly ignore the category information, thus making it more suitable for the task of learning relative attributes. The details of this dataset are described next.

6.1. LFW-10 Dataset

We randomly select 2000 images from LFW dataset [11]. Out of these, 1000 images are used for creating training pairs and the remaining (unseen) 1000 for testing pairs. The annotations are collected for 10 attributes, with 500 training and testing pairs per attribute. In order to minimize the chances of inconsistency in the dataset, each image pair is got annotated from 5 trained annotators, and final annotation is decided based on majority voting. Figure 5 shows example pairs from this dataset.

6.2. Features for Parts

We represent each part using a Bag of Words (BoW) histogram over dense SIFT (DSIFT) [20] features. We consider two settings for learning visual-word vocabulary: (1) In the first setting, we learn a part-specific vocabulary for every part. This is possible since our parts are fixed and known. In practice, we learn a vocabulary of 100 visual words for each part. This gives a 8300-dimensional ($= 83$ parts $\times 100$) (sparse) feature vector per part. (2) In the second setting, we learn a single vocabulary of 100 visual words for all the parts. This again results into a 8300-

Method	Accuracy
Global DSIFT + RSVM [24]	61.28
Global GIST + RGB + RSVM [24]	59.18
SPM (Upto 2 levels) + RSVM [24]	49.60
SPM (Upto 3 levels) + RSVM [24]	49.17
Unweighted parts + Part-specific vocab. (Ours)	62.54
Unweighted parts + Single vocab. (Ours)	62.83
Learned parts + Part-specific vocab. (Ours)	62.67
Learned parts + Single vocab. (Ours)	**63.08**

Table 1. Results on PubFig-29 dataset. Though all the methods give comparable performance, these results are not really indicative of their actual behaviour since the annotations in this dataset are at category-level rather than instance-level.

dimensional (=83 parts $\times$100) feature vector for each part.

6.3. Baselines

We compare with the Ranking SVM method of [24] using the code provided by the authors [1]. We use four features for comparison: (i) BoW histogram over DSIFT features with 1000 visual words, (ii) global 512-dimensional GIST descriptor [23], (iii) global 512-dimensional GIST and 30-dimensional RGB histogram (which was also used in [24]), and (iv) spatial pyramid (SPM) [19] upto two and three levels using DSIFT features and the same vocabulary as in (i).

As another baseline, we compare the quality of our part-learning framework (Sec. 5.2) against human selected parts. For this, we asked a human expert to select a subset of few most representative parts corresponding to every attribute. For a given attribute a_m, all the selected parts are assigned equal weights and the remaining parts are assigned zero weight, and then a ranking model $\mathbf{w}_m$ is learned based on these part weights. The intuition behind this experiment is to analyze the trade-off between the performance obtained using manually selected parts and learned parts.

6.4. Results

Table 1 compares different methods on PubFig-29 dataset. For each attribute, we consider 1500 training pairs from 40 classes, and 1500 testing pairs from the remaining 20 classes. As discussed before, since this dataset has category-level annotations, there exist inconsistencies in instance-level annotations. Due to this, average accuracies of different methods are quite close. Hence, we believe that LFW-10 dataset is more suitable for comparisons.

Table 2 shows the average accuracies over all the attributes obtained by different methods on LFW-10 dataset. Several observations can be made from these results: (1)

[1]The code is available at `https://filebox.ece.vt.edu/~parikh/relative.html`.

Figure 6. For three attributes from LFW-10 dataset ("smiling", "visible-forehead" & "eyes-open" resp.) the first block shows the top five parts and their weights learned using our method, and the second block shows five parts selected by human expert.

Method	Accuracy
Global DSIFT + RSVM [24]	64.61
Global GIST + RSVM [24]	68.89
Global GIST + RGB + RSVM [24]	69.89
SPM (Upto 2 levels) + RSVM [24]	50.73
SPM (Upto 3 levels) + RSVM [24]	50.01
Human selected parts + Part-specific Vocab. (Ours)	80.90
Human selected parts + Single Vocab. (Ours)	80.43
Unweighted parts + Part-specific vocab. (Ours)	80.49
Unweighted parts + Single vocab. (Ours)	80.19
Learned parts + Part-specific vocab. (Ours)	**81.06**
Learned parts + Single vocab. (Ours)	80.71

Table 2. Average relative attribute prediction accuracies using different methods on LFW-10 dataset.

The performance for SPM is comparable to chance accuracy. This is probably because the blocks are big enough to capture minute differences in small parts for learning attributes. This results in learning bigger parts that are not really distinctive with respect to different attributes. (2) The part-based representations always performs significantly better (more than 10% on absolute scale) than [24] with different features. This clearly validates the significance of these representations for learning relative attribute models. (3) Using part-specific vocabulary performs better than single vocabulary. One possible reason for this could be that using vocabularies learned individually for each part results into less confusion than using a single vocabulary learned using all the parts. Investigating the effect of vocabulary size for these two settings could be an interesting

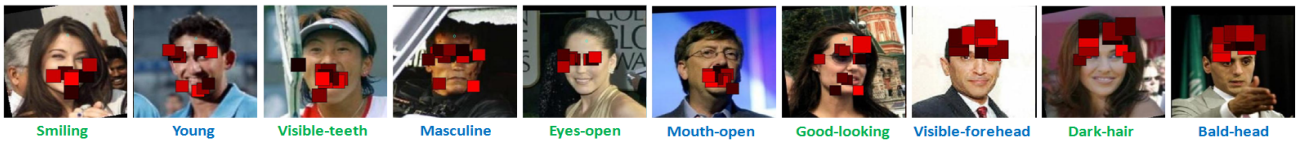

Figure 7. Top 10 parts learned using our method with maximum weights for each of the ten attributes in LFW-10 dataset. Greater is the intensity of red, more important is that part, and vice-versa.

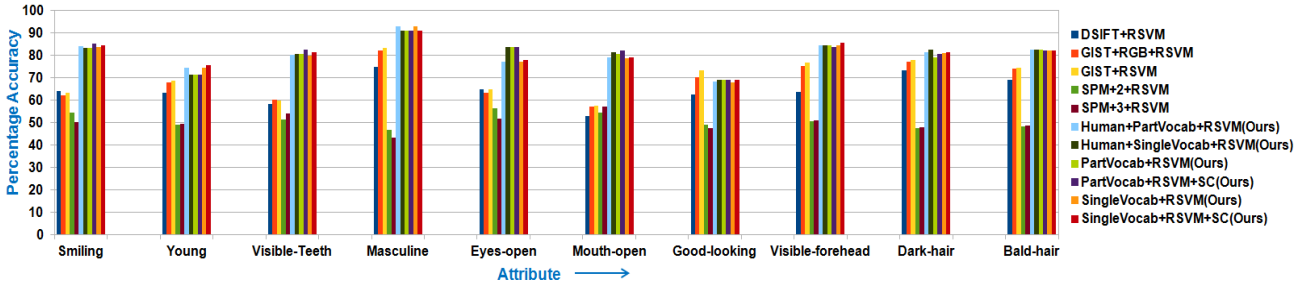

Figure 8. Performance for each of the ten attributes in LFW-10 dataset using different methods and representations.

direction for further research. (4) The performance after combining learned significance-coefficients with parts is always better than unweighted parts (last two blocks of Table 2). This reflects the importance of learning and incorporating part-specific weights into the joint representation. (5) The results obtained using learned parts are better than those using human selected parts. This could be because for humans, it is difficult to precisely assign a weight to every part (hence we used equal weights for all human selected parts). However, this limitation is overcome by our optimization framework ($OP4$) that allows to learn part-specific weights for a given attribute. Figure 6 shows the top five parts with highest significance-coefficients, and (a subset of) five parts selected by human expert for three attributes. Figure 7 shows the top ten learned parts with highest significance-coefficients for all the ten attributes in LFW-10 dataset. These demonstrate that even by using weak associations between image pairs and their annotations, our method can efficiently learn discriminative and semantically representative parts for different attributes.

In Figure 8, we show the performance of different methods for each of the ten attributes in LFW-10 dataset. Here, we can observe that the proposed methods always performs better (sometimes significantly) or comparable to the baseline method of [24]. Also, the performance of our method closely matches with that obtained using human selected parts, thus demonstrating its effectiveness.

6.5. Application to Interactive Image Search

Now, we illustrate the advantage of the proposed method on the task of interactive image search using relative attribute based feedback. Our feedback collection set-up is similar to that of [25]. Given a target image, it needs to be described relative to a few reference images (which are different from the target image) based on relative attributes. For a given attribute's feedback with respect to a reference image, the search set is partitioned into two disjoint sets using that attribute's scores. The rank of all the images in the search set are averaged over all feedbacks over all reference images. To break-up ties, absolute classifier score difference with respect to reference image is used. The intuition behind this set-up is that the images which match maximum with attribute feedback should be ranked towards the top.

The 1000 test images of LFW-10 dataset comprise our search set. We keep number of reference images to be either one or two, and vary the number of attribute-based feedbacks per reference in $\{2, 5, 10\}$. A total of 275 searches are performed for each of the six settings, by collecting feedbacks from 30 human evaluators. Figure 9 shows the performance of different methods for the six settings. For a given rank, we compute how many target images are predicted below that rank. This means that more is the number of search images falling below a specified rank, better is the performance. From the results, we can observe that the performance of all the methods improves with increase in number of feedbacks and/or number of reference images. This is expected since more interactions (feedbacks) result in better describing the target image. These results demonstrate that here also our method consistently outperforms the baseline method, and achieves performance comparable to that using human selected parts, thus validating its efficacy.

7. Conclusion

Inspired from the success of relative attributes, we have presented a novel method that learns relative attribute models using local parts that are shared across categories. Our

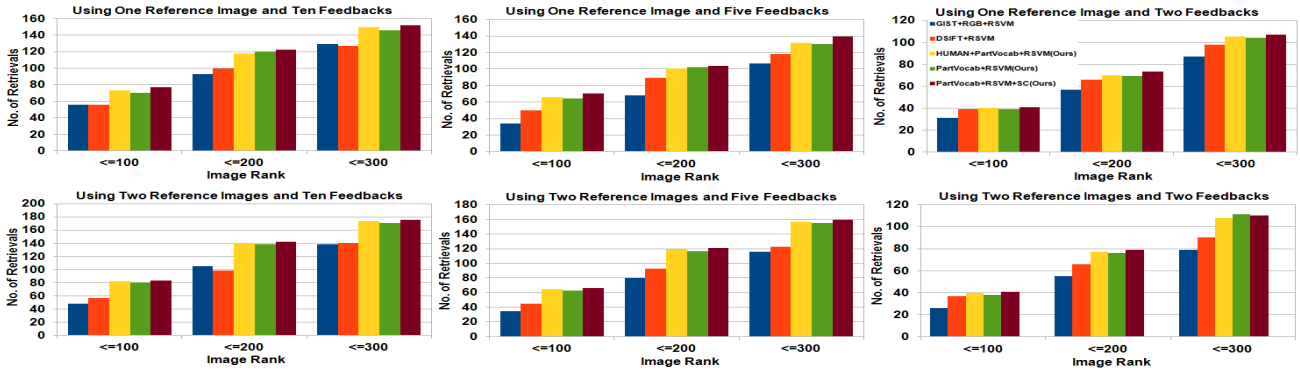

Figure 9. Performance variation of different methods on interactive image search with number of reference images and number of feedbacks. Each plot shows the number of searches in which the target image is ranked below a particular rank. Larger is the number of searches falling below a specified rank, better is the accuracy.

method achieves significant improvements compared to the baseline method. Apart from this, the part-specific weights learned using our method also provide semantic interpretation of different parts for diverse attributes.

Acknowledgement

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

References

[1] T. Berg and P. N. Belhumeur. POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

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

[3] O. Chapelle. Training a support vector machine in the primal. In *Neural Computation*, 2007.

[4] J. Choi, M. Rastegari, A. Farhadi, and L. Davis. Adding unlabeled samples to categories by learned attributes. In *CVPR*, 2013.

[5] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[6] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012.

[7] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[10] V. Ferrair and A. Zisserman. Learning visual attributes. In *NIPS*, 2007.

[11] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[12] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2002.

[13] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout:Distinctive parts for scene classification. In *CVPR*, 2013.

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

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

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

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

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

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[21] S. Maji. Discovering a lexicon of parts and attributes. In *Second International Workshop on Parts and Attributes, ECCV*, 2012.

[22] F. Nie, H. Huang, X. Chai, and C. Ding. Efficient and robust feature selection via join $l_{2,1}$-norms minimization. In *NIPS*, 2010.

[23] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

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

[25] D. Parikh and K. Grauman. Implied feedback: Learning nuances of user behavior in image search. In *ICCV*, 2013.

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

[27] A. Sadovnik, A. C. Gallagher, D. Parikh, and T. Chen. Spoken attributes: Mixing binary and relative attributes to say the right thing. In *ICCV*, 2013.

[28] B. Saleh, A. Farhadi, and A. Elgammal. Object-centric anomaly detection by attribute-based reasoning. In *CVPR*, 2013.

[29] A. Shrivastava, S. Singh, and A. Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *ECCV*, 2012.

[30] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.

[31] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistics Society B*, 58:267–288, 1996.

[32] N. Turakhia and D. Parikh. Attribute dominance: What pops out? In *ICCV*, 2013.

[33] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, 2010.

[34] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

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