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

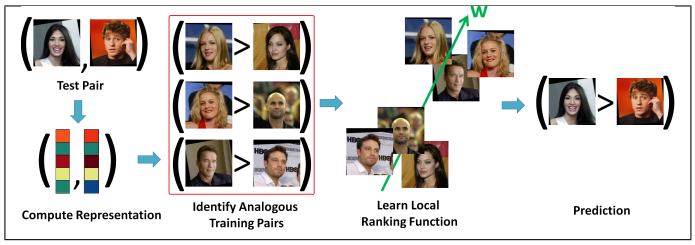

Figure 1: Approach overview: Given a test pair, first its patch-based representation is computed. Then using this representation, its analogous training pairs are identified. These pairs are used to learn a (local) ranking function, which is finally used for relative attribute prediction ("smiling" in this above illustration).

Relative attributes help in comparing two images based on their visual properties [4]. 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 (such as [3, 4, 5, 6]) 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 [1] for this task.

Given an image, we compute HOG descriptors from non-overlapping square patches 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. 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 2). We integrate this representation with two state-of-the-art approaches: (i) "Global" [4] that learns a single, globally trained ranking model (Ranking SVM [2]) for each attribute, and (ii) "LocalPair" [6] that uses a ranking model trained locally using *analogous* training pairs for each test pair. Its another variant, "LocalPair+ML", uses a learned distance metric while computing the analogous pairs. The motivation behind the LocalPair approach is that 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. This approach proposes to consider only the few training pairs for each test pair that are most analogous to it. These can be thought of as the K training pairs that are most similar to the given test pair. In LocalPair+ML, a learned distance metric is used to give more importance to those feature dimensions that are more representative of a particular attribute while computing the analogous pairs. Using the identified pairs, both LocalPair and LocalPair+ML learn a local (specific to the given test pair) ranking model similar to [4]. Note that the "Global" approach can be thought of as a special case of the LocalPair approach where K is the total number of training pairs, and thus all of them are considered while learning a ranking model. This is illustrated in Figure 1.

We refer the above baselines as Global+Hog, LocalPair+Hog and LocalPair+ML+Hog. These baselines are extensively evaluated on three challenging relative attribute datasets: OSR (natural outdoor scenes), LFW-10 (faces) and UT-Zap50K (shoes). While comparing with previous works, we use the representations used by them (wherever applicable). Table 1 summarizes the quantitative results. We can observe that the baselines 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 (note that UT-Zap50K-2 dataset with fine-grained within-pair visual differences is the most challenging among these datasets). For detailed comparisons, please refer to the paper.

	LFW-10	OSR	UTZ-1	UTZ-2
RelativeParts [5]	78.5	–	–	–
Global [4]	63.4	88.0	88.1	61.7
LocalPair [6]	62.4	85.7	87.6	63.1
LocalPair+ML [6]	63.6	90.2	88.7	64.7
Global+Hog	72.9	86.8	90.2	65.8
LocalPair+Hog	72.3	87.9	88.8	67.0
LocalPair+ML+Hog	72.8	88.6	90.0	67.5

Table 1: Average accuracy comparison for all the three datasets. The best results (this work and those of the previous approaches) are underlined.

To analyse the performance gains achieved by Global+Hog on the LFW-10 dataset, we try to visualize what a global ranking model learns using the HOG descriptor. 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 learned by the model. In Figure 2, we show these weights for two attributes. Surprisingly, the top two cells with maximum aggregate weights fall at almost the right place, thus demonstrating the possibility of attribute semantics being encoded in the ranking model.

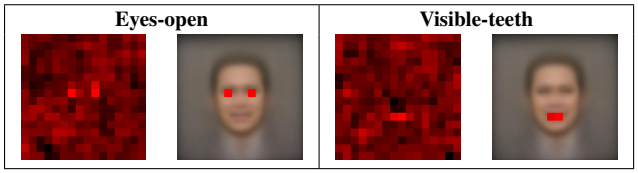

Figure 2: Learned HOG weights using Global+Hog baseline for two attributes from the LFW-10 datasets Left: Normalized distribution of weights. Right: Top few largest weights overlaid on the average image of this dataset (best viewed in color).

In this work, our goal was to develop intuitively simple baselines rather than to create a new method for learning relative attributes. The 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. 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.

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

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

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

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

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

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