Human Attribute Recognition by Rich Appearance Dictionary

Jungseock Joo[1], Shuo Wang[1,2], and Song-Chun Zhu[1]

[1]Center for Vision, Cognition, Learning and Art,
Depts. of Statistics and Computer Science, UCLA
{joo@cs, sczhu@stat}.ucla.edu
[2]School of EECS, Peking University
shuowang@pku.edu.cn

Abstract

We present a part-based approach to the problem of human attribute recognition from a single image of a human body. To recognize the attributes of human from the body parts, it is important to reliably detect the parts. This is a challenging task due to the geometric variation such as articulation and view-point changes as well as the appearance variation of the parts arisen from versatile clothing types. The prior works have primarily focused on handling geometric variation by relying on pre-trained part detectors or pose estimators, which require manual part annotation, but the appearance variation has been relatively neglected in these works. This paper explores the importance of the appearance variation, which is directly related to the main task, attribute recognition. To this end, we propose to learn a rich appearance part dictionary of human with significantly less supervision by decomposing image lattice into overlapping windows at multiscale and iteratively refining local appearance templates. We also present quantitative results in which our proposed method outperforms the existing approaches.

1. Introduction

We present a part-based approach to the problem of human attribute recognition from a single image of a human body. Human attributes, enriched textual descriptions of people such as gender, hair style, clothing types, provide fine-grained semantics. This is a practically important problem which can lead to many applications such as surveillance [19] or image-to-text generation [22].

Since many attributes can be inferred from various body parts (*e.g*., 'legs' → 'jeans'), it is important to reliably detect the parts for accurate attribute recognition. This, detection itself, is a challenging task, as noted in [2], due to the geometric variation such as articulation and view-point changes as well as the appearance variation of the parts arisen from versatile clothing types. The existing approaches [2, 5] have mainly focused on resolving the first issue - geometric variation of parts - by adopting pre-trained part detector or pose estimator.

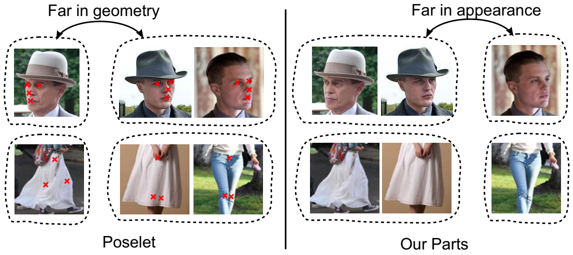

Figure 1. Part learning methods in Poselet-based approach [3] and our method. (*left*) Each poselet is learned from the examples of similar geometric configurations of keypoints (red marks). (*right*) We learn our parts based on appearance to preserve attribute-specific information.

However, there are a few limitations in these approaches. First, appearance variation of parts (hat vs. non-hat) is not taken into account in part learning or detection. The visual part dictionary or part appearance model of pose estimation is usually obtained by geometric constraints and not informative for attribute classification. In other words, these are generic part templates that do not have to distinguish different types of appearance in their learning objectives. Apparently, this is not the case for the problem of attribute recognition because it is the appearance type of body parts that one has to answer. Although prior works also attempt to recognize the appearance type *after* detecting the parts, such approaches might suffer from noisy detections since pose estimation is a still unsolved problem. In addition, it is required to collect keypoint annotation on body parts to train the pose estimators.

This paper explores the other dimension of variation of human parts: the appearance variation. The major source to appearance variation of human parts is a variety of clothings and these different types of clothes or accessories often yield more significant changes in the actual images than articulation or viewpoint changes (see the examples of 'skirt' in Fig. 1). Therefore, it is important to address such variation properly for reliable part detection by learning a rich appearance part dictionary. A rich appearance dictionary means that the dictionary is fluent enough to account for many different appearance part types. To explain appearance type also means to answer given questions in our ultimate task, attribute recognition. We empirically demonstrate the importance of such dictionary for the task of attribute recognition on two publicly available datasets [2, 15] where our method, without using numerous keypoint annotation, outperforms the prior works.

2. Related Works

Human Attribute Recognition from Body Cues. Due to its practical importance, fine-grained human attribute recognition has been studied intensively in the literature. Earlier works used the facial images for classification of gender [9], age group [13], ethnicity [10], and so on, since the face is the most informative part for these attributes. Kumar *et al.*, [12] recently proposed a framework for face verification aggregating a much broader set of general facial attributes. Since frontal face is visually distinct from the other human parts or other objects (*i.e.*, background), detection is relatively reliable. On the other hand, the other body parts such as arms, legs can be also informative for certain types of attributes. Cao *et al.* [4] has shown that the evidences to determine gender can be collected from the whole body and a more general set of attributes (gender, hair style, and clothing types) has been also considered in recent works [2, 5, 15].

In contrast to the face, it is difficult to extract information reliably from the whole body due to huge variation of parts in geometry and appearance. The prior works on attribute recognition can be categorized into two sets by their strategies to handle pose variation. (i) The first class of methods ignore latent pose and use fixed templates [4, 15]. This may not be robust against articulation or viewpoint change which is frequent in real-world examples. (ii) The other methods model the pose with geometric latent variable and rely on pre-trained pose estimator or part detectors to infer it [2, 5]. For example, Bourdev *et al.* [2] proposed a framework for human attribute classification using pre-trained part detectors, 'Poselets' [3]. Chen *et al.* [5] introduced a model based on conditional random field to exploit correlation between attributes explicitly, using the pose estimation of [21].

In the second group of approaches, part detection or pose estimation functions as a pre-processing stage and attribute recognition is performed subsequently. Despite the importance of part localization in the task, there are a few limitations on this strategy. (i) It is a still challenging problem in computer vision to estimate pose, thus it may be risky to rely on potentially noisy output. (ii) The appearance variation of parts (*e.g.* hat vs non-hat) is not taken into account in part learning nor detection. The learned dictionary usually contains *generic* parts mainly constrained in geometry and such parts do not convey attribute-specific information. (iii) Finally, it is expensive to collect keypoint annotation of body parts, which is required to train pose estimators or part detectors.

Weakly-Supervised Part Learning. In this paper, we learn the dictionary of discriminative parts for the task of attribute recognition directly from training images. Our method can be viewed as a weakly-supervised method since we do not require manual annotation of parts used in fully-supervised methods. Fig. 1 illustrates the main difference between our approach and Poselet [3] in part learning. We learn each part by clustering image patches on their appearance (low-level image features) while the poselet approach [3] learns a part from the image patches of similar geometric configurations of keypoints. Intuitively, our parts are more diverse in appearance space and the Poselets are strictly constrained in geometry space. [1]

The recent researches in weakly-supervised part learning suggest two important criteria. First, part learning can be directly guided by overall learning objective of given task (*i.e.*, classification gain) [17, 7]. In particular, Sharma *et al.* propose the expanded parts model [16] to automatically mine disciriminative parts for human attribute and action recognition, which is similar in spirit to our paper but differs in that we learn one part dictionary shared by all attribute categories while [16] learns a separate dictionary for each attribute. Second, it is important to use *flexible* geometric partitioning to incorporate a variety of region primitives [11, 20, 18, 1] rather than a pre-defined and restrictive decomposition which may not capture all necessary parts well. Therefore, we decompose the image lattice into many overlapping image subregions at multiscale and discover useful part candidates after pruning sub-optimal parts with respect to the attribute recognition performance.

3. Part Learning

Now we explain the details of our approach. In general, there are two considerations to be made in part learning, a

[1] More precisely, the poselet approach, after the initial learning stage, filters examples whose appearance is not consistent with the learned detector. This makes each Poselet also tight in the appearance space. However, they are still not diverse.

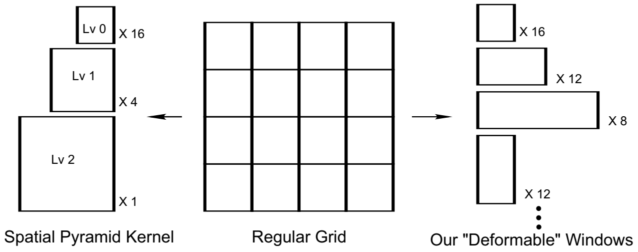

Figure 2. Two region decomposition methods based on the image grid: (left) spatial pyramid [14] and (right) our multiscale overlapping windows. The spatial pyramid subdivides the image region into four quadrants recursively, while we use all rectangular subregions on the grid, which is similar to [20, 18].

geometric and an appearance basis. We first need to specify what kinds of region primitives are allowed to decompose the whole image region into subregions at the part level (Sec. 3.1). Then, we discuss how to learn appearance models to explain the local appearance of each part (Sec. 3.2).

3.1. Geometric Configuration

Given no prior knowledge on human body ontology, our objective is to define a geometric basis (*i.e.*, region decomposition) which is expressive enough to capture the parts of arbitrary shapes, but is still of manageable complexity. While there exist simpler methods such as spatial pyramid [14] or uniform partitioning where all subregions are squares, it is difficult to represent many body parts such as arms and legs in squares, and moreover, we do not know what would be sufficient. Therefore, we examine many possible subregions from which we can learn many part candidates, some of which will be pruned in later stages. We only use rectangular subregions to limit the complexity, but allow them to be of arbitrary aspect ratios and sizes.

Specifically, our method starts by dividing the image lattice into a number of overlapping subregions. In this paper, we refer to each subregion as a **window**. We define a grid of size $W \times H$ [2], and any rectangle on the grid containing one or more number of cells of the grid forms a window. Fig. 2. illustrates our region decomposition method in comparison with the spatial pyramid matching structure (SPM) [14]. Both methods are based on the spatial grid on images. The SPM recursively divides the region into four quadrants and thus, all subregions are squares that do not overlap with each other at the same level. In contrast, we allow more flexibility in shape, size, and location of part window. Another important difference between our approach and SPM is that we treat each window as a template by a set of detectors that can be deformed locally, whereas each region in SPM is used for spatial pooling.

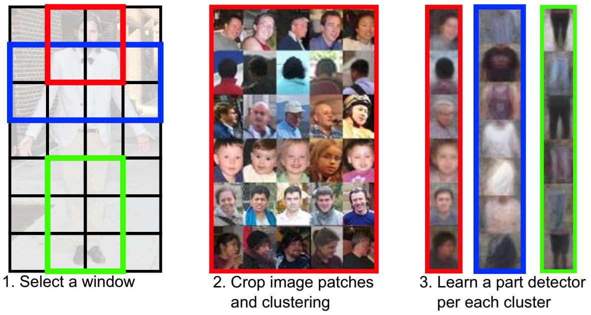
1. Select a window
2. Crop image patches and clustering
3. Learn a part detector per each cluster

Figure 3. Window specific part learning. For every window on the grid, we learn a set of part detectors from clustered image patches in training set. Each learned detector is reapplied to the images and refined.

The advantage of flexible geometric partitioning has been also advocated in the recent literature of scene modeling and object recognition [11, 20, 18, 1]. In particular, our grid decomposition method is very similar to the initial step of [20, 18], which then further attempt to pick up a subset of good windows and represent each image with a small number of non-overlapping windows, which reduces to a single best configuration through explicit parsing. However, we empirically found that it leads to a better performance to allow many number of overlapping windows, therefore we only prune inferior part templates in the later stage but do not eliminate or suppress any windows. In other words, our method allows "all" configurations, each of which is implicitly defined and partially contributes to explain each image.

3.2. Part Appearance Learning

Once we define all windows, we visit each window and learn a set of part detectors that are spatially associated with that particular window. Our motivation is that the human body consists of a number of parts, which are usually spatially constrained, *i.e.* anchored at their canonical positions. Fig. 3. shows the general procedure and examples. For each window, w^i, we first crop all the corresponding image patches from the entire set of training images. Then each patch is represented by the feature descriptor. We use the Histogram of Oriented Gradient (HoG) [6] and color histogram as the low-level features of image patches. On the extracted features, we perform K-means clustering and obtain $K = 50$ clusters, $\{v_1^i, ...v_K^i\}$. Each obtained cluster represents a specific appearance type of a part. Since the initial clusters are noisy, we first train a local part detector for each cluster by logistic regression as a initial detector and then, iteratively refine it by applying it onto the entire set again and updating the best location and scale. We mine the negative patches from the regions outside given bounding boxes. At the initial iteration, we discard the noisy part candidates by cross validation, and limit the maximum number of useful parts to 30 (we will discuss the choice of this quantity in the experimental section). The detection score, g, of

[2] We use $W = 6, H = 9$ and let the unit cell be of aspect ratio of 2:3 in the experiment.

an image I for a part v_k^i can be expressed as follows:

$$g(v_k^i|I^i) = \log \frac{P(v_k^i = +|I^i)}{P(v_k^i = -|I^i)}, \quad (1)$$

where I^i is the image subregion occupied by the window, w^i. We transform this log posterior ratio by logistic function, as in [2]:

$$d(v_k^i|I^i) = P(v_k^i = +|I^i). \quad (2)$$

These detectors are local and only activated at the specific window from which they have been trained. We only allow relatively small range of local deformation (translation and scaling) around the window (20% of deviation from the window). Such assumption may seem to be weak to robustly capture many articulated parts or viewpoint change. For example, a lower-arm can be located virtually everywhere. However, we do not explicitly distinguish the geometric and appearance variation. That is, if a part is articulated and located far from its canonical window *frequently*, we treat this as another appearance part type that is defined at another window. If such arrangement is not frequent, we disregard it. This treatment can be also justified by considering that a part looks differently from the same part in a different pose. Therefore, it may be beneficial to maintain separated part templates for those cases so that each template can explain its own type better.

4. Attribute Classification

Now we explain our method for attribute classification. After learning the parts at multiscale overlapping windows, we mainly follow the strategy for attribute classification proposed in the Poselet-based approach [2]. The key idea is to detect the parts by learned detectors (Poselets in [2]) and then to train a series of part-specific local attribute classifiers. The final prediction is made by aggregating the scores from these local classifiers with the weights given by part detection scores.

Such strategy is effective for the task of fine-grained classification such as human attribute classification. This is because many part elements that are not directly informative about the class to predict (attributes) can still provide a guidance to retrieve the discriminative cues. For example, an upper-body detector itself does not help predict any types of attributes. However, once we locate the upper-body by it, we can run the additional classifier to obtain finer-grained information constrained in the upper-body region.

4.1. Part Specific Attribute Classifiers

Specifically, for each part detector, v_k^i, at each window, w^i, we have a set of image patches, the cluster members used to train the detector. This cluster is a soft cluster, so it is possible that the same image can be included in multiple number of clusters as long as its part detection score is positive ($g(v_k^i|I^i) > 0$). Then for each attribute, we have two disjoint subsets of the image patches in the cluster, the positive and the negative. By using the same image features used for detection, we train an attribute classifier for an individual attribute, a^j, by another logistic regression as follows:

$$f(a^j|v_k^i, I^i) = \log \frac{P(a^j = +|v_k^i, I^i)}{P(a^j = -|v_k^i, I^i)}. \quad (3)$$

Again, we discard the learned classifier whose accuracy is inferior by cross validation so that the number of parts in whole model is limited to 1,000 to address all attributes.

4.2. Aggregating Attribute Scores

We have obtained all part detection scores as well as part-specific attribute classification scores. These are local scores obtained from local parts. Now we use another discriminative classifier to aggregate these local scores to output the global prediction. Again, we use the same strategy used in the Poselet-based approach, which combines the attribute classification scores with the weights given by part detection scores.

Specifically, we form a final feature vector, $\phi(I)$ for each image I and each attribute a as follows:

$$\phi_k^i(I) = d(v_k^i|I^i) \cdot f(a^j|v_k^i, I^i). \quad (4)$$

Therefore, each element is simply the product of two terms that we defined in previous sections. Note that i and k are used to index the window and part type at each window, and we form a 1D vector simply by organizing each part sequentially. We refer to this vector as the part-attribute feature vector. Then we train the logistic regression again with this feature and use it as the final classifier.

In contrast to the Poselet-based approach which leverages an additional layer of classifiers to model contextual information between attributes (*e.g.*, 'long-hair' and 'female'), we do not explicitly model the correlation between attributes in this paper. This is mainly because our dictionary, which preserves appearance type, can be a good basis to implicitly capture such correlation. For example, once we detect a face with 'long-hair', it can immediately inform us that it is more likely to find 'skirt' as well even before proceeding to attribute inference stage. The poselet, however, lacks appearance type inference in detection stage and thus, has to explicitly enforce such constraints in a later stage.

5. Experimental Results

In this section, we present a various experimental results that can support the advantage of our approach.

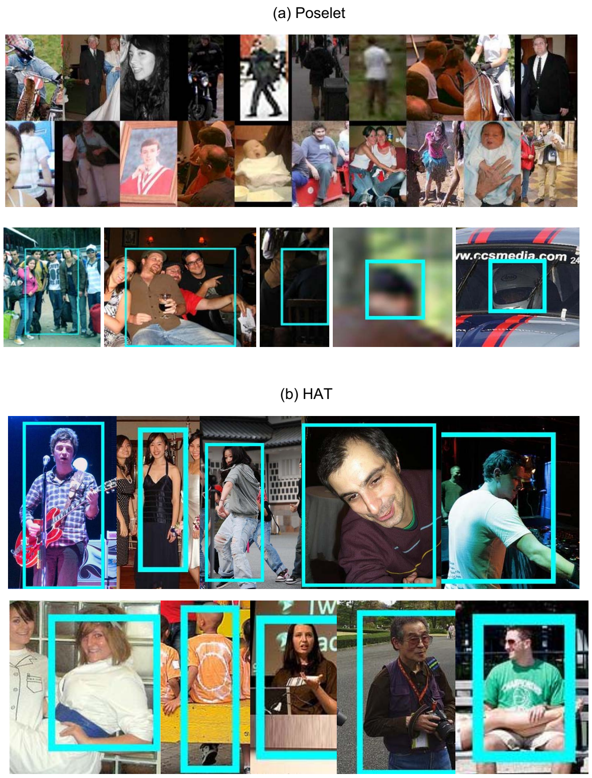

Figure 4. (a) (top) Images from the dataset of Poselet-based approach, copied from [2]. This dataset exhibits a huge variation of pose, viewpoint, and appearance type of people. (a) (middle) A few selected, extremely challenging images from the same set. Either the bounding box (to indicate the person of interest) is ambiguous in cluttered scene, the resolution is too low, or occlusion and truncation is too severe. (b) Randomly selected images from HAT Database [15]. All the bounding boxes have been obtained by person detector [8], therefore the complexity is limited.

Datasets. For evaluation of our approach, we use two publicly available datasets of human images labeled with attributes: the dataset of Poselet [2], and the Database of Human Attributes (HAT) [15]. Fig. 4 shows the examples taken from both sets. Each set has been constructed in a different way. We discuss the details in following subsections.

5.1. Attribute Classification on Dataset of Poselet

The dataset contains 3864 training images and 3600 testing images, each of which is labeled with 9 binary attributes. Fig. 4 shows a few examples from Poselet's dataset. Each image is manually annotated by a visible bounding box of each person. This bounding box is provided at training as well as testing time, *i.e.*, detection is given. Since these boxes that cover visible parts of humans do not provide any alignment, it is very challenging to learn or detect the parts from them. Also, the evaluation may be problematic because the interested person indication is sometimes ambiguous in crowded scenes (Fig. 4) and such box is difficult to obtain in fully automated systems which would typically deploy a person detector prior to attribute inference; such detector would provide the alignment at the level of full-body or upper-body.

Therefore, we first aligned all the training images by using two keypoints on the head and middle-hip and trained upper-body detectors. And by applying it onto the images (while penalizing deviation from the original boxes), we obtained roughly aligned boxes. These aligned boxes are simply enlarged from the outputs of upper-body detectors and share the same aspect ratio.

For fair comparison, we used the same set of image features as [2], HoG and color histogram. While the Poselet-based approach additionally uses the human skin-tone as another feature, we do not use any extra image features in this paper because our goal is weakly-supervised learning which requires no supervision on skin. The total number of parts was set to 1000 while [2] used 1200 parts.

Table 1. shows the full comparison where our full model outperforms the Poselet based approaches in 8 out of 9 attributes. Note that the "full" model indicates the approach using multiscale overlapping windows and the rich appearance part dictionary as we have discussed in this paper. In order to verify the contribution of each factor to the final performance, we conducted two additional tests as follows.

Rich Appearance Dictionary. We have argued that it is important to learn a rich appearance dictionary that can address the appearance variation of parts effectively. We validate this argument by varying the number of parts learned at each window, K, ranging from 1 to 30. However, we still perform clustering to form many clusters and then choose K best part candidates, judged by cross validation detection score.

Table 3. shows the performance improvement according to K and this result can support the importance of rich visual dictionary. In particular, having many parts per window is important for subtle attributes, such as "glasses". Note that, $K = 1$ does not mean that we only have one part template for each true part. Since we have multiscale overlapping windows, and we can still have many other templates learned at the other windows. This can also explain why the gender attribute, whose cues would be more distributed over many subregions as a global attribute, has the least amount of gain from increasing K.

Multiscale Overlapping Windows. We also tested the effect of multiscale overlapping window structure used in our approach. Table 1. (b) shows the performance when we only used a set of non-overlapping windows at single layer, which reduces to a simple grid decomposition, and the row (c) shows the result when we use the windows at two more additional layers as spatial pyramid scheme. Neither method performed as well as the full approach.

		male	long hair	glasses	hat	t-shirt	long sleeves	shorts	jeans	long pants	Mean AP
Base Frequency		.593	.300	.220	.166	.235	.490	.179	.338	.747	.363
Ours (2)	(a) Full	**.880**	**.801**	.560	**.754**	**.535**	**.752**	**.476**	**.693**	**.911**	**.707**
	(b) Uniform Grid	.857	.734	.429	.631	.405	.687	.349	.560	.862	.613
	(c) Spatial Pyramid	.857	.725	.407	.641	.429	.707	.356	.565	.886	.620
Poselet (33)	(d) Full *	.824	.725	.556	.601	.512	.742	.455	.547	.903	.652
	(e) No context *	.829	.700	.489	.537	.430	.743	.392	.533	.878	.615
	(f) No skintone	.825	.732	**.561**	.603	.484	.663	.330	.428	.850	.608

Table 1. The attribute classification performance on the dataset of poselet [2]. The number in parentheses is the number of keypoints used in learning of each method. * indicates the methods to use an additional image feature (skintone).

Attributes	Freq.	Ours	DSR [15]	EPM [16]
Female	.488	**.914**	.820	.859
FrontalPose	.747	**.968**	.913	.936
SidePose	.239	**.772**	.610	.673
TurnedBack	.105	**.898**	.674	.772
UpperBody	.413	.963	.924	**.979**
Standing	.768	.977	.960	**.980**
Run/Walk	.153	.635	.676	**.746**
Crouching	.026	.123	.207	**.240**
Sitting	.117	.593	.546	**.627**
ArmsBent	.765	**.954**	.919	.940
Elderly	.070	.321	.293	**.389**
Mid-Aged	.507	**.700**	.663	.689
Young	.381	**.656**	.594	.642
TeenAged	.159	.335	.291	**.362**
SmallKid	.143	**.535**	.437	.497
SmallBaby	.027	.163	.122	**.243**
TankTop	.081	.370	.332	**.377**
Tee shirt	.352	**.671**	.591	.616
CasualJacket	.088	**.426**	.353	.400
Men'sSuit	.068	**.648**	.482	.571
FemaleLongSkirt	.060	.420	.399	**.448**
FemaleShortSkirt	.058	.301	.337	**.390**
Shorts	.070	**.496**	.427	.468
LowcutTop	.166	**.660**	.556	.613
FemaleSwimsuit	.023	**.467**	.282	.322
WeddingDress	.014	.621	.621	**.642**
Bermuda	.129	.420	.393	**.437**
Mean AP	.230	**.593**	.538	.587

Table 2. The attribute classification performance (average precision) on the dataset of HAT [15]. In addition, EPM [16] achieves .597 in the setting to leverage image context from outside of bounding box region.

K	1	3	5	10	20	30
Male	.858	.860	.862	.870	.880	**.881**
LongHair	.690	.731	.736	.771	.795	**.801**
Glasses	.375	.397	.429	.458	.536	**.560**
Hat	.634	.625	.678	.676	.744	**.754**
Tshirt	.412	.412	.430	.430	.521	**.535**
LongSlv	.715	.719	.726	.738	.745	**.752**
ShortPnts	.373	.386	.407	.470	.474	**.476**
Jeans	.612	.595	.644	.690	**.695**	.693
LongPnts	.891	.906	.905	.910	**.912**	.911
mAP	.618	.630	.654	.679	.693	**.707**

Table 3. The performance improvement according to the maximum number of appearance part types at each window (K).

5.2. Attribute Classification on HAT Database

The HAT database contains 7000 training images and 2344 testing images, labeled with 27 binary attributes. There are two main difference between this dataset and Poselet's dataset. i) The considered attributes includes a set of action or pose categories such as "running" or "side pose". Table 2. shows the full list. ii) The dataset was constructed in a semi-supervised way by a person detector of [8], instead of manual collection. Therefore, the complexity in terms of articulation or occlusion is relatively lower than that of Poselet's dataset. However, such criterion is also meaningful, considering the fully automated real-world system would follow the same procedure - running the person detector and then performing attribute classification. Since the bounding boxes were obtained by the person detector, the people in the images are roughly aligned. Therefore, we did not use any keypoints for the rough alignment in this dataset.

Table 2 shows the performance comparison among our approach, the discriminative spatial representation (DSR) [15], and the expanded part models (EPM) [16]. The DSR, a variant of spatial pyramid representation, seeks for an ideal partitioning for classification, instead of strictly constrained structure (squares) of original spatial pyramid. However, their partitioning, even if it is an optimal, is still fixed. And geometric variation (local deformation) has not been addressed, which is important for human models. On the other hand, the EPM which also attempts to learn the discriminative parts has shown a comparable result to ours in an equivalent setting where recognition is performed solely

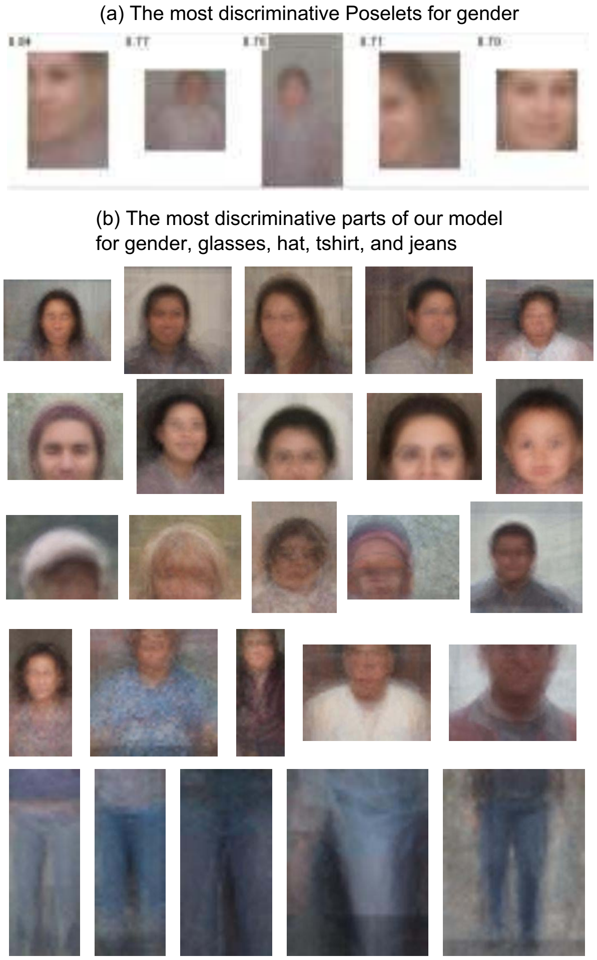
(a) The most discriminative Poselets for gender
0.84
0.77
0.76
0.71
0.70
(b) The most discriminative parts of our model for gender, glasses, hat, tshirt, and jeans

Figure 5. The most discriminative parts in Poselet-based approach [2] and our learned model. Our rich dictionary distinguishes many different appearance part types, which are directly informative for attribute classification, while the selected poselets are generic parts.

from person bounding box region. However, the advantage of our method is to learn a common dictionary shared by all attribute categories whereas the EPM uses a separate dictionary for each category.

5.3. Discriminative Parts for Localization

Finally, we discuss the issue of the discriminative parts by providing qualitative results. The quantitative evaluation is difficult because neither dataset provides required ground-truth annotation. The most discriminative part for an attribute is the part whose contribution to the attribute prediction is the biggest. We measure this from product of the weights of the final logistic regression classifier and the part-attribute feature of each image, $\phi(I)$. Fig. 6. shows the examples in the testing set (from the Poselet's dataset), which output the most positive and negative responses for five attribute categories. We denote the most contributed, most discriminative part window for each image by blue boxes. Although there are some meaningless activation (for example, "NOT JEANS" was inferred from the head), the most parts show reasonable localization.

Fig. 5 shows the most discriminative parts for five selected attributes. We measure this by correlation between attribute labels and the part-attribute feature. As one can easily see, the most discriminative Poselets are unbiased detectors which would respond to both female and male. In contrast, our discriminative parts has distinct polarities.

6. Conclusion

We presented an approach to the problem of human attribute recognition from human body parts. We argue that it is critical to learn a rich appearance visual dictionary to handle appearance variation of parts as well as to use a flexible and expressive geometric basis. While the major focus has been made on appearance learning in this paper, we plan to expand the current model into structured models where we can learn more meaningful geometric representation, as for the future work.

7. Acknowledgement

This work was gratefully supported by NSF CNS 1028381, DARPA MSEE project FA 8650-11-1-7149, ONR MURI N00014-10-1-0933, and NSF IIS 1018751. The first author was also partially supported by Kwanjeong Educational Foundation.

References

[1] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. Seeking the strongest rigid detector. In *CVPR*, 2013.

[2] L. Bourdev, S. Maji, and J. Malik. Describing people: Poselet-based attribute classification. In *ICCV*, 2011.

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[4] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang. Gender recognition from body. In *ACM MM*, 2008.

[5] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] I. Endres, K. J. Shih, J. Jiaa, and D. Hoiem. Learning collections of part models for object recognition. In *CVPR*, 2013.

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

[9] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *NIPS*, 1990.

[10] S. Gutta, H. Wechsler, and P. J. Phillips. Gender and ethnic classification of face images. In *FG*, 1998.

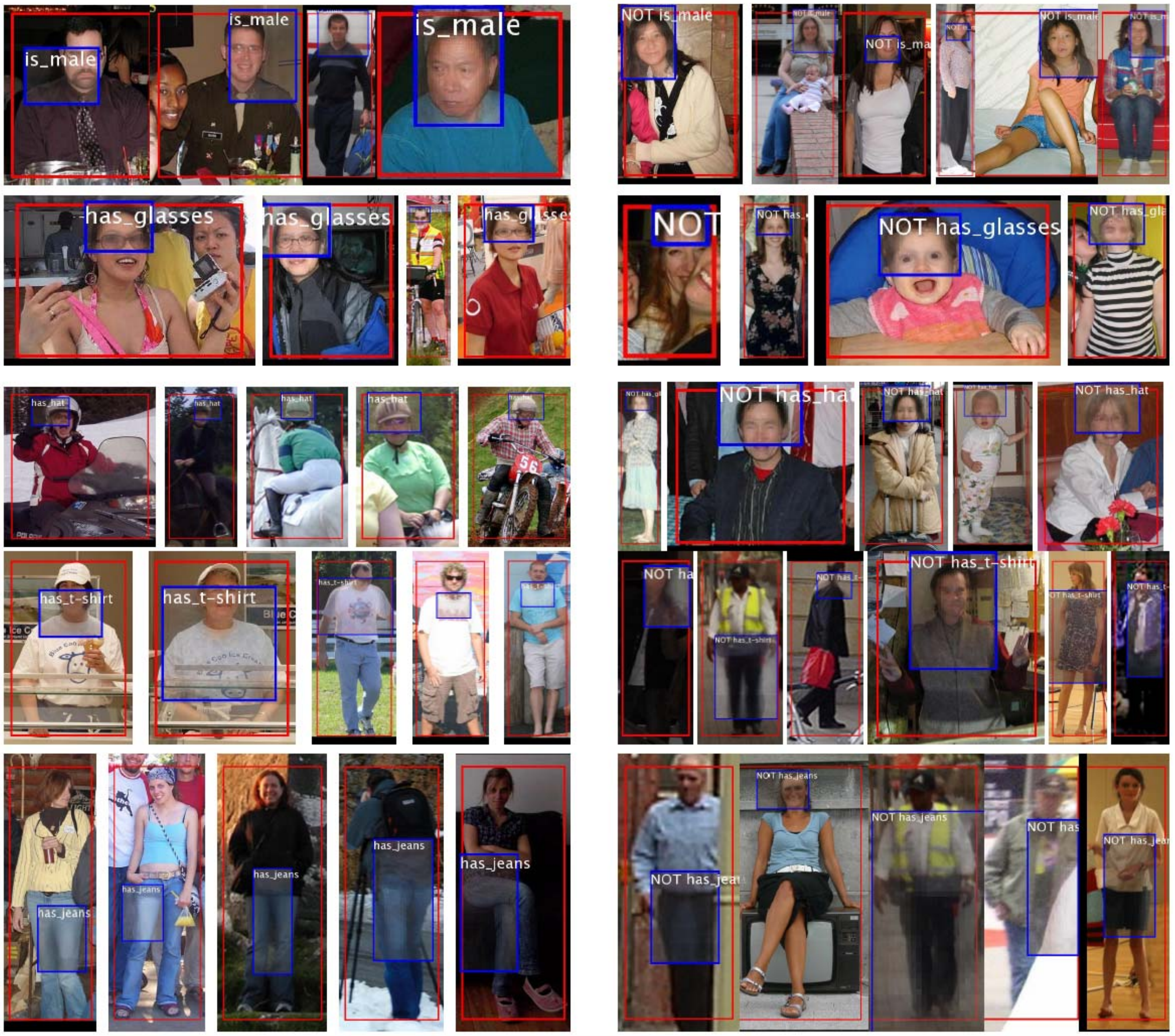

Figure 6. The qualitative results of attribute classification for male, glasses, hat, t-shirt, and jeans categories. (left) The most positive, and (right) the most negative images for each attribute. The red boxes denote the bounding boxes and each blue box represents a part detection whose contribution to prediction is the biggest.

[11] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *CVPR*, 2012.

[12] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *TPAMI*, 33(10):1962–1977, 2011.

[13] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *CVIU*, 1999.

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

[15] G. Sharma and F. Jurie. Learning discriminative spatial representation for image classification. In *BMVC*, 2011.

[16] G. Sharma, F. Jurie, and C. Schmid. Expanded parts model for human attribute and action recognition in still images. In *CVPR*, 2013.

[17] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[18] X. Song, T. Wu, Y. Jia, and S.-C. Zhu. Discriminatively trained and-or tree models for object detection. In *CVPR*, 2013.

[19] D. A. Vaquero, R. S. Feris, D. Tran, L. M. G. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, 2009.

[20] S. Wang, J. Joo, Y. Wang, and S.-C. Zhu. Weakly supervised learning for attribute localization in outdoor scenes. In *CVPR*, 2013.

[21] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[22] B. Yao, X. Yang, L. Lin, M. W. Lee, and S. C. Zhu. I2T: Image parsing to text description. *Proceedings of the IEEE*, pages 1485–1508, 2010.