Describing People: A Poselet-Based Approach to Attribute Classification *

Lubomir Bourdev[1,2], Subhransu Maji[1] and Jitendra Malik[1]
[1]EECS, U.C. Berkeley, Berkeley, CA 94720
[2]Adobe Systems, Inc., 345 Park Ave, San Jose, CA 95110
{lbourdev,smaji,malik}@eecs.berkeley.edu

Abstract

We propose a method for recognizing attributes, such as the gender, hair style and types of clothes of people under large variation in viewpoint, pose, articulation and occlusion typical of personal photo album images. Robust attribute classifiers under such conditions must be invariant to pose, but inferring the pose in itself is a challenging problem. We use a part-based approach based on poselets. Our parts implicitly decompose the aspect (the pose and viewpoint). We train attribute classifiers for each such aspect and we combine them together in a discriminative model. We propose a new dataset of 8000 people with annotated attributes. Our method performs very well on this dataset, significantly outperforming a baseline built on the spatial pyramid match kernel method. On gender recognition we outperform a commercial face recognition system.

1. Introduction

Humans have an impressive ability to reliably recognize the gender of people under arbitrary viewpoint and articulation, even when presented with a cropped part of the image (Figure 1). Clearly we don't rely on the appearance of a single body part; gender can be inferred from the hair style, body proportions, types of clothes and accessories. We use different cues depending on the pose and viewpoint, and the same is true for other attributes, such as the hair style, presence of glasses and types of clothes.

Let us consider how we might build a system for classifying gender and other attributes. If we could somehow isolate image patches corresponding to the same body part from the same viewpoint then attribute classification becomes much easier. If we are not able to detect and align the parts well, however, the effect of nuisance variables, such as the pose, viewpoint and localization will affect the feature vector much more than the relevant signal (Figure 2). The visual cues associated with the attribute "has glasses", for example, are very subtle and different for a person facing the camera *vs*. a person looking sideways. As we show on Table 2, a generic classifier for has-glasses performs only slightly better than chance when trained on the entire person, but works much better when trained on aligned frontal faces.

Localizing body parts, however, is in itself a very hard problem, e.g. [13]. Frontal face is an exception, which is why virtually all state-of-the-art gender recognition approaches rely on carefully aligned frontal faces.

Figure 1. People can easily infer the gender based on the face, the hair style, the body proportions and the types of clothes. A robust gender classifier should take into account all such available cues.

Figure 2. The problem of determining the people wearing hats (top) *vs*. no hats (bottom) is difficult in unconstrained setup (left). If we can detect and align parts from the same view (right) the problem becomes much easier.

*This work was supported by Adobe Systems, Inc., Google, Inc., Intel Corporation, as well as ONR MURI N00014-10-10933

We develop an approach to solve this problem for gender as well as for other attributes, such the hair style, presence of glasses or hat, and the style of clothes. Specifically, we decompose the image into a set of parts, *poselets* [4], each capturing a salient pattern corresponding to a given viewpoint and local pose, such as the one shown in Figure 2 (right). This decomposition allows us to combine evidence from different parts of the body at different scales. **The activations of different poselets give us a robust distributed representation of a person from which attributes can be inferred without explicitly localizing different body parts.**

Prior work on gender recognition has focused on high resolution frontal faces or pedestrians and requires a face detector and alignment modules. Not only do we not need such modules, our method gracefully deals with profiles, back-facing people or even when the face is occluded or at too low a resolution, because we leverage information at multiple scales and aspects. Even though we use standard HOG and color features, on the task of gender recognition we outperform a leading commercial face recognition system that relies on proprietary biometric analysis. Furthermore, the same mechanism allows us to handle not just gender but any other attribute.

We illustrate our approach on the task of determining nine attributes of people – is-male, has-hat, has-t-shirt, has-shorts, has-jeans, has-long-hair, has-glasses, has-long-sleeves, has-long-pants. The training input is a set of images in which the people of interest are specified via their visible bounds and the values of their attributes. We use a three layer feed-forward network (Figure 4). In the first layer we predict the attribute value conditioned on each poselet type, such as the gender given a frontal face. In the second layer we combine the information from all such predictions (such as the gender given the face, the legs and the full body) into a single attribute classification. In the third layer we leverage dependencies between different attributes, such as the fact that gender is correlated with the presence of long hair.

Our second contribution is a new dataset for attribute classification of people in unconstrained settings consisting of 8035 examples labelled with the nine attributes (Section 3). Although attribute recognition of people has been studied for frontal faces [19] and pedestrians [6], our dataset is significantly harder; it exhibits a large variation in viewpoint, pose, occlusion and self-occlusion, close proximity to other people, variable resolution, etc. (Figure 3).

2. Related Work

Prior research on attributes has generally followed two directions. One line of work has used attributes as an intermediate representation layer with the goal of transfer learning as well as describing properties of objects [20, 12]. Farhadi *et al*. propose a method for localizing part-based attributes, such as a head or a wheel [11]. Recognition and localization of low-level attributes in a generative framework has also been proposed by Ferrari and Zisserman [14]. Joint learning of classes and attributes has been explored using Multiple Instance Learning [27] and latent SVMs [29]. Automated discovery of attributes from text and associated images has also been explored [14, 1, 28]. The key advantage of our method is that our parts implicitly model the pose and camera view, which we believe results in more powerful discrimination capabilities.

A second line of work has focused on attributes of people. Gender recognition methods using neural networks date back to the early 1990s [8, 16]. Support vector machines [24] and AdaBoost classifiers on Haar features [25] have been proposed for gender and race recognition. Kumar *et al*. propose using face attributes for the purpose of face recognition [19] as well as visual search [18]. Gallagher and Chen have explored inferring gender and age from visual features combined with names [15]. Gender, age and weight attributes have also been successfully extracted from 3D motion capture data [26]. These approaches generally require careful alignment of the data, and most of them apply to frontal faces only. We leverage the full body under any articulation without the need for alignment.

In our work we are inspired by poselets, which have been used effectively for recognition, segmentation and action classification of people [3, 4, 23, 5]. These problems are similar to ours, because the articulation and camera views are also latent parameters when recognizing and segmenting people. Thus we can think of poselets as a general purpose engine for decomposing the viewpoint and pose from the appearance.

3. The Attributes of People Dataset

There are several existing datasets of attributes of people but we did not find any suitable for the context in which our method is used. FaceTracer [18] uses 15000 faces and full body, but provides only URLs to images and many of the images are no longer available. Other datasets, such as PubFig [19] and the Labeled Faces in the Wild [17] include only frontal faces.

We propose a new dataset of 8035 images, each centered at a full body of a person. The images are collected from the H3D [4] dataset and the PASCAL VOC 2010 [10] training and validation datasets for the person category, but instead of the low-resolution versions used in PASCAL, we collected the full resolution equivalents on Flickr. For each person we cropped the high resolution image around that person, leaving sufficient background around the visible bounds and scaled it so the distance between hips and shoulders is 200 pixels. For each such image we provide the visible bounds of the person in the center and a list of bounds of all other people in the background.

Figure 3. Fifty images drawn at random from our test set and slightly cropped to the same aspect ratio. Each image is centered at a target person. Our dataset is challenging as it has a large variability of viewpoints, poses, and occlusions. In some cases people are close to each other which makes identifying the correct person challenging as well. To aid identification we provide the visible bounds of the target person, as well as the bounds of all other people in the image.

Attribute	True	False	Attribute	True	False
is male	3395	2365	long hair	1456	3361
has hat	1096	5532	glasses	1238	4083
has t-shirt	1019	3350	long sleeves	3045	3099
has shorts	477	2020	long pants	2020	760
has jeans	771	1612			

Table 1. Number of positive and negative labels for our attributes.

We used Amazon Mechanical Turk to provide labels for all attributes on all annotations by five independent annotators [22]. A label was considered as ground truth if at least 4 of the 5 annotators agreed on the value of the label. We discarded 501 annotations in which less than two attributes were specified as ground truths which left us with 8035 images. Table 1 shows the distribution of labels. We split the images into 2003 training, 2010 validation and 4022 test images by ensuring that no cropped images of different set come from the same source image and by maintaining a balanced distribution of the H3D and PASCAL images in each set. Figure 3 shows 50 examples drawn at random from our test set.

4. Algorithm Overview

Our algorithm at test time is shown on Figure 4 and can be summarized as follows:

Step 1 We detect the poselets on the test image and determine which ones are true positives referring to the target person (Section 5). Let q^i denote the probability of poselet type i. q^i is the score of the poselet classifier transformed by a logistic, with zero mean, or 0 if the poselet was not detected.

Step 2 For each poselet type i we extract a feature vector ϕ^i from the image patch of the activation, as described in Section 6. The feature vector consists of HOG cells at three scales, a color histogram and skin-mask features.

Step 3 For each poselet type i and each attribute j we evaluate a classifier r^i_j for attribute j conditioned on the poselet i. We call these the *poselet-level attribute classifiers*. We use a linear SVM followed by a logistic g:

$$r^i_j = g({w^i_j}^T \phi^i + b^i_j) \quad (1)$$

where w^i_j and b^i_j are the weight vector and the bias term of the SVM. These classifiers attempt to determine the presence of an attribute from a given part of the person under a given viewpoint, such as the has-hat classifier for a frontal face shown on Figure 2.

Step 4 We zero-center the outputs of the poselet-level attribute classifiers, modulate them by the poselet detection probabilities q^i and we use them as an input to a second-level classifier for each attribute j, called a *person-level attribute classifier*, whose goal is to combine the evidence from all body parts. It emphasizes poselets from viewpoints that are more frequent and more discriminative. It is also a linear classifier with a logistic g:

$$\psi^i_j = q^i(r^i_j - 0.5) \quad (2)$$

$$s_j = g({w'_j}^T \psi_j + b'_j) \quad (3)$$

Step 5 Finally, for each attribute j, we evaluate a third-level classifier which we call the *context-level attribute classifier*. Its feature vector is the scores of all person-level classifiers for all attributes, s_j. This classifier exploits the correlations between the attributes, such as gender *vs*. the presence of a skirt, or short-sleeves *vs*. short-pants. We use an SVM with quadratic kernel which we found empirically to work best. We denote the score of this classifier with S_j, which is the output of our algorithm.

5. Training and Using Poselets

We use the method of Bourdev *et al*. [3] to train 1200 poselets using images from the training and validation sets. Instead of all poselets having the same aspect ratios, we used four aspect ratios: 96x64, 64x64, 64x96 and 64x128 and trained 300 poselets of each. For each poselet, during training, we build a soft mask for the probability of each body component (such as hair, face, upper clothes, lower clothes, etc) at each location within the normalized poselet patch (Figure 5) using body component annotations on the H3D dataset [4].

We used the method of [3] to detect poselets in an image, cluster them into person detection hypotheses and predict

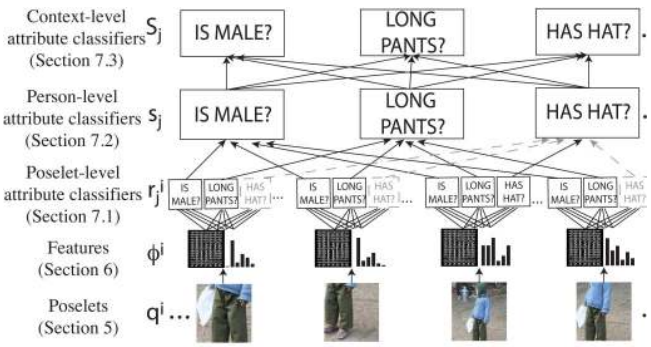

Figure 4. Overview of our algorithm at test time. Poselets are detected on the test image; detection scores q^i are computed and features ϕ^i are extracted. Poselet-level attribute classifiers r_j^i are evaluated for every poselet activation i and attribute j (unless the attribute is part-specific and the poselet does not cover the part, such as the has-hat for three of the four shown poselets). A person-level attribute classifier s_j for every attribute combines the feedback of all poselet-level classifiers. A context-level classifier S_j for the attribute takes into account predictions of the other attributes. This picture uses 4 poselets and 3 attributes, but our system uses 1200 poselets and 9 attributes.

Figure 5. **Left:**Examples of a poselet. **Right:** The poselet soft mask for the hair, face and upper clothes.

the bounds of each person. We now need to decide which cluster of poselets refers to the person in the center of the image and which ones refer to people in the background. Our dataset contains many instances of people very close to each other, so simply picking the bounding box closest to the center of the image is not always correct. Instead it is better to find the optimally global assignment of all hypotheses to all truth bounding boxes by preferring to assign a bounding box to a given truth if its intersection over union is high, and by giving preference to hypotheses with higher scores, which are less likely to be false positives. We formulate this problem as finding the maximum flow in a bipartite graph and we used the Hungarian algorithm to find the optimal matching. The result is a set of poselet activations q^i that refer to the foreground person.

6. Poselet-Level Features ϕ^i

In this section we describe our poselet-level features ϕ^i, which consist of HOG features, color histogram and skin-specific features.

For the HOG features we use the same parameters as described in [9]. In addition to the 8x8 cells we extract HOG at two coarser levels - 16x16 and 32x32. Depending on the patch dimensions this feature is of size between 2124 and 4644. The color histogram is constructed with 10 bins in each of the H, S and B dimensions.

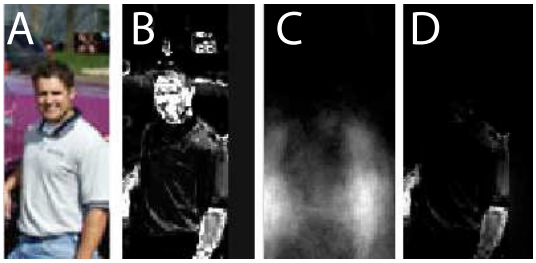

Figure 6. Computing skin-specific features. The skintone classifier is applied to the poselet activation patch (A) to obtain the skintone probability mask (B). The poselet part soft mask (C), in this case, a mask for the hands, is used to modulate the skintone mask and the result is shown in (D). While for this poselet the positions of the hands vary, as evidenced by the widespread hands mask, we are still able to exclude most non-hand skin areas. The hand-skin feature is the fraction of skin pixels in the modulated mask (D). This feature is especially useful for determining if a person wears short or long sleeves.

For the skin-specific features we trained a skin classifier, which is a GMM with 5 components fit from the LAB-transformed patches of skin collected from various skin tones and illuminations. We use three skin features: hands-skin, legs-skin and neck-skin. Each feature is the fraction of skin pixels in the corresponding part. Figure 6 describes how the feature is computed using the hand-skin feature of an upper-body-torso poselet as an example.

7. Classifiers

7.1 Poselet-level attribute classifier r_j^i We train a separate classifier for each of the 1200 poselet types i and for each attribute j. We used the 2003 training images for training these classifiers.

We construct a feature vector from all activations of poselet i on the training set. The label of a given activation is the label associated with the ground truth to which the poselet activation is assigned. We discard any activations on people that don't have a label for the given attribute. Figure 2(right) shows instances of positive (top row) and negative (bottom row) examples for the frontal face poselet and the "has-hat" attribute.

Some attributes have associated parts and poselets in which these parts don't appear are excluded from training of the attribute. For example, as shown on Figure 4 it doesn't make sense to use a legs poselet to train the "has-hat" attribute. To determine if a poselet covers a given part, we check to see if its mask (Figure 5) has presence of that part. This spatial selection reduces the dimensionality of our person-level attribute classifiers and the opportunity for overfitting.

Our classifiers are linear SVMs trained with weighted

examples. The weight of each training example is the probability of the corresponding poselet activation q^i.

7.2 Person-level attribute classifier s_j The person-level attribute classifier for attribute j combines all poselet-level classifiers for the given attribute. The feature vector has one dimension for each poselet type. Our features are zero-centered responses of the poselet-level attribute classifiers, see Equation 2. Our classifier is similar to a linear SVM, except we impose positivity constraints on the weights[1]. Since the input of the classifier is trained on the training set, we use the validation images to train the person-level attribute classifier.

7.3 Context-level attribute classifier S_j There are strong correlations among various attributes: long hair is correlated with gender, short sleeves are correlated with short pants, etc. Other attributes are especially helpful when direct evidence for the attribute is non-salient. We use an SVM with a quadratic kernel for each attribute. The features are the scores of all person-level attribute classifiers for a given person. We trained the context-level classifier on the training + validation sets.

8. Experimental Results

The highest/lowest scoring examples for each attribute on the test set are shown on Figure 7 and the most confused examples are on Figure 8.

8.1. Performance *vs*. baselines

To validate the design choices of our approach we tested the effect of disabling portions of our model. Table 2, columns 7-9 show the effect of disabling the skin features and the context classifier. As expected, skin features are essential for clothes-style attributes (the bottom five on Table 2) and without skin their mean AP drops from 63.18 to 55.10. The other attributes, such as gender and hairstyle are largely unaffected by skin. The context classifiers help on seven of the attributes and decrease performance on two, boosting the overall mean AP from 61.5 to 65.2.

Our baseline method uses Canny-modulated Histogram of Oriented Gradients [2] with Spatial Pyramid Matching kernel [21] which is effective for image classification in Caltech-101 as well as gender classification on MIT pedestrians [7]. The results of training it on the full bounds of the person are in column 6 of Table 2. We handily outperform SPM across all attributes with a mean AP of 65.18 *vs*. 45.91 for the SPM. We believe this is partly due to the fact that the generic spatial model used in the SPM is insufficient and the implicit pose-specific alignment provided by the poselets is necessary. Our examples have large degree of articulation and a generic classifier would suffer from lo-

[1] A negative weight would mean that the SVM takes the opposite of the advice of the poselet-level classifier, which could only happen due to overfitting so we prevent it explicitly.

Figure 7. The six highest and lowest scoring examples of each attribute on our test set. Of the 108 examples, five are classified incorrectly and marked with an X in the upper right corner. Three of them are women wearing hats misclassified as men. The gender attribute is the only one negatively affected by the context classifier and the effect applies only for the lowest recall mode, shown here.

Figure 8. Examples of most confused attributes. Many of the most confused males have long hair and the most confused females hide their hair under a hat. Results are affected by incorrect ground truth labels (has t-shirt, has-shorts), occlusion (has-jeans), and confusion with another person (has-shorts, not has-long-pants).

Figure 9. To help with localization, we provide our baselines the full bounds (left), as well as zoomed and aligned views of the head, upper body and lower body.

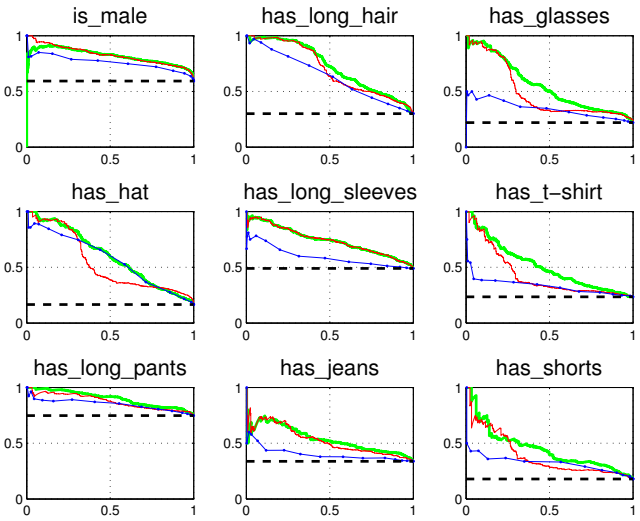

Figure 10. Precision-recall curves of the attribute classifiers on the test set. Our full result (column 9 in Table 2) is shown in thick green. Our performance without context classifiers (column 7) is shown in red; the SPM using the optimal view per attribute (max of columns 3-6) is shown in blue and the frequency of the label (column 2) is the dashed black horizontal line.

calization errors, especially for location-sensitive attributes such as has-glasses. To help SPM with localization we extracted higher resolution views of the people, zoomed on the head, upper body and lower body (Figure 9). Columns 3-5 on Table 2 show the results of using an SPM trained on each of the zoomed views. As expected, the head zoom improves detection of gender, hairstyle, presence of glasses and a hat. However, even if we used the best view for each attribute, we would get a mean AP of 51.87, which, despite the extra supervision, remains substantially lower than our AP of 65.18.

8.2. Performance from different viewpoints

As the examples on Figure 7 show, the classifiers are most confident for people facing the camera. To test the robustness of our method to different viewpoints we partitioned the test set into three partitions – frontal, profile and back-facing people and we tested the performance for each view. To automatically partition the data we made use of the keypoint annotations that come with our images. Specifically, images for which both eyes are present are treated as frontal; if only one eye is present the image is treated as a profile, and if no eyes are present, and the left shoulder is to the left of the right shoulder we assign the image to the back-facing category. Approximately 61% of our test data consists of frontal images, 18% is profile images and 11% is back-facing people. Around 9% of the data did not fall into any of these categories. In some cases this is due to missing annotation data, and in other cases the head is not visible. Table 3 shows the average precision of the attributes on all the data and on each partition. As expected, performance is highest for frontal examples, followed by back-facing and then profile examples.

Attribute	All	Frontal	Profile	Back
is male	82.4	82.9	82.9	83.2
has long hair	72.5	81.3	31.3	47.2
has glasses	55.6	59.8	33.9	18.8
has hat	60.1	66.4	54.8	41.9
has long sleeves	74.2	76.1	70.6	75.1
has t-shirt	51.2	55.7	43.3	46.7
has long pants	90.3	89.9	92.9	94.2
has jeans	54.7	53.0	46.9	70.0
has shorts	45.5	47.8	48.6	45.3
Mean AP	65.18	68.11	56.12	58.05
Num. examples	4022	2449	736	459

Table 3. Average precision for the attributes using all test annotations as well as using frontal-only, profile-only and back facing-only ones. The has-glasses attribute is most affected by the head orientation, and it drops to chance level for the back-facing case.

8.3. Optimal places to look for an attribute

It is not obvious exactly which part of the image is most discriminative for a given attribute. Consider the attribute has-long-hair. Clearly we should look at the face, but what is the optimal zoom level and pose? What if the person is in a profile or back-facing view? Our method automatically determines the optimal location, scale and viewpoint to look for evidence for a given attribute. This is a function of both the frequency of the given pose in the training set and the ease of discrimination given the pose. Specifically, the person-level classifier ranks each poselet type according to its predictive power. Figure 11 shows the top five poselets used for determining the gender, hair length and presence of glasses. Since more than half of the people in our training set are facing the camera, and frontal view is usually more discriminative, the top poselets all come from frontal view.

8.4. Gender recognition performance

Comparison with other methods is challenging because the vast majority of person-specific attribute classification methods operate on frontal faces only [19, 24, 25]. If we applied our method on their datasets, our three-level hierarchy would reduce to a single frontal poselet and the comparison will reduce to the effectiveness of HOG features for gender classification, a problem that is interesting but not directly relevant to our work[2]. In addition, other methods use different attributes, with the exception of gender.

Fortunately we have access to the Cognitec face recognizer, which is the winner of FRVT 2002 and one of the leading commercial face recognizers according to MBE

[2]Our skin features are only useful for attributes not visible from the frontal face

Attribute(1)	Freq(2)	Spatial Pyramid Match				Our Method			Cognitec(10)
		Head(3)	Lower(4)	Upper(5)	BBox(6)	No ctxt(7)	No skin(8)	Full(9)	
is male	59.3	74.9	63.9	71.3	68.1	82.9	82.5	82.4	75.0
has long hair	30.0	60.1	34.0	45.2	40.0	70.0	73.2	72.5	
has glasses	22.0	33.4	22.6	25.5	25.9	48.9	56.1	55.6	
has hat	16.6	53.0	24.3	32.3	35.3	53.7	60.3	60.1	
has t-shirt	23.5	32.2	25.4	30.0	30.6	43.0	48.4	51.2	
has long sleeves	49.0	53.4	52.1	56.6	58.0	74.3	66.3	74.2	
has shorts	17.9	22.9	24.8	22.9	31.4	39.2	33.0	45.5	
has jeans	33.8	38.5	38.5	34.6	39.5	53.3	42.8	54.7	
long pants	74.7	79.9	80.4	76.9	84.3	87.8	85.0	90.3	
Mean AP	36.31	49.81	40.66	43.94	45.91	61.46	60.84	**65.18**	

Table 2. Average precision of baselines relative to our model. **Freq** is the label frequency. We trained separate SPM models on the head (**Head**), lower body (**Lower**), upper body (**Upper**) and full bounding box (**BBox**) as shown on Figure 9. We tested our method by disabling the skin features (**No skin**), the context classifiers (**No ctxt**) and on the full system (**Full**). **Cognitec** is the gender recognition results using the Cognitec engine.

Figure 11. Our algorithm automatically determines the optimal poses and viewpoints to look for evidence of a given attribute. **First row:** The top poselets for is-male. **Second row:** The top poselets for has-long-hair. **Third row:** The top poselets for has-glasses. These three attributes require progressively higher zoom, which is reflected in the choice of poselets. The poselets are drawn by averaging their top ten training examples.

2010, the latest NIST test[3]. Cognitec can also report gender. As with other methods, it operates on frontal faces only. The Cognitec API does not allow for training of gender, so we could not train it on our training set. For optimal performance, we applied the engine on the zoomed head views (Figure 9b). Cognitec failed to find the face in 38.0% of the images (not all of them have frontal faces) and it failed to predict gender of another 20.0%. If we use mean score for the missing predictions we get AP of 75.0% for Cognitec *vs*. our AP of 82.4%. The precision-recall curve is shown on Figure 13. If we restrict the test to the faces for which Cognitec predicts gender, we get AP of 83.72% for Cognitec and 83.74% for our method, essentially equal, even though we aid Cognitec by providing a zoomed centered view of the head. Note that we use simple HOG features

[3]http://www.cognitec-systems.de/FaceVACS-Performance.23.0.html

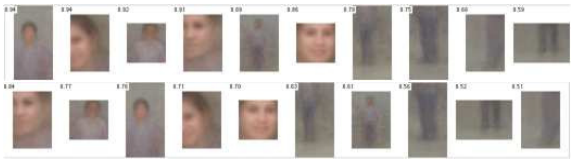

Figure 12. The poselets that performed best (left) to worst (right) for people (top row) and the computer algorithm (bottom).

and linear SVMs and Cognitec uses careful alignment and advanced proprietary biometric analysis. We believe that our method benefits from the power of combining many view-dependent poselet classifiers.

We don't have access to other leading methods, such as [19], but we can give an upper bound to their performance since they all require frontal faces. In our dataset 60.9% of the faces are frontal (i.e. have both eyes visible). If other methods use a perfect face detector, perfect alignment and perfect recognition for frontal faces and perform at chance level for other cases, their AP would be 60.9*1 + 39.1*0.5 = 80.5 *vs*. our AP of 82.4.

8.5. Comparisons to human visual system

Are the cues used by humans similar to the ones exploited by our system? To help answer this question we conducted an experiment using 10 representative poselets chosen to cover various parts of the body at various zoom levels. For each poselet we picked 100 examples, 50 male and 50 female. We flashed a random poselet example for an average of 200ms followed by a random image and asked each of the 8 subjects to immediately choose the gender of the example. We then sorted the 10 poselets using their mean AP averaged over all subjects, and we also sorted them according to their AP of discriminating gender in our system. The results are shown on Figure 12. The figure shows that there is a strong correlation between poselets preferred by humans and those preferred by our system.

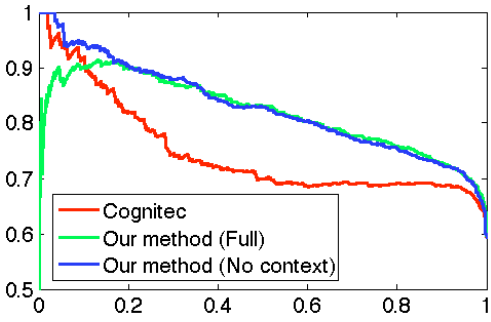

Figure 13. Precision-recall curves on gender recognition using our full method (AP=82.4), our method without context classifiers (AP=82.9) and Cognitec (AP=75.0).

Figure 14. Given a picture of a person our method can generate a natural language description.

8.6. Describing people

We have a simple extension that takes the predicted attributes and generates a natural language description of the person (Figure 14). If the confidence is low, it skips an attribute (or uses "person" instead of "man" or "woman").

9. Conclusion

We are the first to address an important but challenging problem with many practical applications - attribute classification of people "in the wild". Our solution is simple and effective. It is robust to partial occlusion, articulation and camera view. It draws cues from any part of the body at any scale and it leverages the power of alignment without explicitly inferring the pose of the person. While we have demonstrated the technique using nine attributes of people, it trivially extends to other attributes and other visual categories. We provide a large dataset of 8035 people annotated with 9 attributes, which we hope will inspire others to follow with better methods.

References

[1] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 2

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *ACM ICIVR*, 2007. 5

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010. 2, 3

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 2, 3

[5] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR11*. 2

[6] L. Cao, M. Dikmen, Y. Fu, and T. Huang. Gender recognition from body. In *ACM*, 2008. 2

[7] M. Collins, J. Zhang, P. Miller, and H. Wang. Full body image feature representations for gender profiling. In *ICCV Workshop*, 2010. 5

[8] G. W. Cottrell and J. Metcalfe. Empath: face, emotion, and gender recognition using holons. In *NIPS*, 1990. 2

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 2, 2005. 4

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge (2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 2

[11] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010. 2

[12] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. *CVPR*, 2009. 2

[13] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 1

[14] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007. 2

[15] A. Gallagher and T. Chen. Estimating age, gender and identity using first name priors. In *CVPR*, 2008. 2

[16] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *NIPS*, 1990. 2

[17] 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, 2007. 2

[18] N. Kumar, P. N. Belhumeur, and S. K. Nayar. FaceTracer: A Search Engine for Large Collections of Images with Faces. In *ECCV08*. 2

[19] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV*, 2009. 2, 6, 7

[20] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*. IEEE, 2006. 5

[22] S. Maji. Large scale image annotations on amazon mechanical turk. Technical Report UCB/EECS-2011-79, EECS Department, University of California, Berkeley, 2011. 3

[23] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 2

[24] B. Moghaddam and M. hsuan Yang. Learning gender with support faces. *IEEE TPAMI*, 24, 2002. 2, 6

[25] G. Shakhnarovich, P. A. Viola, and B. Moghaddam. A unified learning framework for real time face detection and classification. In *FG*, 2002. 2, 6

[26] L. Sigal, D. J. Fleet, N. F. Troje, and M. Livne. Human attributes from 3d pose tracking. In *ECCV*, 2010. 2

[27] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009. 2

[28] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *BMVC09*. 2

[29] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, 2010. 2