PANDA: Pose Aligned Networks for Deep Attribute Modeling

Ning Zhang\textsuperscript{1,2}, Manohar Paluri\textsuperscript{1}, Marc’Aurelio Ranzato\textsuperscript{1}, Trevor Darrell\textsuperscript{2}, Lubomir Bourdev\textsuperscript{1}

\textsuperscript{1}Facebook AI Research \hspace{1cm} \textsuperscript{2}EECS, UC Berkeley

\{nzhang, trevor\}@eecs.berkeley.edu \hspace{1cm} \{mano, ranzato, lubomir\}@fb.com

Abstract

We propose a method for inferring human attributes (such as gender, hair style, clothes style, expression, action) from images of people under large variation of viewpoint, pose, appearance, articulation and occlusion. Convolutional Neural Nets (CNN) have been shown to perform very well on large scale object recognition problems [15]. In the context of attribute classification, however, the signal is often subtle and it may cover only a small part of the image, while the image is dominated by the effects of pose and viewpoint. Discounting for pose variation would require training on very large labeled datasets which are not presently available. Part-based models, such as poselets [4] and DPM [12] have been shown to perform well for this problem but they are limited by shallow low-level features. We propose a new method which combines part-based models and deep learning by training pose-normalized CNNs. We show substantial improvement vs. state-of-the-art methods on challenging attribute classification tasks in unconstrained settings. Experiments confirm that our method outperforms both the best part-based methods on this problem and conventional CNNs trained on the full bounding box of the person.

1. Introduction

Recognizing human attributes, such as gender, age, hair style, and clothing style, has many applications, such as facial verification, visual search and tagging suggestions. This is, however, a challenging task when dealing with non-frontal facing images, low image quality, occlusion, and pose variations. The signal associated with some attributes is subtle and the image is dominated by the effects of pose and viewpoint. For example, consider the problem of detecting whether a person wears glasses. The signal (glasses wireframe) is weak at the scale of the full person and the appearance varies significantly with the head pose, frame design and occlusion by the hair. Therefore, localizing object parts and establishing their correspondences with model parts can be key to accurately predicting the underlying attributes.

Deep learning methods, and in particular convolutional nets [20], have achieved very good performance on several tasks, from generic object recognition [15] to pedestrian detection [25] and image denoising [6]. Moreover, Donahue et al. [8] show that features extracted from the deep convolutional network trained on large datasets are generic and can help in other visual recognition problems. However, as we report below, they may underperform compared to conventional methods which exploit explicit pose or part-based normalization. We conjecture that available training data, even ImageNet-scale, is presently insufficient for learning pose normalization in a CNN, and propose a new class of deep architectures which explicitly incorporate such representations. We combine a part-based representation with convolutional nets in order to obtain the benefit of both approaches. By decomposing the input image into parts that are pose-specific we make the subsequent training of convolutional nets drastically easier, and therefore, we can learn very powerful pose-normalized features from relatively small datasets.

Part-based methods have gained significant recent attention as a method to deal with pose variation and are the state-of-the-art method for attribute prediction today. For example, spatial pyramid matching [18] incorporates geometric correspondence and spatial correlation for object recognition and scene classification. The DPM model [12] uses a mixture of components with root filter and part filters capturing viewpoint and pose variations. Zhang et al. proposed deformable part descriptors [27], using DPM part boxes as the building block for pose-normalized representations for fine-grained categorization task. Poselets [5, 3] are part detectors trained on positive examples clustered using keypoint annotations; they capture a salient pattern at a specific viewpoint and pose. Several approaches [11, 26] have used poselets as a part localization scheme for fine-grained categorization tasks which are related to attribute prediction. Although part-based methods have been successful on several tasks, they have been limited by the choice of the low-level features applied to the image patches.

In this paper, we propose the PANDA model, Pose
Alignment Networks for Deep Attribute modeling, which augments deep convolutional networks to have input layers based on semantically aligned part patches. Our model learns features that are specific to a certain part under a certain pose. We then combine the features produced by many such networks and construct a pose-normalized deep representation using poselets. Our method can use other parts and we show the performance using DPM [12] as well. We demonstrate the effectiveness of PANDA on attribute classification problems and present state-of-the-art experimental results on four datasets, an attribute dataset from the web, the Berkeley Attributes of People Dataset [4], the Labeled Faces in the Wild dataset [16], and a large-scale gender recognition dataset.

2. Related work

2.1. Attribute classification

Attributes are used as an intermediate representation for knowledge transfer in [17, 10] for object recognition tasks. By representing the image as a list of human selected attributes they recognize unseen objects with few or zero training examples. Other related work on attributes includes that by Parikh et al. [22] exploring the relative strength of attributes by learning a rank function for each attribute, which can be applied to zero-shot learning as well as to generate richer textual descriptions. There is also some related work in automatic attribute discovery: Berg et al. [1] proposed automatic attribute vocabularies discovery by mining unlabeled text and image data sampled from the web. Duan et al. [9] proposed an interactive crowd-sourcing method to discover both localized and discriminative attributes to differentiate bird species.

In [16], facial attributes such as gender, mouth shape, facial expression, are learned for face verification and image search tasks. Some of the attributes used by them are similar to what we evaluate in this work. However, all of their attributes are about human faces and most of images in their dataset are frontal face subjects while our datasets are more challenging in terms of image quality and pose variations.

A very closely related work on attribute prediction is Bourdev et al. [4], which is a three-layer feed forward classification system and the first layer predicts each attribute value for each poselet type. All the predicted scores of first layer are combined as a second layer attribute classifier and the correlations between attributes are leveraged in the third layer. Our method is also built on poselets, from which the part correspondence is obtained to generate a pose-normalized representation.

2.2. Deep learning

The most popular deep learning method for vision, namely the convolutional neural network (CNN), has been pioneered by LeCun and collaborators [20] who initially applied it to OCR [21] and later to generic object recognition tasks [13]. As more labeled data and computational power
has become recently available, convolutional nets have be-

come the most accurate method for generic object category
classification [15] and pedestrian detection [25].

Although very successful when provided very large la-

beled datasets, convolutional nets usually generalize poorly
on smaller datasets because they require the estimation of
millions of parameters. This issue has been addressed
by using unsupervised learning methods leveraging large
amounts of unlabeled data [23, 13, 19]. In this work, we
take instead a different perspective: we make the learning
task easier by providing the network with pose-normalized
inputs.

While there has already been some work on using deep
learning methods for attribute prediction [7], we explore al-
ternative ways to predict attributes, we incorporate the use
of poselets in the deep learning framework and we per-
form a more extensive empirical validation which compares
against conventional baselines and deep CNNs evaluated on
the whole person region.

3. Pose Aligned Networks for Deep Attribute
modeling (PANDA)

We explore part-based models, specifically poselets, and
deep learning, to obtain pose-normalized representations
for attribute classification tasks. Our goal is to use pose-
lets for part localization and incorporate these normalized
parts into deep convolutional nets in order to extract pose-
normalized representations. Towards this goal, we lever-
age both the power of convolutional nets for learning dis-

criminative features from data and the ability of poselets to
simplify the learning task by decomposing the objects into
their canonical poses. We develop Pose Aligned Networks
for Deep Attribute modeling (PANDA), which incorporates
part-based and whole-person deep representations.

While convolutional nets have been successfully applied
to large scale object recognition tasks, they do not general-
ize well when trained on small datasets. Our setup requires
fewer training instances because we are able to augment the
training set size – we consider each poselet activation as a
separate training example.

Specifically, we start from poselet patches, resize them to
64x64 pixels (Figure 3), randomly jitter each patch and flip
it horizontally with probability 0.5 to improve generaliza-
tion, and train a CNN for each poselet. The overall convo-
lutional net architecture is shown in Figure 2. The network
consists of four convolutional, max pooling, local response
normalization layers followed by a fully connected layer
with 576 hidden units. After that, the network branches out
one fully connected layer with 128 hidden units for each at-
tribute and each of the branch outputs a binary classifier of
the attribute. The last two layers are split to let the network
develop customized features for each attribute (e.g., detect-
ing whether a person wears a “dress” or “sunglasses” pre-


Figure 3: Poselet Input Patches from Berkeley Attributes
of People Dataset. For each poselet, we use the detected
patches to train a convolution neural net. Here are some ex-
amples of input poselet patches and we are showing poselet
patches with high scores for poselet 1,16 and 79.

sumably requires different features) while the bottom layers
are shared to a) reduce the number of parameters and b) to
leverage common low-level structure.

The whole network is trained jointly by standard back-
propagation of the error [24] and stochastic gradient de-

scent [2] using as a loss function the sum of the log-losses
of each attribute for each training sample. The details of the
layers are given in Figure 2 and further implementation de-
tails can be found in [15]. To deal with noise and inaccurate
poselet detections, we train on patches with high poselet
detection scores and then we gradually add more low confi-
dence patches.

Different parts of the body may have different signals for
each of the attributes and sometimes signals coming from
one part cannot infer certain attributes accurately. For ex-
ample, deep net trained on person leg patches contains little
information about whether the person wears a hat. There-
fore, we first use deep convolutional nets to generate dis-

criminative image representations for each part separately
and then we combine these representations for the final
classification. Specifically, we extract the activations from
fc_attr layer in Figure 2, which is 576 dimensional, for the
CNN at each poselet, and concatenate the activations of all
poselets together into 576*150 dimensional feature. If a
poselet does not activate for the image, we simply leave the
feature representation to zero.

The part-based deep representation mentioned above
leverages both the discriminative deep convolutional
features and part correspondence. However, poselet detected
parts may not always cover the whole image region and in
Figure 2: **Part-based Convolutional Neural Nets.** For each poselet, one convolutional neural net is trained on patches resized 64x64. The network consists of 4 stages of convolution/pooling/normalization and followed by a fully connected layer. Then, it branches out one fully connected layer with 128 hidden units for each attribute. We concatenate the activation from fc_attr from each poselet network to obtain the pose-normalized representation. The details of filter size, number of filters we used are depicted above.

some degenerate cases, images may have few poselets detected. To deal with that, we also incorporate a deep network covering the whole-person bounding box region as input to our final pose-normalized representation.

Based on our experiments, we find a more complex net is needed for the whole-person region than for the part regions. We extract deep convolutional features from the model trained on Imagenet [15] using the open source package provided by [8] as our deep representation of the full image patch.

As shown in Figure 1, we concatenate the features from the deep representations of the full image patch and the 150 parts and train a linear SVM for each attribute.

### 4. Datasets

#### 4.1. The Berkeley Human Attributes Dataset

We tested our method on the Berkeley Human Attributes Dataset [4]. This dataset consists of 4013 training, and 4022 test images collected from PASCAL and H3D datasets. The dataset is challenging as it includes people with wide variation in pose, viewpoint and occlusion. About 60% of the photos have both eyes visible, so many existing attributes methods that work on frontal faces will not do well on this dataset.

#### 4.2. Attributes 25K Dataset

Unfortunately the training portion of the Berkeley dataset is not large enough for training our deep-net models (they severely overfit when trained just on these images). We collected an additional dataset from Facebook of 24963 people split into 8737 training, 8737 validation and 7489 test examples. We made sure the images do not intersect those in the Berkeley dataset. The statistics of the images are similar, with large variation in viewpoint, pose and occlusions.

We train on our large training set and report results on both the corresponding test set and the Berkeley Attributes test set. We chose to use a subset of the categories from the Berkeley dataset and add a few additional categories. This will allow us to explore the transfer-learning ability of our system.

Not every attribute can be inferred from every image. For example, if the head of the person is not visible, we cannot enter a label for the “wears hat” category. The statistics of ground truth labels are shown on Figure 4.

### 5. Results

In this section, we will present a comparative performance evaluation of the proposed method.

#### 5.1. Results on the Berkeley Attributes of People Dataset

On Table 1 we show the results on applying our system on the publicly available Berkeley Attributes of People dataset. We compare against Poselets, DPD and Joo et al. on that dataset as reported by [4], [27] and [14] respectively. For our method, PANDA, we use Attributes25K train+val dataset to train the poselet-level CNNs of our system, and
Table 1: Attribute classification results on the Berkeley Attributes of People Dataset as compared to the methods of Bourdev et al. [4] and Zhang et al. [27].

<table>
<thead>
<tr>
<th>Attribute</th>
<th>male</th>
<th>long hair</th>
<th>glasses</th>
<th>hat</th>
<th>tshirt</th>
<th>longsleeves</th>
<th>shorts</th>
<th>jeans</th>
<th>long pants</th>
<th>Mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poselets[4]</td>
<td>82.4</td>
<td>72.5</td>
<td>55.6</td>
<td>60.1</td>
<td>51.2</td>
<td>74.2</td>
<td>45.5</td>
<td>54.7</td>
<td>90.3</td>
<td>65.18</td>
</tr>
<tr>
<td>DPD[27]</td>
<td>83.7</td>
<td>70.0</td>
<td>38.1</td>
<td>73.4</td>
<td>49.8</td>
<td>78.1</td>
<td>64.1</td>
<td>78.1</td>
<td>93.5</td>
<td>69.88</td>
</tr>
<tr>
<td>Joo et al. [14]</td>
<td>88.0</td>
<td>80.1</td>
<td>56.0</td>
<td>75.4</td>
<td>53.5</td>
<td>75.2</td>
<td>47.6</td>
<td>69.3</td>
<td>91.1</td>
<td>70.7</td>
</tr>
<tr>
<td>PANDA</td>
<td>91.7</td>
<td>82.7</td>
<td>70.0</td>
<td>74.2</td>
<td>49.8</td>
<td>86.0</td>
<td>79.1</td>
<td>81.0</td>
<td>96.4</td>
<td>78.98</td>
</tr>
</tbody>
</table>

Table 2: Average Precision on the Attributes25K-test dataset.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>male</th>
<th>long hair</th>
<th>hat</th>
<th>glasses</th>
<th>dress</th>
<th>sunglasses</th>
<th>short sleeves</th>
<th>baby</th>
<th>mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poselets150[4]</td>
<td>86.00</td>
<td>75.31</td>
<td>29.03</td>
<td>36.72</td>
<td>34.73</td>
<td>50.16</td>
<td>55.25</td>
<td>41.26</td>
<td>51.06</td>
</tr>
<tr>
<td>DPD[27]</td>
<td>85.84</td>
<td>72.40</td>
<td>27.55</td>
<td>23.94</td>
<td>48.55</td>
<td>34.36</td>
<td>54.75</td>
<td>41.38</td>
<td>48.60</td>
</tr>
<tr>
<td>DeCAF [8]</td>
<td>82.47</td>
<td>65.03</td>
<td>19.15</td>
<td>14.91</td>
<td>44.68</td>
<td>26.91</td>
<td>56.40</td>
<td>50.19</td>
<td>44.97</td>
</tr>
<tr>
<td>DL-DPM</td>
<td>88.27</td>
<td>77.64</td>
<td>43.44</td>
<td>36.70</td>
<td>55.72</td>
<td>55.03</td>
<td>67.95</td>
<td>64.89</td>
<td>61.20</td>
</tr>
<tr>
<td>PANDA</td>
<td>94.10</td>
<td>83.17</td>
<td>39.52</td>
<td>72.25</td>
<td>59.41</td>
<td>66.62</td>
<td>72.09</td>
<td>78.76</td>
<td>70.74</td>
</tr>
</tbody>
</table>

we used validation set to train the SVM classifier.

As the table shows, our system outperforms all the prior methods across most attributes. In the case of t-shirt, [14] performs better, perhaps due to the fact that Attribute25K dataset doesn’t have wearing t-shirt attribute so that the part-level CNNs are not trained on that attribute.

Note that the attributes shorts, jeans and long pants are also not present in the Attributes25K dataset. We don’t have enough labeled data of this dataset to train the convolutional neural nets but the transfer learning is still effective.

We show some examples of highest and lowest scoring PANDA results for some attributes in Figure 6. Figure 5 shows the top failure cases for wearing t-shirts and having short hair on the test dataset. In the case of wearing t-shirt, the top failure cases are picking the sleeveless, which look very similar to tshirts. And for the case of short hair, some failures are due to the person having a pony tail or the long hair is occluded in the image.

5.2. Results on the Attributes25K Dataset

Table 2 shows results on the Attributes25K-test Dataset. Poselets150 shows the performance of our implementation of the three-layer feed-forward network proposed by [4]. Instead of the 1200 poselets in that paper we used the 150 publicly released poselets, and instead of multiple aspect ratios we use 64x64 patches. Our system underperforms [4] and on the Berkeley Attributes of People dataset yields mean AP of 60.6 vs 65.2, but it is faster and simpler and we have adopted the same setup for our CNN-based poselets. This allows us to make more meaningful comparisons between the two methods.

DPD and DeCAF We used the publicly available implementations of [27] based on deformable part models and [8] based on CNN trained on ImageNet.

DL-DPM shows the performance of using DPM parts instead of poselets. We used the pretrained DPM model in [27]. We extracted the patches associated with each of the 8 parts in each of the 6 mixture components for a total of 48 parts (only 8 of which are active at a time). We then used the same setup as PANDA – trained CNN classifiers for each of the 48 parts, combined them with the global model and trained SVM on top. As the table shows, DL-DPM outperforms conventional part-based methods (both DPM and poselets) which do not use deep features as well as DeCAF. However it does not match the performance of the deep poselets in PANDA. The patches from DPM parts have higher noise (because they have to fire even if the pattern is weak or non-existent) and are not well aligned (because they have to satisfy global location constraints).

5.3. Component Evaluation

We now explore the performance of individual components of our system as shown on Table 3 using the Berkeley dataset. Our goal is to get insights into the importance of using deep learning and the importance of using parts.

How well does a conventional deep learning classifier perform? We first explore a simple model of feeding the raw RGB image of the person into a deep network. To help with rough alignment and get signal from two resolutions we split the images into four 64x64 patches – one from the top, center, and bottom part of the person’s bounds, and one from the full bounding box at half the resolution. In total we have 4 concatenated 64x64 square color images as input (12 channels). We train a CNN on this 12x64x64 input on the full Attributes-25K dataset. The structure we used is similar to the CNN in Figure 2 and it consists of two convolution/normalization/pooling stages, followed by a fully connected layer with 512 hidden units followed by nine columns, each composed of one hidden layer with 128
Table 3: Relative performance of baselines and components of our system on the Berkeley Attributes of People test set.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>male</th>
<th>long hair</th>
<th>glasses</th>
<th>hat</th>
<th>tshirt</th>
<th>longsleeves</th>
<th>short</th>
<th>jeans</th>
<th>long pants</th>
<th>Mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL-Pure</td>
<td>80.65</td>
<td>63.23</td>
<td>30.74</td>
<td>57.21</td>
<td>37.99</td>
<td>71.76</td>
<td>35.05</td>
<td>60.18</td>
<td>86.17</td>
<td>58.11</td>
</tr>
<tr>
<td>DeCAF</td>
<td>79.64</td>
<td>62.29</td>
<td>31.29</td>
<td>55.17</td>
<td>41.84</td>
<td>78.77</td>
<td>80.66</td>
<td>81.46</td>
<td>96.32</td>
<td>67.49</td>
</tr>
<tr>
<td>Poselets150 L2</td>
<td>81.70</td>
<td>67.07</td>
<td>44.24</td>
<td>54.01</td>
<td>42.16</td>
<td>71.70</td>
<td>36.71</td>
<td>42.56</td>
<td>87.82</td>
<td>58.62</td>
</tr>
<tr>
<td>DLPoselets</td>
<td>92.10</td>
<td>82.26</td>
<td>76.25</td>
<td>65.55</td>
<td>44.83</td>
<td>77.31</td>
<td>43.71</td>
<td>52.52</td>
<td>87.82</td>
<td>69.15</td>
</tr>
<tr>
<td>PANDA</td>
<td>91.66</td>
<td>82.70</td>
<td>69.95</td>
<td>74.22</td>
<td>49.84</td>
<td>86.01</td>
<td>79.08</td>
<td>80.99</td>
<td>96.37</td>
<td>78.98</td>
</tr>
</tbody>
</table>

Table 4: Performance of PANDA on front-facing, profile-facing and back-facing examples of the Berkeley Attributes of People test set.

<table>
<thead>
<tr>
<th>Partition</th>
<th>male</th>
<th>long hair</th>
<th>glasses</th>
<th>hat</th>
<th>tshirt</th>
<th>longsleeves</th>
<th>shorts</th>
<th>jeans</th>
<th>long pants</th>
<th>Mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal</td>
<td>92.55</td>
<td>88.40</td>
<td>77.09</td>
<td>74.40</td>
<td>51.69</td>
<td>86.84</td>
<td>78.00</td>
<td>79.63</td>
<td>95.70</td>
<td>80.47</td>
</tr>
<tr>
<td>Profile</td>
<td>91.42</td>
<td>59.38</td>
<td>37.06</td>
<td>69.47</td>
<td>49.02</td>
<td>84.61</td>
<td>85.57</td>
<td>82.71</td>
<td>98.10</td>
<td>73.04</td>
</tr>
<tr>
<td>Back-facing</td>
<td>88.65</td>
<td>63.77</td>
<td>72.61</td>
<td>72.19</td>
<td>55.20</td>
<td>84.32</td>
<td>74.01</td>
<td>86.12</td>
<td>96.68</td>
<td>77.06</td>
</tr>
<tr>
<td>All</td>
<td>91.66</td>
<td>82.70</td>
<td>69.95</td>
<td>74.22</td>
<td>49.84</td>
<td>86.01</td>
<td>79.08</td>
<td>80.99</td>
<td>96.37</td>
<td>78.98</td>
</tr>
</tbody>
</table>

hidden units. Each of the 9 branches outputs a single value which is a binary classifier of the attribute. We then use the CNN as a feature extractor on the validation set by using the features produced by the final fully connected layer. We train a logistic regression using these features and report its performance on the ICCV test set as DL-Pure on Table 3.

We also show the results of our second baseline – DeCAF, which is the global component of our system. Even though it is a convolutional neural net originally trained on a completely different problem (ImageNet classification), it has been exposed to millions of images and it outperforms DL-Pure.

How important is deep learning at the part level?. By comparing the results of Poselets150L2 and DLPoselets we can see the effect of deep learning at the part level. Both methods use the same poselets, train poselet-level attribute classifiers and combine them at the person level with a linear SVM. The only difference is that Poselets150L2 uses the features as described in [4] (HOG features, color histogram, skin tone and part masks) whereas DLPoselets uses features trained with a convolutional neural net applied to the poselet image patch. As our table shows, deep-net poselets result in increased performance.

PANDA shows the results of our proposed system which combines DeCAF and DLPoselets. Our part and holistic classifiers use complementary features and combining them together further boosts the performance.

5.4. Robustness to viewpoint variations

In Table 4, we show the performance of our method as a function of the viewpoint of the person. We considered as frontal any image in which both eyes of the person are visible, which includes approximately 60% of the dataset. Profile views are views in which one eye is visible and Back-facing are views where both eyes are not visible. As expected, our method performs best for front-facing people because they are most frequent in our training set. However, the figure shows that PANDA can work well across a wide range of viewpoints.

5.5. Results on the LFW Dataset

We also report results on the Labeled Faces in the Wild dataset [16]. The dataset consists of 13233 images of cropped, centered frontal faces. The registered nature of the data does not leverage the strengths of our system in its ability to deal with viewpoint, pose and partial occlusions. Nevertheless, it provides us another datapoint to compare against other methods. This dataset contains many attributes, but unfortunately the ground truth labels are not released. We used crowd-sourcing to collect ground-truth
(a) Highest scoring results for people wearing glasses.

(b) Highest scoring results for people wearing a hat.

(c) Highest scoring results for people wearing short pants.

(d) Lowest scoring results for men.

Figure 6: Examples of highest and lowest scoring PANDA results on Berkeley Attributes of People Dataset. The images are cropped so the required attribute can be seen better.

Table 5: Average precision of PANDA on the gender recognition of the LFW dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Gender AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simile [16]</td>
<td>95.52</td>
</tr>
<tr>
<td>FrontalFace poselet</td>
<td>96.43</td>
</tr>
<tr>
<td>PANDA</td>
<td>99.54</td>
</tr>
</tbody>
</table>

labels for the gender attribute only. We split the examples randomly into 3042 training and 10101 test examples with the only constraint that the same identity may not appear in both training and test sets. We used our system whose features were trained on Attribute-25K to extract features on the 3042 training examples. Then we trained a linear SVM and applied the classifier on the 10101 test examples. We also used the publicly available gender scores of [16] to compute the average precision of their system on the test subset. The results are shown on Table 5.

PANDA’s AP on LFW is 99.54% using our parts model, a marked improvement over the previous state of the art. Our manual examination of the results shows that roughly 1 in 200 test examples either had the wrong ground truth or we failed to match the detection results with the correct person. Thus PANDA achieves nearly perfect gender recognition performance in LFW and it shows that PANDA is effective even when images are tightly cropped and variation in pose is reduced.

One interesting observation is that, even though the dataset consists of only frontal-face people, the performance of our frontal-face poselet is significantly lower than the performance of the full system. This suggests that our system benefits from combining the signal from multiple redundant classifiers, each of which is trained on slightly different set of images.

5.6. Analysis of dataset size

We also investigate the effects of number of training examples for DL Pure and DL Poselet methods. It is interesting to see if holistic deep learning method trained on the whole bounding box image can deal with pose variations implicitly given a large amount of data. We collected a dataset for gender classification, consisting of ~90K training, ~2.6K validation and ~10K test examples of people from photo albums. The ground truth labels have about 1.5% noise. We trained on the full training set of 90K, and on subsets of 45K and 23.5K. The number of training examples for PANDA (poselet activations) are 5.6 million, 2.8 million and 1.4 million respectively. We followed the same pipeline by using the same set of poselets and same part level convolutional neural nets in the experiments above. The results are shown in Figure 7. The holistic model (DL Pure) has an almost linear improvement over the number of training examples while our pose aligned method outperforms the holistic method but having a smaller improvement as the training size increases. We would like to experiment...
on a larger dataset to see if those two methods can intersect in the future.

6. Conclusion

We presented a method for attribute classification of people that improves performance compared with previously published methods. It is conceptually simple and leverages the strength of convolutional neural nets without requiring datasets of millions of images. It uses poselets to factor out the pose and viewpoint variation which allows the convolutional network to focus on the pose-normalized appearance differences. We concatenate the deep features at each poselet and add a deep representation of the whole input image. Our feature representation is generic and we achieve state-of-the-art results on the Berkeley Attributes of People dataset and on LFW even if we train our CNNs on a different dataset. We believe that our proposed hybrid method using mid-level parts and deep learning classifiers at each part will prove effective not just for attribute classification, but also for problems such as detection, pose estimation, action recognition.

Acknowledgments The first and the fourth authors were supported in part by DARPA Mind’s Eye and MSEE programs, by NSF awards IIS-0905647, IIS-1134072, and IIS-1212798, and by support from Toyota.

References