Human Attribute Recognition by Deep Hierarchical Contexts

Yining Li(✉), Chen Huang, Chen Change Loy, and Xiaoou Tang

Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
{ly015,chuang,ccloy,xtang}@ie.cuhk.edu.hk

Abstract. We present an approach for recognizing human attributes in unconstrained settings. We train a Convolutional Neural Network (CNN) to select the most attribute-descriptive human parts from all poselet detections, and combine them with the whole body as a pose-normalized deep representation. We further improve by using *deep hierarchical contexts* ranging from human-centric level to scene level. Human-centric context captures human relations, which we compute from the nearest neighbor parts of other people on a pyramid of CNN feature maps. The matched parts are then average pooled and they act as a similarity regularization. To utilize the scene context, we re-score human-centric predictions by the global scene classification score jointly learned in our CNN, yielding final scene-aware predictions. To facilitate our study, a large-scale WIDER Attribute dataset(Dataset URL: http://mmlab.ie.cuhk.edu.hk/projects/WIDERAttribute) is introduced with human attribute and image event annotations, and our method surpasses competitive baselines on this dataset and other popular ones.

1 Introduction

Accurate recognition of human attributes such as gender and clothing style can benefit many applications such as person re-identification [1–4] in videos. However, this task still remains challenging in unconstrained settings where images of people exhibit large variation of viewpoint, pose, illumination and occlusion. Consider, for example, Fig. 1 where inferring the attributes "formal suits" and "sunglasses" from only the target person is very difficult, due to the occlusion and low image quality respectively. Fortunately, we have access to the *hierarchical* contexts—from the neighboring similar people to the global image scene wherein the target person appears. Leveraging such contextual cues makes attributes much more recognizable, *e.g.* being aware of a funeral event, we would be more confident about people wearing "formal suits". We build on this intuition to develop a robust method for unconstrained human attribute recognition.

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46466-4_41) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part VI, LNCS 9910, pp. 684–700, 2016.
DOI: 10.1007/978-3-319-46466-4_41

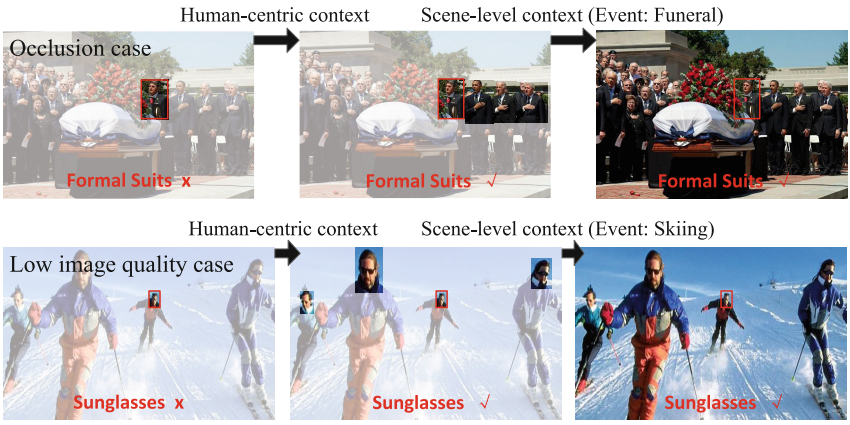

Fig. 1. WIDER Attribute - example images to motivate the use of hierarchical contexts for robust attribute recognition for the target person (red box): the human-centric context and scene-level context help resolve visual ambiguities due to occlusion and low image quality (low resolution/blurring). (Color figure online)

Our method is inspired by recent attribute models using parts, such as Poselets [5], Deformable Part Model (DPM) [6] and window-specific parts [7]. These methods are robust against pose and viewpoint variations. They are also capable of localizing attribute clues at varying scales (*e.g.* small glasses vs. the full body). State-of-the-art studies [8–10] improve by learning CNN features from detected part regions instead of using low-level features, and finally combine them into a pose-normalized deep representation for attribute recognition. The deep part features have also been used in [11,12] for fine-grained categorization tasks. Our method is based on deep parts too, inheriting the aforementioned benefits and end-to-end training.

Our major difference with respect to prior methods lies in the use of *deep hierarchical contexts*. The hierarchical contexts are called 'deep' because they are selected and represented in layers of our deeply trained model. Specifically, at the human-centric level, we compute the nearest neighbor fields between parts of the target person and contextual persons on a pyramid of CNN feature maps. Then we pool all the matched parts together as a multi-scale similarity regularization of our CNN, which proves effective in reducing attribute recognition ambiguities in challenging cases. At the scene level, we fuse our human-centric predictions with the global scene classification score that is jointly learned in CNN, which are finally mapped to scene-aware attribute predictions. We notice that R*CNN [10] also exploits context for human attribute recognition. But this model is limited to using unspecified and potentially insufficient contextual cues from only some bottom-up region proposal [13]. In contrast, we utilize semantically organized contexts from both related human parts and the entire image scene.

To facilitate the study of deep hierarchical contexts, we introduce a large-scale WIDER Attribute dataset with 14 human attribute labels and 30 event class

labels. The dataset consists of 57,524 labeled person bounding boxes in 13,789 event-labeled images collected from the WIDER dataset [14]. It is a new large-scale human attribute dataset with both human attribute and scene annotations. It has more human attribute labels than existing public datasets, *e.g.* Berkeley Attributes of People [5], HAT [15], CRP [16], PARSE-27k [17] and PETA [41].

Contributions: (1) We propose a novel deep model based on adaptively selected human parts for human attribute recognition from unconstrained images. (2) We propose to use deep hierarchical contexts to exploit joint context of humans and scene for more robust attribute recognition. (3) We introduce a large-scale WIDER Attribute dataset with rich human attribute and event class annotations. Our method obtains a mean AP of 81.3 % for all attributes on the test set of WIDER Attribute dataset, 0.8 % higher than the competing R*CNN [10] which suggests the usefulness of hierarchical contexts. Our method achieves state-of-the-art performance on other popular datasets too.

2 Related Work

Attribute Recognition. Attributes have been used as an intermediate representation to describe object properties [18,19] or even unseen object categories [20]. For attributes of people, early works rely on frontal faces only and predict a limited number of attributes. For example, Haar features extracted from the face region can be fed into the SVM [21] and AdaBoost [22] classifiers for gender and race recognition. Kumar *et al.* proposed using the predicted face attributes for face recognition [23] and visual search [24]. More recent works study the problem of recognizing a larger set of attributes, such as gender, hairstyle, and clothing style, from the whole human body image with large variation of viewpoint, pose, and appearance. Part-based methods are the state-of-the-art family of methods nowadays because they can decompose the input image into parts that are pose-specific and allow to combine evidence from different locations and scales. Successful part models include Poselets [5], DPM [6] and window-specific parts [7]. Recent deep part models [8–10] improve attribute recognition performance by training deep CNNs from part regions. Nevertheless, most part models only concern for the target person region, thus miss the opportunity to leverage rich contexts to reduce the attribute recognition ambiguities in challenging cases. An exception is R*CNN [10] that exploits context from adaptive region proposals. Our experiments will show that it is weaker than using hierarchical contextual cues from human-centric relations to global scene event.

Nearest-Neighbor Learning. One of our goals in using hierarchical contexts is to capture human relations or similarities by computing nearest neighbor parts between people. Finding nearest neighbors to define image similarities has a long history for vision tasks like image classification [25,26]. For bird sub-category classification, Zhang *et al.* [27] further proposed a similarity function for poselet neighbor matching. In comparison, our part matching is more adaptive, and is performed at online and multi-scale feature maps in a deep model. A recent deep

learning method [28] for multilabel image annotation also inherits the idea of nearest neighbor matching, and pools the neighbor features for robustness. Our method is related in neighbor pooling, but operates at score level and between different objects on the feature maps of one input image rather than between feature maps of different images, thus can be seen as a self-similarity regularization during CNN training.

Scene Contexts. Oliva and Torralba [29] have studied the role of context in object recognition, and analyzed a rich collection of contextual associations of objects with their scene. Object detection also frequently exploits scene contexts. DPM [30] re-scores each detected object by taking into account the scores of all other classes in an image, thus the object co-occurrences in scenes. Choi *et al.* [31] re-scored by learning the object co-occurrence and spatial priors in a hierarchical tree structure[1]. Mottaghi *et al.* [32] exploited the object class contexts in both local region and global scene. However, our exploited hierarchical contexts are not limited to only object class relations, but also cover the entire background scene. In this respect, our method is closely related to those works *e.g.* [33,34] that use global scene features. This line of work is attractive for not anchoring context analysis to any specific regions or individual objects within a scene, thus has a complete information coverage and also lower computation complexity than *e.g.* the local model [35] that needs to additionally compute an adaptive local contextual region. In our work, to prevent global scene context from enforcing strong properties on some less related objects, we only treat the global scene features as complementary signals and map them into scene classification scores in our CNN, conditioned on which we make probabilistic scene-aware predictions.

3 Human Attributes from Deep Hierarchical Contexts

The proposed human attribute recognition method is part-based, and learns pose-normalized deep feature representations from localized parts in a deep ConvNet. Combining the human parts and deep learning lends us robustness when dealing with unconstrained human images with large variation. We adapt Fast R-CNN [36] to process multiple regions, and the CNN feature maps and attribute scoring modules are trained end-to-end by back-propagation and stochastic gradient descent. This is in contrast to the deep methods in [8,9] that optimize an additional linear SVM for prediction. Figure 2 provides the overview of our network architecture.

Given an input image, each person in it is associated with one bounding box hypothesis and a set of human part detections. The input image and its Gaussian pyramid are passed through the network to obtain multi-scale convolutional feature maps. Then we branch out four attribute scoring paths using different

[1] The term 'Hierarchical Context' is used in this paper to denote the tree-structured organization of object classes in a scene. We use the same term but with a different meaning of (human) object-object and object-scene contextual relations at two semantic levels, which is also more complete in the coverage of image information.

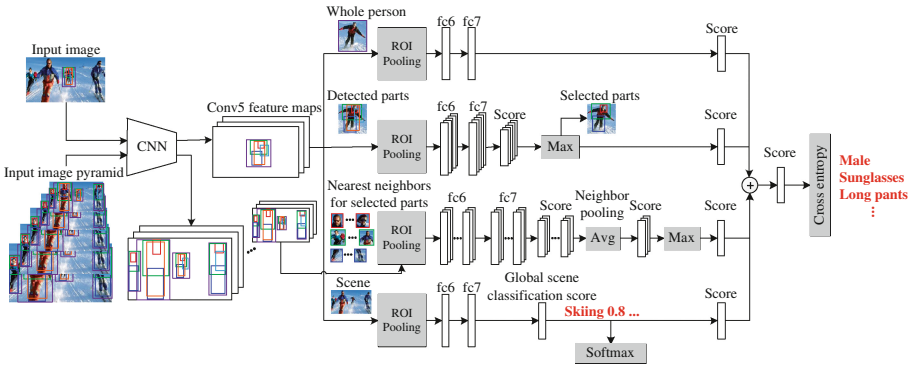

Fig. 2. Network architecture for unconstrained human attribute recognition using deep hierarchical contexts. Given an input image, we compute its Gaussian pyramid and pass them all through the CNN to obtain multi-scale feature maps. From the feature maps we extract features for four sets of bounding box regions: the whole target person, a target's chosen parts (for clarity we only show 3 selected out of 5 detected), nearest neighbor parts from the image pyramid and global image scene. The latter two correspond to hierarchical contexts: human-centric and scene-level contexts. After scoring the four region sets (see texts), we sum up all the scores as the final attribute score.

bounding box regions on the feature maps. On the first two paths, we respectively use the target person's bounding box and part boxes to cover the full scale body and local parts. This representation is widely adopted in many studies [8–10]. We incorporate deep hierarchical contexts on the third and fourth paths. Specifically, the human-centric context is selected from the nearest neighbor parts of other people on the deep feature pyramid; the scene-level context is mapped to the scene classification score as prior probability for re-scoring human attributes in a scene-aware manner. Details will be provided in the following subsections.

3.1 Preliminaries

We first describe backgrounds of the Fast R-CNN framework before delving into our model details. We choose the VGG16 [37] network pre-trained on ImageNet classification [38] for its excellent performance. Fast R-CNN follows the paradigm of generating region proposals first and then classifying them with learned features and classifiers. To ensure computational efficiency, the intense convolutions are only performed at image-level to obtain global *conv5* feature maps, which are reused for each region by ROI Pooling. The ROI Pooling layer functions by superimposing a fixed 7×7 spatial grid over a bounding box region, then performing max pooling within each grid cell to extract a fixed length feature vector from *conv5* feature maps. The feature vector is subsequently passed through two fully connected layers *fc6* and *fc7* to generate prediction scores.

Our task is human centric. We thus adopt the whole person bounding box and Poselet [39] detected regions as region proposals. To detect poselets, strong poselet activations are first localized on the input image in a multi-scale sliding

Fig. 3. Left: example poselet detections on the HAT [15] dataset. Right: example HAT image with only one bounding box annotation (red box) for a person. The yellow and green dashed boxes denote the detections of poselets and new people respectively. (Color figure online)

window fashion. Then following [40], we refine the activation scores by considering the spatial context of each. The refined poselet activations are finally clustered to form consistent person hypotheses. On datasets like HAT [15] and our proposed WIDER Attribute, there may not exist a bounding box annotation for every person in one image. So we associate to an unannotated person the person hypothesis and its related poselets when the hypothesis confidence is above a threshold [40]. We use these new detections to explore human-centric context. For those already annotated people with ground truth bounding boxes, we empirically associate to them the closest poselets whose person hypotheses sufficiently overlap (with IoU larger than 0.6) with them. If such poselets do not exist, we simply associate the nearby poselets that overlap with the ground truth bounding boxes by at least 50 %. Figure 3 shows example detections of poselets on an annotated person, and new detections of poselets and bounding boxes on unannotated people.

3.2 Enriching Human Cues with Deep Hierarchical Contexts

Our goal is to recognize a set of human attributes $\{a \in A\}$ for all the people in an unconstrained image I. Suppose for a target person's bounding box b in I, we have detected a set of parts $\{s \in S\}$. We frame the attribute recognition problem in a probabilistic framework, and estimate the likelihood of the presence of attribute a on the target person given a set of his/her measurements V. We take into account **Human Cues** from both human body b and parts S, and also contextual cues from the remaining background regions in I. Thus we consider measurements $V = \{b, S, N(s_a^*), I\}$, where $N(s_a^*)$ and I are the **Hierarchical Contexts** that will be detailed later.

We evaluate for each attribute a the conditional probability function given measurements V:

$$\begin{aligned} P(a \,|\, V) &= P(a \,|\, b, S, N(s_a^*), I) \\ &\propto P(b, S, N(s_a^*), I \,|\, a) = P(b \,|\, a) \cdot P(S \,|\, a) \cdot P(N(s_a^*) \,|\, a) \cdot P(I \,|\, a), \end{aligned} \quad (1)$$

where we assume uniform distribution for the prior probability $P(a)$ that is hence omitted, and assume conditional independence between different image measurements. Equation 1 can be equivalently solved in a log-linear model. We

implement this model in CNN by directly learning a score function $Score(a;\cdot)$ for each attribute a, and the learned score corresponds to the log probability after normalization. Then we can simply write the attribute score as the sum of four terms:

$$Score(a; b, S, N(s_a^*), I) = \underbrace{\boldsymbol{w}_{a,b}^T \cdot \boldsymbol{\phi}(b; I)}_{person\ bounding\ box} + \underbrace{\max_{s \in S} \boldsymbol{w}_{a,s}^T \cdot \boldsymbol{\phi}(s; I)}_{attribute-specific\ parts}$$
$$+ \underbrace{\frac{1}{|N(s_a^*)|} \sum_{s \in N(s_a^*)} \boldsymbol{w}_{a,s}^T \cdot \boldsymbol{\phi}(s; I)}_{human-centric\ context} + \underbrace{\boldsymbol{w}_{a,sc}^T \cdot \boldsymbol{W}_{sc} \cdot \boldsymbol{\phi}(I)}_{scene-level\ context}. \quad (2)$$

where $\boldsymbol{\phi}(b; I)$ is the extracted $fc7$ features from region b in image I, while $\boldsymbol{w}_{a,\cdot}$ are the scoring weights of attribute a for different regions.

The scoring terms for person bounding box b and parts $\{s \in S\}$ form the basis of our model, and are shown on the upper two paths in Fig. 2. Their sum can be regarded as a pose-normalized deep representation at score level. Such score fusion is found to be more effective than feature fusion (*e.g.* in [8]) in our task, because the latter would generate a very large feature vector from the many parts and overfits easily. In our CNN, the scoring weights and feature vectors are jointly learned for all attributes $a \in A$.

Note for the part set S, we select the most informative part s for each attribute a by a *max* score operation, and only add the maximum to the final attribute score. This is because human attribute signals often reside in different body parts, so not all parts should be responsible for recognizing one particular attribute. For example, the head part can hardly be used to infer the "long pants" attribute. Through the max pooling of part scores, we are now able to capture those distributed attribute signals from the rich part collection.

The third and fourth terms capture deep hierarchical contexts in case the target person contains insufficient information, *e.g.* when he/she appears at a very small scale or occluded (Fig. 1).

Human-centric context. Let $s_a^* = \arg\max_{s \in S} \boldsymbol{w}_{a,s}^T \cdot \boldsymbol{\phi}(s; I)$ be the person's highest scoring part that best describes attribute a, and $N(s_a^*)$ be its part neighbor set searched by computing Euclidean distance between the $fc7$ features of detected parts in the same image. We exploit $N(s_a^*)$ as the *human-centric context* to capture human relations,

Each part neighbor found in human-centric context is scored by the part weights $\boldsymbol{w}_{a,s}$, and then average pooled (see also Fig. 2, third path). By doing so, we hope to accumulate more stable or even stronger signals of attributes from the nearest neighbor fields between contextual people. Indeed, recognizing "sunglasses" from an occluded or low resolution face can become much clearer when considering a lot of similarly looking faces around. Here we choose to define similarities in terms of the human parts instead of whole body because people usually appear quite different globally but very similar at local parts. So it is more reasonable to only transfer the good knowledge locally rather than globally

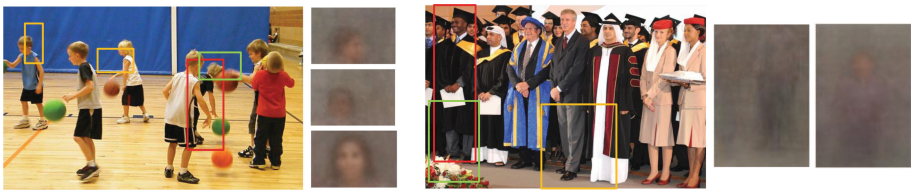

Fig. 4. Example nearest neighbors found for a human part that best describes a particular attribute. In each image, we show the target person's bounding box in red, the attribute-specific part in green, and the part neighbors in yellow. We also show the average images of the found neighbor clusters, which strengthen the signal of attribute. (Color figure online)

from surrounding people. In our approach, the local part matching is made easier by using (1) poselet-based body parts that are pose-specific and well-aligned to match, (2) a compact poselet selection $\{s_a^*\}$ for each person which reduces the computational burden of online neighbor matching.

In practice, our part matching is performed on the CNN feature maps of a Gaussian pyramid of input image (with three scales in our implementation). Such a multi-scale neighbor set $N(s_a^*)$ is able to cover a broader range of part similarities. Its size $K = |N(s_a^*)|$ is determined by experiments. Note we match parts from all detected people including himself in one image. This guarantees our approach can still work on images with only one person. In this case, the use of human-centric context is actually a self-similarity regularization among multi-scale features of the same person.

Figure 4 illustrates some examples of the found part neighbors and their average images. It is observed that the part patterns are strengthened from their multi-scale versions as well as similar patterns from other people. This makes attributes emerge more clearly and their recognition less ambiguous in challenging cases.

Scene-level context. The scene-level context is further exploited by the fourth term in Eq. 2 (see also Fig. 2, fourth path). We propose to reuse the entire scene feature $\boldsymbol{\phi}(I)$ holistically, without the need for explicitly identifying helpful objects or regions within a scene as in [35]. Obviously our method is computationally more appealing, but the downside is that some irrelevant information contained in the global feature $\boldsymbol{\phi}(I)$ may confuse the recognition of attributes.

Therefore, we "filter" $\boldsymbol{\phi}(I) \in \mathbb{R}^D$ by converting it to the scene classification score via $\boldsymbol{W}_{sc} \in \mathbb{R}^{|C| \times D}$, where C refers to all the considered scene types. This way, only the scene-related high level information is preserved, while other variables are marginalized over during the conversion. Then we use the scene score to provide the prior probability for most likely human attributes in the scene, and re-score each attribute a via $\boldsymbol{w}_{a,sc} \in \mathbb{R}^{|C|}$ in a scene-aware manner. This factorization is actually equivalent to applying the Bayes' rule to split the scene

conditional probability function $P(a \mid I)$ in two factors:

$$P(a \mid I) = \sum_{c \in C} P(a \mid c, I) \cdot P(c \mid I), \tag{3}$$

where the latent variable of scene type c is introduced.

Accordingly, the attribute score is learned in our CNN via the total scoring weights $\boldsymbol{w}_{a,sc}^{T} \cdot \boldsymbol{W}_{sc}$. Note in some atypical cases, even the mere scene type can be misleading. When a total mismatch exists between the human attributes and background scene (*e.g.* suitmen on a basketball court), the pure person characteristics is what our model should focus on. So we do not always expect the scene context to have strong re-scoring effects. In our CNN, the weightings between the human- and scene-induced scores are automatically learned in their respective scoring weights.

3.3 Learning Details

We train our CNN together with the four scoring paths from an ImageNet initialization. As shown in Fig. 2, the whole network is trained end-to-end using a *cross entropy* loss over independent logistics to output multi-attribute predictions. Since the first three scoring paths all take human regions as input, we tie their *fc6* and *fc7* layers to reduce the parameter space (but the scoring weights for the whole person and body parts are separated). The fourth path's fully connected layers are not tied with others as they capture semantics of the global scene. Particularly, we attach right after the scene's *fc7* layer a *softmax* scene classification loss, in order to jointly learn the scene context priors.

During training, we augment the data by using bounding boxes of both human body and human parts that have no more than 10 % horizontal and vertical shift from the ground truth. We consider one image per mini-batch, and input with bounding boxes of the whole body and $|S| = 30$ poselet detections for each person. We set the learning rate to 10^{-5}, the momentum to 0.9, and train for 40 K iterations. The running time depends on the person number in one image. On average, training takes about 1 s for all persons in one image per iteration, while testing takes about 0.5 s per image on a NVIDIA Titan X GPU.

4 Datasets

4.1 Existing Human Attribute Datasets

We summarized a few popular human attribute datasets in Table 1. The Berkeley Attributes of People [5] dataset is the most widely used human attribute database. It consists of 2003 training, 2010 validation and 4022 test images, and a total of 17628 bounding boxes and 9 attribute labels such as "is male" and "has hat". Although this dataset is challenging for its wide human variation in pose, viewpoint and occlusion, the number of images is rather small and each is cropped from the original high resolution image and centered at a person's

Table 1. Statistics of the proposed WIDER Attribute dataset and comparison with existing human attribute datasets ('trunc.' denotes truncation; 'fg.' denotes fine-grained).

Dataset	Images	Boxes	Boxes/Img	Attributes	Attribute labels	Scene labels
Berkeley [5]	8,035	17,628	2.2 trunc	9	72,315	-
HAT [15]	9,344	19,872	2.1	27 fg	536,544	-
CRP [16]	20,999	27,454	1.3	4 fg	109,816	-
PARSE-27k [17]	9887	∼27,000	2.7	10 fg	∼270,000	-
WIDER Attribute	13,789	57,524	4.2	14	805,336	13,789

full body, leaving only limited background scene and people (likely truncated). This is not suitable to exploit contexts in our method to attain its full capacity. But we still detect poselets from the few and potentially incomplete neighboring people as described in Sect. 3.1. We treat them as a localized human-centric context to see if they can help in this case.

We also use a larger dataset of HAT [15] with 9344 human images from Flickr that show a considerable variation in pose and resolution. The images are not cropped and of the original full resolution. There are totally 19872 persons with annotated bounding boxes, about 2 full persons per image on average, which is more suitable than Berkeley dataset for our context-based method. To make full use of the human-centric context, we further detect new person bounding boxes and related poselets in HAT images following Sect. 3.1. There are 27 attribute labels for each person, but some refer to human actions (*e.g.* "standing", "sitting", "running") and some are overly-fine-grained (*e.g.*6 age attributes of "baby", "kid", "teen", "young", "middle aged" and "elderly"). We follow the train-val-test split of 3500, 3500 and 2344 images and employ all 27 attributes in our experiments to facilitate comparison.

There are two other video-based human attributes datasets, namely Caltech Roadside Pedestrians (CRP) [16] dataset and PARSE-27k [17] dataset. We summarize these datasets in our supplementary material. We do not employ the two datasets in our experiments since they are either small in terms of images (attributes), or limits (even disables) the exploitation of human-centric context in one image. Also they lack the scene labels to exploit global scene context.

4.2 WIDER Attribute Dataset

We introduce a large-scale WIDER Attribute dataset to overcome all the aforementioned drawbacks of existing public datasets. Our dataset is collected from the 50574 WIDER images [14] that usually contain many people and huge human variations (see Fig. 5). We discard those images full of non-human objects or low quality humans that are hardly attribute-recognizable, ending up with 13789 images. Then we annotate a bounding box for each person in these images, but no more than 20 people (with top resolutions) in a crowd image, resulting in 57524 boxes in total and 4+ boxes per image on average. For each bounding box, we label 14 distinct human attributes (no subcategories as in CRP [16] and PARSE-27k [17]),

Fig. 5. Thirty examples of WIDER Attribute images, each from an image event class.

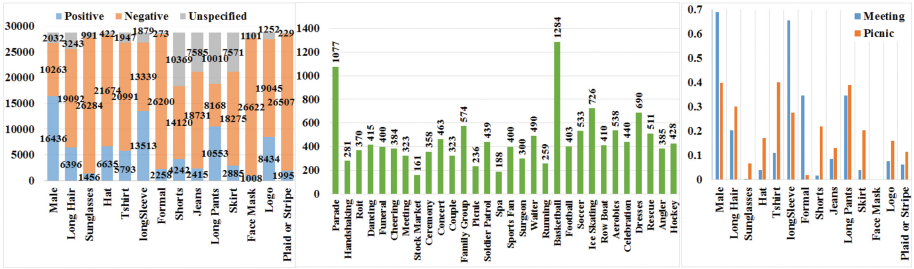

Fig. 6. Statistics of the number of attribute labels (Left), event class labels (Middle), and event-specific attribute labels (Right) on WIDER Attribute dataset.

resulting in a total of 805336 labels. The label statistics are shown in Table. 1 and Fig. 6 (Left). Note that we allow missing annotation since not every attribute can be specified for every person. For example, in the presence of face occlusion, one cannot sensibly label a valid "sunglasses" attribute.

We split our dataset into 5509 training, 1362 validation and 6918 test images. The large quantities of images and human labels permit us to study the benefit of human-centric context. To explore the scene-level context, we further label each image into 30 event classes. Figure 6 (Middle and Right) illustrates the image event distribution and two event-specific attribute distributions. We observe strong correlations between image event and the frequent human attributes in it, *e.g.* "longsleeve" and "formal" are frequent in *meeting*, while "tshirt" is frequent in *picnic*. Such correlations motivate our attribute re-scoring scheme conditioned on scene classification.

Figure 5 shows example images of events. In many cases, humans are small and their attributes are hard to recognize without referring to contexts, *e.g.* in *ceremony*. In other cases where only one person appears in an image (*e.g. angler*) or the person is inconsistent with the background and stands out on his own (*e.g.* a formally dressed man on the *basketball* court), our model should learn to weigh more on the target human features over human-centric or scene-level

contexts. The new WIDER Attribute dataset forms a well-suited testbed for a full examination of using hierarchical contexts.

5 Experiments

We evaluate our method on the test sets of the Berkeley Attributes of People [5] and HAT [15] datasets, where there are many results to compare. We also use the test set of the proposed WIDER Attribute dataset to compare with the baselines that do not fully exploit the joint context of humans and scene. It would be interesting to see the performance of our method when extended to the video datasets of CRP [16] and PARSE-27k [17]. However, these datasets hinder the use of human-centric context as mentioned in Sect. 4.1 and have no scene labels. Hence they are not well-suited for the context where our method is used, and we leave their experiments to future work. We measure performance with the average precision (AP) for each human attribute and the mean average precision (mAP) over all attributes.

Table 2 shows the results on Berkeley dataset where our method is compared against all CNN-based state-of-the-arts. Both PANDA [8] and ACNH [17] achieve relatively low mAPs with 5-layer networks. With a well trained 16-layer network, R-CNN [36] improves the performance for nearly all the attributes, using the holistic human body region only. R*CNN [10] and Gkioxari *et al.* [9] further improve by adding a secondary contextual region and three human parts respectively.

Our baseline that combines the human body and selected attribute-descriptive parts achieves a mean AP of 90.8 %. By searching a different number of parts $K = |N(s_a^*)|$ from other people, our human-centric contextual model obtains consistent gains across attributes, especially those located at small

Table 2. AP on the test set of the Berkeley Attributes of People [5] dataset. Note PANDA [8] and ACNH [17] use 5-layer CNNs while others use 16-layer CNNs. Our baseline combines the human body and attribute-descriptive parts at score level. Our human-centric contextual models further pool scores from $K = 1$ or 2 nearest neighbor parts. Part neighbors searched from single-scale (ss) CNN maps are also evaluated.

AP(%)	Male	Long Hair	Glasses	Hat	Tshirt	LongSleeves	Shorts	Jeans	Long Pants	mAP
PANDA [8]	91.7	82.7	70.0	74.2	49.8	86.0	79.1	81.0	96.4	79.0
ACNH [17]	87.8	81.5	48.8	75.3	64.1	88.1	87.1	89.5	98.1	80.0
R-CNN [36]	91.8	88.9	81.0	90.4	73.1	90.4	88.6	88.9	97.6	87.8
R*CNN [10]	92.8	88.9	82.4	92.2	74.8	91.2	92.9	89.4	97.9	89.2
Gkioxari *et al.* [9]	92.9	90.1	77.7	93.6	72.6	93.2	93.9	**92.1**	98.8	89.5
Ours (baseline)	94.1	90.8	86.8	94.4	76.1	92.9	94.0	90.2	98.2	90.8
Ours ($K = 1$)	**95.0**	**92.4**	**89.3**	95.7	**79.1**	**94.3**	93.7	91.0	99.2	**92.2**
Ours ($K = 2$)	94.8	91.8	88.4	**95.8**	76.6	94.1	**95.4**	91.5	**99.3**	92.0
Ours ($K = 1$, ss)	94.3	91.5	88.0	94.6	77.7	93.9	93.7	90.0	98.7	91.4
Ours ($K = 2$, ss)	94.2	91.3	88.0	94.8	77.6	93.9	92.9	90.2	98.7	91.3

Table 3. Comparing mAP on the test set of the HAT [15] dataset. Note ACNH [17] uses a 5-layer CNN while other deep methods use 16-layer CNNs.

Methods	DSR [15]	EPM [42]	Joo *et al.* [7]	EPM+Context [42]	ACNH [17]	EPM+VGG16 [42]
mAP(%)	53.8	58.7	59.3	59.7	66.2	69.6
Methods	R-CNN [36]	R*CNN [10]	Ours (baseline)	Ours ($K = 1$)	Ours ($K = 2$)	Ours ($K = 5$)
mAP(%)	76.3	76.4	76.7	77.6	**78.0**	77.8

Fig. 7. Visualizing the AP and example image for some hard human attributes in the HAT [15] dataset.

scales *e.g.* "glasses". We attain the highest mean AP of 92.2 % when $K = 1$. This indicates that, even in the case of cropped Berkeley images with rather limited context (2 neighboring people on average with large truncation), the local use of human context can at least help and will not degrade the performance on single-person images. But we found no more gains with more than $K = 1$ part neighbors. When the part neighbors are searched at a single scale (ss) of CNN feature maps instead of their Gaussian pyramid, a performance drop is observed which shows the importance of seeking multi-scale part similarity. In the following experiments, we always use three scales in a pyramid due to the good tradeoff between performance and computational speed.

Table 3 reports the performance in mean AP of our approach and other competitive methods on the HAT dataset. Our method outperforms all others based on CNN or not (the first four). Note the Expanded Parts Model (EPM) [42] uses the immediate context around a person, but only achieves 59.7 % mean AP. Its combination with deep VGG16 features significantly improves to 69.6 % mean AP. We finetune the state-of-the-art R-CNN [36] and R*CNN [10] on this dataset, obtaining substantial margins over others. Whereas our adaptive deep-part and whole body-based baseline performs slightly better. More gains are obtained from using human-centric context, with the best mean AP of 78.0 % using $K = 2$ human part neighbors. This is reasonable considering the HAT dataset contains richer human contexts in full resolution images.

Figure 7 compares the competitive methods with our baseline and best-performing methods in terms of AP of some hard attributes. It is evident that our used human-centric context offers especially large gains for the hard attributes that current competing methods do badly on, *e.g.* "crouching/bent", "teen aged" and "female short skirt". These are the attribute categories that have large pose variation and have small-sized or limited number of human samples.

Table 4. Comparing mAP on the test set of the WIDER Attribute dataset. All methods use 16-layer CNNs. Our full method exploits the scene-level context besides human-centric context. Here 'scene no cl.' means directly using scene features with no scene label conversions to re-score human attributes.

Methods	R-CNN [36]	R*CNN [10]	Ours (baseline)	Ours ($K = 1$)
mAP(%)	80.0	80.5	80.5	80.7
Methods	Ours ($K = 5$)	Ours (scene + $K = 5$)	Ours (scene no cl. + $K = 5$)	
mAP(%)	80.9	**81.3**	81.1	

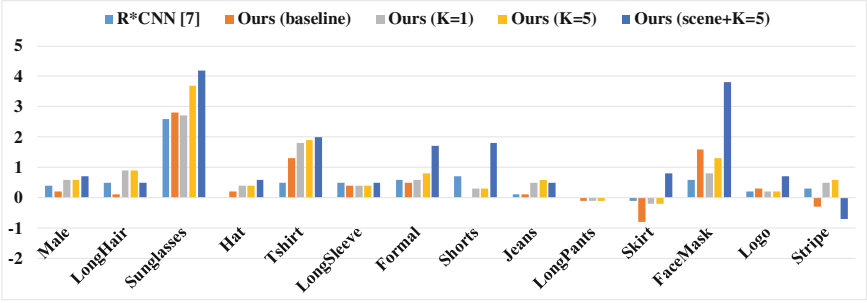

Fig. 8. Visualizing the absolute gains of competing methods over R-CNN [36] in terms of AP(%) on the test set of WIDER Attrbiute

The WIDER Attribute dataset contains richer contexts of humans and event labels. We compare our deep hierarchical context-based method with the state-of-the-art R-CNN [36] and R*CNN [10] in Table 4. Our method performs consistently better for most attributes, achieving the best mean AP of 81.3 % when we use $K = 5$ human part neighbors. The reason is that there are more people (about 4 on average) in the full image of the WIDER Attribute dataset. Our performance also benefits from the use of scene-aware attribute re-scoring, as can be observed from the mAP value difference. We further compare with the result from a direct attribute scoring using global scene features, which is worse than our proposed scheme. Figure 8 shows our absolute improvements in AP over competing methods to validate our advantages. In the supplementary material, the scene classification results as well as more detailed attribute recognition results will be included.

6 Conclusion

We propose a new method for unconstrained human attribute recognition, built on the simple observation that context can unveil more clues to make recognition easier. To this end, we not only learn to score the human body and attribute-specific parts jointly in a deep CNN, but also learn two scoring functions that capture deep hierarchical contexts. Specifically, collaborative part

modeling among humans and global scene re-scoring are performed to respectively capture human-centric and scene-level contexts. We introduce a large-scale WIDER Attribute dataset to enable the exploitation of such hierarchical contexts. Our method achieves state-of-the-art results on this dataset and popular ones as well. We believe our method can be easily extended to the video datasets and other tasks such as human pose estimation.

Acknowledgement. This work is partially supported by SenseTime Group Limited.

References

1. Layne, R., Hospedales, T.M., Gong, S.: Person re-identification by attributes. In: British Machine Vision Conference, pp. 1–11 (2012)
2. Liu, C., Gong, S., Loy, C.C.: On-the-fly feature importance mining for person re-identification. Pattern Recogn. **47**(4), 1602–1615 (2014)
3. Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L.S., Gao, W.: Multi-task learning with low rank attribute embedding for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3739–3747 (2015)
4. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person Re-Identification, vol. 1. Springer, London (2014)
5. Bourdev, L., Maji, S., Malik, J.: Describing people: poselet-based attribute classification. In: IEEE International Conference on Computer Vision, pp. 1543–1550 (2011)
6. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: IEEE International Conference on Computer Vision, pp. 729–736 (2013)
7. Joo, J., Wang, S., Zhu, S.C.: Human attribute recognition by rich appearance dictionary. In: IEEE International Conference on Computer Vision, pp. 721–728 (2013)
8. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.D.: PANDA: pose aligned networks for deep attribute modeling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1637–1644 (2014)
9. Gkioxari, G., Girshick, R., Malik, J.: Actions and attributes from wholes and parts. In: IEEE International Conference on Computer Vision, pp. 2470–2478 (2015)
10. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with R*CNN. In: IEEE International Conference on Computer Vision, pp. 1080–1088 (2015)
11. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: European Conference on Computer Vision, pp. 834–849 (2014)
12. Branson, S., Horn, G.V., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. In: British Machine Vision Conference (2014)
13. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. Int. J. Comput. Vision **104**(2), 154–171 (2013)
14. Xiong, Y., Zhu, K., Lin, D., Tang, X.: Recognize complex events from static images by fusing deep channels. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1600–1609 (2015)
15. Sharma, G., Jurie, F.: Learning discriminative spatial representation for image classification. In: British Machine Vision Conference, pp. 1–11 (2011)

16. Hall, D., Perona, P.: Fine-grained classification of pedestrians in video: benchmark and state of the art. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5482–5491 (2015)
17. Sudowe, P., Spitzer, H., Leibe, B.: Person attribute recognition with a jointly-trained holistic CNN model. In: IEEE International Conference on Computer Vision Workshop, pp. 329–337 (2015)
18. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1785 (2009)
19. Huang, C., Change Loy, C., Tang, X.: Unsupervised learning of discriminative attributes and visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5175–5184 (2016)
20. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 951–958 (2009)
21. Moghaddam, B., Yang, M.H.: Learning gender with support faces. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 707–711 (2002)
22. Shakhnarovich, G., Viola, P.A., Moghaddam, B.: A unified learning framework for real time face detection and classification. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 16–26 (2002)
23. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: IEEE International Conference on Computer Vision, pp. 365–372 (2009)
24. Kumar, N., Belhumeur, P., Nayar, S.: FaceTracer: a search engine for large collections of images with faces. In: European Conference on Computer Vision, pp. 340–353 (2008)
25. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
26. McCann, S., Lowe, D.G.: Local naive bayes nearest neighbor for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3650–3656 (2012)
27. Zhang, N., Farrell, R., Darrell, T.: Pose pooling kernels for sub-category recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3665–3672 (2012)
28. Johnson, J., Ballan, L., Li, F.: Love thy neighbors: Image annotation by exploiting image metadata. In: IEEE International Conference on Computer Vision, pp. 4624–4632 (2015)
29. Oliva, A., Torralba, A.: The role of context in object recognition. Trends Cogn. Sci. **11**(12), 520–527 (2007)
30. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
31. Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S.: Exploiting hierarchical context on a large database of object categories. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 129–136 (2010)
32. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 891–898 (2014)

33. Russell, B., Torralba, A., Liu, C., Fergus, R., Freeman, W.T.: Object recognition by scene alignment. In: Advances in Neural Information Processing Systems, pp. 1241–1248 (2007)
34. Torralba, A.: Contextual priming for object detection. Int. J. Comput. Vision **53**(2), 169–191 (2003)
35. Li, C., Parikh, D., Chen, T.: Extracting adaptive contextual cues from unlabeled regions. In: IEEE International Conference on Computer Vision, pp. 511–518 (2011)
36. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
38. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
39. Bourdev, L., Malik, J.: Poselets: body part detectors trained using 3D human pose annotations. In: IEEE International Conference on Computer Vision, pp. 1365–1372 (2009)
40. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: European Conference on Computer Vision, pp. 168–181 (2010)
41. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 789–792. ACM (2014)
42. Sharma, G., Jurie, F., Schmid, C.: Expanded parts model for semantic description of humans in still images. IEEE Trans. Pattern Anal. Mach. Intell. (2016)