Pattern Recognition Letters 94 (2017) 38–45

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Human attribute recognition by refining attention heat map

Hao Guo[a], Xiaochuan Fan[b], Song Wang[a,*]

[a] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
[b] HERE North America LLC, 425 W. Randolph Street, Chicago, IL 60606, USA

ARTICLE INFO

Article history:
Received 2 December 2016
Available online 13 May 2017

Keywords:
Human attribute recognition
Attention heat map
Exponential loss
Refine

ABSTRACT

Most existing methods of human attribute recognition are part-based, where features are extracted at human body parts corresponding to each human attribute and the part-based features are then fed to classifiers individually or together for recognizing human attributes. The performance of these methods is highly dependent on the accuracy of body-part detection, which is a well known challenging problem in computer vision. Different from these part-based methods, we propose to recognize human attributes by using CAM (Class Activation Map) network and further improve the recognition by refining the attention heat map, which is an intermediate result in CAM and reflects relevant image regions for each attribute. The proposed method does not require the detection of body parts and the prior correspondence between body parts and attributes. In particular, we define a new exponential loss function to measure the appropriateness of the attention heat map. The attribute classifiers are further trained in terms of both the original classification loss function and this new exponential loss function. The proposed method is developed on an end-to-end CNN network with CAM, by adding a new component for refining attention heat map. We conduct experiments on Berkeley Attributes of Human People Dataset and WIDER Attribute Dataset. The proposed methods achieve comparable performance of attribute recognition to the current state-of-the-art methods.

1. Introduction

As important visual concepts, human attributes – such as gender, hair and clothing style, – are highly intuitive, semantic, and informative to describe the appearance of a person. Human attribute recognition can benefit a variety of computer vision applications, such as people search [1], fine-grained recognition [2], object categorization [3], object description [4], face verification [5], and attribute-based classification [6].

Given rich information in an image, only a small portion of it actually contributes most to a specific human attribute. For example, the attribute of "wearing glasses" is highly determined by a small image region around the human eyes. Even if the original image is in high resolution and high quality, the image information associated to each attribute may be in low resolution and low quality in practice. In addition, in different images, the image information/features that reflect the same attribute may be quite different, given the changes of viewpoint and human pose, and possible occlusion of human body parts. As a result, human attribute recognition is still a very challenging research problem in computer vision.

Intuitively, human attributes have strong correspondence with human body parts. For example, hair length corresponds to head, and style of pants corresponds to legs. Based on this observation, many researchers use the correspondence between different attributes and body parts to facilitate the attribute recognition [7–15]. These methods represent the current state-of-the-art in terms of the performance of human attribute recognition.

In these methods, attribute recognition is usually accomplished by taking a two-step procedure. First, a body-part detector or pose estimator is applied to localize important human body parts, such as head, legs, arms, hands, neck, eyes, etc. Second, image features are extracted at each body part and then fed to pre-trained classifiers for attribute recognition. Typically, one classifier is trained for each attribute and this classifier usually takes only the features from the corresponding body parts [9–13]. There is also research work that suggests the combination of features from different body parts for attribute recognition [14]. There are several issues in using part-based methods for human attribute recognition. (1) It requires prior correspondence between body parts and attributes. This may not be trivial for some attributes. (2) It requires an accurate and reliable body-part detector and/or pose estimator, both of which are well known challenging tasks in computer vision. Errors

* Corresponding author.
E-mail addresses: hguo@email.sc.edu (H. Guo), songwang@cec.sc.edu (S. Wang).

http://dx.doi.org/10.1016/j.patrec.2017.05.012
0167-8655/

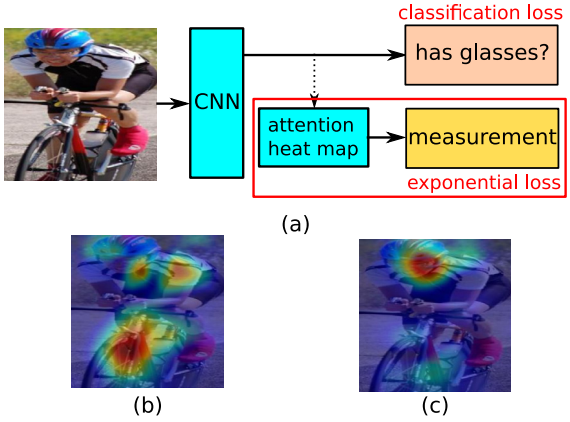

Fig. 1. Motivation of the proposed method. (a) CAM-network based classifier and the added component for refining attention heat map. (b, c) Attention heat map before and after the refining, respectively.

in detecting body parts can seriously hurt the performance of attribute recognition. (3) The training of body-part detector usually requires manual annotations of body parts in large-scale images and this can be highly laborious.

In essence, these part-based attribute recognition methods are just trying to focus on certain image regions that are relevant to each attribute. Instead of using body-part detectors to identify these regions, in this paper we propose to automatically identify relevant regions for each attribute by learning and refining attention heat map. Attention heat map, also called localization score map [16,17] or class activation map (CAM) [18], is a kind of category related saliency map, which highlights the discriminative image regions for an object category and has been widely used for weakly supervised localization [16–18]. It has also been applied to classification tasks [18], where attention heat map is one of the intermediate results. In this paper, we exploit and refine the attention heat map for human attribute recognition.

Specifically, we adapt the CAM network [18], a special kind of Convolutional Neural Network (CNN) [19], for human attribute recognition and further improve it by including an additional component to refine the attention heat map. As shown in Fig. 1(a), we use the CAM-based CNN as the attribute classifiers. We then propose an added component, e.g., the red box in Fig. 1(a), to train CAM by refining the attention heat map, as shown in Fig. 1(b–c). To reduce over-fitting, we use an exponential loss function in this paper to measure the appropriateness of attention heat map. In classifier training, the CAM network is tuned by simultaneously minimizing the exponential loss function and the original classification loss function. Finally, the tuned CAM network is applied to human attribute recognition on test images.

Compared to existing part-based attribute recognition methods, the proposed method does not require prior correspondence between body-parts and attributes. Meanwhile, the proposed method does not need a body detector or pose estimator. As a result, it can well address the three above mentioned issues of the existing methods. Technically, this paper makes four main contributions. (1) We propose to use the CAM network for attribute recognition. (2) We propose a new component to fine-tune the CAM network based on the appropriateness of the intermediate attention heat map. (3) We propose a new exponential loss function for measuring the appropriateness of the heat map. (4) We apply the fine-tuned CAM network to improve the performance of human attribute recognition. In the experiments, we test the proposed method on Berkeley Attributes of Human People Dataset [9] and WIDER Attribute Dataset [15], by comparing with other state-of-the-art attribute recognition methods.

The remainder of the paper is organized as follows. Section 2 overviews the related work. Section 3 introduces the proposed method. Section 4 reports experiment results, followed by a brief conclusion in Section 5.

2. Related work

In this section, we briefly overview the related work on attribute recognition and attention heat map.

2.1. Part-based methods for human attribute recognition

As mentioned above, most of the existing methods on human attribute recognition are part based, by leveraging the correspondence between human body parts and attributes. In [9], image of a person is decomposed into a set of poselets with rich appearance and local pose information and human attributes are then recognized using features extracted from these poselets. In [10], a feature dictionary is built to describe possible appearance variation at each human body part, which can be used to improve part detection and human attribute recognition. In [11], following Deformable Part Models (DPM) [8], Deformable Part Desciptors (DPD) are extracted for part detection and attribute recognition. Recently, based on convolutional neural networks (CNNs) [19], several deep part models are developed for human attribute recognition. In [12], a PANDA system leveraging CNNs trained for each poselet is developed for attribute recognition. Using CNNs, Park and Zhu [13] propose an attribute grammar model to jointly represent both the object parts and their semantic attributes within a unified compositional hierarchy. In [14], Gkioxari et al. suggest the use of deep poselets as part detector to localize human body parts under different poses. Deep-Context [15], another part-based method using deep learning, improves human attribute recognition by using hierarchical contexts.

All these part-based methods require the prior correspondence between body parts and attributes, as well as part detections, for human attribute recognition. As discussed above, these part-based methods have three main issues listed in Section 1, and in this paper, we propose a method that does not require prior part/attribute correspondence and part detections. The R*CNN method proposed by Gkioxari et al. [20] also does not detect body parts and consider the correspondence between body parts and attributes. Instead, in R*CNN, contextual cues in larger regions are exploited and used for facilitating human attribute recognition. The recent Deep-Context [15] combines both part-based and contextual information for attribute recognition. In this paper, we do not rely on such contextual cues – we still try to identify small relevant regions for recognizing each attribute.

2.2. Attention heat map

The attention heat map is a kind of category related saliency map. Different from the traditional bottom-up saliency maps [21–26], which capture the visual saliency that stands out from its neighbors and attract the observer's attention, the attention heat map highlights the area which is discriminative for a particular category of objects. Attention heat map shares certain similarity to the top-down saliency [27–29] – both of them are related to object categories.

The computation of attention heat map is usually formulated as a weakly supervised localization problem [17,30–32]. Several methods [16–18] have been developed in recent years for computing attention heat map. In this paper, we use the attention heat map generated as an intermediate result in the CAM network [18]. We then add a new component to the CAM network to refine the at-

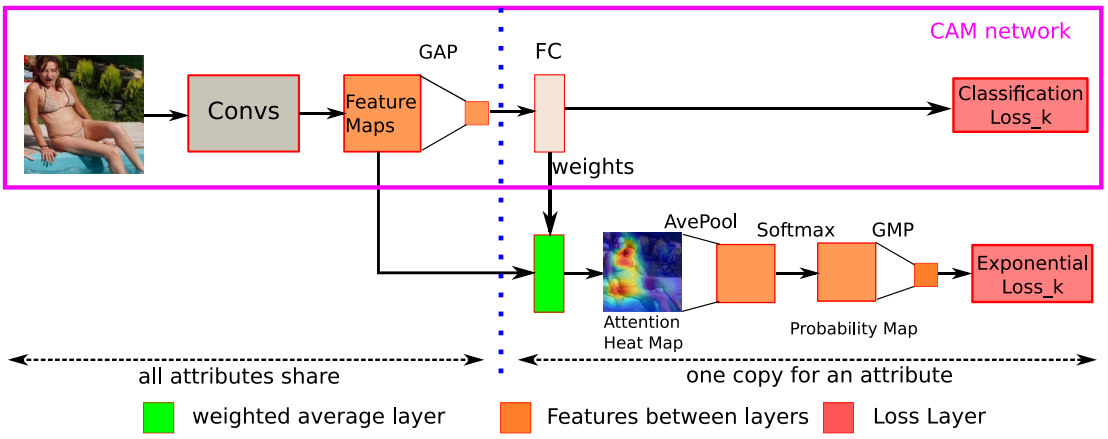

Fig. 2. Framework of the proposed method.

tention heat map by tuning the original network for better human attribute recognition.

3. Proposed method

Human attribute recognition is a binary classification problem: for each attribute, a binary classifier is trained to decide whether the attribute is present in the image or not. The framework of the proposed method is illustrated in Fig. 2. In the following, we first introduce the CAM network for attribute recognition, and then describe the component added to CAM for refining the attention heat map.

3.1. CAM network for attribute recognition

In this paper, we propose to use the CAM network [18], as shown in the pink box in Fig. 2, for human attribute recognition. In the CAM network, an input image first goes through a sequence of convolutional layers ('Convs' in Fig. 2). After that, a Global Average Pooling (GAP) is applied to the feature maps on the last convolutional layer. The GAP output is then taken as the input features for a fully-connected layer ('FC' in Fig. 2). The FC output is then sent to a Softmax layer for deciding whether an attribute is present or not. In this CAM network, the parameters of all the layers before FC are identical for different attribute classifiers. The parameters for the FC layer is categorical – different attribute classifier have different parameters for the FC layer. The attention heat map for each object category can then be computed by using the parameters in the FC layer as weights to linearly combine the feature maps in the last convolutional layer, as shown in Fig. 2.

Given a binary CAM classifier, there are two sets of parameters in the FC layer, one can generate the attention heat map for negative classification, i.e., the category of object is not present, and the other can generate the attention heat map for positive classification, i.e., the category of object is present. We found that the attention heat maps generated using these two sets of parameters are almost identical because the relevant regions for positive and negative classifications are the same. In this paper, we simply take the second set of FC parameters (for positive classification) for generating the attention heat map.

In the ideal case, attention heat map highlights the most relevant regions for the considered attribute. However, in practice, overly small size and low image quality of the actual relevant regions and over-fitting training (mainly due to insufficient training data) may lead to incorrect attention heat maps. Nine examples are shown in Fig. 3, where we show a set of images and their CAM heat maps side by side, with the considered attribute labeled at the bottom left corner of each image. We can see that, the computed attention heat maps may be incorrect by highlighting irrelevant regions. For example, the attention area is not focused on face region when recognizing the attribute of glasses in the image at the center of Fig. 3.

3.2. Attention heat map refinement component

Our basic idea for improving the recognition is to refine the attention heat map by tuning the CAM network. For this purpose, we measure the appropriateness of an attention heat map based on its concentration – the refined attention heat map is expected to highlight a smaller, more concentrated region than those shown in the original heat map, as shown in Fig. 1(c). More specifically, as shown in Fig. 2. a new component consisting of several layers is added to the CAM network. First, this added component includes a weighted average layer to compute the attention heat map as described in Section 3.1. After that, we include an average pooling layer to capture the importance of all the potential relevant regions for recognizing the considered human attribute. Then we perform Softmax to convert the pooled attention heat map to a probability map: Let $z_{x,\,y}$ be the value of point $(x,\,y)$ in the pooled attention heat map and $f(z_{x,\,y})$ is the corresponding value in the probability map, we have

$$f(z_{x,y}) = \frac{e^{z_{x,y}}}{\sum_{p=1}^{H}\sum_{q=1}^{H} e^{z_{p,q}}}, \tag{1}$$

where the size of the pooled attention heat map is $H \times H$. Finally, a Global Maximum Pooling (GMP) layer is included to extract the maximum probability, which reflects the credibility of the identified relevant region. Based on this maximum probability, we define a loss function to reflect the appropriateness of current attention heat map. Two key issues need to be addressed in this component: (1) the definition of the loss function, which measures the appropriateness of the attention heat map, based on the maximum probability, and (2) the tuning of the CAM network to increase the maximum probability and minimize the loss function.

Definition of the loss function. We define a loss function that can measure the appropriateness of the attention heat map by using the maximum probability in the probability map. Given that the summation over the probability map is one, increasing the maximum probability will automatically suppress the regions with smaller probability. This will make the attention area (highlighted region) in the attention heat map more concentrated, which increases the chance of capturing the region relevant to the consid-

Fig. 3. Sample images and their CAM attention heat maps, which may not highlight the relevant regions for the considered attribute.

ered attribute. In this paper, we develop an exponential loss function, independent of any supervised information on the attention heat map, based on the maximum probability on the probability map. Let p_{ij}^M be the maximum probability for image i and attribute j. We define the loss function for jth-attribute as

$$\ell = \frac{1}{N}\sum_{i=1}^{N} e^{\alpha(p_{ij}^M+\beta\mu)}, \tag{2}$$

where α and β are adjustable parameters of the loss function, $\mu = 1/H^2$ is the mean value of the probability map, with $H \times H$ being the size of the attention heat (probability) map, and N is the number of images.

Given that the heat map size $H \times H$ is fixed in a CAM network, μ is also a constant. Since the loss function ℓ is negatively related to the maximum probability p_{ij}^M, α is a negative parameter. Furthermore, β also takes a negative value, making $|\beta\mu|$ a threshold: If the probability p_{ij}^M is less than this threshold, the loss value is large, indicating that the attention heat map is not concentrated.

Fig. 4 shows the curves of this loss function over a single image ($N = 1$). Curves 1, 2, and 3 share a same α but different β. Curves 1, 4, and 5 share a same β, but different α. We can see that α controls the descent rate of the loss value, while β adjusts the impact of the maximum probability by a threshold – the maximum probability p_{ij}^M takes value in the range $[\mu, 1]$. Based on our observation, given $\mu = 1/(14 \times 14)$, the attention area is highly concentrated when p_{ij}^M is above 0.2, Therefore, it is desired that the loss ℓ becomes very small when $p_{ij}^M \geq 0.2$. In our experiments, we use this to guide the selection of α and β.

Tuning of the CAM network. As described above, on the proposed network illustrated in Fig. 2, we have two loss functions – the original *classification loss function* aiming at reducing the attribute recognition error rate and the added *exponential loss function* aiming at refining the heat map. We train the CAM network in two steps: (1) *Pre-training*. In this step, without considering the added component and the exponential loss function, we train the original CAM network, e.g. the component included in the pink box in Fig. 2, by minimizing classification loss function. (2) *Fine-tuning*. In this step, we fine-tune the network parameters by minimizing both loss functions. Note that all the layers in the added component outside the pink box in Fig. 2 do not have free parameters to tune. The exponential loss is back propagated through the added component to fine-tune the parameters of the convolutional layers in the CAM network. In the meantime, the classification loss function is also back propagated through the fully-connected layer and then convolutional layers (in the pink box in Fig. 2) to fine-tune their parameters. This way, the parameters of convolutional layers are actually fine-tuned by minimizing a loss function that combines the exponential loss function and the classification loss

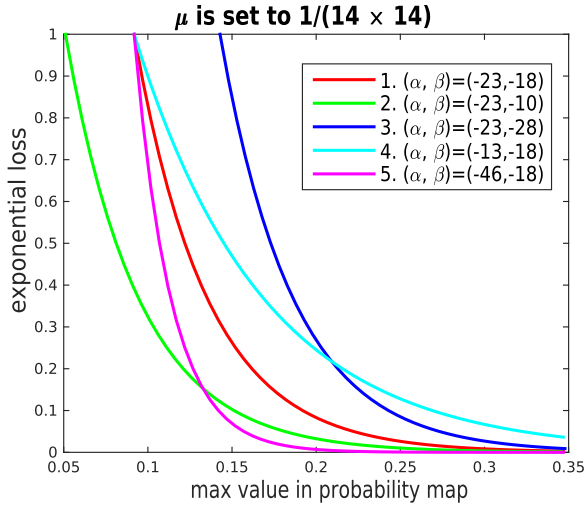

Fig. 4. Curves of the proposed exponential loss function – the loss decreases with the increase of the maximum probability.

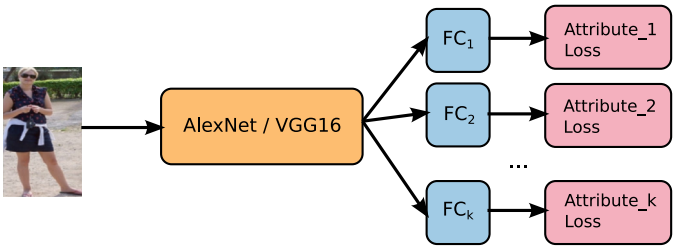

Fig. 5. AlexNet /VGG16 used for recognizing multiple human attributes.

function through back propagation. In the following Section 4, we will discuss the implementation in detail.

4. Experiments

We build our networks using Caffe [33], by customizing *weighted average layer* and *exponential loss layer* and modifying the fully-connected layer. The customization of the weighted average layer is used to embed the attention heat map into the network training and the modification of the fully-connected layer is used to output both features and weights. We use the Berkeley Attributes of Human People Dataset [9] and the WIDER Attribute Dataset [15] to train our networks and evaluate the proposed method.

The Berkeley Attributes of Human People Dataset contains 8035 images, each of which is centered at a full body of a person. Nine human attributes are referred in this dataset, including “is male”, “long hair”, “glasses”, “has hat”, “has t-shirt”, “long sleeves”, “has shorts”, “has jeans” and “long pants”. This dataset is divided into two subsets: the training subset with 4013 images and the testing subset with 4022 images.

The WIDER Attribute Dataset contains 13,789 images with 57,524 annotated persons, each labelled with 14 human attributes, including “male”, “long hair”, “sunglasses”, “hat”, “t-shirt”, “long sleeves”, “formal”, “shorts”, “jeans”, “long pants”, “skirt”, “face mask”, “logo” and “stripe”. It is divided into 5509 training, 1362 validation and 6918 testing images (13,789 in total). We use the training and validation subsets for training and the testing subset for testing.

CAM Pre-Training While CAM networks [18] can be constructed by extending either AlexNet [33], VGG16 [35], GoogLeNet [36] or Network In Network (NIN) [37], for simplicity, we focus on AlexNet and VGG16 in our experiments. Following the previous works, we call the CAMs constructed from AlexNet and VGG16 as AlexNet+CAM and VGG16+CAM, respectively. Since there are multiple attributes, we include the same number of classifiers, one for each attribute, in the final layer, as illustrated in Fig. 5, in both AlexNet and VGG16. We pre-train the CAM using the following three steps. (1) Taking the existing AlexNet and VGG16 models pre-trained on ILSVRC [38]. (2) Further training AlexNet and VGG16 using training samples in Berkeley Attributes of Human People Dataset or WIDER Attribute Dataset, separately. (3) Training AlexNet+CAM and VGG16+CAM (in the pink box in Fig. 2) using training samples in Berkeley Attributes of Human People Dataset or WIDER Attribute Dataset, separately. In Step 2), as shown in Fig. 5, all attributes share the same AlexNet or VGG16 network, but use distinct final FC layers, i.e., FC_1, FC_2, ..., FC_k. In the training, such FC layer of each classifier is updated independently from other classifiers, by simply using the standard back-propagation algorithm. In Step 3), we use base learning rate 0.0001 for both CAMs and the iterative training ends when the classification loss function decreases to an order of magnitude of 10^{-4}.

CAM fine-tuning. After the CAM pre-training, the classification loss is usually low, e.g., 10^{-4}. At this stage, the exponential loss is much higher, e.g., 10^{-1}. We adjust the control parameters α and β in Eq. (2) such that the combined loss is not dominated by any one of them. Empirically, we set $\alpha = -23$ and $\beta = -18$ for all the experiments. In fine-tuning CAMs using the proposed added component for refining attention heat map, gradients resulting from the back propagation of the two loss functions are simply added to update the parameters of the convolutional layers. We call the proposed method AlexNet+CAM+Refine and VGG16+CAM+Refine when using AlexNet+CAM and VGG16+CAM, respectively. After average pooling, the pooled attention heat map is a square matrix, while the Softmax layer requires an input of vector. To address this issue, we convert the square matrix into a vector by concatenating all the rows. After the probabilities are computed, the resulting vector is converted back to a square matrix to form a probability map.

In addition, for each attribute on each image, there are three possible status: ‘positive’, ‘negative’ and ‘non-specified’. ‘Positive’ indicates the presence of the attribute in the considered image, while ‘negative’ indicates the non-presence. ‘Non-specified’ indicates that it is unknown whether the attribute is present or not in the image. In our experiments, if an attribute is ‘non-specified’ for an image, this image will not be included to train the classifier for this attribute. Note that, the added component for refining attention heat map is only used in training the CAM network. In the testing stage, only the CAM network, as shown in the pink box in Fig. 2, is used to determine whether an attribute is present in a testing image or not.

4.1. Results

We first test the effectiveness of the proposed method on the Berkeley Attributes of Human People Dataset. Table 1 shows the mean Average Precision (mAP) of the proposed methods, AlexNet+CAM+Refine and VGG16+CAM+Refine, by adapting CAM with the added heat map refinement component. We also include the mAPs of AlexNet and VGG16 (mAP of VGG16 is cited from [14]), trained using Steps 1) and 2) in the above-mentioned CAM pre-training and mAPs of the AlexNet+CAM and VGG16+CAM, trained using all three steps in the above-mentioned CAM pre-training. We can see that VGG16 performs much better than AlexNet, especially on the attributes of “glasses” and “hat”. Using either AlexNet or VGG16, the constructed CAMs always lead to improved mAP’s and the proposed added component for refining attention heat map can further improve the mAP performance. Fig. 6 shows the original CAM attention heat maps and the refined ones using the proposed method for several sample images. All of them are based on VGG16 networks. We can see that using the proposed method, the obtained attention heat map is more concentrated on the desired relevant region when recognizing an attribute.

We also conduct an experiment to justify the steps in the added heat map refinement component, which consists of (1) weighted average layer, (2) average pooling, (3) softmax, (4) global max pooling and (5) exponential loss, as shown in Fig. 2. Among them, step 1) calculates the attention heat map from the CAM network. Step 3) converts the attention heat map to the probability map to emphasize relevant regions while automatically suppressing the irrelevant regions. Step 4) and 5) specify the loss, which is required for any learning framework. These four steps cannot be removed. We remove step 2) in the proposed method and the performance is shown in Table 1 as ‘AlexNet+CAM+Refine w/o AvePool’ and ‘VGG16+CAM+Refine w/o AvePool’. We can see that the inclusion of average pooling in the attention heat map refinement component does improve the attribute recognition performance.

To further justify the effectiveness of the proposed method, we also compare the performance of the proposed method against eight other state-of-the-art methods for human attribute recog-

Table 1
Attribute recognition performance of AlexNet, VGG16, AlexNet+CAM, VGG16+CAM and the proposed methods on Berkeley Attributes of Human People Dataset.

AP(%)	male	long hair	glasses	hat	t-shirt	long sleeves	shorts	jeans	long pants	mAP
AlexNet	84.9	76.0	46.1	76.1	60.3	86.7	86.9	87.5	97.1	78.0
AlexNet+CAM	88.6	82.4	55.6	83.1	65.7	89.0	88.4	**90.0**	98.0	82.3
AlexNet+CAM+Refine	**88.7**	**83.0**	56.9	**83.8**	**67.7**	**89.2**	**89.7**	89.5	**98.3**	**83.0**
AlexNet+CAM+Refine w/o AvePool	88.3	82.6	**57.1**	83.5	67.6	89.1	89.5	89.8	98.0	82.8
VGG16	93.4	88.7	72.5	91.9	72.1	**94.1**	92.3	91.9	**98.8**	88.4
VGG16+CAM	93.5	90.7	76.7	**93.8**	75.3	92.7	92.1	92.5	98.3	89.5
VGG16+CAM+Refine	**94.1**	**90.8**	**79.6**	93.3	**77.2**	93.2	92.1	**92.8**	98.6	**90.2**
VGG16+CAM+Refine w/o AvePool	93.6	90.7	77.2	93.2	76.6	93.4	**92.8**	92.5	**98.8**	89.9

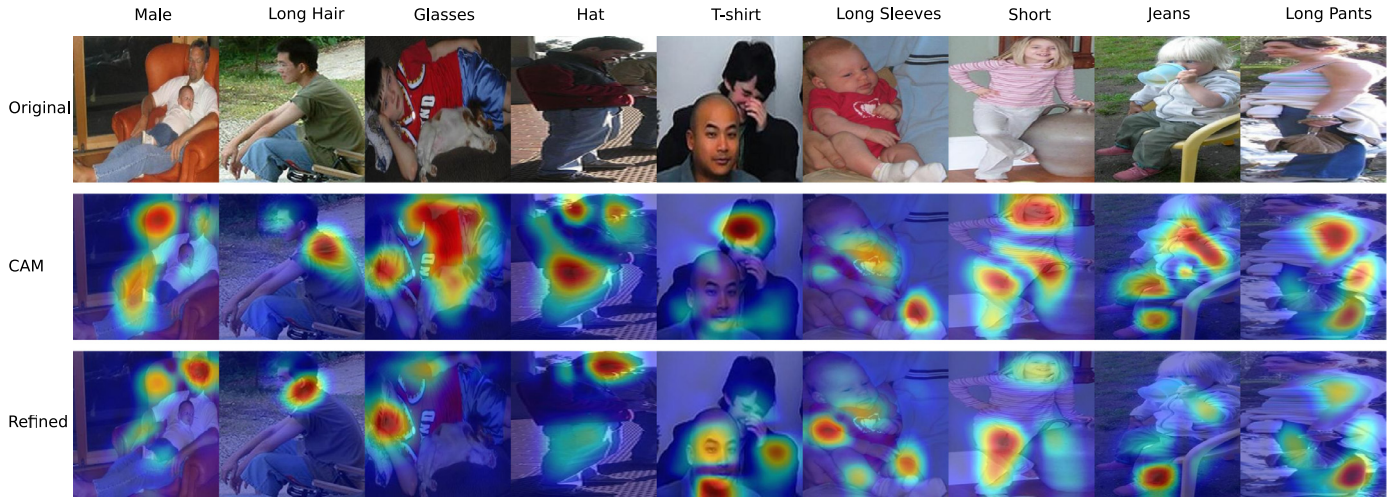

Fig. 6. Sample results of attention heat map refinement. Top row: original images and the considered human attribute. Middle row: the attention heat map from VGG16+CAM. Bottom row: the refined attention heat maps from VGG16+CAM+Refine.

Table 2
The mAP performance of the proposed method and eight comparison methods on Berkeley Attribute of Human People Dataset.

AP(%)	male	long hair	glasses	hat	t-shirt	long sleeves	shorts	jeans	long pants	mAP
Poselet [9]	82.4	72.5	55.6	60.1	51.2	74.2	45.5	54.7	90.3	65.18
DPD [11]	83.7	70.0	38.1	73.4	49.8	78.1	64.1	78.1	93.5	69.88
Joo et al. [10]	88.0	80.1	56.0	75.4	53.5	75.2	47.6	69.3	91.1	70.7
PANDA [12]	91.7	82.7	70.0	74.2	49.8	86.0	79.1	81.0	96.4	79.0
Park et al. [13]	92.1	85.2	69.4	76.2	69.1	84.4	68.2	82.4	94.9	80.2
Gkioxari et al. [14]	92.9	90.1	77.7	93.6	72.6	93.2	**93.9**	92.1	98.8	89.5
Gkioxari et al. [20]	92.8	88.9	82.4	92.2	74.8	91.2	92.9	89.4	97.9	89.2
Deep-Context [15]	**95.0**	**92.4**	**89.3**	**95.8**	**79.1**	**94.3**	93.7	91.0	**99.2**	**92.2**
VGG16+CAM+Refine	94.1	90.8	79.6	93.3	77.2	93.2	92.1	**92.8**	98.6	90.2

nition. Specifically, we choose Poselet [9], Deformable Part Descriptors (DPD) [11], Joo et al. [10], PANDA (Pose Aligned Networks for Deep Attribute) [12], Park et al. [13], Gkioxari et al. [14], Gkioxari et al. [20] and Deep-Context [15] for comparison. Among these eight comparison methods, the first six methods are part based and the seventh one, i.e., Gkioxari et al. [20], is a contextual cues based, while the eighth one, i.e., Deep-Context [15], is both part and context based approach for human attribute recognition. Table 2 summarizes the mAP of these methods and the proposed method (VGG16+CAM+Refine) on the testing data in the Berkeley Attributes of Human People Dataset. For all the comparison methods, their mAP performance are directly copied from their respective papers. We can see that, VGG16+CAM+Refine achieves second best mAP of 90.2%, while the best mAP is 92.2% from Deep-Context [15].

We also test the proposed methods on the WIDER Attribute Dataset introduced by Li et al. [15]. The results are reported in Tables 3 and 4. We can see that the proposed methods achieve better performance than Deep-Context [15] on the WIDER Attribute Dataset. Note that, different from the proposed method, Deep-Context [15] considers the contextual information besides the relevant regions considered in the proposed method. We believe the proposed method can be enhanced by further considering contextual information as in Deep-Context and we will study this problem in our future work.

4.2. Analysis

We use one example to illustrate the importance of using both loss functions, i.e., the classification loss function and the exponential loss function in the proposed method. Fig. 7(a) shows an image where we want to recognize the attribute of “has glasses”. Fig. 7(b) is the attention heat map from VGG16+CAM and this heat map contains multiple highlighted regions, among which there is only one, indicated by a red circle, is the real relevant region for the considered attribute. This is actually the result from minimizing the classification loss function. If we only use the exponential loss function, the fine-tuning is totally unsupervised – the result-

Table 3
Comparing mAP performance on the test set of WIDER Attribute Dataset.

Methods	R-CNN [34]	R*CNN [20]	Deep-Context [15]	VGG16	VGG16+CAM	VGG16+CAM+Refine
mAP(%)	80.0	80.5	81.3	81.7	**82.5**	**82.9**

Table 4
AP performance for each attribute on the test set of WIDER Attribute Dataset.

AP(%)	male	long hair	sungls	hat	tshirt	long sleeves	formal	short	jeans	long pants	skirt	face mask	logo	stripe	mAP
VGG16	94.9	83.8	70.1	92.5	**77.8**	95.0	78.5	**89.3**	72.5	96.2	79.2	70.1	87.6	56.4	81.7
VGG16+CAM	95.0	84.7	69.7	93.5	77.3	95.3	80.6	88.3	74.1	96.2	80.2	72.3	**88.0**	**60.0**	82.5
VGG16+CAM+Refine	**95.3**	**85.2**	**71.3**	**93.6**	77.7	**95.5**	**80.7**	88.9	**74.9**	**96.3**	**80.7**	**72.6**	87.5	**60.0**	**82.9**

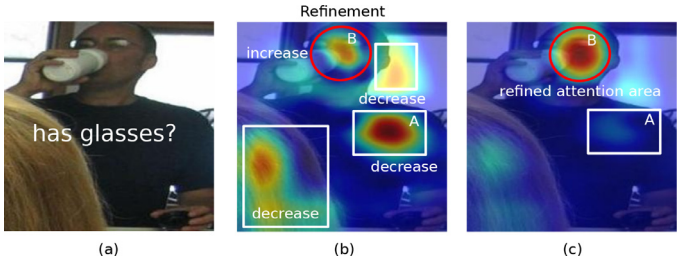

Fig. 7. An example for illustrating the importance of the two loss functions in the proposed method. (a) An image for recognizing the attribute of "has glasses"; (b) Attention heat map extracted by VGG16+CAM; (c) Refined attention heat map extracted by VGG16+CAM+Refine.

ing attention heat map will be highly concentrated on one small region, but this region may not be relevant to the considered attribute. By using both loss function, we can see that, the refined attention heat map, as shown in Fig. 7(c), can get not only more concentrated but also more attribute relevant.

Additionally, in some examples in Fig. 6, the most highlighted area completely "moves" from one region to another region based on the attention heat map refinement. This is due to the change of the heat map value. As shown in Fig. 7(b), both regions A and B are highlighted, although A has the highest heat map value, before the attention heat map refinement. With the refinement, the heat map value of region B increases to be the most highlighted region, while the heat map value in region A is decreased by the suppression. This leads to a visual effect of the moving of the highlighted area, Fig. 7(c).

Finally, CAM network was initially developed for object localization. An interesting question would be whether the proposed method with added heat map refinement component can be used for improving object localization. Based on the results in Fig. 6, we believe the proposed method can help improve the CAM-based object localization. However, it requires that only a single object of interest exists in the image, since the proposed loss function and attention heat map refinement component could not focus on multiple objects simultaneously.

5. Conclusion

In this paper, we proposed to use Class Activation Map (CAM) networks for human attribute recognition, without requiring prior correspondence between the human body parts and the attributes. Based on the CAM networks, we further introduced a new component to extract and refine the attention heat map for each training image. A new exponential loss function was defined to measure the appropriateness of the attention heat map. Considering this new loss function and the original classification loss function, the proposed method can highlight the attribute-relevant regions with higher concentration in the CAM training. We compared the performance of the proposed method with state-of-the-art attribute recognition methods on the Berkeley Attributes of Human People Dataset and WIDER Attribute Dataset. The proposed method achieved the second best mAP performance of 90.2% on Berkeley Attribute of Human People Dataset and the best mAP performance of 82.9% on WIDER Attribute Dataset.

Acknowledgments

This work was supported in part by UES Inc./AFRL-S-901-486-002, NSF-1658987, NSFC-61672376 and NCPTT-P16AP00373.

References

[1] D. Vaquero, R. Feris, D. Tran, L. Brown, A. Hampapur, M. Turk, Attribute-based people search in surveillance environments, in: IEEE Workshop on Applications of Computer Vision, 2009, pp. 1–8.
[2] K. Duan, D. Parikh, D. Crandall, K. Grauman, Discovering localized attributes for fine-grained recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3474–3481.
[3] C. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 453–465.
[4] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1778–1785.
[5] N. Kumar, A. Berg, P. Belhumeur, S. Nayar, Attribute and simile classifiers for face verification, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 365–372.
[6] C. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 951–958.
[7] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2, 2006, pp. 2169–2178.
[8] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[9] L. Bourdev, S. Maji, J. Malik, Describing people: a poselet-based approach to attribute classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1543–1550.
[10] J. Joo, S. Wang, S.-C. Zhu, Human attribute recognition by rich appearance dictionary, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 721–728.
[11] N. Zhang, R. Farrell, F. Iandola, T. Darrell, Deformable part descriptors for fine–grained recognition and attribute prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 729–736.
[12] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, Panda: pose aligned networks for deep attribute modeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1637–1644.
[13] S. Park, S.-C. Zhu, Attributed grammars for joint estimation of human attributes, part and pose, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2372–2380.
[14] G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2470–2478.
[15] Y. Li, C. Huang, C.C. Loy, X. Tang, Human attribute recognition by deep hierarchical contexts, in: European Conference on Computer Vision, Springer, 2016, pp. 684–700.
[16] X. Fan, K. Zheng, Y. Lin, S. Wang, Combining local appearance and holistic view: dual-source deep neural networks for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1347–1355.

[17] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Is object localization for free?-weakly-supervised learning with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 685–694.
[18] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in Neural Information Processing Systems, 2014, pp. 487–495.
[19] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[20] G. Gkioxari, R. Girshick, J. Malik, Contextual action recognition with r* cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1080–1088.
[21] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[22] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1597–1604.
[23] K.-Y. Chang, T.-L. Liu, H.-T. Chen, S.-H. Lai, Fusing generic objectness and visual saliency for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 914–921.
[24] X. Shen, Y. Wu, A unified approach to salient object detection via low rank matrix recovery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 853–860.
[25] Q. Wang, Y. Yuan, P. Yan, X. Li, Saliency detection by multiple-instance learning, IEEE Trans. Cybern. 43 (2) (2013) 660–672.
[26] M.-M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[27] C. Kanan, M. Tong, L. Zhang, G. Cottrell, Sun: top-down saliency using natural statistics, Vis. Cogn. 17 (6–7) (2009) 979–1003.
[28] J. Yang, M.-H. Yang, Top-down visual saliency via joint crf and dictionary learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2296–2303.
[29] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[30] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1717–1724.
[31] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[32] R. Cinbis, J. Verbeek, C. Schmid, Weakly supervised object localization with multi-fold multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell 2015 (2015).
[33] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.
[34] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014 (2014).
[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[37] M. Lin, Q. Chen, S. Yan, Network in network, in: Proceedings of the International Conference on Learning Representations, 2013.
[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int J. Comput. Vis. 115 (3) (2015) 211–252.