Attribute Recognition from Adaptive Parts

Luwei Yang[1]
luweiy@sfu.ca
Ligeng Zhu[2]
zhuligeng@zju.edu.cn
Yichen Wei[3]
yichenw@microsoft.com
Shuang Liang[4]
shuangliang@tongji.edu.cn
Ping Tan[1]
pingtan@sfu.ca

[1] Simon Fraser University
Vancouver, Canada
[2] Zhejiang University
Hangzhou, China
[3] Microsoft Research Asia
Beijing, China
[4] Tongji University
Shanghai, China

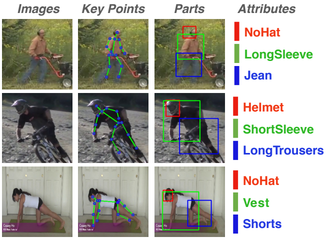

Figure 1: Given an image, we estimate the key points, generate object parts accordingly, and predict attributes of each part. The learning is end-to-end in a single deep neural network.

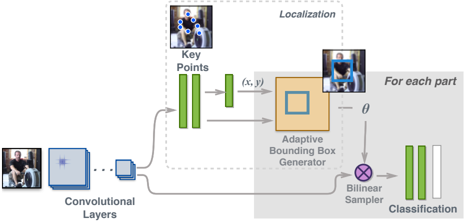

Figure 2: Overview of network architecture. It consists of initial convolutional feature exaction layers, a key point localization, an adaptive bounding box generator for each part, and the final attribute classification network for each part.

We focus on the problem of object attribute recognition. Previous part-based attribute recognition approaches perform part detection and attribute recognition in separate steps. The parts are not optimized for attribute recognition and therefore could be sub-optimal. In this paper, we present an end-to-end deep learning approach to overcome the limitation.

In our network architecture, instead of training part detector and attribute classifier separately, both key point estimation and attribute recognition are learned jointly in a multi-task setting. Figure 1 shows the example of our pipeline. We firstly estimate object key points as an auxiliary task. Because the definition of key point is clear, their annotation is less ambiguous than part bounding boxes. From the key points, the object parts are generated adaptively, with free parameters to be learned for adjusting its spatial extent. This adaptive part generation is inspired by the recent spatial transformer network [1], which can learn image spatial transformation from the image classification goal. Instead of applying the transform to the whole image [1], we apply a spatial transform for each part and use the bilinear sampler [1] to warp the image features for subsequent attribute recognition. The whole network is learned end-to-end in a multi-task setting, with attribute classification as the main task and key point prediction as the auxiliary one.

The network framework is illustrated in Figure 2. It consists of a convolutional network for feature extraction, a localization network for key point estimation, an adaptive bounding box generator for each part, and part based feature sampler and attribute classifier.

Key Point Estimation As shown in Figure 2, three fully-connected layers are appended on the top of the convolutional features as the regression network, with the output dimensions $2N$ (N is the number of key points). We use L2 distance loss for key point estimation, $\sum_i^N \|\hat{p}_i - p_i\|_2^2$, where $\hat{p}_i, p_i \in \mathbb{R}^2$ are the normalized ground truth and estimation for key point i.

Adaptive Part Generation Some attributes are clearly associated with certain object parts. We specify a subset of key points $\mathcal{P}_t$ for each part t. The initial part bounding box $b_t = [w_t, h_t, x_t, y_t]$ encodes the origin and size of a rectangle area that covers all key points in $\mathcal{P}_t$, and can be obtained by finding the maximum and minimum of the subset. The final bounding box is defined as $[w_t(1+\Delta_w), h_t(1+\Delta_h), x_t+\Delta_x, y_t+\Delta_y]$, with additional adjustment parameters $\Delta = [\Delta_w, \Delta_h, \Delta_x, \Delta_y]$ to be learned adaptively.

Bilinear Sampling and Attribute Recognition For each part bounding box, the convolutional feature maps are warped accordingly. We use the Bilinear Sampler in [1] that warps the local feature via bilinear interpolation, and it serves as a bridge that allows the gradient of attribute classification flow into the localization network. The warping employs a 2×3 affine transformation, parameterized as

$$\theta_t = \begin{bmatrix} w_t(1+\Delta_w) & 0 & x_t+\Delta_x \\ 0 & h_t(1+\Delta_h) & y_t+\Delta_y \end{bmatrix}. \tag{1}$$

The transformation θ_t warps the local coordinates, the corresponding content can be sampled by bilinear interpolation subsequently. For attribute parsing, the Softmax multi-class classifier is adopted.

Our approach is validated on human attribute recognition on two datasets, via extensive experiment comparison. The comparable results show the effectiveness of jointly training of localization and classification task.

[1] Max Jaderberg, Karen Simonyan, and Andrew Zisserman. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.