Image Parsing with Stochastic Grammar: The Lotus Hill Dataset and Inference Scheme

Benjamin Yao[1,2], Xiong Yang[1] and Tianfu Wu[1,2]
[1]Lotus Hill Institute for Computer Vision and Information Science, Ezhou, China
[2]Department of Statistics University of California, Los Angeles

We present the LHI dataset, a large-scale ground truth image dataset, and a top-down/bottom-up scheme for scheduling the inference processes in stochastic image grammar (SIG). Development of stochastic image grammar needs ground truth image data for diverse training and evaluation purposes, which can only be collected through manual annotation of thousands of images for a variety of object categories. This is too time-consuming a task for each research lab to do independently and a centralized general purpose ground truth dataset is much needed. In response to this need, the Lotus Hill Institute (LHI), an independent non-profit research institute in China, is founded in the summer of 2005. It has a full time annotation team for parsing the image structures and a development team for the annotation tools and database construction. Each image or object is parsed, semi-automatically, into a parse graph where the relations are specified and objects are names using the WordNet standard. The Lotus Hill Institute has now over 500,000 images (or video frames) parsed, covering 280 object categories. Fig. 1 shows an example parse tree of car. Since this ground truth annotation is aimed at stochastic image grammar researches, it has more hierarchic structures, finer annotation and broader scope than other datasets collected in various groups, such as Berkeley, MSRC, Caltech, and MIT. Fig. 2 shows a comparison of segmentation labelmap between MSRC dataset and LHI dataset.

In computing, we present a method for scheduling bottom-up and top-down processes in image parsing with And-Or graph (AoG) for advancing performance and speeding up on-line computation. For each node in an AoG, two types of bottom-up computing processes and one kind of top-down computing process are identified: (1) inference based on features directly computed from raw image data, (2) inference based on binding the hypotheses of children nodes, and (3) inference based on predicting from hypotheses of parent nodes. The three computing processes are called α, β and γ processes, respectively. At the off-line training stage, we individually train the three processes and numerically measure their information contributions as heuristics for inference. At the on-line inference stage, we schedule the three computing processes of different nodes in the AoG to parse an input image, and the scheduling is adaptively guided by minimizing the conditional entropy, and is a recursive algorithm. Through scheduling, the performance is improved because the explicit integration of α, β and γ processes does not leave unturned any possible computing ways for each node in the AoG, and the computing time is speeded up because the explicit ordering of α, β and γ processes does not turn any node twice. We demonstrate the scheduling inference on human faces.

Fig.1 Parse tree in the Lotus Hill Institute dataset

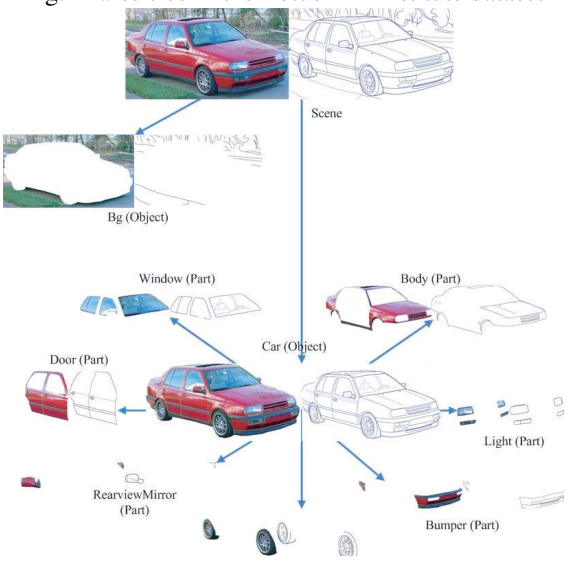

Fig.2 Segmentation labelmap of MSRC and LHI dataset

(a) Original Image (b) MSRC annotation (c) LHI annotation