Understanding Image Structure via Hierarchical Shape Parsing

Xianming Liu[†] Rongrong Ji[‡] Changhu Wang[§] Wei Liu[♮] Bineng Zhong[♯] Thomas S. Huang[†]
[†]University of Illinois at Urbana-Champaign [‡]Xiamen University
[§]Microsoft Research [♮]IBM T. J. Watson Research Center [♯]Huaqiao University

Understanding structure of images is one of fundamental challenges in the computer vision community and beyond [4][7]. It is commonly agreed in the cognitive research [4] that such structure is hierarchical in general, and visual appearances along the structure range from coarse to fine configurations. These evidences result in multi-scale image representation [5][9].

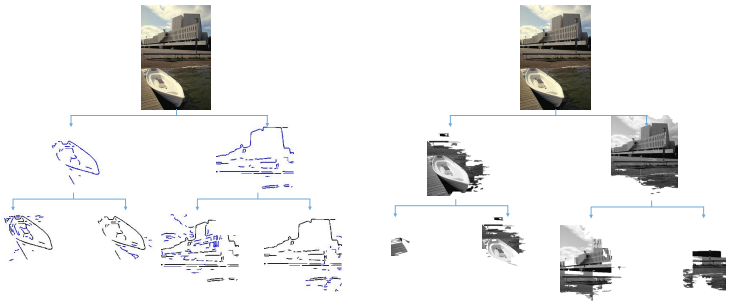

Figure 1: Example of the hierarchical shape parsing. Left: the original image and the hierarchical edge parsing tree. For better visualization, the edge segments of each node are in blue, while the ones of its ancestor nodes are in black. Right: the results after appearance bundling. To illustrate the coarse-to-fine phenomenon, the appearances of child nodes are integrated to the parent node.

Inspired by our previous work on the scale of edges [6], we are motivated to parse hierarchical structure of image components according to their scale distributions and shapes, as the example shown in Figure 1. By "parsing", we mean to detect visual components (such as parts of objects) indicated by shapes, which will be organized into hierarchical structure according to the coarse-to-fine cognitive rule (such as "part of", and "outline-and-details" relations) to generate a multi-scale representation. To further improve the discrimination of parsed visual components, appearances of image regions are embed correspondingly in a statistical way.

As for visual search and matching, human brains recognize objects based on not only visual appearances, but also heavily relying on structure, according to recent advance in cognitive study [3]. With such a structure, we simulate the human cognitive mechanism in visual search as a *conditional matching process*: matchings are formulated as Markov Process, with dependencies defined by the structural "visual ontology" along the hierarchical parsing tree. By simple statistical inference, we derive a *hierarchical structural pooling strategy* to approximate the above process when building region descriptions.

Our approach starts with building a *Hierarchical Edge Tree* in a top-down manner. Given such a coarse-to-fine shape structure, local regions are further appended onto corresponding tree nodes to increase the discriminative ability. When matching two parsing trees / subtrees, an structural appearance pooling operation is performed based on a *Markov Process*, in which the parent-children dependency is forced. By this pooling operation, the tree hierarchy is encoded together with the appended appearance information, and it avoids the time-consuming recursive subtree alignment schemes in existing works.

Successfully parsing the image structure at low level can benefit a wide variety of computer vision tasks, such as feature designing, scene understanding, object recognition and detection. In this paper, we show two exemplar applications about our scheme, including unsupervised objectness detection [1, 8]. Quantitative experiments with comparisons to the state-of-the-arts show advantages of our algorithm. Figure 2 shows the performance

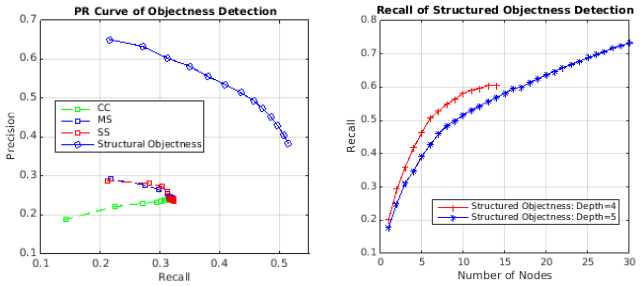

Figure 2: (a) Performance comparison: PR-curves of different methods on the task of objectness detection, by varying the thresholds. (b) Recall of (*Structural Objectness*) on object proposals. Tested on VOC 2007 test dataset.

of the proposed method being applied to objectness detection compared with state-of-the-arts [2][1][8]. With competitive and even better performance, we dramatically reduce the number of windows for object proposal, as shown in Table 1.

As for future work, we are planning to involve structural inference into current algorithm to perform supervised learning, and employ random algorithms to add perturbations in the scale splitting to improve the robustness.

Table 1: The number of windows (candidate regions) used in [1] and the structural objectness

	HoG[2]	Objectness [1]
#Win	1,000	1,000
	Selective Search[8]	Proposed
#Win	395	14

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 73–80. IEEE, 2010.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[3] J J DiCarlo, D Zoccolan, and N C Rust. How does the brain solve visual object recognition? *Neuron*, 2012.

[4] David H Hubel. *Eye, brain, and vision*. Scientific American Library New York, 1988.

[5] Jan J Koenderink. The structure of images. *Biological cybernetics*, 50 (5):363–370, 1984.

[6] Xian-Ming Liu, Changhu Wang, Hongxun Yao, and Lei Zhang. The scale of edges. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 462–469. IEEE, 2012.

[7] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, 63(2):113–140, 2005.

[8] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[9] Andrew Witkin. Scale-space filtering: A new approach to multi-scale description. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 150–153. IEEE, 1984.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.