

Semantic Object Segmentation

*King N. Ngan (IEEE Fellow), The Chinese University of Hong Kong, China
Hongliang Li, University of Electronic Science and Technology China, China
knngan@ee.cuhk.edu.hk*

We often say “a picture is worth a thousand words”. This is because the semantic content of a picture cannot be easily described by a few words. Have you ever wondered when you look at a picture, how do your eyes and brain extract information from the scene? What is the basic unit for the image understanding when a person observes an image or video scene? The possible answer may lie in the semantic objects, which can provide the meaningful cues for finding the scene content. However, in order to understand the semantics of the scene content, we need to separate or segment the content into its constituent parts. This has posed many challenges to vision researchers over the past decade and so far arbitrary object segmentation from images or videos is still an unsolved problem.

The extraction of semantic objects plays an important role in digital image/video processing, pattern recognition, and computer vision. It provides an efficient way to bridge the primary image data and semantic (and thus meaningful) content in image understanding. The task of segmenting a semantic object can be found in many fields, such as video surveillance, content-based video summarization, content-based coding application, computer vision, video conferencing, and digital entertainment [1]. A semantic object can be exploited to provide the user with the flexibility of content-based access and manipulation such as fast indexing from video databases and efficient coding of regions of interest [2]. An interesting example can be illustrated by the cartoon face in real-time video, where the extracted face region is used to construct the basic face contour, and helps to perform the animation of face expression.

Earlier techniques of image segmentation can be traced back over forty years, which aimed to extract some uniform and homogeneous regions based on the edge, color, or other low-level features. These methods are usually called non-semantic modes. Among them, a boundary-based method is mostly used in segmenting image

regions, which first employs the gradient operator to extract the edges, and then groups them into object contour. The main drawback of this method is the lack of robustness during the contour closure due to the difficult computation of the region's boundaries.

In order to satisfy the future content-based multimedia analysis, more and more researchers seek for the semantic object segmentation by discovering the meaningful regions. However, this method usually needs to first identify the position of semantic objects from images by performing the spatial classification based on a certain prior knowledge. An intrinsic problem of this progress is that there is no unified method that is available for detecting all semantic objects. To avoid the complicated object recognition process, an interactive image segmentation approach based on graph-cut optimization has been developed which extracts a semantic object at the cost of interactive effort on the part of the user [3]–[5]. These methods can provide users with much better segmentation performance than automatic ways.

Fortunately, some specific objects of interest (e.g., face) can be segmented in an unsupervised manner by designing appropriate detectors based on physical model or training scheme [6]. Most work on this topic can be divided into two classes. The first class focuses on the design of robust detector for identifying the specific semantic object, which aims to provide the best candidates for the object of interest. The second class is to present the efficient clustering algorithm for improving the quality of extracting similar pixels. Recently, some joint methods were proposed to perform the semantic object detection and segmentation simultaneously from images [7]. Generally, the final object regions need further modification due to the coarse object segmentation.

Object segmentation based on attention model is another important approach for segmenting semantic object, which allow us to find interesting objects by simulating human

IEEE COMSOC MMTC E-Letter

visual characteristic. Unlike the previous methods, attention-based scheme aims to segment the meaningful physical entities that are more likely to attract viewers' attention than other objects. Most objects of interest tend to be the attention objects that have distinctive features from their surroundings [1]. A saliency-based visual attention model for rapid scene analysis was first presented in [8], which combined multi-scale image features into a single topographical saliency map. The application of this model on object extraction from color images was reported in [9], which formulated the attention objects as a Markov random field by integrating computational visual attention mechanisms with attention object growing techniques. Based on the visual attention idea, several saliency models are successfully constructed recently to extract the object of interest in different video sequences, such as the facial saliency model [10] and focused saliency model [11].

Although some achievements in semantic object segmentation have been obtained, the limitations of this method are quite evident. It is still a challenge to provide an efficient segmentation solution to extract the semantic object accurately in unsupervised manner, such as the perfect video object planes in MPEG-4 video standard. More work should be carried out to improve the segmentation schemes for meaningful semantic objects.

REFERENCES

- [1] H. Li, and King N. Ngan, "Automatic video segmentation and tracking for content-based applications", *IEEE Communications Magazine*, vol. 45, no. 1, pp. 27 - 33, 2007.
- [2] T. Meier and K.N. Ngan, "Automatic segmentation of moving objects for video objects plane generation," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 525-538, September 1998.
- [3] C. Rother, V. Kolmogorov, and A. Blake, "GrabCutl interactive foreground extraction using iterated graph cuts," in *Proc. SIGGRAPH 2004*, pp. 309-314, 2004.
- [4] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," in *Proc. SIGGRAPH*, pp. 303-308, 2004.
- [5] J. Wang, P. Bhat, R. A. Colburn, M. Agrawala, and M. F. Cohen, "Interactive video cutout," in *Proc. SIGGRAPH*, pp. 585-594, 2005.
- [6] H. Li, King N. Ngan, and Qiang Liu, "FaceSeg: automatic face segmentation for real-time video", *IEEE Transactions on Multimedia*, vol.11, no.1, pp.77-88, 2009.
- [7] B. Wu, and R. Nevatia, "Simultaneous object detection and segmentation by boosting local shape feature based classifier", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 17-22 June 2007.
- [8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1254-1259, Nov. 1998.
- [9] J. Han, King N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised Extraction of Visual Attention Objects in Color Images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 141-145, Jan. 2006.
- [10] H. Li, and King N. Ngan, "Saliency model based face segmentation in head-and-shoulder video sequences", *Journal of Visual Communication and Image Representation*, Elsevier Science, vol. 19, no. 5, pp 320-333, 2008.
- [11] H. Li, and King N. Ngan, "Unsupervised video segmentation with low depth of field", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no.12, pp. 1742-1751, 2007.



King N. Ngan received the Ph.D. degree in Electrical Engineering from the Loughborough University in U.K. He is currently a chair professor at the Department of Electronic

IEEE COMSOC MMTC E-Letter

Engineering, Chinese University of Hong Kong, and was previously a full professor at the Nanyang Technological University, Singapore, and the University of Western Australia, Australia.

Professor Ngan is an associate editor of the Journal on Visual Communications and Image Representation, U.S.A., as well as an area editor of EURASIP Journal of Signal Processing: Image Communication, and served as an associate editor of IEEE Transactions on Circuits and Systems for Video Technology and Journal of Applied Signal Processing. He chaired a number of prestigious international conferences on video signal processing and communications and served on the advisory and technical committees of numerous professional organizations. He is a general co-chair of the IEEE International Conference on Image Processing (ICIP) to be held in Hong Kong in 2010. He has published extensively including 3 authored books, 5 edited volumes and over 200 refereed technical papers in the areas of image/video coding and communications.

Professor Ngan is a Fellow of IEEE (U.S.A.), IET (U.K.), and IEAust (Australia), and an IEEE Distinguished Lecturer in 2006-2007.



Hongliang Li (M'06) received his Ph.D. degree in Electronics and Information Engineering from Xi'an Jiaotong University, China, in 2005. From 2005 to 2006, he joined the visual signal processing and communication laboratory (VSPC) of the Chinese University of Hong Kong (CUHK) as a Research Associate. From 2006 to 2008, he was a Postdoctoral Fellow at the same laboratory in CUHK. He is currently a Professor in the School of Electronic Engineering, University of Electronic Science and Technology of China. His research interests include image segmentation, object detection and tracking, image and video coding, and multimedia communication system.