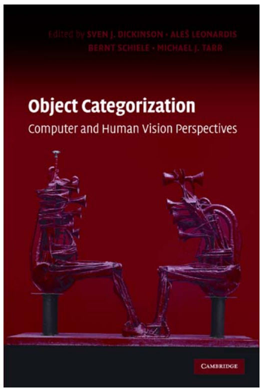
Book Review

Object Categorization, Computer and Human Vision Perspectives

Sven J. Dickinson, Ales Leonardis, Bernt Schiele, and Michael J. Tarr, Eds., 536 pp., ISBN 978-0-521-88738-0, Cambridge University Press (2009), $125.00 hardcover.

Reviewed by Jie Yu and Dhiraj Joshi Kodak Research Labs, Eastman Kodak Company, Rochester, New York


Edited by SVEN J. DICKINSON · ALEŠ LEONARDIS
BERNT SCHIELE · MICHAEL J. TARR
Object Categorization
Computer and Human Vision Perspectives
CAMBRIDGE

This book is a collection of articles written by some of the leading experts in vision research and approaches object categorization from two distinct (human and computer vision) perspectives. The preface provides a historical background of how the editors organized generic object recognition workshops at IEEE Conference on Computer Vision and Pattern Recognition and IEEE Conference on Computer Vision over a span of ten years and how that effort resulted in the preparation of this book. The editors also discuss their rationale for having a flat organization with interleaved computer and human vision articles in order to preserve an overall integrative theme binding different pieces of work into a comprehensive book. Contrary to the organization of the book, the authors of this review decided to separate discussion on human vision and computer vision chapters to maintain coherency in their review.

Chapter 1 offers an excellent overview of the changing paradigms, challenges, emphasis, and scope of object-recognition research over the last several decades. The author takes the readers on a historical trip while discussing evolution of approaches from coarse 3-D geometrical shapes (1970s), more sophisticated CAD models (1980s), appearance-based recognition (1990s), and local representations (2000s). The chapter also discusses the evolving roles of shape, structure, visual grouping, object-to-scene mapping, and 3-D models with respect to object recognition.

The discussion on human vision is marked by a few defining themes that focus on: (i) framework for human vision, (ii) detailed working of the human visual system, (iii) adaptation of the visual system with age (and differential development in people with disabilities), and (iv) roles of senses, conditioning, and memory in visual recognition. In our review, we attempt to group discussion of chapters by themes wherever possible.

Chapter 4 gives an epistemological view of object recognition and discusses the roles of context and conceptual knowledge in vision. The author's main criticism is the inherent assumption in vision algorithms that an image can be completely interpreted by a finite set of concepts or objects. Chapter 8 introduces interface theory, an interesting theory inspired from Darwin's natural selection. The author argues that unlike conventional vision theories, perception is best treated as a user interface between an organism and the objective world. The author elaborates on the role of vision in animal and insect lives in the light of the proposed theory.

Chapter 2 presents a neuro-computational discussion on object understanding. It is interesting to note that relatively primitive biologically inspired models (mimicking visual cortex V1) perform as well as the state-of-the-art vision algorithms on Caltech 101 dataset. In Chapter 6, the reader finds a thorough presentation of the functional organization of the human visual system. The authors lead us through human neural responses to several diverse object categories and discuss biological implications for algorithmic object recognition. They also demonstrate how the human neural response with respect to recognition adapts with age. Chapter 10 presents evidence for parts-based structural coding in the human brain and emphasizes the importance of reconciliation between biological and computer vision in terms of structural representation of objects. Chapter 12 offers yet greater insights into the working of biological visual systems with discussions of the neuro-computational architecture, categorization evidence from human and monkey experimental data, and a general positivism about productive collaboration between human and computer vision researchers. Chapter 14 links neurological representation of objects and faces with complex natural scenes. In particular, the chapter describes strategies used by the human brain to computationally solve the object invariance problem.

Chapter 16 takes a step back and discusses relevant work in object discovery and primate visual processing results to propose a theoretical framework for object recognition. Of particular interest is the discussion on the apparent ease with which infants learn to recognize the world in spite of scarce and unannotated training data, little or no feedback, and poor-quality imagery (as a result of the still-developing visual system). Chapter 18 explores the problem of a face-related visual recognition in humans with a focus on differential development of a face-related cortex in people with autism. Authors point out that although children exhibit almost adult-like neural patterns with respect to perceiving places and objects, face-related neural activation develops slowly with age.

Chapter 20 underlines the importance of acquired knowledge and proposes a memory-based human recognition model. The chapter discusses the presence of coarse top-down processing in the brain which can exert a biasing influence on bottom-up recognition. On a similar note, Chapter 22 emphasizes the combination of sensory information and previous knowledge about the environment for recognition; the underlying focus being an experience-dependent differential processing for optimal decision making. In Chapter 24, the authors delve deeper and link neural behavior with correlation structures in natural scenes. The discovery of positive and negative neural interactions based on prior tunings of neurons is intriguing.

Chapter 26 concludes the human vision discussion by emphasizing the synergy between perceptual research, computer graphics, and computer vision. The chapter draws upon the human ability to perceive material properties of objects in pictures and hypothesizes that the role of nonvisual senses (touch and hearing) in object understanding is significant. More specifically, the authors discuss studies that show how human visuo-haptic memory guides visual recognition.

Chapter 3 serves as starting point of the computer vision part of this book. It summarizes the fundamental problems of visual recognition areas, which include what to recognize, why visual recognition is needed, how it can be categorized, and how fast an ideal algorithm should be. It further surveys the recent attempts on Caltech 101 and 256 data sets to solve those problems and points out some future research directions.

Chapters 5 and 27 discuss the methods that leverage 3-D models to facilitate 2-D image matching from different aspects. Chapter 5 proposes a 3-D volumetric part representation for generic object-recognition tasks. Compared with surface-based schemes, the volumetric method enjoys several advantages, such as better view-point tolerance and more natural shape abstraction. The general process of extracting the part-based descriptions involve hierarchical grouping of low-level features such as edge descriptors to mid-level features such as generalized cylinders, which leads to 3-D part descriptors and the corresponding connectivity graph. Chapter 27 summarizes several techniques proposed by the authors to solve the 2-D image-matching problem by considering the 2-D image as an embedding of a 3-D model. To solve the posed variation problem, the authors propose to use stereo matching to find image correspondence with minimal cost scores. To handle the deformation and articulation problems, geodesic distance-based features have been found to be very robust. It also points out that gradient direction is less sensitive to lighting conditions on nonisotropic surfaces.

Chapter 7 offers an extensive survey on object recognition via functionality reasoning. In areas such as computer vision, the objects are often represented as features or structures. However, the high-level analysis on functionality of the objects discovers that reasoning through functionality may lead to more general object recognition. This chapter summarizes the major contributions in this multidisciplinary area, especially in artificial intelligence, computer vision, and robotics.

Chapter 9 reviews the research on images and their associated text in the form of title, annotations, and surrounding text. It starts with a review of research works that analyze the relationship of images and their surrounding information. Then, it introduces two new research results from the coauthors: one trying to find word classifiers with low dimensional representations for scene annotation tasks and the other leveraging online ontologies such as Wikipedia to improve text query-based image retrieval. The remainder of the chapter discusses the interesting problem of how “depiction” and “modifier” in the text describes the images.

The representation of object structure has been a crucial problem that is relevant to both recognition generality and scalability. Different hierarchical representations have successfully been applied to object-recognition tasks, some of which are highlighted in Chapters 11, 13, 15 and 21. Chapter 11 argues that in principle a hierarchical composition should be learned in a bottom-up manner. Furthermore, this chapter briefly discusses an unsupervised solution that realizes these design principles. The key idea of this approach is to construct the hierarchy by considering both the parts’ similarity and their co-occurrence in close spatial proximity. Chapter 13 proposes a hierarchical compositional model for learning object structures from a small training set. The aim is to learn a general model using an AND-OR graph. The relationships of the nodes in different layers are defined by AND or OR operators, which allows a more general model of the object and is expected to handle occlusion or misdetection better. Chapter 15 proposes to learn hierarchical representation from an information theory point of view. In contrast to the unsupervised bottom-up method described in Chapter 11, it detects the most informative features in a supervised top-down fashion. First, the informativeness of top-level parts is determined by its correlation with images within and outside of a given category of images. Then, the hierarchy grows by breaking the informative parts into small pieces and constructing the next layer by selecting the informative ones among the smaller pieces. Finally, another layer of features called semantic features are learned to discover the visually dissimilar parts that are related to the same semantic concept by analyzing their context similarity. Chapter 21 introduces the well-known spatial pyramid representation for image classification. Different from many local features used in object-detection tasks, the authors have designed a holistic feature specifically for classification of an entire image without segmentation. In addition to revisiting the original method proposed in 2006, a survey discusses recent extensions and applications of the method and suggests several directions for future improvement.

It has been observed that different paradigms exist in almost every stage of the object-recognition process, e.g. local versus global feature extraction, generative versus discriminative modeling, and supervised versus unsupervised learning. Combining the advantages of these paradigms has been a long-sought-after goal of researchers. Chapter 17 presents a new attempt at integrating these paradigms. Specifically, they combine a generative topic model for localized gradients across categories and discriminative training of support vector machine models for object detection.

Unlike many research works using local appearance-based features in a bag-of-words representation, Chapter 19 aims at extracting features that depict pair-wise geometric relationships of feature points without considering the local appearance. Using weakly supervised samples, the proposed method is able to learn an abstract representation of the contours of the objects. Interestingly, it even outperforms a pure appearance-based feature, which suggests that the geometric relationship may be more robust for some object-detection tasks.

Feature extraction methods related to medial loci are discussed in Chapters 23 and 25. Chapter 23 introduces shock graphs to describe the topology of shapes. Compared with point- and contour-based features, these exploit the additional geometric information from adjacent contours, which makes them a powerful midlevel representation for object recognition. Chapter 25 reports certain findings that reveal the critical role of medial loci in human vision systems. It further proposes a new method to extract useful medial points using average outward flux calculation.

Overall, the book is very well written with expert insights into the development and current state of the art in human and computer vision research. Each chapter, though written independently, is quite complete and rich in content, background, and visions for the future. This book is an excellent attempt at presenting and understanding different object recognition paradigms and philosophies in both human and computer vision and can serve as a comprehensive resource for students and researchers alike.

Jie Yu joined Kodak Research Laboratories in 2007 as a research scientist. He received his PhD in computer science at the University of Texas at San Antonio. His research interests include multimedia information retrieval, machine learning, computer vision, and pattern recognition. He has published over 20 journal articles, conference papers, and book chapters in these fields. He received the Best Poster Paper Award of ACM CIVR 2008, the Student Paper Contest Winner Award of IEEE ICASSP 2006 and Presidential Dissertation Award of UTSA in 2006. He is a member of IEEE and ACM.

Dhiraj Joshi completed his PhD in computer science from Penn State University and joined Kodak Research Laboratory in 2007. His research interests include contextual inference-based image understanding, large-scale image retrieval, and content analysis in multimedia. He has been a research intern at IBM T.J. Watson Research Labs, New York, and the IDIAP Research Institute, Switzerland. In 2006, he was selected as an emerging leader in multimedia research to present at the Watson Emerging Leaders in Multimedia Workshop. He co-organized a special session on Image Aesthetics, Moods, and Emotions at the IEEE International Conference on Image Processing, 2008. He is a member of IEEE and currently serves as a Rochester chapter vice-chair of IEEE Signal Processing Society.