Real-World Image Annotation and Retrieval: An Introduction to the Special Section

James Z. Wang, *Senior Member*, *IEEE*, Donald Geman, *Senior Member*, *IEEE*, Jiebo Luo, *Senior Member*, *IEEE*, and Robert M. Gray, *Fellow*, *IEEE*

1 Introduction

Indexing and retrieving large quantities of image data is an extremely challenging and increasingly topical problem for both industry and academia. Massive volumes of image data are all around us—in personal and commercial collections and on public websites accessible via the Internet. According to a recent study by the market researcher IDC, digital camera sales rose 15 percent in 2006 to 105.7 million units worldwide [1]. A four-year old online photo sharing website, Flickr, has more than 40 million monthly visitors and 2 billion photos uploaded; in fact, in a single day, a few million photos are uploaded [2]. These developments have spurred enormous interest in digital images and a corresponding demand, both from the public and from industry, for better ways of cataloging, annotating, and accessing these data. This in turn has motivated researchers in pattern analysis and machine intelligence to address these tasks. Indeed, in a recent survey of the field of image annotation and retrieval, Wang et al. noticed an exponential growth over the last 10 years in the number of publications [3] arising from researchers in computer vision, database management, machine learning, mathematical statistics, and signal and image processing.

Technological advances have certainly been made. In this special section, we showcase a variety of these, with an emphasis on solving specific problems stemming from areas as diverse as art retrieval, biomedical information management, education, sensor networks, large-scale online personal and professional photo sharing, classification and filtering of Web images, computer forensics, and threat assessment.

The groundwork for many of these developments had already been laid by work in the computer vision community over the past decades. In particular, image retrieval has been an active area for almost two decades and image interpretation, especially object recognition, for even longer. The difference between recognition and what is referred to here as "image annotation" is basically one of scale and generality. Work in object recognition has largely focused on detecting instances of either specific examples or from a limited number of generic categories and often under controlled conditions; in contrast, in image annotation, we are concerned with supplying appropriate semantic labels for whatever might appear in the database, independently of the conditions under which the images were captured. Manual tagging at this scale is clearly hopeless. However, by leveraging advances in pattern analysis, machine learning, and statistical modeling, and exploiting massive computing power and massive learning sets, large-scale annotation, both in terms of image numbers and semantic concepts, seems to be underway.

To mirror the variety of topics and approaches, the Guest Editors represent expertise in diverse areas: James Z. Wang in the study and development of intelligent content and context-based image indexing, annotation, and retrieval systems; Donald Geman in statistics, machine learning, and computer vision; Jiebo Luo with extensive industrial experience in digital photo analysis; and Robert M. Gray in information theory and signal and image processing.

The special section proposal was submitted by the Guest Editors to the EIC on 9 January 2007, reviewed by three anonymous referees, and approved on 22 March 2007. The reviewers and the Editor-in-Chief, David Kriegman, regarded the topic as timely and important for the *TPAMI* community. The Guest Editors were not permitted to author a submission.

2 The Review Process

The Call for Papers solicited contributions in the following areas of interest:

- Web-scale machine-assisted annotation and retrieval of images;
- intelligent summarization and visualization for large-scale image retrieval;
- statistical and relevance feedback models for interactive search;
- annotation and retrieval for scientific discovery, biodiversity, and the arts; and
- image retrieval in computer forensics, threat assessment, and other security areas.

- *J.Z. Wang is with the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16803. E-mail: jwang@ist.psu.edu.*
- *D. Geman is with the Department of Applied Mathematics and Statistics, Johns Hopkins University, Clark Hall 302A, 3400 N. Charles Street, Baltimore, MD 21218. E-mail: geman@jhu.edu.*
- *J. Luo is with Kodak Research Laboratories, Eastman Kodak Company, Rochester, NY 14650. E-mail: jiebo.luo@kodak.com.*
- *R.M. Gray is with the Department of Electrical Engineering, Stanford University, 261 Packard Building, 350 Serra Mall, Stanford, CA 94305. E-mail: rmgray@stanford.edu.*

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org.

Submissions involving any real-world image data, including video data, 3D model data, and scanned document images, were welcomed. Ideally, a paper would apply novel ideas, or integrate existing technologies, in order to solve the problems faced by real users of specific image collections. Work of a purely methodological nature was considered less attractive.

We received a diverse pool of more than 50 submissions. Geographically, the contact authors were from the United States (17 papers), China (14 papers: mainland 9, Hong Kong 4, and Taiwan 1), Brazil (2), Canada (2), India (2), Singapore (2), Spain (2), The Netherlands (2), Algeria (1), France (1), Germany (1), Iceland (1), Israel (1), Italy (1), Japan (1), South Korea (1), and Switzerland (1). The authors proposed a very wide range of approaches and scenarios and represented many disciplines.

Since most manuscripts were assigned to at least three suitably knowledgeable reviewers, it was necessary to contact approximately 200 individuals, many of whom have published in *TPAMI* or other high-impact journals. Whereas more than 100 people accepted our invitations, some reviewers were asked to evaluate two or more papers. Most of the reviews were timely and detailed and we are very grateful to the reviewers for their diligence. In addition, all four Guest Editors examined, and often discussed at length, the manuscripts and reviews in order to reach a consensus for each paper. In the end, only 14 submissions were recommended for revision and a nonperfunctory second round of reviews to verify that satisfactory improvements were made. Of these 14, 11 were eventually accepted. The contact authors are from the United States (5), China (2; mainland 1, Hong Kong 1), Singapore (2), Israel (1), and The Netherlands (1).

Under current IEEE and *TPAMI* policy, substantially new concepts, ideas, implementations, and experiments are expected; some submissions were rejected on these grounds. For example, some papers were seen as minor variations on previously published conference papers and others merely tweaked the application presented in a previously published journal paper. Many reviewers noted a tendency to squeeze multiple publications out of a single idea. In particular, the Guest Editors feel that the interface with conference proceedings is ill-defined and that journals should not play second fiddle to conferences (see, e.g., [4]).

The standards for scholarship were also high. It became clear that researchers with different backgrounds frequently do not know about advances on the same problem by researchers in other fields or even in the same field at earlier times. In some cases, similar ideas which had been discovered many years earlier were not properly credited. This can be partially excused by the scope of the field and, hopefully, the most egregious errors were corrected during the review process. Finally, the Guest Editors felt that a solid paper should provide readers with a comprehensive analysis of an approach, including the disadvantages and failure scenarios.

3 Accepted Papers

The Guest Editors are excited about the pace of the development in this research field, as is evident from the accepted papers. Among the 11 papers selected for publication, there are four papers on image retrieval, four on image annotation, and three on video. The key ideas and contributions of these papers are now summarized.

3.1 Image Retrieval

In the paper titled "VisualRank: Applying PageRank to Large-Scale Image Search," Yushi Jing and Shumeet Baluja address the fundamental question of whether or not image-based features can significantly improve the performance of image retrieval systems which currently rely mainly on textual annotation. Human knowledge about linking relevant images to webpages is significantly exploited by inferring a visual similarity graph and ranking the visual link structures among images, thereby pooling information. Large-scale experiments are conducted to retrieve images for common queries, such as those seeking 2,000 popular products, and yield results which compare very favorably with Google Image Search.

The paper "Homotopic Image Pseudo-Invariants for Openset Object Recognition and Image Retrieval" by Yoshihisa Shinagawa treats the problem of openset face recognition: the identification of a face in an image or the decision that the target face is not in the image. First, a candidate match is selected using a similarity measure based on a weighted L^2 distance between wavelet decompositions of the best homotopy (e.g., smooth warping) of the exemplar and an observed image. The relation of the best similarity and the distance between the unwarped exemplar and the observation yields a "homotopic pseudo-invariant" that forms the basis of success or failure of the identification.

A content-based image retrieval (CBIR) system seeks images in a data set with content similar to that of a target image. A *localized* CBIR system seeks only specific regions of interest in a target image or collection of target images exemplifying relevant and not relevant content. Retrieved images must indicate the relevant regions. In the paper titled "Localized Content-Based Image Retrieval," Rouhollah Rahmani, Sally A. Goldman, Hui Zhang, Sharath R. Cholleti, and Jason E. Fritts report their localized CBIR methods based on segmentation and salient points. The work has been evaluated against benchmarks.

A valuable application area for image annotation and retrieval is scanned document databases. In "Document Image Retrieval through Word Shape Coding" by Shijian Lu, Linlin Li, and Chew L. Tan, a technique is reported for searching document images without optical character recognition. A set of shape features is used to index the words in a document. The retrieval of documents can then be performed by specifying query keywords or a query document image. The authors demonstrate efficiency and tolerance to document degradations.

3.2 Image Annotation

Xin-Jing Wang, Lei Zhang, Xirong Li, and Wei-Ying Ma tackle the challenging problem of real-time image annotation in their paper titled "Annotating Images by Mining Image Search Results." Different from recently published generative and discriminative modeling approaches for image annotation, the work represents a new dimension because it relies on searching in a very large collection of

images with textual descriptions. The approach has three main steps: a search process, a mining process, and a filtering process. A large number of real-world images have been used to test the method and promising results are reported.

Since most images posted on the Web are not indexed semantically, e.g., by keywords, "concept-based image retrieval" has depended on low-level signatures. The paper "Automatic Semantic Annotation of Real World Web Images" by R.C.F. Wong and C.H.C. Leung addresses this "semantic gap" with a novel method for automatic semantic annotation aimed at retrieving appropriate images in response to user-generated queries about the image content. The main idea is to cluster images based on embedded image-capture metadata, including acquisition parameters such as camera properties and GPS information. A learning framework is developed using decision trees based on components of the acquisition parameter vector and the method is validated on over 100,000 web images from flickr.com and elsewhere.

In the paper titled "Handwritten-Word Spotting Using Biologically Inspired Features," authors Tijn van der Zant, Lambert Schomaker, and Koen Haak tackle the difficult problem of recognizing words from connected cursive handwritten text in images of historical books. Standard machine learning methods perform poorly unless large amounts of training data are available, at least several hundred instances per word for a given writing style. In contrast, the methodology here, which utilizes hierarchical models inspired by the neuro-physiology of vision, as well as a relatively small feature vector, achieves impressive accuracy when applied to recognizing 2,100 different words in a text containing abstracts of royal decrees for which only two or three training instances are available for about half the classes.

To illustrate the benefit of a data-driven approach in object and scene understanding, Antonio Torralba, Rob Fergus, and William T. Freeman leverage an exceptionally large amount of Web image data in their paper titled "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition." Each image is stored as a 32×32 color thumbnail and is loosely labeled with a nonabstract noun in English. The semantic relationships between the words, as recorded in Wordnet, are used together with nearest-neighbor methods to perform object classification. For some object classes (e.g., people), the authors have shown a recognition performance comparable to some class-specific methods.

3.3 Video Annotation and Retrieval

Real-world video poses both challenges and opportunities for annotation and retrieval. The work reported in the paper "Nonchronological Video Synopsis and Indexing" by Yael Pritch, Alex Rav-Acha, and Shmuel Peleg aims at handling massive video data captured by surveillance cameras. The video synopsis system provides a short video representation to summarize the main activities. To process continuous video streams, the system performs online conversion of the stream into a collection of objects and activities. The video synopsis can then be generated for queries like "show in one minute the synopsis of this camera broadcast during the past day."

In the paper titled "Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment" by Dong Xu and Shih-Fu Chang, event recognition in unconstrained news video sequences is accomplished by using the "bag-of-words" approach to object recognition to extract collections of unordered descriptors from video clips, a form of the half-century old Kantorovich distance at the heart of linear programming (also known as the "earth mover's distance") between the descriptor bags of video clips to measure their similarity, and support vector machine kernel classifiers to detect events. Event evolution is characterized by a multiresolution extension to match temporally aligned pyramids.

Finally, biomedicine has become an important application area for image annotation and retrieval due to the large volume of data generated by modern imaging devices used in biology labs and hospitals. Leandro Cortés and Yali Amit, the authors of "Efficient Annotation of Vesicle Dynamics Video Microscopy," present an algorithm for detecting events of interest in video microscopy. The relatively blurry video quality makes it a great challenge to detect and track multiple possibly overlapping vesicles. Statistical modeling has been used to weigh various hypotheses online in order to find the most likely trajectories. The computational efficiency primarily depends on the number of vesicles to be tracked.

Acknowledgments

The hard work of more than 100 reviewers and all of the contributing authors is gratefully acknowledged. We would also like to thank the anonymous reviewers of our proposal and Editor-in-Chief David Kriegman for sharing our vision. The editorial staff of *TPAMI*, especially Elaine Stephenson, has been extremely responsive to our many inquiries. J.Z. Wang would like to thank Carnegie Mellon University, Takeo Kanade, and The Pennsylvania State University for being supportive of his work. Johns Hopkins University, Eastman Kodak Company, and Stanford University have supported the work of D. Geman, J. Luo, and R.M. Gray, respectively.

James Z. Wang
Donald Geman
Jiebo Luo
Robert M. Gray
Guest Editors

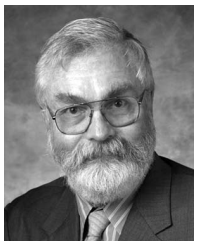
References

[1] "Shipments of Digital Cameras Rose 15% last Year, IDC Says," *The Wall Street J.*, 3 Apr. 2007.
[2] "Flickr Maps the World's Photos," *Reuters*, 19 Nov. 2007.
[3] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Computing Surveys*, vol. 40, no. 2, article 5, 2008.
[4] D. Geman, "Ten Reasons Why Conference Papers Should Be Abolished," Nov. 2007, http://www.cis.jhu.edu/publications/Ten_Reasons.pdf.

James Z. Wang received the bachelor's degree in mathematics and computer science summa cum laude from the University of Minnesota, the MS degree in mathematics and the MS degree in computer science, both from Stanford University, and the PhD degree in medical information sciences from Stanford University. He has been a faculty member in the College of Information Sciences and Technology, the Department of Computer Science and Engineering, and the Integrative Biosciences Program at The Pennsylvania State University since 2000. His main research interests are automatic image tagging, image retrieval, computational aesthetics, and computerized analysis of paintings. He was a visiting professor at the Robotics Institute at Carnegie Mellon University (2007-2008). He has also held visiting positions at SRI International, IBM Almaden Research Center, NEC Computer and Communications Research Lab, and Academia Sinica. He has been a recipient of a US National Science Foundation Career award and the endowed PNC Technologies Career Development Professorship. He is a senior member of the IEEE.

Donald Geman received the BA degree in literature from the University of Illinois and the PhD degree in mathematics from Northwestern University. He was a Distinguished Professor at the University of Massachusetts until 2001, when he joined the Department of Applied Mathematics and Statistics at The Johns Hopkins University, where he is a member of the Center for Imaging Science and the Institute for Computational Medicine. He also has an ongoing affiliation with the École Normale Supérieure de Cachan in France. He works at the intersection of applied mathematics and computer science, specializing in statistical learning, computer vision, and computational biology. Current research projects include mental image retrieval, semantic scene interpretation, molecular cancer diagnosis, and modeling protein-protein interaction networks. He is a senior member of the IEEE.

Jiebo Luo received the BS and MS degrees from the University of Science and Technology of China in 1989 and 1992, respectively, and the PhD degree from the University of Rochester in 1995, all in electrical engineering. He is a senior principal scientist with Kodak Research Laboratories, Rochester, New York. His research interests include image processing, pattern recognition, computer vision, computational photography, medical imaging, and multimedia communication. He is the author of more than 120 technical papers and holds more than 40 granted US patents. He currently serves on the editorial boards of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*TPAMI*), the *IEEE Transactions on Multimedia* (*TMM*), *Pattern Recognition* (*PR*), and the *Journal of Electronic Imaging*. He is a guest editor for a few special issues, including Image Understanding for Digital Photos (*PR*, 2005), Real-World Image Annotation and Retrieval (*TPAMI*, 2008), Event Analysis (*TCSVT*, 2008), and Integration of Content and Context for Multimedia Management (*TMM*, 2009). He is a Kodak Distinguished Inventor and a winner of the 2004 Eastman Innovation Award. He has also been an organizer of numerous technical conferences, including the general chair of the 2008 ACM International Conference on Image and Video Retrieval (CIVR), an area chair of the 2008 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), a program cochair of the 2007 SPIE International Symposium on Visual Communication and Image Processing (VCIP), and a special sessions cochair of the 2006 IEEE International Conference on Multimedia and Expo (ICME). He is a senior member of the IEEE and a fellow of the SPIE.



Robert M. Gray received the BS and MS degrees from the Massachusetts Institue of Technology in 1966 and the PhD degree from the University of Southern California in 1969, all in electrical engineering. Since 1969, he has been with Stanford University, where he is currently the Alcatel Lucent Technologies Professor of Engineering and a professor of electrical engineering. His primary research interests are quantization, compression, and statistical classification. He is a fellow of the IEEE and the Institute of Mathematical Statistics and was a fellow of the John Simon Guggenheim Foundation in 1982. He was an associate editor for *Source Coding* (1977-1980), an editor (1980-1983) of the *IEEE Transactions on Information Theory*, and cochair of the 1993 IEEE International Symposium on Information Theory. He was a corecipient of the 1976 IEEE Information Theory Group Paper Award and the 1983 IEEE ASSP Senior Award. He received the IEEE Signal Processing Society 1993 Society Award, 1997 Technical Achievement Award, and 2005 Meritorious Service Award. He received IEEE Centennial and Third Millennium Medals. In 1998, he received a Golden Jubilee Award for Technological Innovation from the IEEE Information Theory Society. He received a 2002 Presidential Award for Excellence in Science, Mathematics, and Engineering Mentoring (PAESMEM). He is the recipient of the 2008 Claude E. Shannon Award of the IEEE Information Theory Society and the 2008 IEEE Jack S. Kilby Signal Processing Medal. He is a member of the National Academy of Engineering.