Multimed Tools Appl
DOI 10.1007/s11042-008-0248-6

GUEST EDITORIAL

Introduction to the special issue on "metadata mining for image understanding"

Gabriela Csurka · Katerina Pastra

The number of digital images stored, managed and shared through the internet is growing at a phenomenal rate. Press and photo agencies, photo-sharing networks and image search engines face the challenge of managing effectively billions of images. Millions of people register to photo-sharing networks or simply visit such sites through a variety of devices ranging from mobile phones to broadband connected digital video recorders. High quality cameras come part and parcel with mobile devices boosting consumer-generated visual content, as well as the exchange of such content with others through photo-messaging. These trends point to a shift in *user groups*, *user needs* and *content type* in digital imaging: it is not only professionals but also laymen who generate, access and interact with digital images for professional, educational or entertainment-related purposes; subsequently, the content itself is not always of the highest quality, or of professionally staged scenes, it now carries all idiosyncrasies of consumer-generated content.

Within such conditions, finding the most relevant or the most appealing image for a given task (e.g. to illustrate a story) has become an extremely difficult process, a process that requires one to take advantage of any pieces of information related to the images. For example, metadata related to image capture such as date, location, camera settings or name of photographer is often available from the digital camera used to take the photograph. The owner can further add a relevant title, filename or/and descriptive caption or any other textual reference. If the image is uploaded to a shared photo collection, additional comments are frequently added to the image by other users. On the other hand, images used in documents, i.e. web pages, frequently have captions and surrounding text. All this information can be considered *image metadata* and is of value for organizing, sharing, and processing images.

G. Csurka
XEROX Research Centre Europe, Meylan, France
e-mail: Gabriela.Csurka@xrce.xerox.com
URL: http://www.xrce.xerox.com/people/csurka/home.html

K. Pastra (✉)
Institute for Language and Speech Processing, Athens, Greece
e-mail: kpastra@ilsp.gr
URL: http://www.ilsp.gr/homepages/pastra_eng.html

However, *how could one exploit the information contained in such metadata in an intelligent, generic or task-specific way*? Linking this information with the actual image content is still an open challenge. The aim of this special issue is to present research on bridging textual and visual data for developing technology that shows a more "advanced understanding" of image contents. To this end, related work from two major research communities is brought together: computer vision and computational linguistics. The papers presented in this volume have been reviewed by researchers from both communities, as a further step in activating the dialogue between the communities on such topics of common interest.

Actually, this special issue sprung out of the First International Workshop on Metadata Mining for Image Understanding (MMIU 2008), that took place in January 2008 in Madeira, Portugal, as a satellite event of the Computer Vision Theory and Applications Conference (VISAPP 2008). The event was an opportunity for researchers, content providers and related user-service providers to elaborate on the needs and practices of digital image management, to share ideas that point to new directions on using metadata for image understanding and to demonstrate related technology representative of the state of the art and beyond. The workshop was successful in attracting interdisciplinary and international interest, as evidenced in the diverse programme committee, as well as the papers presented. Computer vision and computational linguistics researchers from the academia and the industry showed that the time is ripe for the two communities to interact more and gain from the resulting cross-fertilization of ideas for addressing several multimedia application challenges. The best papers of the workshop were selected, extended and revised and now form the main volume of this special issue.

In particular, the papers that we have selected to include in this volume introduce methods for combining visual and textual metadata for a variety of multimedia applications. Half of them focus on methods that could be used in a variety of multimedia applications, while the other half present methods that are more tuned to the idiosyncrasies of specific applications:

Battiato et al., *Using Visual and Text Features for Direct Marketing in the Multimedia Messaging Services Domain*, combine visual and textual features in a cascade of regression methods for learning which advertisements in the form of mobile multimedia messages are more likely to appeal to the users. The work demonstrates the usefulness of such technology in Direct Marketing. On their turn, *Ah-Pine et al*., *Crossing textual and visual content in different application scenarios*, present two "trans-media" feedback metrics for automatic image annotation, text illustration and multimedia retrieval and clustering. The metrics have been used within a travel blog assistant system and a tool for browsing the Wikipedia. In *Kludas et al.*, *Can Feature Information Interaction help for Information Fusion in Multimedia Problems*?, the authors present an information fusion method that takes into account feature interaction in multivariate settings. The method is compared to bivariate dependence measures on both artificial and real world data, the latter comprising of a captioned image database.

Turning to the idiosyncrasies of personal photo-collections, *Carvalho et al*., *Attributing Semantics to Personal Photographs*, present work towards automatic propagation of image tags, in personal photo-albums. They introduce a hybrid information extraction method for extracting person, object and location information from image captions and implement their suggestion for combining visual content and time-capture metadata for clustering photographs in personal photo-albums and propagating their location-related tags. *Lindstaedt et al.*, *Automatic Image Annotation using Visual Content and Folksonomies*, present automatic image classification and similarity methods for a tag recommendation system within the context of collaborative annotation. The system relies on visual content analysis, tag

association and user preferences. Last, the needs of professional image cataloguers in large image libraries are addressed in *Klavans et al*., *Computational Linguistics for Metadata Building* (*CLIMB*): *Using Text Mining for the Automatic Identification*, *Categorization*, *and Disambiguation of Subject Terms for Image Metadata*. The authors provide an overview of a toolkit for image cataloguers which augments image metadata by associating web-based text segments to image captions and categorizes this metadata in terms of the type of information it reveals (e.g. historical context). Word sense disambiguation takes also place in these textual resources for more accurate indexing and retrieval.

We believe that in these papers the readers will find valuable information and ideas not only on methods to be used for combining visual and textual metadata, but also for the multimedia applications that benefit from such research and affect the specifics of the methods to be followed. Vision–language integration has a long past in Artificial Intelligence, and a very active present; we hope that the current volume will contribute to a more vivid future.

Acknowledgements We would like to thank all members of the programme committee of the MMIU workshop for reviewing not only the original workshop papers, but also the extended and revised versions of the selected papers included in this volume; their feedback has been more than constructive: *Suzanne Boll* (University of Oldenburg, Germany), *Paul Clough* (University of Sheffield, UK), *Christophe Garcia* (France Telecom Research, France), *Daniel Gatica-Perez* (Idiap, Martigny, Switzerland), *Benoit Huet* (Institut Eurécom, Sophia-Antipolis, France), *Gareth Jones* (Dublin City University, Ireland), *Franciska de Jong* (University of Twente, The Netherlands), *Ales Leonardis* (University of Ljubljana, Slovenia), *Jiebo Luo* (Kodak Research Lab, Rochester, NY, USA), *Georges Quenot* (LIG, Grenoble, France), *Rahul Nair* (Yahoo! Research Berkeley, USA), *Stefan Rueger* (KMI, Open University, UK), *Horacio Saggion* (University of Sheffield, UK), *Tamás Szirányi* (SZTAKI, Hungary), *Tinne Tuytelaars* (University of Leuven, Belgium), *Xin-Jing Wang* (Microsoft Research, China), *Geoffrey Woolfe* (XRCW, NY, USA), *Lei Zhang* (Microsoft Research, China), and *Roelof van Zwol* (Yahoo! Research Barcelona, Spain).

Gabriela Csurka is a research scientist in the Textual and Visual Pattern Analysis team at Xerox Research Centre Europe (XRCE). She obtained her Ph.D. degree (1996) in Computer Science from University of Nice Sophia-Antipolis. Before joining XRCE in 2002, she worked in fields such as stereo vision and projective reconstruction at INRIA (Sophia Antipolis, Rhone Alpes and IRISA) and image and video watermarking at University of Geneva and Institute Eurécom, Sophia Antipolis. Author of several publications in main journals and international conferences, she is also an active reviewer both for journals and conferences. Her current research interest concerns the exploration of new technologies for image content and aesthetic analysis, cross-modal image categorization and semantic based image segmentation.

Katerina Pastra is a Research Fellow, currently working on language technologies and multimedia. She is the coordinator of the recently funded project POETICON (FP7-ICT), which aims at developing grounding resources and mechanisms for artificial agents. She holds a BA in Greek Literature and Linguistics (University of Athens), an M.Sc. in Machine Translation (UMIST, UK) and a Ph.D. in Artificial Intelligence (University of Sheffield, UK) in which she explored the integration of vision and language within artificial agents engaged in everyday interaction. She has worked on information extraction and automatic text-based image/audiovisual indexing and retrieval within semantics-based indexing and retrieval projects (CONCERTO/IST, SOCIS/EPSRC, REVEAL THIS/IST). She has lectured on Human–Computer/Human–Robot Interaction and the use of cognitive and psychological methods in software engineering (University of Sheffield) and has organized international workshops on multimedia processing. She is the author of a number of publications on the above topics, one of which has won a distinction by the British Computer Society.