Vision and Language Integration Meets Multimedia Fusion

Proceedings of ACM Multimedia 2016 Workshop

Marie-Francine Moens
KU Leuven
Heverlee, Belgium
sien.moens@cs.kuleuven.be

Kate Saenko
UMass Lowell
Lowell, MA, USA
saenko@cs.uml.edu

Katerina Pastra
CSRI / ILSP / ATHENA
Athens, Greece
kpastra@csri.gr

Tinne Tuytelaars
KU Leuven
Heverlee, Belgium
tinne.tuytelaars@esat.kuleuven.be

ABSTRACT

Multimodal information fusion both at the signal and the semantics levels is a core part in most multimedia applications, including multimedia indexing, retrieval, summarization and others. Early or late fusion of modality-specific processing results has been addressed in multimedia prototypes since their very early days, through various methodologies including rule-based approaches, information-theoretic models and machine learning. Vision and Language are two of the predominant modalities that are being fused and which have attracted special attention in international challenges with a long history of results, such as TRECVid, ImageClef and others. During the last decade, vision-language semantic integration has attracted attention from traditionally non-interdisciplinary research communities, such as Computer Vision and Natural Language Processing. This is due to the fact that one modality can greatly assist the processing of another providing cues for disambiguation, complementary information and noise/error filtering. The latest boom of deep learning methods has opened up new directions in joint modelling of visual and co-occurring verbal information in multimedia discourse. The workshop on Vision and Language Integration Meets Multimedia Fusion has been held during the workshop weekend of the ACM Multimedia 2016 Conference and the European Conference on Computer Vision (ECCV 2016) on October 16, 2016 in Amsterdam, the Netherlands. The proceedings contain seven selected long

papers, which have been orally presented at the workshop, and three abstracts of the invited keynote speeches. The papers and abstracts discuss data collection, representation learning, deep learning approaches, matrix and tensor factorization methods and graph based clustering with regard to the fusion of multimedia data. A variety of applications is presented including image captioning, summarization of news, video hyperlinking, sub-shot segmentation of user generated video, cross-modal classification, cross-modal question-answering, and the detection of misleading metadata of user generated video.

The workshop is organized and supported by the EU COST action iV&L Net, the European Network on Integrating Vision and Language: Combining Computer Vision and Language Processing for Advanced Search, Retrieval, Annotation and Description of Visual Data (IC 1307, 2014-2018).

CCS Concepts

•Information systems → Multimedia and multimodal retrieval; *Web mining*; •Computing methodologies → Natural language processing; Computer vision;

Keywords

Cross-modal and multimodal processing of visual and language data

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Copyright held by the owner/author(s)

MM '16 October 15-19, 2016, Amsterdam, Netherlands

ACM ISBN 978-1-4503-3603-1/16/10.

DOI: <http://dx.doi.org/10.1145/2964284.2980537>