Massive-Scale Multimedia Semantic Modeling

John R. Smith
IBM T. J. Watson Research Center
Yorktown Heights
New York, USA
jsmith@us.ibm.com

Liangliang Cao
IBM T. J. Watson Research Center
Yorktown Heights
New York, USA
liangliang.cao@us.ibm.com

ABSTRACT

Visual data is exploding! 500 billion consumer photos are taken each year world-wide, 633 million photos taken per year in NYC alone. 120 new video-hours are uploaded on YouTube per minute. The explosion of digital multimedia data is creating a valuable open source for insights. However, the unconstrained nature of Şimage/video in the wildŤ makes it very challenging for automated computer-based analysis. Furthermore, the most interesting content in the multimedia files is often complex in nature reflecting a diversity of human behaviors, scenes, activities and events. To address these challenges, this tutorial will provide a unified overview of the two emerging techniques: Semantic modeling and Massive scale visual recognition, with a goal of both introducing people from different backgrounds to this exciting field and reviewing state of the art research in the new computational era.

Categories and Subject Descriptors

H.2.4 [**Systems**]: Multimedia databases; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*; I.2.4 [**Knowledge Representation Formalisms and Methods**]: [Semantic networks]; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*

Keywords

Multimedia information retrieval, video analysis, content-based search, machine learning, semantic modeling

1. INTRODUCTION

Across multiple generations of information technology that have dealt with structured and unstructured data, the explosion of multimedia data is creating the biggest wave of all. Huge volumes of multimedia – images, video and audio are being generated and consumed daily. Currently, multimedia makes up 60% of internet traffic, 70% of mobile phone traffic and 70% of all available unstructured data. To give specific examples, Web users are uploading 72 video-hours to YouTube per minute, and on an average day, social media users post 300 hundred million photos to Facebook. Consumers using mobile phones and digital cameras are taking 500 billion photos per year, or 78 per person on the planet [4]. Specialized domains are participating too. Medical institutions are acquiring one billion radiological images per year, and cities are installing hundreds of millions of video cameras worldwide for safety, security and law enforcement. Industries across life sciences, petroleum exploration, astronomy, insurance, retail and many others are faced with huge and growing volumes of multimedia data.

Multimedia is "big data" not just because there is a lot of it. Multimedia is big data because increasingly it is becoming a valuable source for insights and information. Multimedia data can tell us about things happening in the world, point out places, events or topics of interest, give clues about a person's preferences and even capture a rolling log of human history [4, 7]. However, the challenge with multimedia big data is that images, video and audio require much more sophisticated algorithms for content analysis than previous waves of structured and unstructured data. This is spurring on a tremendous amount of research on efficient and effective techniques for "bridging the semantic gap" to enable large-scale multimedia information extraction and retrieval [5, 2].

2. TUTORIAL OVERVIEW

The instructors will share their knowledge in developing the leading industrial systems as well the winning experience in prestigious visual recognition competitions in the ImageNet Large Scale Visual Recognition Challenges, ImageCLEF recognition and TRECVID challenges. This tutorial will cover both the semantic and massive computing aspects:

- Semantic modeling: Many recent research works have recognized that detection of semantic concepts would significantly improve the effectiveness of image and video retrieval. One of the first large taxonomy is LSCOM [3], which defines more than 1,000 semantic concepts. The collection of ImageNet has collected more than 20K synsets. The efforts of building large scale semantic datasets have also motivated a lot of research on ontology design, web search and social media mining. Many research efforts have been paid on active learning and transfer learning to organize the large amount of unstructured data on the Web, of which the power will become more and more significant with the increasing amount of Web data.

MM'13, October 21–25, 2013, Barcelona, Spain.
ACM 978-1-4503-2404-5/13/10.
http://dx.doi.org/10.1145/2502081.2502235.

- Massive scale visual recognition: Both the recent develop of computational resource (CPU, GPU, high performance computers) and the increasing amount of multimedia content have contributed to the success of massive scale visual recognition. In the past decade, we have witnessed many exciting techniques for massive visual recognition, including SIFT-like local features, sparse coding, hierarchical vocabulary tree search, supervector-based recognition, attribute learning, locality sensitive hashing, average stochastic gradient descent, and etc. These new techniques are being combined with the new parallel computing frameworks such as MapReduce and Stream computing.

In this talk we present a perspective across multiple industry problems, including safety and security, medical, Web, social and mobile media, and motivate the need for large-scale analysis and retrieval of multimedia data. We describe a multi-layer architecture that incorporates capabilities for audio-visual feature extraction, machine learning and semantic modeling and provides a powerful framework for learning and classifying contents of multimedia data. We discuss the role semantic Ontologies for representing audio-visual concepts and relationships, which are essential for training semantic classifiers [3]. We discuss the importance of using faceted classification schemes in particular for organizing multimedia semantic concepts in order to achieve effective learning and retrieval [6]. We also show how training and scoring of multimedia semantics can be implemented on big data distributed computing platforms to address both massive-scale analysis and low-latency processing [8]. We describe multiple efforts [1, 9] at IBM on image and video analysis and retrieval, including IBM Multimedia Analysis and Retrieval System (IMARS), and show recent results for semantic-based classification and retrieval. We conclude with future directions for improving analysis of multimedia through interactive and curriculum-based techniques for multimedia semantics-based learning and retrieval.

3. ABOUT THE PRESENTERS

John R. Smith is Senior Manager of the Intelligent Information Management Department at IBM T. J. Watson Research Center. He received his M. Phil and Ph.D. degrees in Electrical Engineering from Columbia University in 1994 and 1997, respectively. He currently leads R&D across multiple areas at IBM Research including multimedia, image/video analytics, biometrics, exploratory computer vision and machine learning. Dr. Smith is also principal investigator for the IBM Multimedia Analysis and Retrieval System (IMARS) project. Previously, Dr. Smith led IBM's participation in MPEG-7 / MPEG-21 standards and served as a Chair of the MPEG Multimedia Description Schemes Group and co-project Editor of MPEG-7 Standard. Dr. Smith is currently Editor-in-Chief of IEEE Multimedia and Fellow of IEEE.

Dr. Smith has given numerous tutorials at major conferences including ACM Multimedia, IEEE Intl. Conf. on Multimedia and Expo (ICME), World Wide Web (WWW), ACM Intl. Conference on Management of Data (SIGMOD), ACM International Conference on Conceptual Modeling (ER). Dr. Smith has published more than 100 papers at top conferences and journals. His papers have received more than 13,000 citations and have an h-index of 55 and i10-index of 164.

Liangliang Cao is a Research Staff Member in IBM T. J. Watson Research Center, and also an adjunct assistant professor at Columbia University. His research lies in the intersection of computer vision, new media and multimedia big data. His work has won three prestigious visual recognition competitions, include ImageCLEF Medical Image Classification (2012, 2013), ImageNet Large Scale Visual Recognition Challenge (2010), and TRECVID Airport Surveillance Competition (2008). He received many awards including the IBM Outstanding Accomplishment for multimedia group (2012), the Best Paper Award in the First International Workshop on Big Data Mining (2012), IBM Watson Emerging Leader in Multimedia and Signal Processing (2010), Facebook Fellowship Finalist (2010), and UIUC Computational Science and Engineering Fellowship (2009-2010).

Dr. Cao has authored more than 40 papers in top conferences and journals, including ACM Multimedia, ICCV, CVPR, ECCV, NIPS, WWW, TPAMI, and PIEEE. Dr. Cao is an area chair of ACM Multimedia 2012 and IEEE WACV 2014. He fulfills review duties for more than 15 journals and various conferences. He is a general chair of New York Area Multimedia and Vision Meeting in Greater in 2012 and 2013. He is a guest editor of ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP), Computer Vision and Image Understanding (CVIU) Journal, and also IEEE MultiMedia.

4. REFERENCES

[1] L. Cao, L. Gong, J. R. Kender, N. C. Codella, and J. R. Smith. Learning by focusing: A new framework for concept recognition and feature selection. *Proc. of IEEE Conference on Multimedia and Expo*, 2013.

[2] A. Hanjalic, R. Lienhart, W.-Y. Ma, and J. R. Smith. The holy grail of multimedia information retrieval: So close or yet so far away? *Proc. IEEE*, 94(4):541–547, 2008.

[3] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kenendy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3), July-September 2006.

[4] J. R. Smith. History made everyday. *IEEE MultiMedia*, 18(2), July-September 2011.

[5] J. R. Smith. Minding the gap. *IEEE MultiMedia*, 19(2):53–62, January-March 2012.

[6] J. R. Smith. Just the facets. *IEEE MultiMedia*, 20(1), January-March 2013.

[7] L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith. Visual memes in social media: Tracking real-world news in youtube videos. *Proc. of the 19th ACM Intl. Conf. on Multimedia*, pages 53–62, November 2011.

[8] R. Yan, M. O. Fleury, M. Merler, A. Natsev, and J. R. Smith. Large-scale multimedia semantic concept modeling using robust subspace bagging and map-reduce. *Proc. of the First ACM Workshop on Large-Scale Multimedia Retrieval*, 2009.

[9] F. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.