What speech researchers should know about video technology!

Koichi Shinoda and Florian Metze

Importance

Internet media sharing sites and the one-click upload capability of smartphones have led to a deluge of multimedia content becoming available online. Thousands of videos are constantly being uploaded to the web, creating a vast resource, and an ever-growing demand for methods to make them easier to retrieve, search, and index. While visual information is a very important part of a video, acoustic and speech information often complements it – and the INTERSPEECH community needs to rise to the challenge. By facilitating access to large amounts of data, the text-based Internet gave a huge boost to the field of natural language processing. The rising tide of consumer produced online video will do the same for video processing, and – once suitable tools for computational analysis will have been developed – for fields such as human computer interaction, media analytics, robotics, social sciences, etc. This data is a treasure trove also for the speech research community, which has lots of experience, tools, and ideas to contribute, but it does come with its own set of challenges, success criteria, technical terms, etc., which one needs to understand, in order to fully benefit from it.

Objectives

Similar to the experience in the speech research community, consumer-created (as opposed to Broadcast-News style) multimedia material offers a great opportunity for research on all aspects of Human-to-human as well as Man-machine interaction. Speech and audio is naturally an important part of these interactions, able to link visual objects, people, and other observations across modalities; research results gathered on this offline data will inform future experiments in interactive settings (robotics, interactive agents, etc.).

State of the art (mainstream) "content-based video retrieval (CBVR) research relies largely on so-called "low-level features", which can be extracted from the video's image using various transformations. The situation reminds us of speech/ speaker recognition in the 90's, when speech researchers were faced with a very similar problem. The solution we found was a robust data-driven approach, which heavily relied on probability theory. Thanks to the advancement of computation technology, the same approach can now be readily used for video information retrieval. Large-scale multi-modal analysis of audio-visual material is central to a number of multi-site research projects around the world, driven by multiple communities including information retrieval, video search, copyright protection, etc. While each of these have slightly different targets, they are facing largely the same challenges: how to robustly and efficiently process large amounts of data, how to represent and then fuse information across modalities, how to train classifiers and segmenters on unlabeled data, how to include human feedback, etc.

This tutorial aims to present to the speech community the state of the art in video processing, by discussing the most relevant tasks at NIST's TREC Video Retrieval Evaluation (TRECVID) evaluation and workshop series, which has been going on since 2001, to promote CBVR research and development. We liken TRECVID's "Semantic Indexing" (SIN) task, in which a system must identify occurrences of concepts such as "desk", or "dancing" in a video to the word spotting approach, which will be familiar to speech researchers. We then proceed to explain more recent, and challenging tasks, such as "Multimedia Event Detection" (MED), and "Multimedia Event Recounting" (MER), which can be compared to meeting transcription and summarization tasks, which will again be familiar to our audience. We will explain the particular challenges of dealing with consumer produced multimedia material, and associated audio tracks; we will discuss evaluation metrics and explain ways of achieving good performance in standard tasks and problems – enabling speech researchers to participate in discussions on “multimedia” problems. We will then proceed to lay out how the speech and language community can contribute, given its own vast body of experience, and identify opportunities for advancing speech-centric research on these datasets.

Target Audience

The tutorial will be suitable for any graduate student in the speech and language community who is familiar with the basic concepts of decision theory and classifier design. An understanding of machine learning principles, applied data processing (particularly text and audio), and experiment design will be required, but no in-depth knowledge will be assumed in areas that INTERSPEECH attendees would not normally be proficient in.

Outline

1. Introduction (Shinoda/ Metze)
    - 1-1. What is video information retrieval
    - 1-2. Speech methodology and technology for video information technology
    - 1-3. The role of semantics and the “semantic gap”
    - 1-4. On multi-modality: how speech and audio can help
    - 1-5. TRECVID workshop
    - 1-6. TRECVID task overview: SIN, MED, MER, KIS, …
2. Semantic indexing (SIN, Shinoda)
    - 2-1. What is semantic indexing?
    - 2-2. The nature of consumer videos
    - 2-3. Effective low-level features: SIFT, HOG, MFCC, Dense features
    - 2-4. Overview of the conventional approaches: Bag of Words, Feature fusion, SVM
    - 2-5. Multimodal, Multi-frame, Multi kernel
    - 2-6. Speaker recognition/adaptation methodology is effective in semantic indexing
    - 2-7. GMM supervectors and Fisher kernels
    - 2-8. Robust MAP estimation for semantic modeling
    - 2-9. Fast computation techniques
    - 2-10. Evaluation criteria and performance
    - 2-11. Demonstration of the state-of-the-art semantic indexing systems
    - 2-12. Future directions: video is a communication tool
3. Multimedia Event Detection (MED, Metze)
    - 3-1. Multimedia Event Detection vs video retrieval and semantic indexing
    - 3-2. Overview to approaches and state of the art
    - 3-3. The labeling and annotation problem, speech diarization
    - 3-4. Methods and techniques for audio event classification
    - 3-5. “Low-level” vs “High-level” features
    - 3-6. (Multi-modal) fusion for event detection
    - 3-7. Strengths and weaknesses of audio and video – contributions to multimedia
    - 3-8. Semi-supervised and active learning
    - 3-9. How speech and language can contribute: semantics for “high-level features”
    - 3-10. Research goals: developing accurate “high-level” features
4. Multimedia Event Recounting (MER, Metze)
    - 4-1. Overview and goal of MER
    - 4-2. Justifying and explaining a classification

4-3. Intrinsic vs. extrinsic summarization tasks
4-4. Natural language generation: grammar-based and statistical
4-5. How to select information to present: learning discriminative concepts
4-6. Statistical methods for information fusion and presentation
4-7. Evaluation and user studies
4-8. Human-in-the-loop system development
4-9. Where are we and where are we going?

5. Summary and Future Problems (Shinoda/ Metze)

5-1. Computational Requirements
5-2. Other resources and tools
5-3. Demonstration/ hands-on experience
5-4. Opportunities for participation
5-5. Speculations & discussion on future problems and approaches
5-6. The Speech Recognition Virtual Kitchen

Biography

Koichi Shinoda

Koichi Shinoda received his B.S. in 1987 and his M.S. in 1989, both in physics, from the University of Tokyo. He received his D.Eng. in computer science from Tokyo Institute of Technology in 2001. In 1989, he joined NEC Corporation and was involved in research on automatic speech recognition. From 1997 to 1998, he was a visiting scholar with Bell Labs, Lucent Technologies. From 2001, he was an Associate Professor with the University of Tokyo. He is currently Professor at the Tokyo Institute of Technology. His research interests include speech recognition, video information retrieval, and human interfaces. Dr. Shinoda received the Awaya Prize from the Acoustic Society of Japan in 1997 and the Excellent Paper Award from the Institute of Electronics, Information, and Communication Engineers (IEICE) in 1998. He is an Associate Editor of Computer Speech and Language and a subject editor of Speech Communication. He is a member of IEEE, ACM, ASJ, IEICE, IPSJ, and JSAI.
Web page: http://www.ks.cs.titech.ac.jp/english/index.html

Florian Metze

Florian Metze received his PhD from Universität Karlsruhe (TH) for his work on "Articulatory Features for Conversational Speech Recognition" in 2005. He since worked at Deutsche Telekom Laboratories (T-Labs) and joined Carnegie Mellon University's research faculty in 2009 where he regularly teaches several classes and labs.

Dr. Metze has worked on a wide range of topics in the field of speech and audio processing as well as user interfaces. Recent work includes non-textual aspects of speech (such as the perception of personality of speech and emotional speech synthesis) as well as retrieval and summarization of speech and multi-media material. This work shows how audio and speech technologies can be applied to multi-media material, and provides textual summaries of video content, rather than just video skims.

Dr. Metze has published more than 100 papers in his research areas and is a member of the IEEE Speech and Language Technical Committee. He chaired multiple conferences and workshops in the speech and audio domain, and recently co-taught a tutorial on para-linguistic aspects of speech at ICASSP 2012.
Web page: http://www.cs.cmu.edu/~fmetze/interACT/Home.html