Semantics for Large-Scale Multimedia: New Challenges for NLP

Florian Metze
Carnegie Mellon University
fmetze@cs.cmu.edu

Koichi Shinoda
Tokyo Institute of Technology
shinoda@cs.titech.ac.jp

1 Description

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. As it becomes feasible to extract both low-level as well as high-level (symbolic) audio, speech, and video features from this data, these need to be processed further, in order to learn and extract meaningful relations between these. The language processing community has made huge process in analyzing the vast amounts of very noisy text data that is available on the Internet. While it is very difficult to create semantic units of low-level image descriptors or non-speech sounds by themselves, it is comparatively easy to ground semantics in the word output of a speech recognizer, or text data that is loosely associated with a video. This creates an opportunity for NLP researchers to use their unique skills, and make significant contributions to solve tasks on data that is even noisier than web text, but (we argue) even more interesting and challenging.

This tutorial aims to present to the NLP community the state of the art in audio and video processing, by discussing the most relevant tasks at NIST's TREC Video Retrieval Evaluation (TRECVID) workshop series. We liken "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. We then proceed to explain more recent, and challenging tasks, "Multimedia Event Detection" (MED) and "Multimedia Event Recounting" (MER), which can be compared to transcription and summarization tasks. Finally, we will present an easy way to get started in multi-media analysis using Virtual Machines from the "Speech Recognition Virtual Kitchen", which will enable tutorial participants to perform hands-on experiments during the tutorial, and at home.

2 Outline

1. Introduction
   - Content based video retrieval
   - What is the "Semantic Gap"?
   - The TRECVid workshop and its tasks
2. Semantic Indexing
   - State-of-the art frameworks
   - Extension of Bag-of-Word model
   - Multi-modality
3. Multimedia Event Detection & Recounting
   - State-of-the art frameworks
   - Multimodal fusion
   - Semi-supervised and active learning
   - Video Summarization
4. Challenges for NLP
   - How to design visual concepts?
   - Intermediate representations?
   - Are there any grammars in video?
5. Practice session
   - Virtual Machines in the Speech Recognition Virtual Kitchen (http://speechkitchen.org/)

3 Instructors

Florian Metze received his PhD from Universitat Karlsruhe (TH) in 2005. He worked as a Senior Research Scientist at Deutsche Telekom Laboratories (T-Labs) and joined Carnegie Mellon University's faculty in 2009. His interests includes speech and audio processing, and user interfaces.
Koichi Shinoda received his D. Eng. from Tokyo Institute of Technology in 2001. In 1989, he joined NEC Corporation. From 1997 to 1998, he was a visiting scholar with Bell Labs, Lucent Technologies. He is currently a Professor at the Tokyo Institute of Technology. His research interests include speech recognition, video information retrieval, and human interfaces.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials, page 6,
Baltimore, Maryland, USA, 22 June 2014.