Generating Natural Language Summaries for Multimedia

Duo Ding, Florian Metze, Shourabh Rawat, Peter F. Schulam, Susanne Burger
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA, USA 15213
{dding, fmetze, srawat, pschulam, sburger}@cs.cmu.edu

Abstract

In this paper we introduce an automatic system that generates textual summaries of Internet-style video clips by first identifying suitable high-level descriptive features that have been detected in the video (e.g. visual concepts, recognized speech, actions, objects, persons, etc.). Then a natural language generator is constructed using SimpleNLG to compile the high-level features into a textual form. The generated summary contains information from both visual and acoustic sources, intending to give a general review and summary of the video. To reduce the complexity of the task, we restrict ourselves to work with videos that show a limited number of "events". In this demo paper, we describe the design of the system and present example outputs generated by the video summarization system.

1 Introduction

The Internet allows us to browse millions of videos. For some of them, the content is well organized with human-generated tags and labels (e.g. wedding ceremony, birthday party, etc.), but the rate at which content is uploaded daily makes it unrealistic to expect that user-provided labels will be sufficient for organizing this information in the future. We believe that automatically generating a brief summary (or "abstract") of videos is both an attractive solution to this problem and an exciting challenge for the natural language generation community. Converting audio and video output into natural language to create a human readable summary that facilitates effective browsing, supports classification decisions, or helps differentiating videos from one another without having to watch them in their entirety has both academic and practical value.

In this paper, we introduce an automatic video summary generation system that uses a natural language realization engine (Gatt and Reiter, 2009) to create sentences based on state-of-the-art video classification features. These features are computed on a large corpus from the TrecVID evaluation (Bao, et al. 2011). In a recent user study (Ding, et al. 2012), we compared automatically generated and manually generated summaries with respect to several tasks. The study shows, for example, that more specific information (e.g. "food" instead of "some object") and temporal information (something happened first and then…) is helpful in improving the quality of machine-generated summaries. This is a first step to implement an automatic system which is not only able to describe videos using natural language, but accomplishes more sophisticated tasks such as differentiating videos, finding supporting evidence for video classification and other tasks.

2 Related Work

Significant work has been done in the field of video summarization. A large part of it is based on the idea that the summarization should be a graphical representation such as visually rich storyboards. These storyboards intended to help users to efficiently browse the videos, e.g. in the Open-Video Archive (Marchionini, Song et al. 2009). Christel, et al. (2006) are mainly focusing on the research in user interface designs for video browsing and summarization. Li, et al. (2010) introduced a maximal marginal relevance algorithm working across video genres to improve the quality of the informative summary for a video, which exploits both audio and video information. Truong et al. (2007) worked on techniques targeting video data from various domains that were developed to summarize and organize the information and present surrogates to the users. Tan et al. (2011) recently have

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 128–130,
Utica, May 2012.

worked on using recognition techniques to obtain audio-visual concept classifiers to generate textual descriptions of videos. They manually defined a template for each concept and built a rule-based language generation system to create textual descriptions. But the template approach, which is directly related to specific events, cannot be adapted to new events. In our work, we use SimpleNLG to generate video-specific summaries, which can be applied to any new event.

3 System Description

3.1 Architecture

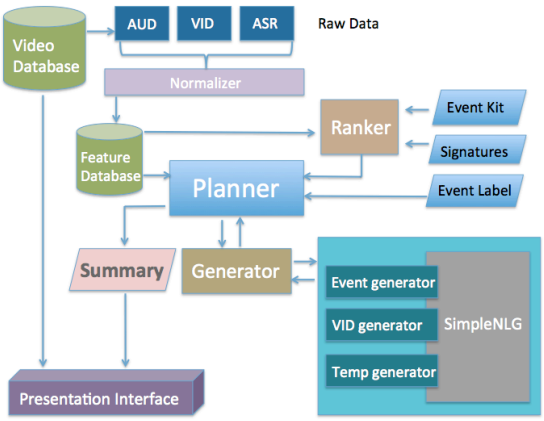

Figure 1. System Architecture.

Figure 1 shows the overall system architecture. The raw data of the videos is extracted and normalized to a format that can be read by the feature database, which stores all the features from the videos. The ranker contains a set of algorithms that rank the features from a video and conduct content determination. For example, when there is a long list of visual conceptual features, the ranker will sort all the features based on their relevance to the specific event's signature and return a ranked list to the planner. The planner is the "commander" of the system; it receives the ranked features and passes them to the language generator. For each set of the features, the language generator uses SimpleNLG to compile a sentence stating the scenario of the video. Eventually, the planner combines all the sentences into a summarization passage presenting the information detected from the video.

3.2 Feature Extraction

High-level features are extracted from the video using the techniques described in (Bao, et al. 2011). Visual conceptual features are detected with SVM classifiers trained on the SIN task in TRECVID 2011 using MOSIFT and CSIFT features to describe keyframes. Other features are also extracted, including event labels, event signatures and the event kit, etc. Event signatures are relevant features describing a certain event, similar to a fingerprint, and the event kit is a textual description of important objects and actions that make up the event. For features that make use of temporal information, we use a GMM based segmenter to cut the audio of each video into small clips (1-3 seconds) and give a label to each clip.

3.3 Language Generation

Taking a series of features, each of the sentence generators composes these features into a human readable sentence using the SimpleNLG generation tool. We use SimpleNLG at the lexical level (i.e. orthography, morphology and simple grammar) and at the phrase and sentence level (i.e. phrase element coordination, clause subordinates). For each set of features, the system generates a sentence specifically mentioning these features.

The VID generator deals with visual concepts, i.e. the probabilities of the occurrence of 346 visual concepts extracted from the video. A list of visual features (e.g. food, people, room) will be processed as follows:

SPhraseSpec p = nlgFactory.createClause();
p.setSubject("the system");
p.setVerb("observe");
p.setObject("food, people, room");
p.setFeature(Feature.TENSE, Tense.PAST);
p.addComplement("in the video")

to generate a sentence like:

The system observed food, people and room in the video.

Another sentence generator is the "temporal information generator", which takes the temporal information and produces a sentence describing what is happening in the video. We first segment the audio into small clips lasting three seconds each, and assign an audio semantic label to each clip (e.g. music, crowd, cheer, speech). Using temporal information, we generate a sentence like:

From the video, the system heard the sound of music at first, then cheer, and then speech.

When the system generates several sentences, we compose them into a summary paragraph of the video. For example, we combine the subordinate clauses using the conjunction "*because*":

The video summarization system thinks this video is about Birthday Party because it found 3 Or More People Meeting in Room.

In this sentence, "Birthday Party" is the event label for the given video, and "3 Or More People", "Meeting", "Room" are the visual concepts extracted from the video.

4 Demo System Interface

We demonstrate the video summarization system in a dynamic web page. A screen shot of the demo page can be seen in Figure 2. The top gallery shows several videos for selection. The user can choose a video by clicking on it, and the selected video will play in the main area of the page. Once a video is selected and playing, a summary paragraph will be automatically generated and displayed underneath the video, presenting the video's information in natural language.

Figure 2. A screen shot of the user interface.

The demo and the interface are currently being tested internally, in order to stabilize and improve all components, and to prepare for task-based and free-form evaluations on platforms such as Amazon Mechanical Turk, which will serve to further develop the NLG system. While the NLG is currently mostly hard-coded, the availability of an evaluation framework will allow us to learn parameters from data, and increase the amount of automation successively. In future work we will also explore and extend the feature sets by extracting additional visual, acoustic, textual features from the video. We also plan to employ more sophisticated NLG techniques (e.g. microplanning and document structuring) to generate more complex and authentic natural language sentences.

Acknowledgments

This work is partly supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Lei Bao, Shoou-I Yu, Zhen-zhong Lan, Arnold Overwijk, Qin Jin, Brian Langner, Michael Garbus, Susanne Burger, Florian Metze, Alexander Hauptmann. Informedia @TRECVID2011. *TRECVID2011, NIST.*

Michael G. Christel. 2006. Evaluation and User Studies with Respect to Video Summarization and Browsing. In *Proc. "Multimedia Content Analysis, Management, and Retrieval", part of the IS&T/SPIE Symposium on Electronic Imaging.*

Duo Ding, Florian Metze, Shourabh Rawat, Peter Franz Schulam, Susanne Burger, Ehsan Younessian, Lei Bao, Michael G. Christel, Alexander Hauptmann. 2012. Beyond Audio and Video Retrieval: Towards Multimedia Summarization. In *Proc. 2012 International Conference on Multimedia Retrieval.*

Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proc. 12th European Workshop on Natural Language Gneratoin-2009,* pages 90-93.

Yingbo Li, Bernardo Merialdo. 2010.Multi-video Summarization Based on AV-MMR. In *Proc. 2010 Int'l Workshop on Content-Based Multimedia Indexing*.

Gary Marchionini, Yaxiao Song, and Robert Ferrell. 2009. Multimedia Surrogates for Video Gisting: Toward Combining Spoken Words and Imagery. *Information Processing & Management* 45(6), 615-630.

Ba Tu Truong and Svetha Venkatesh. 2007. Video Abstraction: A Systematic Review and Classification. *ACM Trans. (TOMCCAP)* 3(1), 1-37.

Chun Chet Tan, Yu-Gang Jiang, Chong-Wah Ngo. 2011. Towards Textually Describing Complex Video Contents with Audio-Visual Concepts Classifiers. In *Proc. ACM Multimedia-2011.*