A Dataset for Movie Description

Xuewen Yang

June 4 2018

Abstract

Audio Description (AD) provides linguistic descriptions of movies and allows visually impaired people to follow a movie along with their peers.The author propose a novel dataset which contains transcribed ADs in this paper, which are temporally aligned to full length HD movies. In addition they also collected the aligned movie scripts which have been used in prior work and compare the two different sources of descriptions. In total the MPII Movie Description dataset (MPII-MD) contains a parallel corpus of over 68K sentences and video snippets from 94 HD movies.The author characterize the dataset by benchmarking different approaches for generating video descriptions.Comparing ADs to scripts,they find that ADs are far more visual and describe precisely what is shown rather than what should happen according to the scripts created prior to movie production.

1. Introduction

Audio descriptions (ADs) make movies accessible to millions of blind or visually impaired people1. AD provides an audio narrative of the most important aspects of the visual information, namely actions, gestures, scenes,and character appearance as can be seen in Figures 1.AD is prepared by trained describers and read by professional narrators. More and more movies are audio transcribed,but it may take up to 60 person-hours to describe a 2-hour movie, resulting in the fact that only a small subset of movies and TV programs are available for the blind. Consequently, automating this would be a noble task. In this section the author present a novel dataset which provides transcribed ADs, which are aligned to full length HD movies. For this they retrieve audio streams from Blu-ray HD disks, segment out the sections of the AD audio and transcribe them via a crowd-sourced transcription service. As the ADs are not fully aligned to the activities in the video,they manually align each sentence to the movie.Therefore, in contrast to [5], our dataset provides alignment to the actions in the video, rather than just to the audio track of the description.

AD: Abby gets in the basket.

Script: After a moment a frazzled Abby pops up in his place.

Mike leans over and sees how high they are.

Mike looks down to see – they are now fifteen feet above the ground.

Abby clasps her hands around his face and kisses him passionately. For the first time in her life, she stops thinking and grabs Mike and kisses the hell out of him.

Figure 1. Audio description (AD) and movie script samples from the movie Ugly Truth.

In addition the author also mine existing movie scripts, prealign them automatically, similar to [1] [2] and then manually align the sentences to the movie.

2. Related Work

In this section,the author first discuss recent approaches to video description and then the existing works using movie scripts and ADs.In recent years there has been an increased interest in automatically describing images and videos with natural language. While recent works on image description show impressive results by learning the relations between images and sentences and generating novel sentences [4], the video description works typically rely on retrieval or templates and frequently use a separate language corpus to model the linguistic statistics. A few exceptions exist: it uses a pre-trained model for image-description and adapts it to video description. [4] learn a translation model, however, the approaches rely on a strongly annotated corpus with aligned videos, annotated labels and sentences. The main reason for video description lacking behind image description seems to be a missing corpus to learn and understand the problem of video description. they aim to address this limitation by collecting a large, aligned corpus of video snippets and descriptions. To handle the setting of having only videos and sentences without annotated labels for each

video snippet, they propose an approach which adapts, by extracting labels from the sentences.Their extraction of labels has similarities, but they aim to extract the senses of the words automatically by using semantic parsing.

3. The MPII Movie Description dataset

Despite the potential benefit of ADs for computer vision, they have not been used so far apart from as well as who study how to automate AD production. We believe the main reason for this is that they are not available in the text format, i.e. transcribed. We tried to get access to AD transcripts from description services as well as movie and TV production companies, but they were not ready to provide or sell them. While script data is easier to obtain, large parts of it do not match the movie, and they have to be cleaned up.

3.1. Collection of ADs

In this section,the author search for Blu-ray movies with ADs in the AudioDescription section of the British Amazon and select a set of 55 movies of diverse genres. As ADs are only available in audio format,they first retrieve the audio stream from Blu-ray HD disk.Then they semi-automatically segment out the sections of the AD audio (which is mixed with the original audio stream) with the approach described below.The audio segments are then transcribed by a crowdsourced transcription service that also provides the time-stamps for each spoken sentence. As the AD is added to the original audio stream between the dialogs, there might be a small misalignment between the time of speech and the corresponding visual content. Therefore,the author manually align each sentence to the movie in-house.

3.2. Collection of script data

In addition to the ADs they mine script web resources and select 39 movie scripts. As starting point the author use the movies scripts from Hollywood2 [3] that have highest alignment scores to the movie.The author are also interested in comparing the two sources (movie scripts and ADs), so they are looking for the scripts labeled as Final, Shooting, or Production Draft where ADs are also available.They found that the overlap is quite narrow, so they analyze 11 such movies in our dataset. This way they end up with 50 movie scripts in total. They follow existing approaches [2] to automatically align scripts to movies. First they parse the scripts, extending the method of [2] to handle scripts which deviate from the default format. Second, they extract the subtitles from the Blu-ray disks. Then they use the dynamic programming method of [2] to align scripts to subtitles and infer the time-stamps for the description sentences.The author select the sentences with a reliable alignment score (the ratio of matched words in the near-by monologues) of at least 0.5. The obtained sentences are then manually aligned to video in-house.

References

[1] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *ECCV*, 2008. 1

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 2

[3] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2

[4] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013. 1

[5] A. Torabi, C. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. *Computer Science*, 2015. 1