A Dataset for Movie Description

Anna Rohrbach[1], Marcus Rohrbach[2] Niket Tandon[1] Bernt Schiele[1]
[1]Max Planck Institute for Informatics, Saarbrücken, Germany. [2]UC Berkeley EECS and ICSI, Berkeley, CA, United States.

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".

	Unique movies	Total sentences	Total clips	Average length	Total length
AD	55	37,272	37,266	4.1 sec.	42.5 h.
Movie script	50	31,103	31,071	3.6 sec.	31.1 h.
Total	94	68,375	68,337	3.9 sec.	73.6 h.

Table 1: MPII Movie Description dataset statistics.

	Correctness	Relevance
ADs	66.1 (35.7)	66.6 (44.9)
Movie scripts	33.9 (11.2)	33.4 (16.8)

Table 2: Human evaluation of ADs and movie scripts: which sentence is more correct/relevant with respect to the video. Majority vote of 5 judges in %. In brackets: at least 4 of 5 judges agree.

Audio Descriptions (ADs) provide linguistic descriptions of movies and allow visually impaired people to follow a movie along with their peers. Such descriptions are by design mainly visual and naturally form an interesting data source for computer vision and computational linguistics. ADs are 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. Consequently, only a small subset of movies and TV programs are available for the blind, thus automating AD would be a noble task. In addition to the benefits for the blind, generating descriptions for video is an interesting task in itself requiring to understand and combine core techniques of computer vision and computational linguistics.

To be able to learn how to generate descriptions of visual content, parallel datasets of visual content paired with descriptions are indispensable. While recently several large datasets have been released which provide images with descriptions [3, 4], video description datasets focus on short video snippets only and are limited in size [1] or not publicly available. TACoS Multi-Level [5] and YouCook [2] are exceptions by providing multiple sentence descriptions and longer videos, however they are restricted to the cooking scenario. In this work we propose a novel dataset which contains transcribed ADs, which are temporally aligned to full length HD movies. We also collected the aligned movie scripts which have been used in prior work and compare the two different sources of descriptions (see Figure 1). Movies are open domain and realistic, even though, as any other video source (e.g. YouTube or surveillance videos), have their specific characteristics. ADs and scripts associated with movies provide rich multiple sentence descriptions. This opens new possibilities to understand stories and plots across multiple sentences in an open domain large scale scenario.

In total the *MPII Movie Description* dataset (MPII-MD) contains a parallel corpus of over 68K sentences and video snippets from 94 HD movies (see Table 1). 55 movies are aligned to ADs, while 39 are aligned only to scripts and 11 to both scripts and ADs. We characterize the dataset by benchmarking different approaches for generating video descriptions. First are nearest neighbour retrieval using state-of-the-art visual features which do not require any additional labels, but retrieve sentences from the training data. Second, we adapt the approach of [6] by automatically extracting the semantic representation from the sentences using semantic parsing.

Figure 1 shows examples of ADs and compares them to movie scripts. Scripts have been used for various tasks, but so far not for video description. The main reason for this is that automatic alignment frequently fails due to the discrepancy between the movie and the script. Even when perfectly aligned to the movie it frequently is not as precise as ADs because it is usually produced prior to the shooting of the movie. A typical case is that part of a sentence is correct, while another part contains irrelevant information. We compare ADs and script data using 11 movies from our dataset where both are available. For these movies we select the overlapping time intervals with the intersection over union overlap of at least 75%. We ask humans via Amazon Mechanical Turk (AMT) to compare the sentences with respect to their correctness and relevance to the video, using both video intervals as references (one at a time). Each task was completed by 5 different human subjects. Table 2 presents the results of this evaluation. ADs are ranked as more correct and relevant in about 2/3 of the cases, which supports our intuition that scrips contain mistakes and irrelevant content even after being cleaned up and manually aligned.

Our dataset is available at www.mpii.de/movie-description. We will release sentences, temporal alignments, video snippets (with audio), semantic parser outputs, and intermediate computed features. We hope that this will foster research in different areas including video description, activity recognition, visual grounding, and understanding of plots.

[1] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011.

[2] Pradipto Das, Chenliang Xu, Richard Doell, and Jason Corso. Thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[3] Peter Hodosh, Alice Young, Micah Lai, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics (TACL)*, 2014.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[5] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Proceedings of the German Confeence on Pattern Recognition (GCPR)*, September 2014.

[6] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.