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

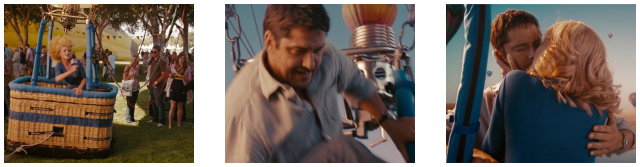
Abstract

Audio Description (AD) provides linguistic descriptions of movies and allows visually impaired people to follow a movie along with their peers. Such descriptions are by design mainly visual and thus naturally form an interesting data source for computer vision and computational linguistics. In this work we propose a novel dataset which contains transcribed ADs, which are temporally aligned to full length HD movies. In addition we 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. We characterize the dataset by benchmarking different approaches for generating video descriptions. Comparing ADs to scripts, we 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 people[1]. AD provides an audio narrative of the "most important aspects of the visual information" [61], namely actions, gestures, scenes, and character appearance as can be seen in Figures 1 and 2. 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 [44], 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 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 understand the visual input one has to reliably recognize scenes, human activities, and participating objects. To generate a good description one has to decide what part of the visual information to verbalize, i.e. recognize what is salient.



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.

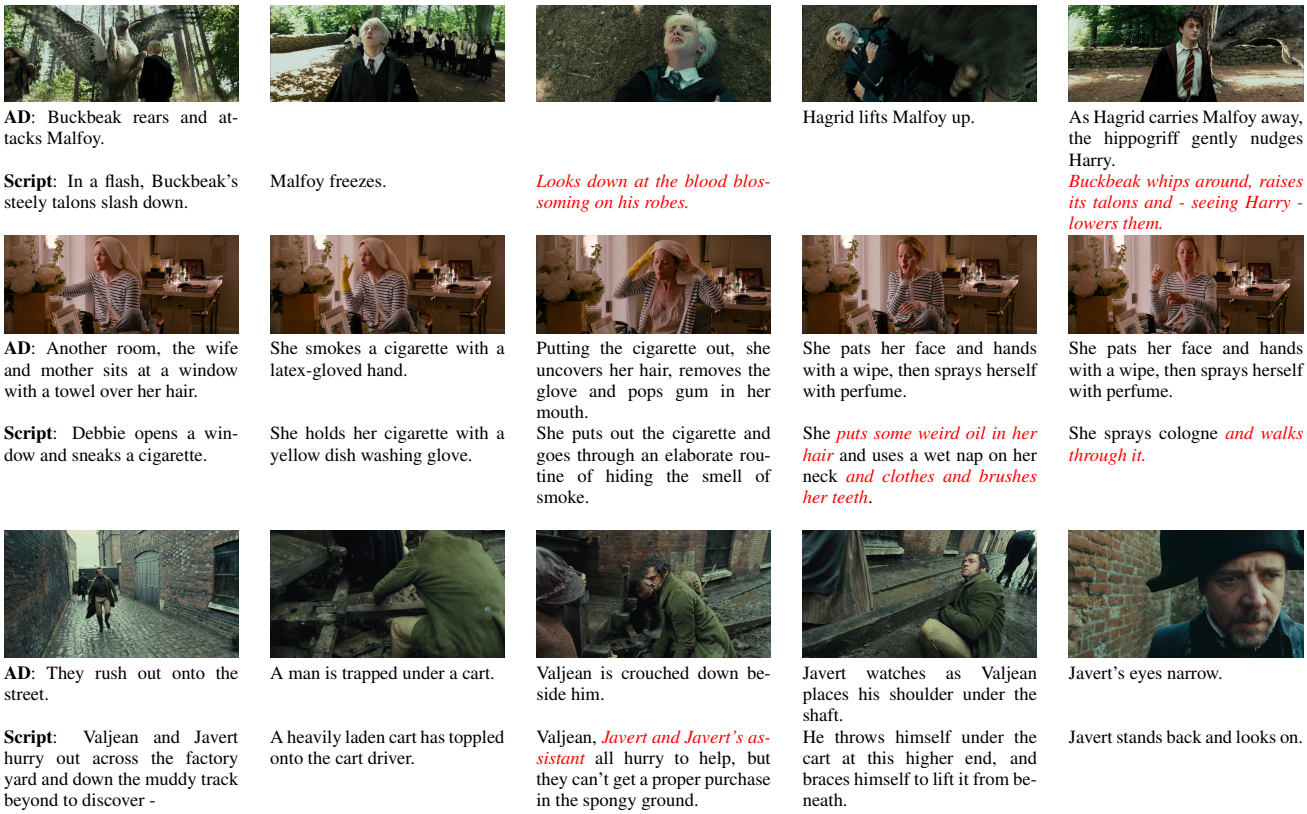
Mike looks down to see – they are now fifteen feet above the ground.

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".

Large datasets of objects [19] and scenes [73, 76] had an important impact in the field and significantly improved our ability to recognize objects and scenes in combination with CNNs [40]. To be able to learn how to generate descriptions of visual content, parallel datasets of visual content paired with descriptions are indispensable [59]. While recently several large datasets have been released which provide images with descriptions [31, 49, 53], video description datasets focus on short video snippets only and are limited in size [12] or not publicly available [54]. TACoS Multi-Level [58] and YouCook [17] are exceptions by providing multiple sentence descriptions and longer videos. While these corpora pose challenges in terms of fine-grained recognition, they are restricted to the cooking scenario. In contrast, movies are open domain and realistic, even though, as any other video sources (e.g. YouTube or surveillance videos), have their specific characteristics. ADs and scripts associated with movies provide rich multiple sentence descriptions. They even go beyond this by telling a story which means they allow to study how to extract plots and understand long term semantic dependencies and human interactions from the visual and textual data.

[1] In this work we refer for simplicity to "the blind" to account for all blind and visually impaired people which benefit from AD, knowing of the variety of visually impaired and that AD is not accessible to all.



AD: Buckbeak rears and attacks Malfoy.

Script: In a flash, Buckbeak's steely talons slash down.

Malfoy freezes.

Looks down at the blood blossoming on his robes.

Hagrid lifts Malfoy up.

As Hagrid carries Malfoy away, the hippogriff gently nudges Harry.
Buckbeak whips around, raises its talons and - seeing Harry - lowers them.

AD: Another room, the wife and mother sits at a window with a towel over her hair.

Script: Debbie opens a window and sneaks a cigarette.

She smokes a cigarette with a latex-gloved hand.

She holds her cigarette with a yellow dish washing glove.

Putting the cigarette out, she uncovers her hair, removes the glove and pops gum in her mouth.
She puts out the cigarette and goes through an elaborate routine of hiding the smell of smoke.

She pats her face and hands with a wipe, then sprays herself with perfume.

She *puts some weird oil in her hair* and uses a wet nap on her neck *and clothes and brushes her teeth*.

She pats her face and hands with a wipe, then sprays herself with perfume.

She sprays cologne *and walks through it.*

AD: They rush out onto the street.

Script: Valjean and Javert hurry out across the factory yard and down the muddy track beyond to discover -

A man is trapped under a cart.

A heavily laden cart has toppled onto the cart driver.

Valjean is crouched down beside him.

Valjean, *Javert and Javert's assistant* all hurry to help, but they can't get a proper purchase in the spongy ground.

Javert watches as Valjean places his shoulder under the shaft.
He throws himself under the cart at this higher end, and braces himself to lift it from beneath.

Javert's eyes narrow.

Javert stands back and looks on.

Figure 2: Audio description (AD) and movie script samples from the movies "Harry Potter and the Prisoner of Azkaban", "This is 40", "Les Miserables". Typical mistakes contained in scripts marked in *red italic*.

Figures 1 and 2 show examples of ADs and compare them to movie scripts. Scripts have been used for various tasks [14, 21, 45, 48, 51], 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, the script frequently is not as precise as the AD because it is typically produced prior to the shooting of the movie. E.g. in Figure 2 we note mistakes marked in red. A typical case is that part of the sentence is correct, while another part contains irrelevant information.

In this work we present a novel dataset which provides transcribed ADs, which are aligned to full length HD movies. For this we 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 [2]. As the ADs are not fully aligned to the activities in the video, we manually align each sentence to the movie. Therefore, in contrast to [61, 62, 69], our dataset provides alignment to the actions in the video, rather than just to the audio track of the description. In addition we also mine existing movie scripts, pre-align them automatically, similar to [14, 45] and then manually align the sentences to the movie.

As a first study on our dataset we benchmark several approaches for movie description. First are nearest neighbour retrieval using state-of-the-art visual features [32, 72, 76] which do not require any additional labels, but retrieve sentences from the training data. Second, we adapt the approach of [59] by automatically extracting the semantic representation from the sentences using semantic parsing. This approach achieves competitive performance on the TACoS Multi-Level corpus [58] without using the annotations and outperforms the retrieval approaches on our novel MPII Movie Description dataset.

The main contribution of this work is our novel MPII Movie Description dataset (MPII-MD) which provides transcribed and aligned AD and script data sentences. We provide access to our dataset on our web page. We hope that our dataset will foster research in different areas including video description, activity recognition, visual grounding, and understanding of plots. Additionally we present an approach to semi-automatically collect and align AD data and analyse the differences between ADs and movie scripts.

2. Related Work

We 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 [24, 25, 36, 41, 42, 43, 47, 52, 65] and videos [8, 17, 28, 29, 30, 34, 39, 58, 66, 70] with natural language. While recent works on image description show impressive results by *learning* the relations between images and sentences and generating novel sentences [13, 20, 33, 37, 43, 50, 59, 71], the video description works typically rely on retrieval or templates [17, 28, 29, 39, 41, 66, 68] and frequently use a separate language corpus to model the linguistic statistics. A few exceptions exist: [70] uses a pre-trained model for image-description and adapts it to video description. [20, 59] 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. We 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, we propose an approach which adapts [59], by extracting labels from the sentences. Our extraction of labels has similarities to [68], but we aim to extract the senses of the words automatically by using semantic parsing as discussed in Section 5.

Movie scripts have been used for automatic discovery and annotation of scenes and human actions in videos [21, 45, 51]. We rely on the approach presented in [45] to align movie scripts using subtitles. [10] attacks the problem of learning a joint model of actors and actions in movies using weak supervision provided by scripts. They also rely on a semantic parser (SEMAFOR [16]) trained on the FrameNet database [7], however, they limit the recognition only to two frames. [11] aims to localize individual short actions in longer clips by exploiting the ordering constrains as weak supervision. [10, 11, 21, 45, 51] proposed datasets focused on extracting several activities from movies. Most of them are part of the "Hollywood2" dataset [51] which contains 69 movies and 3669 clips. Another line of works, [15, 22, 56, 64, 67], proposed datasets for character identification targeting TV shows. All the mentioned datasets rely on alignment to movie/TV scripts and none uses ADs.

Recently, [74] proposed a method to generate audio descriptions (ADs) from video using recurrent neural networks and incorporating a soft-attention mechanism. The method is evaluated on a new corpus of audio described movies M-VAD [69], collected in parallel and independent from our corpus. We provide more detailed comparison of their corpus to ours in Section 3.3. ADs have also been used to understand which characters interact with each other [62]. There are some initial works to support AD production using scripts as source [44] and automatically finding scene boundaries [27]. [61] analyses the linguistic properties on a non-public corpus of ADs from 91 movies. Their corpus is based on the original sources to create the ADs and contains different kinds of artifacts not present in actual description, such as dialogs and production notes. In contrast our text corpus is much cleaner as it consists only of the actual ADs.

Semantic parsing has received much attention in computational linguistics recently, see, for example, the tutorial [6] and references given there. Although aiming at general-purpose applicability, it has so far been successful rather for specific use-cases such as natural-language question answering [9, 23] or understanding temporal expressions [46].

3. The MPII Movie Description dataset

Despite the potential benefit of ADs for computer vision, they have not been used so far apart from [74] as well as [27, 44] 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". In the following we describe our semi-automatic approach to obtain AD and scripts and align them to the video.

3.1. Collection of ADs

We search for Blu-ray movies with ADs in the "Audio Description" section of the British Amazon [1] and select a set of 55 movies of diverse genres. As ADs are only available in audio format, we first retrieve the audio stream from Blu-ray HD disk[2]. Then we 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 crowd-sourced transcription service [2] that also provides us 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, we manually align each sentence to the movie in-house.

Semi-Automatic segmentation of ADs. We first estimate the temporal alignment difference between the ADs and the original audio (which is part of the ADs), as they might

[2]We use [3] to extract a Blu-ray in the .mkv file, then [5] to select and extract the audio streams from it.

	Unique Movies	Before alignment Words	After alignment Words	Sentences	Clips	Avg. length	Total length
AD	55	346,557	332,846	37,272	37,266	4.1 sec.	42.5 h.
Movie script	50	398,072	320,621	31,103	31,071	3.6 sec.	31.1 h.
Total	94	744,629	653,467	68,375	68,337	3.9 sec.	73.6 h.

Table 1: MPII Movie Description dataset statistics. Discussion see Section 3.3.

Dataset	multi-sentence	domain	sentence source	videos	clips	sentences
YouCook [28]	x	cooking	crowd	88	-	2,668
TACoS [57, 59]	x	cooking	crowd	127	7,206	18,227
TACoS Multi-Level [58]	x	cooking	crowd	185	14,105	52,593
MSVD [12]		open	crowd	-	1,970	70,028
M-VAD [69]	x	open	professional	92	48,986	55,904
MPII-MD (ours)	x	open	professional	94	68,337	68,375

Table 2: Comparison of video description datasets. Discussion see Section 3.3.

be off a few time frames. The precise alignment is important to compute the similarity of both streams. Both steps (alignment and similarity) are computed using the spectograms of the audio stream, which is computed using Fast Fourier Transform (FFT). If the difference between both audio streams is larger than a given threshold we assume the AD contains audio description at that point in time. We smooth this decision over time using a minimum segment length of 1 second. The threshold was picked on a few sample movies, but has to be adjusted for each movie due to different mixing of the audio description stream, different narrator voice level, and movie sound.

3.2. Collection of script data

In addition to the ADs we mine script web resources[3] and select 39 movie scripts. As starting point we use the movies scripts from "Hollywood2" [51] that have highest alignment scores to the movie. We are also interested in comparing the two sources (movie scripts and ADs), so we are looking for the scripts labeled as "Final", "Shooting", or "Production Draft" where ADs are also available. We found that the "overlap" is quite narrow, so we analyze 11 such movies in our dataset. This way we end up with 50 movie scripts in total. We follow existing approaches [14, 45] to automatically align scripts to movies. First we parse the scripts, extending the method of [45] to handle scripts which deviate from the default format. Second, we extract the subtitles from the Blu-ray disks[4]. Then we use the dynamic programming method of [45] to align scripts to subtitles and infer the time-stamps for the description sentences.

[3]http://www.weeklyscript.com, http://www.simplyscripts.com, http://www.dailyscript.com, http://www.imsdb.com

[4]We extract .srt from .mkv with [4]. It also allows for subtitle alignment and spellchecking.

We 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.

3.3. Statistics and comparison to other datasets

During the manual alignment we filter out: a) sentences describing movie introduction/ending (production logo, cast etc); b) texts read from the screen; c) irrelevant sentences describing something not present in the video; d) sentences related to audio/sounds/music. Table 1 presents statistics on the number of words before and after the aligment to video. One can see that for the movie scripts reduction in number of words is about 19.5%, while for ADs it is 3.9%. In case of ADs filtering mainly happens due to inital/ending movie intervals and transcribed dialogs (when shown as text). For the scripts it is mainly attributed to irrelevant sentences. Note, that in cases when the sentences are "alignable" but have minor mistakes we still keep them. As sometimes multiple sentences migh refer to the same video snippet, the number of clips (68,337) is slightly smaller than the total number of sentences (68,375).

We compare our corpus to other existing parallel video corpora in Table 2. The main limitations of existing datasets are single domain [17, 57, 58] or limited number of video clips [28]. Recently, in parallel with our work, [69] proposed a similar dataset M-VAD of movies with ADs. There are three differences between our and their corpus. First, our corpus consists both of movie scripts and ADs, while they only use ADs. Second, we manually align every sentence to the corresponding activity in the video, while they rely on automatic AD detection and use its timestamps, leading to less precise alignment. Last, we use Blu-ray HD movies, while they use DVDs. Both datasets contribute with

a large number of realistic open domain videos, provide high quality (professional) sentences and allow for multi-sentence description. We end up with the largest (as of now) parallel corpus of over 68K video-sentence pairs and a total length over 73 hours.

3.4. Visual features

We extract video snippets from the full movie based on the aligned sentence intervals. We also uniformly extract 10 frames from each video snippet. As discussed above AD and scripts describe activities, object, and scenes (as well as emotions which we do not explicitly handle with these features, but they might still be captured, e.g. by the context or activities). In the following we briefly introduce the visual features computed on our data which we will also make publicly available.

DT We extract the improved dense trajectories compensated for camera motion [72]. For each feature (Trajectory, HOG, HOF, MBH) we create a codebook with 4000 clusters and compute the corresponding histograms. We apply L1 normalization to the obtained histograms and use them as features.

LSDA We use the recent large scale object detection CNN [32] which distinguishes 7604 ImageNet [19] classes. We run the detector on every second extracted frame (due to computational constraints). Within each frame we max-pool the network responses for all classes, then do mean-pooling over the frames within a video snippet and use the result as a feature.

PLACES and HYBRID Finally, we use the recent scene classification CNNs [76] featuring 205 scene classes. We use both available networks: *Places-CNN* and *Hybrid-CNN*, where the first is trained on the Places dataset [76] only, while the second is additionally trained on the 1.2 million images of ImageNet (ILSVRC 2012) [60]. We run the classifiers on all the extracted frames of our dataset. We mean-pool over the frames of each video snippet, using the result as a feature.

4. Approaches to video description

In this section we discuss different approaches to the video description task that we benchmark on our proposed dataset. Given a training corpus of aligned videos and sentences we want to describe a new unseen test video.

Nearest neighbor We retrieve the closest sentence from the training corpus using the L1-normalized visual features introduced in Section 3.4 and the intersection distance.

SMT We adapt the two-step translation approach of [59] which uses an intermediate semantic representation (SR), modeled as a tuple, e.g. $\langle cut, knife, tomato \rangle$. As the first step it learns a mapping from the visual input to the semantic representation (SR), modeling pairwise dependencies in a CRF using visual classifiers as unaries. The unaries are trained using an SVM on dense trajectories [72]. In the second step [59] translates the SR to a sentence using Statistical Machine Translation (SMT) [38]. For this the approach uses a concatenated SR as input language, e.g. *cut knife tomato*, and natural sentence as output language, e.g. *The person slices the tomato.* While we cannot rely on an annotated SR as in [59], we automatically mine the SR from sentences using semantic parsing which we introduce in the next section. In addition to dense trajectories we use the features described in Section 3.4.

Phrase	WordNet Mapping	VerbNet Mapping	Expected Frame
the man	man#1	Agent.animate	Agent: man#1
begin to shoot	shoot#4	shoot#vn#3	Action: shoot#4
a video	video#1	Patient.solid	Patient: video#1
in	in	PP.in	
the moving bus	bus#1	NP.Location. solid	Location: moving bus#1

Table 3: Semantic parse for "*He began to shoot a video in the moving bus*". Discussion see Section 5.1

SMT Visual words As an alternative to potentially noisy labels extracted from the sentences, we try to directly translate visual classifiers and visual words to a sentence. We model the essential components by relying on activity, object, and scene recognition. For objects and scenes we employ on the pre-trained models LSDA and PLACES. For activities we rely on the state-of-the-art activity recognition feature DT. We cluster the DT histograms to 300 visual words using k-means. The index of the closest cluster center DT_i from our activity category is chosen as label. To build our tuple we obtain the highest scoring class labels of the object detector and scene classifier. More specifically for the object detector we consider two highest scoring classes: for subject and object. Thus we obtain the tuple $\langle SUBJECT, ACTIVITY, OBJECT, SCENE \rangle = \langle argmax(LSDA), DT_i, argmax2(LSDA), argmax(PLACES) \rangle$, for which we learn a translation model to a natural sentence using the SMT approach discussed above.

5. Semantic parsing

Learning from a parallel corpus of videos and natural language sentences is challenging when no annotated intermediate representation is available. In this section we introduce our approach to exploit the sentences using semantic parsing. The proposed method extracts intermediate semantic representations (SRs) from the natural sentences. Later in the section we perform an evaluation of our method on

a corpus where annotated SRs are available in context of a video description task.

5.1. Semantic parsing approach

We lift the words in a sentence to a semantic space of roles and WordNet [26, 55] senses by performing SRL (Semantic Role Labeling) and WSD (Word Sense Disambiguation). For an example, refer to Table 3, the expected outcome of semantic parsing on the input sentence "*He shot a video in the moving bus*" is "`Agent: man, Action: shoot, Patient: video, Location: bus`". Additionally, the role fillers are disambiguated.

We use the ClausIE tool [18] to decompose sentences into their respective clauses. For example, "*he shot and modified the video*" is split into two phrases "*he shot the video*" and "*he modified the video*"). We then use the OpenNLP tool suite[5] for chunking the text of each clause. In order to provide the linking of words in the sentence to their WordNet sense mappings, we rely on a state-of-the-art WSD system, IMS [75]. The WSD system, however, works at a word level. We enable it to work at a phrase level. For every noun phrase, we identify and disambiguate its head word (e.g. `the moving bus` to "*bus#1*", where "*bus#1*" refers to the first sense of the word `bus`). We link verb phrases to the proper sense of its head word in WordNet (e.g. `begin to shoot` to "*shoot#4*"). The phrasal verbs such as e.g. *"pick up"* or *"turn off"* are preserved as long as they exist in WordNet.

In order to obtain word role labels, we link verbs to VerbNet [35, 63], a manually curated high-quality linguistic resource for English verbs. VerbNet is already mapped to WordNet, thus we map to VerbNet via WordNet. We perform two levels of matches in order to obtain role labels. First is the syntactic match. Every VerbNet verb sense comes with a syntactic frame e.g. for `shoot`, the syntactic frame is `NP V NP`. We first match the sentence's verb against the VerbNet frames. These become candidates for the next step. Second we perform the semantic match: VerbNet also provides a role restriction on the arguments of the roles e.g. for `shoot` (sense killing), the role restriction is `Agent.animate V Patient.`**`animate`** `PP Instrument.solid`. For the other sense for `shoot` (sense snap), the semantic restriction is `Agent.animate V Patient.`**`solid`**. We only accept candidates from the syntactic match that satisfy the semantic restriction.

VerbNet contains over 20 roles and not all of them are general or can be recognized reliably. Therefore, we group them to get the SUBJECT, VERB, OBJECT and LOCATION roles. We explore two approaches to obtain the labels based on the output of the semantic parser. First is to use the extracted text chunks directly as labels. Second is to use the corresponding senses as labels (and therefore group multiple text labels). In the following we refer to these as *text-* and *sense-labels*. Thus from each sentence we extract a semantic representation in a form of (SUBJECT, VERB, OBJECT, LOCATION).

[5] http://opennlp.sourceforge.net/

Approach	BLEU
SMT [59]	24.9
SMT [58]	26.9
SMT with our text-labels	22.3
SMT with our sense-labels	24.0

Table 4: Video description performance (BLEU@4 in %) on Detailed Descriptions from the TACoS Multi-Level [58], see Section 5.2.

Labels	activity	tool	object	source	target
Manual [58]	78	53	138	69	49
	verb	object		location	
Our text-labels	145	260		85	
Our sense-labels	158	215		85	

Table 5: Comparing manual labels from TACoS Multi-Level [58] and automatic labels obtained by our semantic parser, see Section 5.2.

5.2. Applying parsing to TACoS Multi-Level corpus

We apply the proposed semantic parsing to the TACoS Multi-Level [58] parallel corpus. We extract the SR from the sentences as described above and use those as annotations/labels. Note, that this corpus is annotated with the tuples (ACTIVITY, OBJECT, TOOL, SOURCE, TARGET) and the subject is always the person. Therefore we drop the SUBJECT role and only use (VERB, OBJECT, LOCATION) as our SR. After extracting the labels for VERBs, OBJECTs, LOCATIONs with our parser we only use those that appear at least 30 times. For them we train the visual classifiers as in [58]. Next we train a CRF with 3 nodes for verbs, objects and locations, using the visual classifier responses as unaries. We follow the translation approach of [59] and train the SMT on the Detailed Descriptions part of the corpus using our labels. Finally, we translate the SR predicted by our CRF to generate the sentences. Table 4 shows the results comparing our method to [59] and [58] who use manual annotations to train their models. As we can see the sense-labels perform better than the text-labels as they provide better grouping of the labels. Our method produces competitive result which is only 0.9% below the result of [59]. At the same time [58] uses more training data, additional color SIFT features and recognizes the dish prepared in the video. All these points, if added to our approach, would also improve the performance.

We analyze the labels selected by our method in Table

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

Table 6: Human evaluation of movie scripts and ADs: which sentence is more correct/relevant with respect to the video (forced choice). Majority vote of 5 judges in %. In brackets: at least 4 of 5 judges agree. See also Section 6.1.

Corpus	Clause	NLP	Labels	WSD
TACoS Multi-Level [58]	0.96	0.86	0.91	0.75
MPII-MD	0.89	0.62	0.86	0.7

Table 7: Semantic parser accuracy on TACoS Multi-Level and MPII-MD. Discussion in Section 6.2.

	Correctness	Grammar	Relevance
Nearest neighbor			
1: DT	8.7	6.9	8.7
2: LSDA	8.4	6.5	8.5
3: PLACES	8.1	6.4	8.2
4: HYBRID	7.8	6.3	7.8
5: SMT Visual words	6.7	8.5	6.9
SMT with our text-labels			
6: DT 30	6.4	6.4	6.2
7: DT 100	6.4	6.4	6.4
8: Combi 100	6.1	6.7	6.2
SMT with our sense-labels			
9: DT 30	5.4	5.8	5.4
10: DT 100	5.3	6.1	5.4
11: Combi 100	5.7	6.3	5.7
12: Scripts/ADs	2.9	5.7	2.5
13: Movie scripts*	3.7	5.9	3.1
14: ADs*	2.3	5.5	2.1

Table 8: Video description performance of different methods on MPII-MD. Mean Ranking (1-12), lower is better. Movie scripts only and ADs Discussion in Section 6.3. * Mean ranks computed only on the corresponding subset of the data.

5. We see that comparing to the manual labels the number nearly doubles. This is due to two reasons. First, different labels might still be assigned to very similar concepts. Second, the manual annotations were created prior to the sentence collection, so some words used by humans in sentences might not be present in the annotations.

From this experiment we conclude that the output of our automatic parsing approach can serve as a replacement of manual annotations and allows to achieve competitive results. In the following we apply this approach to our Movie Description dataset.

6. Evaluation

In this section we provide more insights about our movie description dataset. First we compare ADs to movie scripts and then benchmark the approaches to video description introduced in Section 4.

6.1. Comparison of AD vs script data

We compare the AD and script data using 11 movies from our dataset where both are available (see Section 3.2). For these movies we select the overlapping time intervals with the intersection over union overlap of at least 75%, which results in 279 sentence pairs, we remove 2 pairs which have idendical sentences. 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 a reference (one at a time). Each task was completed by 5 different human subjects, covering 2770 tasks done in total. Table 6 presents the results of this evaluation. AD is 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.

6.2. Semantic parser evaluation

Table 7 reports the accuracy of individual components of the semantic parsing pipeline. The components are clause splitting (Clause), POS tagging and chunking (NLP), semantic role labeling (Labels) and word sense disambiguation (WSD). We manually evaluate the correctness on a randomly sampled set of sentences using human judges. It is evident that the poorest performing parts are the NLP and the WSD components. Some of the NLP mistakes arise due to incorrect POS tagging. WSD is considered a hard problem and when the dataset contains less frequent words, the performance is severely affected. Overall we see that the MPII-MD dataset is more challanging than TACoS Multi-Level but the drop in performance is reasonable compared to the siginificantly larger variability.

6.3. Video description

As the collected text data comes from the movie context, it contains a lot of information specific to the plot, such as names of the characters. We pre-process each sentence in the corpus, transforming the names to “Someone” or “people” (in case of plural). The transformed version of the corpus is used in all the experiments below. We provide the transformed and the original corpus.

Labels	subject	verb	object	location
text-labels 30	94	534	387	153
sense-labels 30	89	608	385	159
text-labels 100	16	186	88	50
sense-labels 100	16	193	80	51

Table 9: Statistics for the labels obtained by our semantic parser on MPII-MD. 30 and 100 is the minimum number of label occurrences, see Section 6.3.

For the video description task we split the 11 movies with associated scripts and ADs (in total 22 alignments, 2 for each movie) from Section 3.2) into validation set (8) and test set (14). The other 83 movies are used for training. Human judges are asked to rank multiple sentence outputs with respect to their correctness, grammar and relevance to the video (as in [58]).

Table 8 summarizes results of the human evaluation from 373[6] randomly selected test video snippets, showing the mean rank, where lower is better. In the top part of the table we show the nearest neighbor results based on multiple visual features. When comparing the different features, we notice that the pre-trained features (LSDA, PLACES, HYBRID) perform better than DT, where HYBRID performing best. Next is the translation approach with the visual words as labels, performing better than the nearest neighbors in terms of correctness and relevance, however loosing to human written sentences in terms of grammar. The next two blocks correspond to the translation approach when using the labels from our semantic parser. After extracting the labels we select the ones which appear at least 30 or 100 times as our visual attributes, see Table 9 for details. "Combi 100" refers to combining DT, HYBRID, and PLACES as unaries in the CRF. We did not add LSDA as we found that it reduces the performance of the CRF. Finally, the last "Reference" block refers to the human written test sentences from the corpus and not surprisingly ranks best.

Overall we can observe the following tendencies: (1) Using our parsing with SMT outperforms nearest neighbor baselines and SMT Visual words. This is also reflected in Figure 3, showing example outputs of all the evaluated approaches for a single movie snippet. (2) The mean rank of the actual movie script/AD is significantly lower, i.e. better, than any of the automatic approaches. (3) When computing ranks only on the AD (line 14), the mean rank for correctness/relevance is 1.4/1.0 lower than for Movie scripts (line 13). This confirms the observation made in Table 6.

[6]From 500 submitted tasks to AMT we had to remove 3 workers who obviously did not follow the instructions, leaving us with 373 evaluated snippets.


Nearest neighbor	
1: DT	Someone fetches someone, who's holding her resume.
2: LSDA	Someone watches the man place his suit in a locker as the guard rubs someone's butt.
3: PLACES	Someone holds up two ties for her approval.
4: HYBRID	She looks round on hearing the noise and sees a cupboard door opening.
5: SMT Visual W.	Someone gets out of the desk.
SMT with our text-labels	
6: DT 30	Someone opens the door.
7: DT 100	Someone opens the door.
8: Combi 100	Someone enters and closes the door.
SMT with our sense-labels	
9: DT 30	Someone opens the door behind her.
10: DT 100	Someone comes in.
11: Combi 100	Someone walks up to the door.
12: Script/AD	Someone decides to go and investigate.

Figure 3: Qualitative comparison of different video description methods. Discussion in Section 6.3.

7. Conclusions

In this work we presented a novel dataset of movies with aligned descriptions sourced from movie scripts and ADs (audio descriptions for the blind). We present first experiments on this dataset using state-of-the art visual features, combined with a recent movie description approach from [59]. We adapt the approach for this dataset to work without annotations, but rely on semantic parsing of labels. We show competitive performance on the TACoS Multi-Level dataset and promising results on our movie description data. We compare AD with previously used script data and find that AD tends to be more correct and relevant to the movie than script sentences. Beyond our first study on single sentences, the dataset opens new possibilities to understand stories and plots across multiple sentences in an open domain scenario on large scale.

Acknowledgements. Marcus Rohrbach was supported by a fellowship within the FITweltweit-Program of the German Academic Exchange Service (DAAD).

References

[1] British amazon. http://www.amazon.co.uk/, 2014. 3

[2] Castingwords transcription service. http://castingwords.com/, 2014. 2, 3

[3] Makemkv. http://www.makemkv.com/, 2014. 3

[4] Subtitle edit. http://www.nikse.dk/SubtitleEdit/, 2014. 4

[5] Xmedia recode. http://www.xmedia-recode.de/, 2014. 3

[6] Y. Artzi, N. FitzGerald, and L. S. Zettlemoyer. Semantic parsing with combinatory categorial grammars. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013. 3

[7] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley framenet project. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1998. 3

[8] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. Video in sentences out. In *Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI)*, 2012. 3

[9] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013. 3

[10] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 3

[11] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Weakly supervised action labeling in videos under ordering constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 3

[12] D. Chen and W. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011. 1, 4

[13] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv:1411.5654*, 2014. 3

[14] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008. 2, 4

[15] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 3

[16] D. Das, A. F. Martins, and N. A. Smith. An exact dual decomposition algorithm for shallow semantic parsing with constraints. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012. 3

[17] P. Das, C. Xu, R. Doell, and J. 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. 1, 3, 4

[18] L. Del Corro and R. Gemulla. Clausie: Clause-based open information extraction. In *Proceedings of the International World Wide Web Conference (WWW)*, 2013. 6

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1, 5

[20] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[21] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009. 2, 3

[22] M. Everingham, J. Sivic, and A. Zisserman. "hello! my name is... buffy" - automatic naming of characters in tv video. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2006. 3

[23] A. Fader, L. Zettlemoyer, and O. Etzioni. Open question answering over curated and extracted knowledge bases. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014. 3

[24] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. *arXiv:1411.4952*, 2014. 3

[25] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010. 3

[26] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998. 6

[27] L. Gagnon, C. Chapdelaine, D. Byrns, S. Foucher, M. Heritier, and V. Gupta. A computer-vision-assisted system for videodescription scripting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, 2010. 3

[28] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 3, 4

[29] A. Gupta, P. Srinivasan, J. Shi, and L. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 3

[30] P. Hanckmann, K. Schutte, and G. J. Burghouts. Automated textual descriptions for a wide range of video events with 48

human actions. In *Proceedings of the European Conference on Computer Vision Workshops (ECCV Workshops)*, 2012. 3

[31] P. Hodosh, A. Young, M. Lai, and J. 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. 1

[32] J. Hoffman, S. Guadarrama, E. Tzeng, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. LSDA: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2, 5

[33] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv:1412.2306*, 2014. 3

[34] M. U. G. Khan, L. Zhang, and Y. Gotoh. Human focused video description. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011. 3

[35] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. Extending verbnet with novel verb classes. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2006. 6

[36] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014. 3

[37] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539*, 2014. 3

[38] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007. 5

[39] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision (IJCV)*, 2002. 3

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. 1

[41] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 3

[42] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012. 3

[43] P. Kuznetsova, V. Ordonez, T. L. Berg, U. C. Hill, and Y. Choi. Treetalk: Composition and compression of trees for image descriptions. In *Transactions of the Association for Computational Linguistics (TACL)*, 2014. 3

[44] Lakritz and Salway. The semi-automatic generation of audio description from screenplays. Technical report, Dept. of Computing Technical Report, University of Surrey, 2006. 1, 3

[45] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2, 3, 4

[46] K. Lee, Y. Artzi, J. Dodge, and L. Zettlemoyer. Context-dependent semantic parsing for time expressions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014. 3

[47] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi. Composing simple image descriptions using web-scale N-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 2011. 3

[48] C. Liang, C. Xu, J. Cheng, and H. Lu. Tvparser: An automatic tv video parsing method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 1

[50] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv:1412.6632*, 2014. 3

[51] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2, 3, 4

[52] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. C. Berg, T. L. Berg, and H. D. III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2012. 3

[53] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems (NIPS)*, 2011. 1

[54] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, A. F. Smeaton, and G. Quéenot. Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012. 1

[55] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, 2004. 6

[56] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 3

[57] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1, 2013. 4

[58] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description

with variable level of detail. In *Proceedings of the German Confeence on Pattern Recognition (GCPR)*, September 2014. 1, 2, 3, 4, 6, 7, 8

[59] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 1, 2, 3, 4, 5, 6, 8

[60] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014. 5

[61] A. Salway. A corpus-based analysis of audio description. *Media for all: Subtitling for the deaf, audio description and sign language*, 2007. 1, 2, 3

[62] A. Salway, B. Lehane, and N. E. O'Connor. Associating characters with events in films. In *Proceedings of the ACM international conference on Image and video retrieval (CIVR)*, 2007. 2, 3

[63] K. K. Schuler, A. Korhonen, and S. W. Brown. Verbnet overview, extensions, mappings and applications. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2009. 6

[64] J. Sivic, M. Everingham, and A. Zisserman. "who are you?"-learning person specific classifiers from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 3

[65] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics (TACL)*. 3

[66] C. C. Tan, Y.-G. Jiang, and C.-W. Ngo. Towards textually describing complex video contents with audio-visual concept classifiers. In *Proceedings of the ACM international conference on Multimedia (MM)*, 2011. 3

[67] M. Tapaswi, M. Baeuml, and R. Stiefelhagen. "knock! knock! who is it?" probabilistic person identification in tv-series. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[68] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2014. 3

[69] A. Torabi, C. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv:1503.01070v1*, 2015. 2, 3, 4

[70] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015. 3

[71] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv:1411.4555*, 2014. 3

[72] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2, 5

[73] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 1

[74] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Video description generation incorporating spatio-temporal features and a soft-attention mechanism. *arXiv:1502.08029v3*, 2015. 3

[75] Z. Zhong and H. T. Ng. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, 2010. 6

[76] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *Advances in Neural Information Processing Systems (NIPS)*, 2014. 1, 2, 5