Grounding Action Descriptions in Videos

Michaela Regneri *, **Marcus Rohrbach** ◇, **Dominikus Wetzel** *,
Stefan Thater *, **Bernt Schiele** ◇ and **Manfred Pinkal** *

* Department of Computational Linguistics, Saarland University, Saarbrücken, Germany
(regneri|dwetzel|stth|pinkal)@coli.uni-saarland.de

◇ Max Planck Institute for Informatics, Saarbrücken, Germany
(rohrbach|schiele)@mpi-inf.mpg.de

Abstract

Recent work has shown that the integration of visual information into text-based models can substantially improve model predictions, but so far only visual information extracted from static images has been used. In this paper, we consider the problem of grounding *sentences describing actions* in visual information extracted from *videos*. We present a general purpose corpus that aligns high quality videos with multiple natural language descriptions of the actions portrayed in the videos, together with an annotation of how similar the action descriptions are to each other. Experimental results demonstrate that a text-based model of similarity between actions improves substantially when combined with visual information from videos depicting the described actions.

1 Introduction

The estimation of semantic similarity between words and phrases is a basic task in computational semantics. Vector-space models of meaning are one standard approach. Following the distributional hypothesis, frequencies of context words are recorded in vectors, and semantic similarity is computed as a proximity measure in the underlying vector space. Such distributional models are attractive because they are conceptually simple, easy to implement and relevant for various NLP tasks (Turney and Pantel, 2010). At the same time, they provide a substantially incomplete picture of word meaning, since they ignore the relation between language and extralinguistic information, which is constitutive for linguistic meaning. In the last few years, a growing amount of work has been devoted to the task of grounding meaning in visual information, in particular by extending the distributional approach to jointly cover texts and images (Feng and Lapata, 2010; Bruni et al., 2011). As a clear result, visual information improves the quality of distributional models. Bruni et al. (2011) show that visual information drawn from images is particularly relevant for concrete common nouns and adjectives.

A natural next step is to integrate visual information from *videos* into a semantic model of event and action verbs. Psychological studies have shown the connection between action semantics and videos (Glenberg, 2002; Howell et al., 2005), but to our knowledge, we are the first to provide a suitable data source and to implement such a model.

The contribution of this paper is three-fold:

- We present a *multimodal corpus* containing textual descriptions aligned with high-quality videos. Starting from the video corpus of Rohrbach et al. (2012b), which contains high-resolution video recordings of basic cooking tasks, we collected multiple textual descriptions of each video via Mechanical Turk. We also provide an accurate sentence-level alignment of the descriptions with their respective videos. We expect the corpus to be a valuable resource for computational semantics, and moreover helpful for a variety of purposes, including video understanding and generation of text from videos.

- We provide a *gold-standard dataset* for the evaluation of similarity models for action verbs and phrases. The dataset has been designed as analogous to the Usage Similarity dataset of

Transactions of the Association for Computational Linguistics, 1 (2013) 25–36. Action Editor: Hal Daumé III.
Submitted 10/2012; Published 3/2013.

Erk et al. (2009) and contains pairs of natural-language action descriptions plus their associated video segments. Each of the pairs is annotated with a similarity score based on several manual annotations.

- We report an experiment on *similarity modeling of action descriptions* based on the video corpus and the gold standard annotation, which demonstrates the impact of scene information from videos. Visual similarity models outperform text-based models; the performance of combined models approaches the upper bound indicated by inter-annotator agreement.

The paper is structured as follows: We first place ourselves in the landscape of related work (Sec. 2), then we introduce our corpus (Sec. 3). Sec. 4 reports our action similarity annotation experiment and Sec. 5 introduces the similarity measures we apply to the annotated data. We outline the results of our evaluation in Sec. 6, and conclude the paper with a summary and directions for future work (Sec. 7).

2 Related Work

A large multimodal resource combining language and visual information resulted from the ESP game (von Ahn and Dabbish, 2004). The dataset contains many images tagged with several one-word labels.

The Microsoft Video Description Corpus (Chen and Dolan, 2011, MSVD) is a resource providing textual descriptions of videos. It consists of multiple crowd-sourced textual descriptions of short video snippets. The MSVD corpus is much larger than our corpus, but most of the videos are of relatively low quality and therefore too challenging for state-of-the-art video processing to extract relevant information. The videos are typically short and summarized with a single sentence. Our corpus contains coherent textual descriptions of longer video sequences, where each sentence is associated with a timeframe.

Gupta et al. (2009) present another useful resource: their model learns the alignment of predicate-argument structures with videos and uses the result for action recognition in videos. However, the corpus contains no natural language texts.

The connection between natural language sentences and videos has so far been mostly explored by the computer vision community, where different methods for improving action recognition by exploiting linguistic data have been proposed (Gupta and Mooney, 2010; Motwani and Mooney, 2012; Cour et al., 2008; Tzoukermann et al., 2011; Rohrbach et al., 2012b, among others). Our resource is intended to be used for action recognition as well, but in this paper, we focus on the inverse effect of visual data on language processing.

Feng and Lapata (2010) were the first to enrich topic models for newspaper articles with visual information, by incorporating features from article illustrations. They achieve better results when incorporating the visual information, providing an enriched model that pairs a single text with a picture.

Bruni et al. (2011) used the ESP game data to create a visually grounded semantic model. Their results outperform purely text-based models using visual information from pictures for the task of modeling noun similarities. They model single words, and mostly visual features lead only to moderate improvements, which might be due to the mixed quality and random choice of the images. Dodge et al. (2012) recently investigated which words can actually be grounded in images at all, producing an automatic classifier for visual words.

An interesting in-depth study by Mathe et al. (2008) automatically learnt the semantics of motion verbs as abstract features from videos. The study captures 4 actions with 8-10 videos for each of the actions, and would need a perfect object recognition from a visual classifier to scale up.

Steyvers (2010) and later Silberer and Lapata (2012) present an alternative approach to incorporating visual information directly: they use so-called *feature norms*, which consist of human associations for many given words, as a proxy for general perceptual information. Because this model is trained and evaluated on those feature norms, it is not directly comparable to our approach.

The *Restaurant Game* by Orkin and Roy (2009) grounds written chat dialogues in actions carried out in a computer game. While this work is outstanding from the social learning perspective, the actions that ground the dialogues are clicks on a screen rather than real-world actions. The dataset has successfully been used to model determiner meaning (Reckman et al., 2011) in the context of the *Restaurant Game*,

but it is unclear how this approach could scale up to content words and other domains.

3 The TACOS Corpus

We build our corpus on top of the "MPII Cooking Composite Activities" video corpus (Rohrbach et al., 2012b, *MPII Composites*), which contains videos of different activities in the cooking domain, e.g., *preparing carrots* or *separating eggs*. We extend the existing corpus with multiple textual descriptions collected by crowd-sourcing via Amazon Mechanical Turk[1] *(MTurk)*. To facilitate the alignment of sentences describing activities with their proper video segments, we also obtained approximate timestamps, as described in Sec. 3.2.

MPII Composites comes with timed gold-standard annotation of low-level activities and participating objects (e.g. OPEN [HAND,DRAWER] or TAKE OUT [HAND,KNIFE,DRAWER]). By adding textual descriptions (e.g., *The person takes a knife from the drawer*) and aligning them on the sentence level with videos and low-level annotations, we provide a rich multimodal resource (cf. Fig. 2), the "Saarbrücken Corpus of Textually Annotated Cooking Scenes" (TACOS). In particular, the TACOS corpus provides:

- A collection of coherent *textual descriptions for video recordings* of activities of medium complexity, as as a basis for empirical discourse-related research, e.g., the selection and granularity of action descriptions in context
- A high-quality *alignment of sentences with video segments*, supporting the grounding of action descriptions in visual information
- *Collections of paraphrases* describing the same scene, which result as a by-product from the text-video alignment and can be useful for text generation from videos (among other things)
- The alignment of textual activity descriptions with *sequences of low-level activities*, which may be used to study the decomposition of action verbs into basic activity predicates

We expect that our corpus will encourage and enable future work on various topics in natural language and video processing. In this paper, we will make use of the second aspect only, demonstrating the usefulness of the corpus for the grounding task.

After a more detailed description of the basic video corpus and its annotation (Sec. 3.1) we describe the collection of textual descriptions with MTurk (Sec. 3.2), and finally show the assembly and some benchmarks of the final corpus (Sec. 3.3).

3.1 The video corpus

MPII Composites contains 212 high resolution video recordings of 1-23 minutes length (4.5 min. on average). 41 basic cooking tasks such as *cutting a cucumber* were recorded, each between 4 and 8 times. The selection of cooking tasks is based on those proposed at "Jamie's Home Cooking Skills".[2] The corpus is recorded in a kitchen environment with a total of 22 subjects. Each video depicts a single task executed by an individual subject.

The dataset contains expert annotations of low-level activity tags. Annotations are provided for segments containing a semantically meaningful cooking related movement pattern. The action must go beyond single body part movements (such as *move arm up*) and must have the goal of changing the state or location of an object. 60 different activity labels are used for annotation (e.g. PEEL, STIR, TRASH). Each low-level activity tag consists of an activity label (PEEL), a set of associated objects (CARROT, DRAWER,...), and the associated timeframe (starting and ending points of the activity). Associated objects are the participants of an activity, namely tools (e.g. KNIFE), patient (CARROT) and location (CUTTING-BOARD). We provide the coarse-grained role information for *patient*, *location* and *tool* in the corpus data, but we did not use this information in our experiments. The dataset contains a total of 8818 annotated segments, on average 42 per video.

3.2 Collecting textual video descriptions

We collected textual descriptions for a subset of the videos in *MPII Composites*, restricting collection to tasks that involve manipulation of cooking ingredients. We also excluded tasks with fewer than four

[1] mturk.com

[2] www.jamieshomecookingskills.com

video recordings in the corpus, leaving 26 tasks to be described. We randomly selected five videos from each task, except the three tasks for which only four videos are available. This resulted in a total of 127 videos. For each video, we collected 20 different textual descriptions, leading to 2540 annotation assignments. We published these assignments (HITs) on MTurk, using an adapted version[3] of the annotation tool Vatic (Vondrick et al., 2012).

In each assignment, the subject saw one video specified with the task title (e.g. *How to prepare an onion*), and then was asked to enter at least five and at most 15 complete English sentences to describe the events in the video. The annotation instructions contained example annotations from a kitchen task not contained in our actual dataset.

Annotators were encouraged to watch each video several times, skipping backward and forward as they wished. They were also asked to take notes while watching, and to sketch the annotation before entering it. Once familiarized with the video, subjects did the final annotation by watching the entire video from beginning to end, without the possibility of further non-sequential viewing. Subjects were asked to enter each sentence as soon as the action described by the sentence was completed. The video playback paused automatically at the beginning of the sentence input. We recorded pause onset for each sentence annotation as an approximate ending timestamp of the described action. The annotators resumed the video manually.

The tasks required a HIT approval rate of 75% and were open only to workers in the US, in order to increase the general language quality of the English annotations. Each task paid 1.20 USD. Before paying we randomly inspected the annotations and manually checked for quality. The total costs of collecting the annotations amounted to 3,353 USD. The data was obtained within a time frame of 3.5 weeks.

3.3 Putting the TACOS corpus together

Our corpus is a combination of the MTurk data and MPII Composites, created by filtering out inappropriate material and computing a high-quality alignment of sentences and video segments. The alignment is done by matching the approximate timestamps of the MTurk data to the accurate timestamps in MPII Composites.

[3] github.com/marcovzla/vatic/tree/bolt

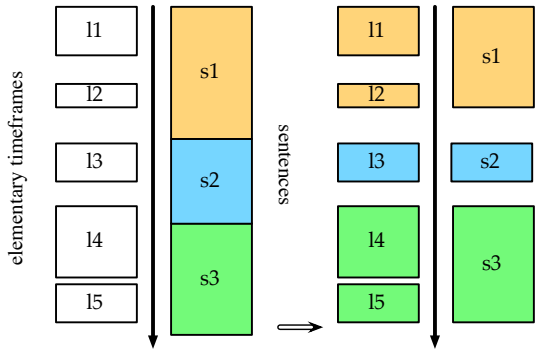

Figure 1: Aligning action descriptions with the video.

We discarded text instances if people did not time the sentences properly, taking the association of several (or even all) sentences to a single timestamp as an indicator. Whenever we found a timestamp associated with two or more sentences, we discarded the whole instance. Overall, we had to filter out 13% of the text instances, which left us with 2206 textual video descriptions.

For the alignment of sentence annotations and video segments, we assign a precise timeframe to each sentence in the following way: We take the timeframes given by the low-level annotation in MPII Composites as a gold standard micro-event segmentation of the video, because they mark all distinct frames that contain activities of interest. We call them *elementary frames*. The sequence of elementary frames is not necessarily continuous, because idle time is not annotated.

The MTurk sentences have end points that constitute a coarse-grained, noisy video segmentation, assuming that each sentence spans the time between the end of the previous sentence and its own ending point. We refine those noisy timeframes to gold frames as shown in Fig. 1: Each elementary frame (*l1-l5*) is mapped to a sentence (*s1-s3*) if its noisy timeframe covers at least half of the elementary frame. We define the final gold sentence frame then as the timespan between the starting point of the first and the ending point of the last elementary frame.

The alignment of descriptions with low-level activities results in a table as given in Fig. 3. Columns contain the textual descriptions of the videos; rows

Top 10 Verbs	cut, take, get, put, wash, place, rinse, remove, **pan*, peel
Top 10 Activities	move, take out, cut, wash, take apart, add, shake, screw, put in, peel

Figure 4: 10 most frequent verbs and low-level actions in the TACOS corpus. *pan* is probably often mis-tagged.

correspond to low-level actions, and each sentence is aligned with the last of its associated low-level actions. As a side effect, we also obtain multiple paraphrases for each sentence, by considering all sentences with the same associated time frame as equivalent realizations of the same action.

The corpus contains 17,334 action descriptions (tokens), realizing 11,796 different sentences (types). It consists of 146,771 words (tokens), 75,210 of which are content word instances (i.e. nouns, verbs and adjectives). The verb vocabulary comprises 28,292 verb tokens, realizing 435 lemmas. Since verbs occurring in the corpus typically describe actions, we can note that the linguistic variance for the 58 different low-level activities is quite large. Fig. 4 gives an impression of the action realizations in the corpus, listing the most frequent verbs from the textual data, and the most frequent low-level activities.

On average, each description covers 2.7 low-level activities, which indicates a clear difference in granularity. 38% of the descriptions correspond to exactly one low-level activity, about a quarter (23%) covers two of them; 16% have 5 or more low-level elements, 2% more than 10. The corpus shows how humans vary the granularity of their descriptions, measured in time or number of low-level activities, and it shows how they vary the linguistic realization of the same action. For example, Fig. 3 contains *dice* and *chop into small pieces* as alternative realizations of the low-level activity sequence SLICE - SCRATCH OFF - SLICE.

The descriptions are of varying length (9 words on average), reaching from two-word phrases to detailed descriptions of 65 words. Most sentences are short, consisting of a reference to the person in the video, a participant and an action verb (*The person rinses the carrot*, *He cuts off the two edges*). People often specified an instrument (*from the faucet*), or the resulting state of the action (*chop the carrots in small pieces*). Occasionally, we find more complex constructions (support verbs, coordinations).

As Fig. 3 indicates, the timestamp-based alignment is pretty accurate; occasional errors occur like *He starts chopping the carrot...* in NL Sequence 3. The data contains some typos and ungrammatical sentences (*He washed carrot*), but for our own experiments, the small number of such errors did not lead to any processing problems.

4 The Action Similarity Dataset

In this section, we present a gold standard dataset, as a basis for the evaluation of visually grounded models of action similarity. We call it the "Action Similarity Dataset" (ASim) in analogy to the Usage Similarity dataset (USim) of Erk et al. (2009) and Erk et al. (2012). Similarly to USim, ASim contains a collection of sentence pairs with numerical similarity scores assigned by human annotators. We asked the annotators to focus on the similarity of the activities described rather than on assessing semantic similarity in general. We use sentences from the TACOS corpus and record their timestamps. Thus each sentence comes with the video segment which it describes (these were not shown to the annotators).

4.1 Selecting action description pairs

Random selection of annotated sentences from the corpus would lead to a large majority of pairs which are completely dissimilar, or difficult to grade (e.g., *He opens the drawer – The person cuts off the ends of the carrot*). We constrained the selection process in two ways: First, we consider only sentences describing activities of manipulating an ingredient. The low-level annotation of the video corpus helps us identify candidate descriptions. We exclude rare and special activities, ending up with CUT, SLICE, CHOP, PEEL, TAKE APART, and WASH, which occur reasonably frequently, with a wide distribution over different scenarios. We restrict the candidate set to those sentences whose timespan includes one of these activities. This results in a conceptually more focussed repertoire of descriptions, and at the same time admits full linguistic variation (*wash an apple under the faucet – rinse an apple*, *slice the cucumber – cut the cucumber into slices*).

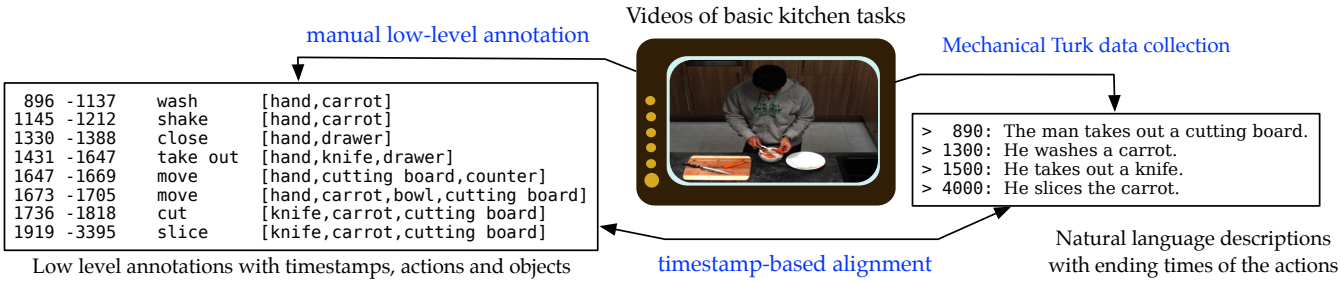

Figure 2: Corpus Overview

Sample frame	Start	End	Action	Participants	NL Sequence 1	NL Sequence 2	NL Sequence 3
	743	911	wash	hand, carrot	He washed carrot	The person rinses the carrot.	He rinses the carrot from the faucet.
	982	1090	cut	knife, carrot, cutting board	He cut off ends of carrots	The person cuts off the ends of the carrot.	He cuts off the two edges.
	1164	1257	open	hand, drawer			
	1679	1718	close	hand, drawer			He searches for something in the drawer, failed attempt, he throws away the edges in trash.
	1746	1799	trash	hand, carrot		The person searches for the trash can, then throws the ends of the carrot away.	
	1854	2011	wash	hand, carrot			He rinses the carrot again.
	2011	2045	shake	hand, carrot	He washed carrot	The person rinses the carrot again.	He starts chopping the carrot in small pieces.
	2083	2924	slice	knife, carrot, cutting board			
	2924	2959	scratch off	hand, carrot, knife, cutting board			
	3000	3696	slice	knife, carrot, cutting board	He diced carrots		He finished chopping the carrots in small pieces.

Figure 3: Excerpt from the corpus for a video on PREPARING A CARROT. Example frames, low-level annotation (*Action* and *Participants*) is shown along with three of the MTurk sequences (*NL Sequence 1-3*).

Second, we required the pairs to share some lexical material, either the head verb or the manipulated ingredient (or both).[4] More precisely, we composed the ASim dataset from three different subsets:

Different activity, same object: This subset contains pairs describing different types of actions carried out on the same type of object (e.g. *The man washes the carrot. – She dices the carrot.*). Its focus is on the central task of modeling the semantic relation between *actions* (rather than the objects involved in the activity), since the object head nouns in the descriptions are the same, and the respective video segments show the same type of object.

Same activity, same object: Description pairs of this subset will in many cases, but not always, agree in their head verbs. The dataset is useful for exploring the degree to which action descriptions are underspecified with respect to the precise manner of their practical realization. For example, peeling an onion will mostly be done in a rather uniform way, while *cut* applied to *carrot* can mean that the carrot is chopped up, or sliced, or cut in halves.

Same activity & verb, different object: Description pairs in this subset share head verb and low-level activity, but have different objects (e.g. *The man washes the carrot. – A girl washes an apple under the faucet.*). This dataset enables the exploration of the objects' meaning contribution to the complete action, established by the variation of equivalent actions that are done to different objects.

We assembled 900 action description pairs for annotation: 480 pairs share the object; 240 of which have different activities, and the other 240 pairs share the same activity. We included paraphrases describing the same video segment, but we excluded pairs of identical sentences. 420 additional pairs share their head verb, but have different objects.

4.2 Manual annotation

Three native speakers of English were asked to judge the similarity of the action pairs with respect to *how they are carried out*, rating each sentence pair with a score from 1 (not similar at all) to 5 (the same or nearly the same). They did not see the respective videos, but we noted the relevant kitchen task (i.e. which vegetable was prepared). We asked the annotators explicitly to ignore the actor of the action (e.g. whether it is a man or a woman) and score the similarities of the underlying actions rather than their verbalizations. Each subject rated all 900 pairs, which were shown to them in completely random order, with a different order for each subject.

Part of Gold Standard	**Sim**	σ	ρ
DIFF. ACTIVITY, SAME OBJECT	2.20	1.07	0.73
SAME ACTIVITY, SAME OBJECT	4.19	1.04	0.73
ALL WITH SAME OBJECT	3.20	1.44	0.84
SAME VERB, DIFF. OBJECT	3.34	0.69	0.43
COMPLETE DATASET	3.27	1.15	0.73

Figure 5: Average similarity ratings (*Sim*), their standard deviation (σ) and annotator agreement (ρ) for ASim.

We compute inter-annotator agreement (and the forthcoming evaluation scores) using Spearman's rank correlation coefficient (ρ), a non-parametric test which is widely used for similar evaluation tasks (Mitchell and Lapata, 2008; Bruni et al., 2011; Erk and McCarthy, 2009). Spearman's ρ evaluates how the samples are ranked relative to each other rather than the numerical distance between the rankings.

Fig. 5 shows the average similarity ratings in the different settings and the inter-annotator agreement. The average inter-rater agreement was $\rho = 0.73$ (averaged over pairwise rater agreements), with pairwise results of $\rho = 0.77$, 0.72, and 0.69, respectively, which are all highly significant at $p < 0.001$.

As expected, pairs with the same activity and object are rated very similar (4.19) on average, while the similarity of different activities on the same object is the lowest (2.2). For both subsets, inter-rater agreement is high ($\rho = 0.73$), and even higher for both SAME OBJECT subsets together (0.84).

Pairs with identical head verbs and different objects have a small standard deviation, at 0.69. The inter-annotator agreement on this set is much lower than for pairs from the SAME OBJECT set. This indicates that similarity assessment for different variants of the same activity is a hard task even for humans.

[4] We refer to the latter with the term *object*; we don't require the ingredient term to be the actual grammatical object in the action descriptions, we rather use "object" in its semantic role sense as the entity affected by an action.

5 Models of Action Similarity

In the following, we demonstrate that visual information contained in videos of the kind provided by the TACOS corpus (Sec. 3) substantially contributes to the semantic modeling of action-denoting expressions. In Sec. 6, we evaluate several methods for predicting action similarity on the task provided by the ASim dataset. In this section, we describe the models considered in the evaluation. We use two different models based on visual information, and in addition two text based models. We will also explore the effect of combining linguistic and visual information and investigate which mode is most suitable for which kinds of similarity.

5.1 Text-based models

We use two different models of textual similarity to predict action similarity: a simple word-overlap measure (Jaccard coefficient) and a state-of-the-art model based on "contextualized" vector representations of word meaning (Thater et al., 2011).

Jaccard coefficient. The Jaccard coefficient gives the ratio between the number of (distinct) words common to two input sentences and the total number of (distinct) words in the two sentences. Such simple surface-oriented measures of textual similarity are often used as baselines in related tasks such as recognizing textual entailment (Dagan et al., 2005) and are known to deliver relatively strong results.

Vector model. We use the vector model of Thater et al. (2011), which "contextualizes" vector representations for individual words based on the particular sentence context in which the target word occurs. The basic intuition behind this approach is that the words in the syntactic context of the target word in a given input sentence can be used to refine or disambiguate its vector. Intuitively, this allows us to discriminate between different actions that a verb can refer to, based on the different objects of the action.

We first experimented with a version of this vector model which predicts action similarity scores of two input sentences by computing the cosine similarity of the contextualized vectors of the verbs in the two sentences only. We achieved better performance with a variant of this model which computes vectors for the two sentences by summing over the contextualized vectors of all constituent content words.

In the experiments reported below, we only use the second variant. We use the same experimental setup as Thater et al. (2011), as well as the parameter settings that are reported to work best in that paper.

5.2 Video-based models

We distinguish two approaches to compute the similarity between two video segments. In the first, unsupervised approach we extract a video descriptor and compute similarities between these raw features (Wang et al., 2011). The second approach builds upon the first by additionally learning higher level attribute classifiers (Rohrbach et al., 2012b) on a held out training set. The similarity between two segments is then computed between the classifier responses. In the following we detail both approaches:

Raw visual features. We use the state-of-the-art video descriptor *Dense Trajectories* (Wang et al., 2011) which extracts visual video features, namely histograms of oriented gradients, flow, and motion boundary histograms, around densely sampled and tracked points.

This approach is especially suited for this data as it ignores non-moving parts in the video: we are interested in activities and manipulation of objects, and this type of feature implicitly uses only information in relevant image locations. For our setting this feature representation has been shown to be superior to human pose-based approaches (Rohrbach et al., 2012a). Using a bag-of-words representation we encode the features using a 16,000 dimensional codebook. Features and codebook are provided with the publicly available video dataset.

We compute the similarity between two encoded features by computing the intersection of the two (normalized) histograms.

Visual classifiers. Visual raw features tend to have several dimensions in the feature space which provide unreliable, noisy values and thus degrade the strength of the similarity measure. Intermediate level attribute classifiers can learn which feature dimensions are distinctive and thus significantly improve performance over raw features. Rohrbach et al. (2012b) showed that using such an attribute classifier representation can significantly improve per-

	Model	same object	same verb	overall
text	jaccard	0.28	0.25	0.25
	textual vectors	0.30	0.25	0.27
	text combined	0.39	**0.35**	0.36
video	visual raw vectors	0.53	-0.08	0.35
	visual classifier	0.60	0.03	0.44
	video combined	0.61	-0.04	0.44
mix	all unsupervised	0.58	0.32	0.48
	all combined	**0.67**	0.28	**0.55**
upper bound		0.84	0.43	0.73

Figure 6: Evaluation results in Spearman's ρ. All values > 0.11 are significant at $p < 0.001$.

formance for composite activity recognition. The relevant attributes are all activities and objects annotated in the video data (cf. Section 3.1). For the experiments reported below we use the same setup as Rohrbach et al. (2012b) and use all videos in MPII Composites and MPII Cooking (Rohrbach et al., 2012a), excluding the 127 videos used during evaluation. The real-valued SVM-classifier output provides a confidence how likely a certain attribute appeared in a given video segment. This results in a 218-dimensional vector of classifier outputs for each video segment. To compute the similarity between two vectors we compute the cosine between them.

6 Evaluation

We evaluate the different similarity models introduced in Sec. 5 by calculating their correlation with the gold-standard similarity annotations of ASim (cf. Sec. 4). For all correlations, we use Spearman's ρ as a measure. We consider the two textual measures (JACCARD and TEXTUAL VECTORS) and their combination, as well as the two visual models (VISUAL RAW VECTORS and VISUAL CLASSIFIER) and their combination. We also combined textual and visual features, in two variants: The first includes all models (ALL COMBINED), the second only the unsupervised components, omitting the visual classifier (ALL UNSUPERVISED). To combine multiple similarity measures, we simply average their normalized scores (using z-scores).

Figure 6 shows the scores for all of these measures on the complete ASim dataset (OVERALL), along with the two subparts, where description pairs share either the object (SAME OBJECT) or the head verb (SAME VERB). In addition to the model results, the table also shows the average human inter-annotator agreement as UPPER BOUND.

On the complete set, both visual and textual measures have a highly significant correlation with the gold standard, whereas the combination of both clearly leads to the best performance (0.55). The results on the SAME OBJECT and SAME VERB subsets shed light on the division of labor between the two information sources. While the textual measures show a comparable performance over the two subsets, there is a dramatic difference in the contribution of visual information: On the SAME OBJECT set, the visual models clearly outperform the textual ones, whereas the visual information has no positive effect on the SAME VERB set. This is clear evidence that the visual model does not capture the similarity of the participating objects but rather genuine action similarity, which the visual features (Wang et al., 2011) we employ were designed for. A direction for future work is to learn dedicated visual object detectors to recognize and capture similarities between objects more precisely.

The numbers shown in Figure 7 support this hypothesis, showing the two groups in the SAME OBJECT class: For sentence pairs that share the same activity, the textual models seem to be much more suitable than the visual ones. In general, visual models perform better on actions with different activity types, textual models on closely related activities.

	Model (Same Object)	*same action*	*diff. action*
TEXT	JACCARD	0.44	0.14
	TEXT VECTORS	0.42	0.05
	TEXT COMBINED	**0.52**	0.14
VIDEO	VIS. RAW VECTORS	0.21	0.23
	VIS. CLASSIFIER	0.21	**0.45**
	VIDEO COMBINED	0.26	0.38
MIX	ALL UNSUPERVISED	0.49	0.24
	ALL COMBINED	0.48	0.41
UPPER BOUND		0.73	0.73

Figure 7: Results for sentences with the same object, with either the same or different low-level activity.

Overall, the supervised classifier contributes a good part to the final results. However, the supervision is not strictly necessary to arrive at a significant correlation; the raw visual features alone are sufficient for the main performance gain seen with the integration of visual information.

7 Conclusion

We presented the TACOS corpus, which provides coherent textual descriptions for high-quality video recordings, plus accurate alignments of text and video on the sentence level. We expect the corpus to be beneficial for a variety of research activities in natural-language and visual processing.

In this paper, we focused on the task of grounding the meaning of action verbs and phrases. We designed the ASim dataset as a gold standard and evaluated several text- and video-based semantic similarity models on the dataset, both individually and in different combinations.

We are the first to provide semantic models for action-describing expressions, which are based on information extracted from videos. Our experimental results show that these models are of considerable quality, and that predictions based on a combination of visual and textual information even approach the upper bound given by the agreement of human annotators.

In this work we used existing similarity models that had been developed for different applications. We applied these models without any special training or optimization for the current task, and we combined them in the most straightforward way. There is room for improvement by tuning the models to the task, or by using more sophisticated approaches to combine modality-specific information (Silberer and Lapata, 2012).

We built our work on an existing corpus of high-quality video material, which is restricted to the cooking domain. As a consequence, the corpus covers only a limited inventory of activity types and action verbs. Note, however, that our models are fully unsupervised (except the Visual Classifier model), and thus can be applied without modification to arbitrary domains and action verbs, given that they are about observable activities. Also, corpora containing information comparable to the TACOS corpus but with wider coverage (and perhaps a bit noisier) can be obtained with a moderate amount of effort. One needs videos of reasonable quality and some sort of alignment with action descriptions. In some cases such alignments even come for free, e.g. via subtitles, or descriptions of short video clips that depict just a single action.

For future work, we will further investigate the compositionality of action-describing phrases. We also want to leverage the multimodal information provided by the TACOS corpus for the improvement of high-level video understanding, as well as for generation of natural-language text from videos.

The TACOS corpus and all other data described in this paper (videos, low-level annotation, aligned textual descriptions, the ASim-Dataset and visual features) are publicly available. [5]

Acknowledgements

We'd like to thank Asad Sayeed, Alexis Palmer and Prashant Rao for their help with the annotations. We're indebted to Carl Vondrick and Marco Antonio Valenzuela Escrcega for their extensive support with the video annotation tool. Further we thank Alexis Palmer and in particular three anonymous reviewers for their helpful comments on this paper. – This work was funded by the Cluster of Excellence "Multimodal Computing and Interaction" of the German Excellence Initiative and the DFG project SCHI989/2-2.

[5] http://www.coli.uni-saarland.de/projects/smile/page.php?id=tacos

References

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of SIGCHI 2004*.

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. In *Proceedings of GEMS 2011*.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of ACL 2011*.

Timothee Cour, Chris Jordan, Eleni Miltsakaki, and Ben Taskar. 2008. Movie/script: Alignment and parsing of video and text transcription. In *Computer Vision – ECCV 2008*, volume 5305 of *Lecture Notes in Computer Science*, pages 158–171. Springer Berlin Heidelberg.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of MLCW 2005*.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C. Berg, and Tamara L. Berg. 2012. Detecting visual text. In *HLT-NAACL*, pages 762–772.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of EMNLP 2009*.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of ACL/AFNLP 2009*.

Katrin Erk, Diana McCarthy, and Nick Gaylord. 2012. Measuring word meaning in context. *CL*.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Proceedings of HLT-NAACL 2010*.

A. M. Glenberg. 2002. Grounding language in action. *Psychonomic Bulletin & Review*.

Sonal Gupta and Raymond J. Mooney. 2010. Using closed captions as supervision for video activity recognition. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2010)*, pages 1083–1088, Atlanta, GA, July.

Abhinav Gupta, Praveen Srinivasan, Jianbo Shi, and Larry S. Davis. 2009. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *Proceedings of CVPR 2009*.

Steve R. Howell, Damian Jankowicz, and Suzanna Becker. 2005. A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *JML*.

S. Mathe, A. Fazly, S. Dickinson, and S. Stevenson. 2008. Learning the abstract motion semantics of verbs from captioned videos. pages 1–8.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL 2008*.

Tanvi S. Motwani and Raymond J. Mooney. 2012. Improving video activity recognition using object recognition and text mining. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI-2012)*, pages 600–605, August.

Jeff Orkin and Deb Roy. 2009. Automatic learning and generation of social behavior from collective human gameplay. In *Proceedings of AAMAS 2009*.

Hilke Reckman, Jeff Orkin, and Deb Roy. 2011. Extracting aspects of determiner meaning from dialogue in a virtual world environment. In *Proceedings of CCS 2011*, IWCS '11.

Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012a. A database for fine grained activity detection of cooking activities. In *Proceedings of CVPR 2012*.

Marcus Rohrbach, Michaela Regneri, Micha Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012b. Script data for attribute-based recognition of composite activities. In *Proceedings of ECCV 2012*.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of EMNLP-CoNLL 2012*.

Mark Steyvers. 2010. Combining feature norms and text data with topic models. *Acta Psychologica*, 133(3):234 – 243. ¡ce:title¿Formal modeling of semantic concepts¡/ce:title¿.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of IJCNLP 2011*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning. vector space models for semantics. *JAIR*.

E. Tzoukermann, J. Neumann, J. Kosecka, C. Fermuller, I. Perera, F. Ferraro, B. Sapp, R. Chaudhry, and G. Singh. 2011. Language models for semantic extraction and filtering in video action recognition. In *AAAI Workshop on Language-Action Tools for Cognitive Artificial Agents*.

Carl Vondrick, Donald Patterson, and Deva Ramanan. 2012. Efficiently scaling up crowdsourced video annotation. *IJCV*.

Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action Recognition by Dense Trajectories. In *Proceedings of CVPR 2011*.