Translating Video Content to Natural Language Descriptions

Marcus Rohrbach[1] Wei Qiu[1,2] Ivan Titov[3]
Stefan Thater[2] Manfred Pinkal[2] Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Department of Computational Linguistics, Saarland University, Saarbrücken, Germany
[3]University of Amsterdam, Amsterdam, the Netherlands

Abstract

Humans use rich natural language to describe and communicate visual perceptions. In order to provide natural language descriptions for visual content, this paper combines two important ingredients. First, we generate a rich semantic representation of the visual content including e.g. object and activity labels. To predict the semantic representation we learn a CRF to model the relationships between different components of the visual input. And second, we propose to formulate the generation of natural language as a machine translation problem using the semantic representation as source language and the generated sentences as target language. For this we exploit the power of a parallel corpus of videos and textual descriptions and adapt statistical machine translation to translate between our two languages. We evaluate our video descriptions on the TACoS dataset [23], which contains video snippets aligned with sentence descriptions. Using automatic evaluation and human judgments we show significant improvements over several baseline approaches, motivated by prior work. Our translation approach also shows improvements over related work on an image description task.

1. Introduction

Computer vision has advanced to detect people, classify their actions, or to distinguish between a large number of objects and specify their attributes. The output is often a semantic representation encoding activities and objects categories. While such representations can be well processed by automated systems, the natural way to communicate this information with humans is natural language. Thus, this work addresses the problem of generating textual descriptions for videos. This task has a wide range of applications in the domain of human-computer/robot interaction, generating summary descriptions of (web-)videos, and automating movie descriptions for visually impaired people. Furthermore, being able to convert visual content to language is an important step in understanding the relationship between visual and linguistic information which are the richest interaction modalities available to humans.

Generating natural language descriptions of visual content is an intriguing task but requires combining the fundamental research problems of visual recognition and natural language generation (NLG). While for descriptions of images, recent approaches have proposed to statistically model the conversion from images to text [5, 15, 16, 18], most approaches for video description use rules and templates to generated video descriptions [14, 9, 2, 10, 11, 26, 3, 8]. Although these works have started exploring the domain of describing visual content, important research questions remain: (1) How to best approach the conversion from visual information to linguistic expressions? (2) Which part of the visual information is verbalized by humans and what is verbalized even though it is not directly present in the visual information? (3) What is a good semantic representation (SR) of visual content and what is the limit of such a representation given perfect visual recognition?

Answering these questions is clearly beyond the scope of a single paper but we aim to address them jointly here. To address the first question we suggest to learn the conversion from video to language descriptions in a two-step approach. In the first step we learn an intermediate SR using a probabilistic model, following ideas used to generate image descriptions [5, 15]. Then, given the SR, we propose to phrase the problem of NLG as a *translation problem*, that is translating the SRs to natural language descriptions. In contrast to related work on video description, we learn both the SR as well as the language descriptions from an aligned parallel corpus containing videos, semantic annotations and textual descriptions. We compare our approach to related work and baselines using no intermediate SR and/or language model.

Second, we do not want to define manually the right level of verbalization. Instead we learn from a parallel training corpus the most relevant information to verbalize and how to verbalize it. For this we employ the methods from statis-

To appear in the Proc. of the *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, December 2013.

tical machine translation [12]. (a) We learn the correct ordering of words and phrases, referred to as surface realization in NLG. (b) We can learn which SR should be realized in language. When describing a video, using "cooking" as a running example, the visually recognized object PEELER would normally not be mentioned when describing that *a person is peeling a carrot* but can still contribute to the verbalization of *peeling*. (c) We learn the proper correspondence between semantic concepts and verbalization, i.e. we do not have to define how semantic concepts are realized. E.g. the concepts ⟨MOVE, PAN, COUNTER, HOB⟩ could be realized as *He puts the frying pan on the stove* rather than being limited to *He moves the pan from the counter to the hob* when just adding function words.

Although NLG can be defined purely by rules and templates which might provide a more robust approach for limited domains, we believe that learning these parameters from data is a much more attractive approach. For any sufficiently rich domain, the required complexity of rules and templates is likely to make the rule engineering task either infeasible or prohibitively expensive. This has been shown for language translation, where statistical machine translation has generally replaced rule-based approaches [12].

To address the third question of the right visual input we compare three different visual representations, namely a raw video descriptor [27], an attribute based representation [24], and our CRF model. To understand the limits of our SR we also run the translation on ground truth annotations.

The main contributions are as follows. First, we phrase video description as a translation problem from video content to natural language descriptions (Sec. 3). As intermediate step we employ a SR of the video content. Second, we evaluate our approach on the TACoS [23] video-description dataset (Sec. 5.1). Using automatic as well as human evaluation, the proposed approach outperforms several baseline methods inspired by previous work. The SR, when using ground truth annotations, allows generating language that is close to human performance. Additionally our approach also compares favorably to [5] on the Pascal-sentence dataset for an image description task (Sec. 6). Third, annotations as well as intermediate outputs and final descriptions to allow for comparisons to our work or building on our SR are released on our website.

2. Related work

Statistical machine translation (SMT). Machine translation aims to translate from one natural language to another. SMT formulates this problem as data-driven machine learning problem. SMT is a mature field with existing approaches achieving respectable results across many language pairs, see e.g. [17] for a review and tutorial. Based on sentence-aligned corpora of source and target language a translation model is estimated. Additionally, a model for the target language is learnt to generate a fluent and grammatical output. The open source Moses [13] toolkit optimizes this pipeline on a training set (see Sec 3.2). [4] propose to approach object recognition in analogy to machine translation by learning a lexicon from images segments to associated keywords from images with keywords. Rather than translating to words or labels we translate from a SR to full descriptions.

NLG from images and video. Generating descriptions of visual content can be roughly divided in four different directions according to: (1) generating descriptions for (test) images or videos which already contain some associated text, (2) generating descriptions by using manually defined rules or templates, (3) retrieving existing descriptions from similar visual content, or (4) learning a language model from a training corpus to generate descriptions.

(1) Assuming the availability of text associated with the image at test time one can effectively use summarization techniques [1, 7] which benefit from visual content. This setting is different from ours as we want to generate descriptions at test time from visual content only.

(2) Given a SR extracted from visual content it is possible to generate language using manually defined rules and templates. To describe images, [15] extracts objects and their attributes as well as their spatial prepositions from images. These entities are modeled in a Conditional Random Field (CRF). From the CRF predictions they generate descriptions based on simple templates (or n-gram model, which falls into (4)). We also use a CRF to predict an intermediate SR but we show that our translation system generates descriptions more similar to human descriptions. For videos, [14] builds a concept hierarchy of actions which is manually defined and associated with different body, hand and head movements. Our setting is visually more challenging and varied making manual definitions challenging. [26] learns audio-visual concepts and generates a video description for three different activities using rules to combine action, scene, and audio concepts with glue words. [9] extracts an AND-OR graph from sports videos to model causal relationships. Using the graph, sentences can then be constructed using simple templates. [10, 2] extract actions, body-pose, objects and their tracks on the DARPA Mind's eye corpus which depict 48 different verbs. Using a set of templates they generate text for their SR. Similarly, [11] uses templates to describe videos on the TREC Video summarization task. [3] follows a different route and uses a topic model to jointly model textual and visual words and a tripartite graph based on object/concept detectors. Text generation is done with manually defined templates. The recent work of [8] predicts multiple subject-verb-object triples for a video snippet. These are reweighed according to the confidence along a classifier hierarchy and a language model. The best suited triple is used to generate multiple sentences

based on a template which are again scored against a n-gram language model. Similarly, our translation approach weights resulting sentences according to a language model. However, using templates limits the natural flexibility of language, as noted by [16].

(3) The third group of approaches reduces the generation process to retrieving sentences from a training corpus based on locally [20] or globally [5] similar images. [5] learns an intermediate SR of object, action, and scenes using a Markov Random Field. We compare to their retrieval results by applying our translation approach to their SR.

(4) The fourth line of work, which also includes this work, goes beyond retrieving existing descriptions by learning a language model to compose novel descriptions. [15] learns an n-gram language model to predict function words for their SR. One of our baselines is based on this idea (Sec. 4.2). Two recent approaches use an aligned corpus of images and descriptions as a basis for generating novel descriptions for images using state-of-the art language generation techniques. [16] retrieves candidate phrases from an image-caption database based on object, scene, and region recognition. Using an Integer Linear Programming formulation for content planning and surface realization they construct the most relevant and linguistically coherent descriptions. While they hand craft constraints to translate from the image, we learn a statistical translation model. [18] uses a corpus of 700,000 Flickr images with associated descriptions. Based on the visual recognition system of [15] they learn to predict sets of nouns and their order and add necessary prepositions, predicates, and determiners to form syntactically well-formed phrases. In contrast to their Tree-adjoining-grammar (TAG)-like natural language generation approach we use flat, cooccurrence based techniques from SMT.

3. Video description as a translation problem

In this section we present a two-step approach which describes video content with natural language. We assume that for training we have a parallel corpus which contains a set of video snippets and sentences. Video snippets represented by the video descriptor x_i are aligned with a sentence z_i, i.e we have (x_i, z_i). In case there is an extra description for the same video snippet we treat it as an independent alignment (x_k, z_k) with $x_k = x_i$. Additionally we introduce an intermediate level semantic representation (SR) in form of labels y_i.

At test time we first predict the SR y^* for a new video (descriptor) x^* and then generate a sentence z^* from y^*.

In the following we present our proposed approach using human-activity videos in a kitchen scenario based on the TACoS corpus, where people are recorded preparing different kinds of ingredients. However, we show in section 6 that this can also be applied to translate images to descriptions.

We build the SR based on the annotations provided with TACoS. It distinguishes *activities*, *tools*, *ingredients/objects*, *(source) location/container*, and *(target) location/container*. This directly converts to our SR y in the form of $\langle$ACTIVITY, TOOL, OBJECT, SOURCE, TARGET$\rangle$. As a tool, object or location can be missing, we represent this with an additional NULL label for the respective node.

The SR annotations in TACoS have sometimes a finer granularity than the sentences, i.e. $(y_i^1, \ldots, y_i^{l_i}, \ldots, y_i^{L_i}, z_i)$ where L_i is the number of SR annotations for sentence z_i. For learning the SR we just extract the corresponding video snippet for the SR, i.e. $(x_i^{l_i}, y_i^{l_i})$. As there are no annotations at test time, there exist no alignment problem when predicting y^*. In Sec. 3.2 we discuss several variants how to handle the different granularity of the SR and the sentences.

3.1. Predicting a SR from visual content

In the first step we extract a SR from the visual content. Typically different visual information is highly correlated with each other. E.g. for cooking activities, the activity *slice* is more correlated with the object *carrot* and tool *knife* than with *milk* and *spoon*. We model these relationships with a CRF where the visual entities are modeled as nodes n_j observing the video descriptors x as unaries. In our case we use a fully connected graph and learn linear pairwise (p) and unary (u) weights, using the following standard energy formulation for the structured model:

$$E(n_1, ..., n_N; x_i) = \sum_{j=1}^{N} E^u(n_j; x_i) + \sum_{j \sim k} E^p(n_j, n_k) \quad (1)$$

with $E^u(n_j; x_i) = \langle w_j^u, x_i \rangle$, where w_j^u is a vector of the size of the video representation x_i and $E^p(n_j, n_k) = w_{j,k}^p$.

We learn the model with training videos $x_i^{l_i}$ and SR labels $y_i^{l_i} = \langle n_1, n_2, \ldots, n_N \rangle$ using loopy belief propagation (LBP) implemented in [25]. We model the five SR categories as nodes ($N = 5$), the different states are based on the provided labels of TACoS (for samples see Table 1).

3.2. Translating from a SR to a description

Converting a SR to descriptions ($SR \rightarrow D$) has many similarities to translating from a source to a target language ($L_S \rightarrow L_T$) in machine translation.

1. For $SR \rightarrow D$ we have to find the verbalization of a label n_i, e.g. HOB$\rightarrow$*stove*, similar to translating a word from L_S to L_T.
2. For $SR \rightarrow D$ we have to determine the ordering of the concepts of the SR in D, which is similar to finding the alignment between two languages.
3. In a natural description of video not necessarily all semantic concepts are verbalized, e.g. KNIFE might not be verbalized when we describe *He cuts a carrot*.

There exists a similar problem for $L_S \rightarrow L_T$, where certain words in L_S, e.g. articles, are either not represented in L_T or multiple ones are combined to one.

4. The inverse problem also exist, e.g. adding function words to the SR to form a full sentence, e.g. CUT, CARROT→*He cuts the carrots*.
5. When translating $L_S \rightarrow L_T$ a language model of L_T is used to achieve a grammatically correct and fluent target sentence, same for D in $SR \rightarrow D$.

Motivated by these similarities, we propose to use established techniques for statistical machine translation (SMT) to learn a translation model from a parallel corpus of SRs and descriptions. We use the widely used Moses toolkit [13] to learn a translation model and in the following shortly layout the steps taken.

First we have to build a parallel corpus. In TACoS we encounter the problem that one sentence can be aligned to multiple SRs, i.e. $(y_i^1, \ldots, y_i^{L_i}, z_i)$. However, the input for SMT is aligned single sentences. We propose the following variants to handle the different granularity levels of SRs and descriptions:

All. For all SR annotations aligned to a sentence we create a separate training example, i.e. $(y_i^1, z_i), \ldots, (y_i^{L_i}, z_i)$.

Last. We only use the last SR as this frequently is the most important one, which is an artifact of the recording of the TACoS dataset, where users indicate only the ending time of their description in the video, i.e. $(y_i^{L_i}, z_i)$.

Semantic overlap. We estimate the highest word overlap between the sentence and the string of the SR: $\frac{|y_i \cap Lemma(z_i)|}{|y_i|}$, where $Lemma$ refers to lemmatizing, i.e. reducing to base forms, e.g. *took* to *take*, *knives* to *knife*.

Sentence level prediction. While we do not have an annotated SR for the sentence level, we can predict one SR for each sentence, i.e. y_i^* for z_i. While this will be noisier during training time it also reflects better the situation at test time where we also have predictions at sentence level as annotations are unavailable.

SMT expects an input string as source language expression. We convert our SR ⟨ACTIVITY, TOOL, OBJECT, SOURCE, TARGET⟩ in a string by concatenating the concepts using spaces as delimiters to indicate word boundaries, i.e. *activity tool object source target*, where NULL states are converted to empty strings.

Next we use giza++ [19] to learn a word-level alignment, i.e. in our case concepts-word alignment. This is the basis for the phrase-based translation model learned by Moses, which does not look at single words but tries to find multiple words (phrases) which correspond to each other and the corresponding probability. Additionally a reordering model is learned based on the training data alignment statistics [13].

To estimate the fluency of the descriptions we use IRSTLM [6] which is based on n-gram statistics of TACoS.

The final step involves optimizing a linear model between the probabilities from the language model, phrase tables, and reordering model, as well as word, phrase, and rule counts [13]. For this we use 10% of the training data as a validation set. In the optimization, the BLEU@4 score is used to compute the difference between predicted and provided reference descriptions.

For testing, we apply our translation model to the SR y^* predicted by the CRF for a given input video x^*. This decoding results in the description z^*.

4. Baselines

In the following we describe baselines which are motivated by related work and which fully or partially replace our translation approach. For all these variants we use the same setup as for our translation system, see Sec. 5.

4.1. Sentence retrieval

An alternative to generating novel descriptions is to retrieve the most likely sentence from a training corpus [5]. Given a test video x^* we search for the closest training video x_i and output the sentence $z^* = z_i$ (in case there are several we choose the first). To measure the distance between videos we distinguish three variants:

Raw video features. We use the L2-distance between BoW quantized dense trajectory representations [27]. This requires no intermediate level annotation of the data.

Attribute classifiers. While the raw video features tend to be too noisy to compute reliable distances, it has been shown that using the vector of attribute classifier outputs instead of the raw video features improves similarity estimates between videos [23].

CRF predictions. We use the estimated configuration to find the most similar SR in training data using hamming distance. This is the most similar variant to [5] which also use a probabilistic graphical model to represent the intermediate representation.

4.2. NLG with N-grams

While we keep the same SR we replace the SMT pipeline by learning a n-gram language model on the training set of the descriptions. It predicts function words between the content words from the SR-labels, similar to one of the approaches discussed in [15]. For the n-gram model to work we have do manually define the following steps: 1) the order of the content words has to be identical to the ones in the target sentence; 2) for our corpus, tool and location is frequently not verbalized, thus our model could only find a sensible string when we reduced it to ACTIVITY and OBJECT; 3) to further improve performance we only use the verb in the activity, e.g. CUT DICE→*cut*, and the root word for noun phrases, e.g. PLASTIC BAG→*bag*.

Node	states	Example states	SVM	LBP
ACTIVITY	66	cut dice, pour, stir, peel	58.7	**60.8**
TOOL	43	fork, hand, knife, towel	81.6	**82.0**
OBJECT	109	bread, carrot, salt, pot	32.5	**33.2**
SOURCE	51	fridge, plate, cup, pot	**76.0**	71.0
TARGET	35	counter, plate, hook	**74.9**	70.3
All nodes correct			18.7	**21.6**

Table 1: CRF nodes of our SR. SVM vs. LBP inference: Node accuracy in % over all test sentences.

5. Evaluation: Translating video to text

We evaluate our video description approach on the TACoS dataset [23] which contains videos with aligned SR annotations and sentence descriptions. We use an updated version of TACoS with a total of 18,227 video/sentence pairs on 7,206 unique time intervals. There are 5609 intermediate level annotations, which form our semantic representation (SR) and consists of the tuple ⟨ACTIVITY, TOOL, OBJECT, SOURCE, TARGET⟩.

To describe the video we use the dense trajectory features [27] which extract trajectory information, HOG, HOF, and MBH to form a descriptor which has shown state-of-the art performance on many activity recognition datasets, including the one we use [24]. As our final video descriptor and input for the CRF we use our attribute-classifier representation from [24] which includes both actions and objects on top of the dense trajectory features.

We test our approach on a subset of 490 video snippet / sentence pairs. There is no overlap in the human subjects to the training data. The CRF and Moses are trained on the remaining TACoS corpus, using 10% as a validation set for parameter estimation. The attribute classifiers are trained on the remaining videos of the MPII Cooking Composite Activity dataset [24], which is a superset of TACoS. We preprocess all text data by substituting gender specific identifiers with "the person" as we do not distinguish male and female with our visual system.

We evaluate automatically using the BLEU score which is widely used to evaluate machine translations against reference translations [21]. It computes the geometric mean of n-gram word overlaps for n=1,...,N, weighted by a brevity penalty. While BLEU@4 (N=4) has shown to provide the best correlation with human judgments, we also provide BLEU@1 to comply with results reported in [16, 15]. For manual evaluation, we follow [16] and ask 10 human subjects to rate grammatical correctness (independent of video content), correctness, and relevance (latter two independent of grammatical correctness). Correctness rates if the sentences are correct with respect to the video, and relevance judges if the sentence describes the most salient activity and objects. We additionally ask the judges to separately rate the correctness of the activity, objects (tools and ingredients), and locations described. We ask to rate on a scale from 1 to 5 with 5: perfect, 4: almost perfect, 3:70-80% good, 2: 50-70% good, 1: totally bad [16].

We present the human judges with different sentences of our systems in a random order for each video and ask explicitly to make consistent relative judgment between different sentences. If needed, continuous scores (e.g. 3.5) can be assigned. We limit our human evaluation to the best and most discriminant approaches.

In Table 1 we evaluate our visual recognition system, reporting accuracy over all test sentences for the different nodes.

5.1. Results: Translating video to text

Results of the various baselines and from our translation system are provided in Table 2 and typical sample outputs of our approach and baseline systems are shown in Table 4. We start by comparing the evaluation according to BLEU scores which is available for all approaches. We first examine the baseline approaches. When retrieving the closest sentence from the training data based on the raw video features (first row in Table 2), we obtain BLEU@4 of 6.0%. By replacing the raw features with the higher level representations of attribute classifier outputs and the CRF prediction we improve to 12.0% and 13.0% @4 respectively, where the latter one is similar to the concept presented in [5] for image description. Modeling the language statistics with a n-gram model to fill function words between predicted keywords of the SR leads to a further improvement to 16% with $n = 3$ and a search span of up to 10 words. Other n-gram models with smaller search span or different n perform worse.

Next we compare the baselines to our translation system. We first notice that most variants improve over the various baseline approaches, up to 22.1% BLEU@4. This is a significant improvement over the best baseline achieving 16.0% which uses a 3-gram language model. From this we can conclude two things. First, with respect to the SR, it seems that the CRF provides a strong intermediate representation, compared to representing the video with only raw or attribute features. Second, using our translation approach clearly improves over sentence retrieval (+9.1%) or a pure n-gram model (+6.1%). We note that the n-gram model could not be applied directly to the SR, but we had to manually select a subset of the SR and preprossess the data (see Sec. 4.2) which can be learned from data using SMT.

Comparing our different variants it is interesting to see that it is important how to match a SR with descriptions during training SMT model. When a sentence is aligned to multiple SRs, just matching all SRs to it leads to a noisy model (11.2%). It is better to use the last SR (16.9%), or the largest semantic overlap between a SR and training sentence (18.9%). Best is training on the predictions rather

Approach	BLEU in % @4	@1	Human judgments Grammar	Correctness	Relevance
Baselines					
Sentence retrieval (raw video features)	6.0	32.3			
Sentence retrieval (attributes classifiers)	12.0	39.9	4.6	2.3 (3.1/2.0/2.7)	2.1
Sentence retrieval (CRF predictions)	13.0	40.0	4.6	2.8 (3.7/2.5/3.0)	2.6
CRF + N-gram generation	16.0	56.2	4.7	2.9 (3.9/2.6/2.7)	2.5
Translation (this work)					
CRF + Training on annotations (All)	11.2	38.5			
CRF + Training on annotations (Last)	16.9	44.5			
CRF + Training on annotations (Semantic overlap)	18.9	48.1	4.6	2.9 (3.7/2.6/3.2)	2.6
CRF + Training on sentence level predictions	22.1	49.6	4.6	3.1 (3.9/2.9/3.3)	2.8
Upper Bounds					
CRF + Training & test on annotation (Last)	27.7	58.2			
CRF + Training & test on annotation (Semantic overlap)	34.2	66.9	4.8	4.5 (4.5/4.7/4.0)	4.1
Human descriptions	36.0[1]	66.9[1]	4.6	4.6 (4.6/4.7/3.7)	4.3

Table 2: Evaluating generated descriptions on TACoS video-description corpus. Human judgments from 1-5, where 5 is best. For correctness judgments we additionally report correctness of activity, objects, and location.

than ground truth SRs (22.1%) which is impressive given that it is learned on noisy predictions. In contrast to the SRs based on annotations, the predictions are on sentence intervals. This indicates that a SR on the same level of the sentence granularity is most powerful.

To answer the question what is the limit of our SR, we test on the ground truth SR, i.e. we model perfect visual recognition. This results in 27.7% / 34.2% for the last/overlap variant. This is a significant improvement and can be explained by the noisy visual predictions (see Table 1). As an upper bound we report the BLEU score for the human descriptions which is 36.0%[1].

While BLEU is a good indicator for performance, it cannot level with human judgments summarized in the last three columns of Table 2. Starting with the last column (relevance, 6th column) the two main trends suggested by the BLEU scores are confirmed: our proposed approach using *training on sentence level predictions* outperforms all baselines; and using our SR based on annotations is encouragingly close to human performance (4.1 vs. 4.3, on a scale from 1 to 5, where 5 is best). The human judgments about correctness (5th column) show scores for overall correctness (first number) followed by the scores for activities, objects (including tools and ingredients), and location (covering source and target location, see Table 1). Again the two main trends are confirmed. All approaches based on CRF perform similar (2.8-2.9), only our *training on sentence level predictions* performs higher with a average score of 3.1 as it can recover from errors by learning typical errors by the CRF during training (see also examples in Table 4). It is interesting to look at the 4th column which judges the grammatical correctness of the produced sentences disregarding the visual input. Training and testing on annotations (score 4.8) outperforms the score for human descriptions (4.6), indicating that our system learned a better language model than most human descriptions have. Our translation system achieves the same score as human descriptions. The n-gram generation receives a slightly better score of 4.7 which is however due to the shorter sentences produced by this model, leading to less grammatical errors.

[1]Computed only on a 272 sentence subset where the corpus contains more than a single reference sentence for the same video. This reduces the number of references by one which leads to a lower BLEU score.

6. Evaluation: Translating images to text

We perform a second evaluation to compare with related work and show that our approach for video description can also be applied for image description. For our evaluation we choose the Pascal sentence dataset [5] which consist of 1,000 images, each paired with 5 different descriptions of one sentence. Rather than building our own SR (SR) we use the predictions provided by [5]

The SR consists of object-activity-scene triples which we annotate for the training set as they are not provided. We learn our translation approach on the training set of triples and image descriptions. We evaluate on a subset of 323 images where there are predicted descriptions available for both related approaches [5, 15]. We use the first predicted triple (with highest score) from [5]. [18] also predicts sentences for this dataset but only example sentences were available to us.

6.1. Results

We start by comparing our computed results to numbers reported by related work. [15] reports 15% BLEU@1

Approach	BLEU @4	BLEU @1
Related Work		
Template-based generation [15]	0.0	14.9
MRF + sentence retrieval [5]	1.1	25.6
Translation (this work)		
MRF + translation	4.6	34.6
MRF + adjective extension + translation	5.2	32.7
Upper Bound		
Human descriptions	15.2	56.7

Table 3: Evaluating generated descriptions on the Pascal Sentence dataset.

for their template-based generation and 50% for human descriptions. On our test subset we receive 14.9% and 56.7%, respectively, indicating that the results on the different subsets are comparable. Next we compare the two baselines with our approach shown in Table 3. For BLEU@4 the template approach [15] achieves 0.0 as the 4-gram precision is 0 (n-gram precision for 2- and 3-gram are very low (0.2%, 1.4%). This is not surprising as the templates produce very different text compared to descriptions by humans.

The sentences retrieved by [5] achieve a higher BLEU@4 of 1.1% and BLEU@1 of 25.6%. As these are sentences produced by humans this improvement is not surprising, but indicates that errors in the prediction cannot be recovered. Using the predicted triples from [5] together with our translation approach significantly improves performance to 4.6% @4 and 34.6% @1. Still, we found the SR not to be rich enough to produce good predictions. Adding adjectives and counts from the SR predicted by [15] could slightly increase to 5.2% @4 but decreasing to 32.7% @1. The BLEU@4 of only 15.2% for humans indicates the difficulty and diversity of the dataset. Never-the-less we outperform the best reported BLEU-score result on this dataset of 30% @1 by 5% (note the not identical test set) for language model based generation or meaning representation [15]. In this case [15] allows synonyms which our translation system determines automatically from the training data.

7. Conclusion

Automatically describing videos with natural language is both a compelling as well as a challenging task. This work proposes to learn the conversion from visual content to natural descriptions from a parallel corpus of videos and textual descriptions rather than using rules and templates to generate language. Our model is a two-step approach, first learning an intermediate representation of semantic labels from the video, and then translating it to natural language adopting techniques from statistical machine translation. This allows training which part of the visual content to verbalize and in which order. In order to form a natural description of the content as humans would give it our model learns which words should be added although they are not directly present in the visual content.

In an extensive experimental evaluation we show improvements of our approach compared to retrieval and n-gram based sentence generation used in prior work. The improvements are consistent across automatic evaluation with BLEU scores and human judgments of correctness and relevance. The application of our approach to sentence descriptions shows clear improvements over [15] and [5] using BLEU sore evaluation, indicating that we produce descriptions more similar to human descriptions.

To handle the different levels of granularity in the SR compared to the description we compare different variants of our model, showing that an estimation of the largest semantic overlap between the SR and the description during training performs best.

While we show the benefits of phrasing video description as a translation problem, there are many possibilities to improve our work. Further directions include modeling temporal dependencies in both the SR and the language generation, as well as modeling the uncertainty of the visual input explicitly in the generation process, which has similarities to translating from uncertain speech input. This work could be combined with approaches which automatically extract a semantic representation from a text description, which has recently been proposed in [22] for activities.

Acknowledgements. This work was partially funded by the DFG project SCHI989/2-2.

References

[1] A. Aker and R. J. Gaizauskas. Generating image descriptions using dependency relational patterns. In *ACL*, 2010. 2

[2] 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 *UAI*, 2012. 1, 2

[3] J. Corso, C. Xu, P. Das, R. F. Doell, and P. Rosebrough. Thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013. 1, 2

[4] P. Duygulu, K. Barnard, N. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 2

[5] 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 *ECCV*, 2010. 1, 2, 3, 4, 5, 6, 7

[6] M. Federico, N. Bertoldi, and M. Cettolo. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech*. ISCA, 2008. 4

[7] Y. Feng and M. Lapata. How many words is a picture worth? Automatic caption generation for news images. ACL'10. 2

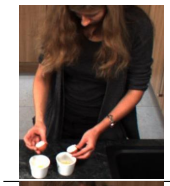
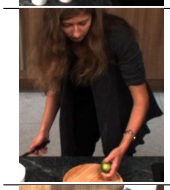
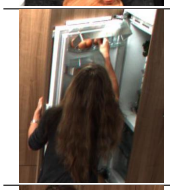
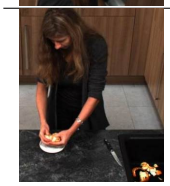
	**(1)** SR predicted by CRF	⟨ OPEN EGG, HAND, EGG, BOWL, NULL ⟩
	Sentence retrieval (CRF predictions)	the person slices the avocado
	CRF + N-gram generation	the person opens up egg over
	CRF+Train on annotations (Overlap)	the person cracks the eggs into the bowl
	CRF+Train on sentence level predictions	the person cracks the eggs
	Human description	the person dumps any remaining whites of the eggs from the shells into the cup with the egg whites
	**(2)** SR predicted by CRF	⟨ TAKE OUT, HAND, PLASTIC-BAG, FRIDGE, CUTTING-BOARD⟩
	Sentence retrieval (CRF predictions)	the person took out cucumber
	CRF + N-gram generation	the person takes out a bag of chilies
	CRF+Train on annotations (Overlap)	the person gets out a package of limes from the fridge and places it on the cutting board
	CRF+Train on sentence level predictions	the person gets out a cutting board from the loaf of bread from the fridge
	Human description	the person gets the lime, a knife and a cutting board
	**(3)** SR predicted by CRF	⟨ PUT IN, HAND, WRAPPING-PAPER, NULL, FRIDGE⟩
	Sentence retrieval (CRF predictions)	person then places cucumber on plate
	CRF + N-gram generation	the person puts the bread with existing plastic paper
	CRF+Train on annotations (Overlap)	the person rinses and puts away the butter back in the fridge
	CRF+Train on sentence level predictions	the person takes out a carrot from the fridge
	Human description	the person procures an egg from the fridge
	**(4)** SR predicted by CRF	⟨ REMOVE FROM PACKAGE, KIWI, HAND, PLASTIC-BAG, NULL⟩
	Sentence retrieval (CRF predictions)	the person selects five broad beans from the package
	CRF + N-gram generation	the person removes a kiwi
	CRF+Train on annotations (Overlap)	the person takes the package of beans out of the kiwi
	CRF+Train on sentence level predictions	the person goes to the refrigerator and takes out the half kiwi
	Human description	using her hands, the person splits the orange in hald over the saucer

Table 4: Example output of our system (blue) compared to baseline approaches and human descriptions, errors in red. (1, 2) our system provides the best output; (2, 3) our system partially recovers from a wrong SR; (4) failure case.

[8] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition. In *ICCV*, 2013. 1, 2

[9] A. Gupta, P. Srinivasan, J. B. Shi, and L. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009. 1, 2

[10] P. Hanckmann, K. Schutte, and G. J. Burghouts. Automated textual descriptions for a wide range of video events with 48 human actions. In *ECCV Workshops*, 2012. 1, 2

[11] M. U. G. Khan, L. Zhang, and Y. Gotoh. Human focused video description. In *ICCV Workshops*, 2011. 1, 2

[12] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010. 2

[13] 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 *ACL demo*, 2007. 2, 4

[14] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 2002. 1, 2

[15] 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 *CVPR*, 2011. 1, 2, 3, 4, 5, 6, 7

[16] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *ACL*, 2012. 1, 3, 5

[17] A. Lopez. Statistical machine translation. *ACM*, 2008. 2

[18] 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 *EACL*, 2012. 1, 3, 6

[19] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *CL*, 2003. 4

[20] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 3

[21] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5

[22] V. Ramanathan, P. Liang, and L. Fei-Fei. Video event understanding using natural language descriptions. In *ICCV*, 2013. 7

[23] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *TACL*, 2013. 1, 2, 4, 5

[24] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *ECCV*, 2012. 2, 5

[25] M. Schmidt. UGM: Matlab code for undirected graphical models. di.ens.fr/∼mschmidt/Software/UGM.html, 2013. 3

[26] C. C. Tan, Y.-G. Jiang, and C.-W. Ngo. Towards textually describing complex video contents with audio-visual concept classifiers. In *ACM Multimedia*, 2011. 1, 2

[27] H. Wang, A. Kläser, C. Schmid, and C. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 2013. 2, 4, 5