

# Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild

**Jesse Thomason\***

University of Texas at Austin  
jesse@cs.utexas.edu

**Subhashini Venugopalan\***

University of Texas at Austin  
vsub@cs.utexas.edu

**Sergio Guadarrama**

University of California Berkeley  
sguada@eecs.berkeley.edu

**Kate Saenko**

University of Massachusetts Lowell  
saenko@cs.uml.edu

**Raymond Mooney**

University of Texas at Austin  
mooney@cs.utexas.edu

## Abstract

This paper integrates techniques in natural language processing and computer vision to improve recognition and description of entities and activities in real-world videos. We propose a strategy for generating textual descriptions of videos by using a factor graph to combine visual detections with language statistics. We use state-of-the-art visual recognition systems to obtain confidences on entities, activities, and scenes present in the video. Our factor graph model combines these detection confidences with probabilistic knowledge mined from text corpora to estimate the most likely subject, verb, object, and place. Results on YouTube videos show that our approach improves both the joint detection of these latent, diverse sentence components and the detection of some individual components when compared to using the vision system alone, as well as over a previous  $n$ -gram language-modeling approach. The joint detection allows us to automatically generate more accurate, richer sentential descriptions of videos with a wide array of possible content.

## 1 Introduction

Integrating language and vision is a topic that is attracting increasing attention in computational linguistics (Berg and Hockenmaier, 2013). Although there is a fair bit of research on generating natural-language descriptions of images (Feng and Lapata, 2013; Yang et al., 2011; Li et al., 2011; Ordonez et al., 2011), there is significantly less work on describing videos (Barbu et al., 2012; Guadarrama et al., 2013; Das et al., 2013; Rohrbach et al., 2013; Senina et al., 2014). In particular, much of the research on videos utilizes artificially constructed videos with prescribed sets of objects and actions (Barbu et al., 2012; Yu and Siskind, 2013). Generating natural-language descriptions of videos *in the wild*, such as those posted on YouTube, is a very challenging task.

In this paper, we focus on selecting content for generating sentences to describe videos. Due to the large numbers of video actions and objects and scarcity of training data, we introduce a graphical model for integrating statistical linguistic knowledge mined from large text corpora with noisy computer vision detections. This integration allows us to infer which vision detections to trust given prior linguistic knowledge. Using a large, realistic collection of YouTube videos, we demonstrate that this model effectively exploits linguistic knowledge to improve visual interpretation, producing more accurate descriptions compared to relying solely on visual information. For example, consider the frames of the video in Figure 1. Instead of generating the inaccurate description “A person is playing on the keyboard in the kitchen” using purely visual information, our system generates the more correct “A person is playing the piano in the house” by using statistics mined from parsed corpora to improve the interpretation of the uncertain visual detections, such as the presence of both a computer keyboard and a piano in the video.

## 2 Background and Related Work

Several recent projects have integrated linguistic and visual information to aid description of images and videos. The most related work on image description is Baby Talk (Kulkarni et al., 2011), which uses

---

\*Indicates equal contribution

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: <http://creativecommons.org/licenses/by/4.0/>

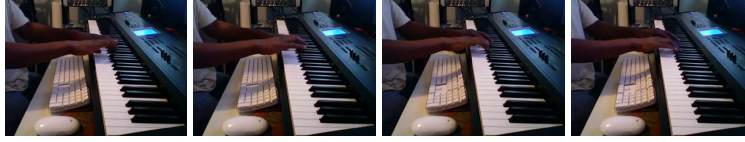


Figure 1: Frames which depict a person playing a piano in front of a keyboard from one of the videos in our dataset. Purely visual information is more confident in the computer keyboard’s presence than the piano’s, while our model can correctly determine that the person is more likely to be playing the piano than the computer keyboard.

a Conditional Random Field (CRF) to integrate visual detections with statistical linguistic knowledge mined from parsed image descriptions and Google queries, and the work of Yang et al. (2011) which uses corpus statistics to aid the description of objects and scenes. We go beyond the scope of these previous works by also selecting verbs through the integration of activity recognition from video and statistics from parsed corpora.

With regard to video description, the work of Barbu et al. (2012) uses a small, hand-coded grammar to describe a sparse set of prescribed activities. In contrast, we utilize corpus statistics to aid the description of a wide range of naturally-occurring videos. The most similar work is (Krishnamoorthy et al., 2013; Guadarrama et al., 2013) which uses an  $n$ -gram language model to help determine the best subject-verb-object for describing a video. Krishnamoorthy et al. (2013) used a limited set of videos containing a small set of 20 entities, and the work of Guadarrama et al. (2013) showed an advantage of using linguistic knowledge only for the case of “zero shot activity recognition,” in which the appropriate verb for describing the activity was never seen during training. Compared to this prior work, we explore a much larger set of entities and activities (see Section 3.2) and add scene recognition (see Section 3.3) to further enrich the descriptions. Our experiments demonstrate that our graphical model produces a more accurate subject-verb-object-place description than these simpler  $n$ -gram language modeling approaches.

### Our Contributions:

- We present a new method, a *Factor Graph Model* (FGM), to perform content selection by integrating visual and linguistic information to select the best subject-verb-object-place description of a video.
- Our model includes scene (location) information which has not been addressed by previous video description works (Barbu et al., 2012; Krishnamoorthy et al., 2013; Guadarrama et al., 2013).
- We demonstrate the scalability of our model by evaluating it on a large dataset of naturally occurring videos (1297 training, 670 testing), recognizing sentential subjects out of 45 candidate entities, objects out of 218 candidate objects, verbs out of 218 candidate activities, and places out of 12 candidate scenes.

## 3 Approach

Our overall approach uses a probabilistic graphical model to integrate the visual detection of entities, activities, and scenes with language statistics to determine the best subject, verb, object, and place to describe a given video. A descriptive English sentence is generated from the selected sentential components.

### 3.1 Video Dataset

We use the video dataset collected by Chen and Dolan (2011). The dataset contains 1,967 short YouTube video clips paired with multiple human-generated natural-language descriptions. The video clips are 10 to 25 seconds in duration and typically consist of a single activity. Portions of this dataset have been used in previous work on video description (Motwani and Mooney, 2012; Krishnamoorthy et al., 2013; Guadarrama et al., 2013). We use 1,297 randomly selected videos for training and evaluate predictions on the remaining 670 test videos.

### 3.2 Visual Recognition of Subject, Verb, and Object

We utilize the visual recognition techniques employed by Guadarrama et al. (2013) to process the videos and produce probabilistic detections of grammatical subjects, verbs, and objects. In our data-set there are 45 candidate entities for the grammatical subject (such as *animal*, *baby*, *cat*, *chef*, and *person*) and 241 for the grammatical object (such as *flute*, *motorbike*, *shrimp*, *person*, and *tv*). There are 218 candidate activities for the grammatical verb, including *climb*, *cut*, *play*, *ride*, and *walk*.

**Entity Related Features** From each video two frames per second are extracted and passed to pre-trained visual object classifiers and detectors. As in Guadarrama et al. (2013), we compute representations based on detected objects using ObjectBank (Li et al., 2010) and the 20 PASCAL (Everingham et al., 2010) object classes for each frame. We use the PASCAL scores and ObjectBank scores with max pooling over the set of frames as the entity descriptors for the video clip. Additionally, to be able to recognize more objects, we use the LLC-10k proposed by Deng et al. (2012) which was trained on ImageNet 2011 object dataset with 10k categories. LLC-10K uses a bank of linear SVM classifiers over pooled local vector-quantized features learned from the 7K bottom level synsets of the 10K ImageNet database. We aggregate the 10K classifier scores obtained for each frame by doing max pooling across frames.

**Activity Related Features** We use the activity recognizers described in Guadarrama et al. (2013) to produce probabilistic verb detections. They extract Dense Trajectories developed by Wang et al. (2011) and compute HoG (Histogram of Gradients), HoF (Histograms of Optical Flow) and MBH (Motion Boundary Histogram) features over space time volumes around the trajectories. We used the default parameters proposed in Wang et al. (2011) ( $N = 32$ ,  $n_\sigma = 2$ ,  $n_r = 3$ ) and adopted a standard bag-of-features representation. We construct a codebook for each descriptor (Trajectory, HoG, HoF, MBH) separately. For each descriptor we randomly sampled 100K points and clustered them using K-means into a codebook of 4000 words. Descriptors are assigned to their closest vocabulary word using Euclidean distance. Each video is then represented as a histogram over these clusters.

**Multi-channel SVM** To allow object and activity features inform one another, we combine all the features extracted using a multi-channel approach inspired by Zhang et al. (2007) to build three non-linear SVM (Chang and Lin, 2011) classifiers for the subject, verb, and object, as described in Guadarrama et al. (2013). Note that we do not employ the hierarchical semantic model of Guadarrama et al. (2013) to augment our object or activity recognition. In addition, each SVM learns a Platt scaling (Platt, 1999) to predict the label and a visual confidence value,  $C(t) \in [0, 1]$ , for each entity or activity  $t$ . The output of the SVMs constitute the visual confidences on subject, verb, and object in all the models described henceforth.

### 3.3 Visual Scene Recognition

In addition to the techniques employed by Guadarrama et al. (2013) used to obtain probabilistic detections of grammatical subjects, verbs, and objects, we developed a novel scene detector based on state-of-the-art computer vision methods.

We examined the description of all the 1,967 videos in the YouTube dataset and extracted scene words from the dependency parses as described in Section 3.4. With the help of WordNet<sup>1</sup> we grouped the list of scene words and their synonyms into distinct scene classes. Based on the frequency of mentions and the coverage of scenes in the dataset, we shortlisted a set of 12 final scenes (*mountain*, *pool*, *beach*, *road*, *kitchen*, *field*, *snow*, *forest*, *house*, *stage*, *track*, and *sky*).

For the detection itself, we follow Xiao et al. (2010) and select several state-of-the-art features that are potentially useful for scene recognition. We extract GIST, HOG2x2, SSIM (self-similarity) and Dense SIFT descriptors. We also extract LBP (Local Binary Patterns), Sparse SIFT Histograms, Line features, Color Histograms, Texton Histograms, Tiny Images, Geometric Probability Map and Geometric specific histograms. The code for extracting the features and computing kernels for the features is taken from

---

<sup>1</sup><http://wordnet.princeton.edu>

the original papers as described in Xiao et al. (2010). Using the features and kernels, we train one-vs-all SVMs (Chang and Lin, 2011) to classify images into scene categories. As in Xiao et al. (2010), this gave us 51 different SVM classifiers with different feature and kernel choices. We use the images from the UIUC 15 scene dataset (Lazebnik et al., 2006) and the SUN 397 scene dataset (Xiao et al., 2010) for training the scene classifiers for all scenes except *kitchen*. The training images for *kitchen* were obtained by selecting 100 frames from about 15 training videos, since the classifier trained on images from the existing scene datasets performed extremely poorly on the videos. We use all the classifiers to detect scenes for each frame. We then average the scene detection scores over all the classifiers across all the frames of the video. This gives us visual confidence values,  $C(t)$ , over all scene categories  $t$  for the video.

### 3.4 Language Statistics

A key aspect of our approach is the use of language statistics mined from English text corpora to bias visual interpretation. Like Krishnamoorthy et al. (2013), we use dependency-parsed text from four large “out of domain” corpora: English Gigaword, British National Corpus (BNC), ukWac and WaCkypedia\_EN. We also use a small, specialized “in domain” corpus: dependency parsed sentences from the human-generated, English descriptions for the YouTube training videos mentioned in Section 3.1. We extract SVOP (subject, verb, object, place) tuples from the dependency parses. The subject-verb relationships are identified using *nsubj* dependencies, the verb-object relationships using *doobj* and *prep* dependencies. Object-place relationships are identified using the *prep* dependency, checking that the noun modified by the preposition is one of our recognizable places (or synonyms of the recognizable scenes as indicated by WordNet). We then extract co-occurring SV, VO, and OP bigram statistics from the resulting SVOP tuples to inform our factor-graph model, which uses both the out-of-domain ( $p_o$ ) and in-domain ( $p_i$ ) bigram probabilities.

### 3.5 Content Selection Using Factor Graphs

In order to combine visual and linguistic evidence, we use the probabilistic factor-graph model shown in Figure 2. This model integrates the uncertain visual detections described in Sections 3.2 and 3.3 with the language statistics described in Section 3.4 to predict the best words for describing the subject (S), verb (V), object (O), and place (P) for each test video. After instantiating the potential functions for this model, we perform a maximum a posteriori (MAP) estimation (via the max-product algorithm) to determine the most probable joint set of values for these latent variables.

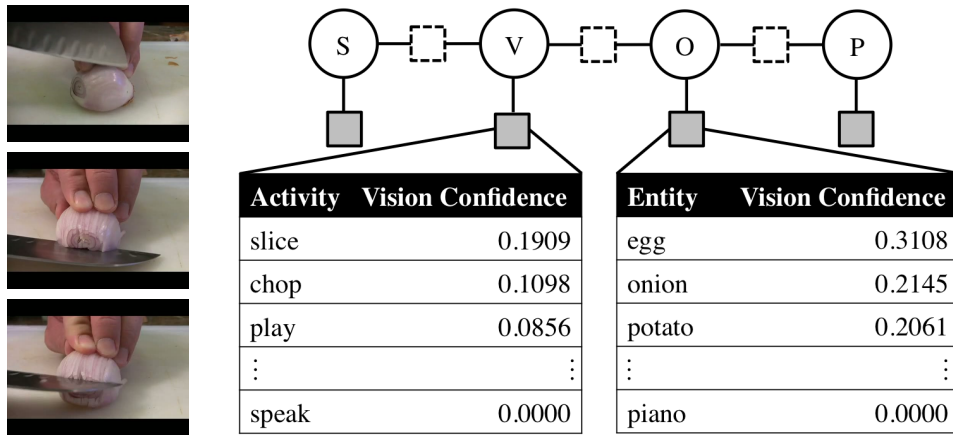


Figure 2: The factor graph model used for content selection (right), and sample frames from a video to be described (left). Visual confidence values are observed (gray potentials) and inform sentence components. Language potentials (dashed) connect latent words between sentence components. Samples of the vision confidence values used as observations for the verb and object are shown for the example test video.

**Observation Potentials.** The observations in our model take the form of confidence scores from the visual detectors described in Sections 3.2 and 3.3. That is, the potential for each sentence component  $k \in \{S, V, O, P\}$ ,  $\phi_k(t) = C_k(t)$  is the detection confidence that the classifier for component  $k$  ( $C_k$ ) gives to the word  $t$ .

**Language Potentials.** Language statistics were gathered as described in Section 3.4 and used to determine the language potentials as follows:

$$\phi_{k,l}(t, s) := p(l = s | k = t) := \alpha p_o(l = s | k = t) + (1 - \alpha) p_i(l = s | k = t)$$

Where  $k$  and  $l$  are two contiguous components in the SVOP sequence and  $t$  and  $s$  are words that are possible values for these two components, respectively. We would expect

$$\phi_{V,O}(\text{ride}, \text{motorbike}) := p(O=\text{motorbike} | V=\text{ride})$$

to be relatively high, since *motorbike* is a likely object of the verb *ride*. The potential between two sequential components  $k$  and  $l$  in the SVOP sequence is computed by linearly interpolating the bigram probability observed in the out-of-domain corpus of general text ( $p_o$ ) and the in-domain corpus of video descriptions ( $p_i$ ). The interpolation parameter  $\alpha$  adjusts the importance of these two corpora in determining the bigram probability. We optimized performance by fixing  $\alpha = 0.25$  when cross-validating on the training data. This weighting effectively allows general text corpora to be used to smooth the probability estimates for video descriptions. We note that meaningful information would likely be captured by non-contiguous language potentials such as  $\phi_{V,P}$ , but that the resulting factor graphs would contain cycles, preventing us from performing exact inference tractably.

### 3.6 Sentence Generation

Finally, we use the SVOP tuple chosen by our model to generate an English sentence using the following template: “*Determiner (A, The) - Subject - Verb (Present, Present Continuous) - Preposition (optional) - Determiner (A, The) - Object (optional) - Preposition - Determiner (A, The) - Place (optional)*” The most probable prepositions are identified using preposition-object and preposition-place bigram statistics mined from the dependency parsed corpora described in Section 3.4. Given an SVOP tuple, our objective is to generate a rich sentence using the subject, verb, object, and place information. However, it is not prudent to add the object and place to the description of all videos since some verbs may be intransitive and the place information may be redundant. In order to achieve the best set of components to include, we use the above template to first generate a set of candidate sentences based on the SVO triple, SVP triple and the SVOP quadruple. Then, each sentence type (SVO, SVP, and SVOP) is ranked using the BerkeleyLM language model (Pauls and Klein, 2011) trained on the GoogleNgram corpus. Finally, we output the sentence with the highest average 5-gram probability in order to normalize for sentence length.

## 4 Experimental Results

We compared using the vision system alone to our model, which augments that system with linguistic knowledge. Specifically, we consider the *Highest Vision Confidence* (HVC) model, which takes for each sentence component the word with the highest confidence from the state-of-the-art vision detectors described in Sections 3.2 and 3.3. We compare the results of this model on the 670 test videos to those of our *Factor Graph Model* (FGM), as discussed in Section 3.5.

### 4.1 N-gram Baseline

Additionally, we compare both models against the existing, baseline *n-gram* model of Krishnamoorthy et al. (2013) by extending their best *n-gram* model to support places. To be specific, we build a quadr-gram model, similar to the trigram model of Krishnamoorthy et al. (2013). We first extract SVOP tuples from the dependency parses as described in Section 3.4. We then train a backoff language model with Kneyser-Ney smoothing (Chen and Goodman, 1996) for estimating the likelihood of the SVOP quadruple. On quadruples that are not seen during training, this quadr-gram language model backs off to SVO

<b>Most</b>	S%	V%	O%	[P]%	SVO%	SVO[P]%
n-gram	76.57	11.04	11.19	18.30	2.39	1.86
HVC	76.57	+22.24	11.94	17.24	+4.33	+2.92
FGM	76.42	+21.34	12.39	19.89	<b>+5.67</b>	+3.71
<b>Any</b>						
n-gram	86.87	19.25	21.94	21.75	5.67	2.65
HVC	86.57	+38.66	22.09	21.22	+10.15	+4.24
FGM	86.27	+37.16	<b>+24.63</b>	<b>24.67</b>	+10.45	+6.10

Table 1: Average binary accuracy of predicting the **most** common word (top) and of predicting **any** given word (bottom). **Bold** entries are statistically significantly ( $p < 0.05$ ) greater than the HVC model, while + entries are significantly greater than the n-gram model. No model scored significantly higher than FGM on any metric. [P] indicates that the score ranges only over the subset of videos for which any annotator provided a place.

triple and subject-verb, verb-object, object-place bigrams to estimate the probability of the quadruple. As in the case of the factor graph model, we consider the effect of learning from a domain specific text corpus. We build quadragram language models for both out-of-domain and in-domain text-corpora described in Section 3.4. The probability of a quadragram in the language model is computed by linearly interpolating the probabilities from the in-domain and out-of-domain corpus. We experiment with different number of top subjects, objects, verbs, and places to estimate the most likely SVOP quadruple from the quadragram language model. We report the results for the best performing n-gram model that considers the top 5 subjects, 5 objects, 10 verbs, and 3 places based on the vision confidences and an out-of-domain corpus weight of 1. This model also incorporates *verb expansion* as described in the original work (Krishnamoorthy et al., 2013).

## 4.2 Content Evaluation

Table 1 shows the accuracy of the models when their prediction for each sentence component is considered correct only if it is the word *most commonly* used by human annotators to describe the video, as well as the accuracy of the models when the prediction is considered correct if used by *any* of the annotators to describe the video. We evaluate the accuracy of each component (S,V,O,P) individually, and for complete SVO and SVOP tuples, where *all* components must be correct in order for a complete tuple to be judged correct. Because only about half (56.3%) of test videos were described with a place by some annotator, accuracies involving places (“[P]”) are averaged only over the subset of videos for which any annotator provided a place. Significance was determined using a paired t-test which compared the distributions of the binary correctness of each model’s prediction on each video for the specified component(s).

We also use the WUP metric from Wordnet::Similarity<sup>2</sup> to measure the quality of the predicted words to account for semantically similar words. For example, where the binary metric would mark “slice” as an incorrect substitute for “cut”, the WUP metric will provide “partial credit” for such predictions. The results using WUP similarity metrics for the most common word and any valid word (maximum WUP similarity is chosen from among valid words) are presented in Table 2. Since WUP provides scores are in the range [0,1], we view the scores as “percent relevance,” and we obtain tuple scores for each sentence by taking the product of the component WUP scores.

## 5 Discussion

It is clear from the results in Table 1 that both the HVC and the FGM outperform the n-gram language model approach used in the most-similar previous work (Krishnamoorthy et al., 2013; Guadarrama et al., 2013). Note that while Krishnamoorthy et al. (2013) showed an improvement with an n-gram model considering only the top few vision detections, the FGM considers vision confidences over the entire set

<sup>2</sup><http://wn-similarity.sourceforge.net/>

<b>Most</b>	S%	V%	O%	[P]%	SVO%	SVO[P]%
n-gram	89.00	41.56	44.01	<b>57.62</b>	17.53	10.83
HVC	89.09	+*48.85	43.99	56.00	+20.82	+12.95
FGM	89.01	+47.05	+ <b>45.29</b>	+ <b>59.64</b>	+21.54	+ <b>14.50</b>
<b>Any</b>						
n-gram	96.60	55.08	65.52	<b>61.98</b>	35.70	22.84
HVC	96.54	+*65.61	65.32	60.67	+42.53	+27.75
FGM	96.32	+63.49	+ <b>67.52</b>	+ <b>64.68</b>	+42.43	+29.34

Table 2: Average WUP score of the predicted word against the **most** common word (top) and the maximum score against **any** given word (bottom). **Bold** entries are statistically significantly ( $p < 0.05$ ) greater than the HVC model; + entries are significantly greater than the n-gram model; \* entries are significantly greater than the FGM. [P] indicates that the score ranges only over the subset of videos for which any annotator provided a place.

of grammatical objects. Additionally, our models are evaluated on a much more diverse set of videos while Krishnamoorthy et al. (2013) evaluate the n-gram model on 185 videos (a small subset of the 1,967 videos containing the 20 grammatical objects that their system recognized).

The performance differences between the vision system (HVC) and our integrated model (FGM) are modest but significant in important places. Specifically, the FGM makes improvements to SVO (Table 1, top) and SVOP (Table 2, top) tuple accuracies. FGM also significantly improves both the O and [P] (Table 1, bottom, and Table 2) component accuracies, suggesting that it can help clean up some noise from the vision systems even at the component level by considering related bigram probabilities. FGM causes no significant losses under the binary metric, but performs worse than the HVC model on predicting a verb component semantically similar to the correct verb under the WUP metric (Table 2). This loss on the verb component is worth the gains in tuple accuracy, since tuple prediction is the more difficult and most central part of the content selection task. Additionally, experiments by the authors of Guadarrama et al. (2013) on Amazon Mechanical Turk have shown that humans tend to heavily penalize tuples and descriptions even if they have most of the components correct.

Table 3 shows frames from some test videos and the sentence components chosen by the models to describe them. In the top four videos we see the FGM improving raw vision results. For example, it determines that a person is more likely slicing an onion than an egg. Some specific confidence values for the HVC can be seen for this video in Figure 2. In the bottom two videos of Table 3 we see the HVC performing better without linguistic information. For example, the FGM intuits that a person is more likely to be driving a car than lifting it, and steers the prediction away from the correct verb. This may be part of a larger phenomenon in which YouTube videos often depict unusual actions, and consequently general language knowledge can sometimes hurt performance by selecting more common activities.

## 6 Future Work

Compared to the human gold standard descriptions, there appears to be room for improvement in detecting activities, objects, and scenes with high precision. Visual recognition of entities and activities in diverse real-world videos is extremely challenging, partially due to lack of training data. As a result our current model is faced with large amounts of noise in the vision potentials, especially for objects. Going forward, we believe that improving visual recognition will allow the language statistics to be even more useful. We are currently exploring deep image feature representations (Donahue et al., 2013) to improve object and verb recognition, as well as model transfer from large labeled object ontologies (Deng et al., 2009).

From the generation perspective, there is scope to move beyond the template based sentence generation. This becomes particularly relevant if we detect multiple grammatical objects such as adjectives or adverbs. We need to decide whether additional grammatical objects would enrich the sentence de-



























FGM improves over HVC					
“A person is slicing the onion in the kitchen”					Gold: person, slice, onion, ( <i>none</i> ) HVC: person, slice, egg, kitchen FGM: person, slice, onion, kitchen
					
“A person is running a race on the road”					Gold: person, run, race, ( <i>none</i> ) HVC: person, ride, race, ground FGM: person, run, race, road
					
“A person is playing the guitar on the stage”					Gold: person, play, guitar, tree HVC: person, play, water, kitchen FGM: person, play, guitar, stage
					
“A person is playing a guitar in the house”					Gold: person, play, guitar, ( <i>none</i> ) HVC: person, pour, chili, kitchen FGM: person, play, guitar, house
					
HVC better alone					
“A person is lifting a car on the road”					Gold: person, lift, car, ground HVC: person, lift, car, road FGM: person, drive, car, road
					
“A person is pouring the egg in the kitchen”					Gold: person, pour, mushroom, kitchen HVC: person, pour, egg, kitchen FGM: person, play, egg, kitchen
					

Table 3: Example videos and: (Gold) the most common SVOP provided by annotators; (HVC) the highest vision confidence selections; (FGM) the selections from our factor graph model. The top section shows videos where the FGM improved over HVC; the bottom shows videos where the HVC did better alone. For each video, the sentence generated from the components chosen from the more successful system is shown.

scription and identify when to add them appropriately. With increasing applications for such systems in automatic video surveillance and video retrieval, generating richer and more diverse sentences for longer videos is an area for future research. In comparison to previous approaches (Krishnamoorthy et al., 2013; Yang et al., 2011) the factor graph model can be easily extended to support this. Additional nodes can be attached suitably to the graph to enable the prediction of adjectives and adverbs to enrich the base SVOP tuple.

## 7 Conclusions

This work introduces a new framework to generate simple descriptions of short videos by integrating visual detection confidences with language statistics obtained from large textual corpora. Experimental results show that our approach achieves modest improvements over a pure vision system and significantly improves over previous methods in predicting the complete subject-verb-object and subject-verb-object-place tuples. Our work has a broad coverage of objects and verbs and extends previous works by predicting place information.



There are instances where our model fails to predict the correct verb when compared to the HVC model. This could partially be because the SVM classifiers that detect activity already leverage entity information during training, and adding external language does not appear to improve verb prediction significantly. Further detracting from performance, our model occasionally propagates, rather than correcting, errors from the HVC. For example, when the HVC predicts the correct verb and incorrect object, such as in “person ride car” when the video truly depicts a person riding a motorbike, our model selects the more likely verb pairing “person drive car”, extending the error from the object to the verb as well.

Despite these drawbacks, our approach predicts complete subject-verb-object-place tuples more closely related to the most commonly used human descriptions than vision alone (Table 2), and in general improves both object and place recognition accuracies (Tables 1, 2).

## Acknowledgements

This work was funded by NSF grant IIS1016312, DARPA Minds Eye grant W911NF-10-9-0059, and NSF ONR ATL grant N00014-11-1-0105. Some of our experiments were run on the Mastodon Cluster (NSF grant EIA-0303609).

## References

- Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell Waggoner, Song Wang, Jinlian Wei, Yifan Yin, and Zhiqi Zhang. 2012. Video in sentences out. In *Association for Uncertainty in Artificial Intelligence (UAI)*.
- Tamara Berg and Julia Hockenmaier. 2013. Workshop on vision and language. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. NAACL.
- Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.
- David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics.
- Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL)*, pages 310–318. Association for Computational Linguistics.
- Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jia Deng, Kai Li, Minh Do, Hao Su, and Li Fei-Fei. 2009. Construction and analysis of a large scale image ontology. Vision Sciences Society.
- Jia Deng, Jonathan Krause, Alex Berg, and Li Fei-Fei. 2012. Hedging Your Bets: Optimizing Accuracy-Specificity Trade-offs in Large Scale Visual Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*.
- Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, June.
- Yansong Feng and Mirella Lapata. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(4):797–812.

- Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *IEEE International Conference on Computer Vision (ICCV)*, December.
- Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond J. Mooney, Kate Saenko, and Sergio Guadarrama. 2013. Generating natural-language video descriptions using text-mined knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 541–547.
- Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Alexander Berg, Yejin Choi, and Tamara Berg. 2011. Baby talk: Understanding and generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178. IEEE.
- Li-Jia Li, Hao Su, Eric Xing, and Li Fei-Fei. 2010. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Advances in Neural Information Processing Systems (NIPS)*.
- Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 220–228, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Tanvi S. Motwani and Raymond J. Mooney. 2012. Improving video activity recognition using object recognition and text mining. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 600–605.
- Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, pages 1143–1151.
- Adam Pauls and Dan Klein. 2011. Faster and smaller n-gram language models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 258–267. Association for Computational Linguistics.
- John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances In Large Margin Classifiers*, pages 61–74. MIT Press.
- Marcus Rohrbach, Qiu Wei, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. 2013. Translating video content to natural language descriptions. In *IEEE International Conference on Computer Vision (ICCV)*.
- Anna Senina, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. *arXiv preprint arXiv:1403.6173*.
- Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3169–3176. IEEE.
- Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492.
- Yezhou Yang, Ching Lik Teo, Hal Daumé, III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 444–454.
- Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 53–63.
- Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision (IJCV)*, 73(2):213–238.