Generating Natural-Language Video Descriptions Using Text-Mined Knowledge

Niveda Krishnamoorthy *
UT Austin
niveda@cs.utexas.edu

Girish Malkarnenkar *
UT Austin
girish@cs.utexas.edu

Raymond Mooney
UT Austin
mooney@cs.utexas.edu

Kate Saenko
UMass Lowell
saenko@cs.uml.edu

Sergio Guadarrama
UC Berkeley
sguada@eecs.berkeley.edu

Abstract

We present a holistic data-driven technique that generates natural-language descriptions for videos. We combine the output of state-of-the-art object and activity detectors with "real-world" knowledge to select the most probable subject-verb-object triplet for describing a video. We show that this knowledge, automatically mined from web-scale text corpora, enhances the triplet selection algorithm by providing it contextual information and leads to a four-fold increase in activity identification. Unlike previous methods, our approach can annotate arbitrary videos without requiring the expensive collection and annotation of a similar training video corpus. We evaluate our technique against a baseline that does not use text-mined knowledge and show that humans prefer our descriptions 61% of the time.

<person, ride, horse>
A person is riding a horse.

Figure 1: Content planning and surface realization

Introduction

Combining natural-language processing (NLP) with computer vision to to generate English descriptions of visual data is an important area of active research (Motwani and Mooney 2012; Farhadi et al. 2010; Yang et al. 2011). We present a novel approach to generating a simple sentence for describing a short video that:

1. Identifies the most likely subject, verb and object (SVO) using a combination of visual object and activity detectors and text-mined knowledge to judge the likelihood of SVO triplets. From a natural-language generation (NLG) perspective, this is the *content planning* stage.
2. Given the selected SVO triplet, it uses a simple template-based approach to generate candidate sentences which are then ranked using a statistical language model trained on web-scale data to obtain the best overall description. This is the *surface realization* stage.

Figure 1 shows sample system output. Our approach can be viewed as a holistic data-driven three-step process where we first detect objects and activities using state-of-the-art visual recognition algorithms. Next, we combine these often noisy detections with an estimate of real-world likelihood, which we obtain by mining SVO triplets from large-scale web corpora. Finally, these triplets are used to generate candidate sentences which are then ranked for plausibility and grammaticality. The resulting natural-language descriptions can be usefully employed in applications such as semantic video search and summarization, and providing video interpretations for the visually impaired.

Using vision models alone to predict the best subject and object for a given activity is problematic, especially while dealing with challenging real-world YouTube videos as shown in Figures 4 and 5, as it requires a large annotated video corpus of similar SVO triplets (Packer, Saenko, and Koller 2012). We are interested in annotating arbitrary short videos using off-the-shelf visual detectors, without the engineering effort required to build domain-specific activity models. Our main contribution is incorporating the pragmatics of various entities' likelihood of being the subject/object of a given activity, learned from web-scale text corpora. For example, animate objects like people and dogs are more likely to be subjects compared to inanimate objects like balls or TV monitors. Likewise, certain objects are more likely to function as subjects/objects of certain activities, e.g., "riding a horse" vs. "riding a house."

Selecting the best verb may also require recognizing activities for which no explicit training data has been provided. For example, consider a video with a man walking his dog. The object detectors might identify the man and dog; however the action detectors may only have the more general activity, "move," in their training data. In such cases, real-

*Indicates equal contribution

world pragmatics is very helpful in suggesting that "walk" is best used to describe a man "moving" with his dog. We refer to this process as *verb expansion*.

After describing the details of our approach, we present experiments evaluating it on a real-world corpus of YouTube videos. Using a variety of methods for judging the output of the system, we demonstrate that it frequently generates useful descriptions of videos and outperforms a purely vision-based approach that does not utilize text-mined knowledge.

Background and Related Work

Most prior work on natural-language description of visual data has focused on static images (Felzenszwalb, McAllester, and Ramanan 2008; Laptev et al. 2008; Yao et al. 2010; Kulkarni et al. 2011). The small amount of existing work on videos (Khan and Gotoh 2012; Lee et al. 2008; Kojima, Tamura, and Fukunaga 2002; Ding et al. 2012; Yao and Fei-Fei 2010) uses hand-crafted templates or rule-based systems, works in constrained domains, and does not exploit text mining. Barbu et al. (2012) produce sentential descriptions for short video clips by using an interesting dynamic programming approach combined with Hidden Markov Models for obtaining verb labels for each video. However, they make use of extensive hand-engineered templates.

Our work differs in that we make extensive use of text-mined knowledge to select the best SVO triple and generate coherent sentences. We also evaluate our approach on a generic, large and diverse set of challenging YouTube videos that cover a wide range of activities. Motwani and Mooney (2012) explore how object detection and text mining can aid activity recognition in videos; however, they do not determine a complete SVO triple for describing a video nor generate a full sentential description.

With respect to static image description, Li et al. (2011) generate sentences given visual detections of objects, visual attributes and spatial relationships; however, they do not consider actions. Farhadi et al. (2010) propose a system that maps images and the corresponding textual descriptions to a "meaning" space which consists of an object, action and scene triplet. However, they assume a single object per image and do not use text-mining to determine the likelihood of objects matching different verbs. Yang et al. (2011) is the most similar to our approach in that it uses text-mined knowledge to generate sentential descriptions of static images after performing object and scene detection. However, they do not perform activity recognition nor use text-mining to select the best verb.

Approach

Our overall approach is illustrated in Figure 2 and consists of visual object and activity recognition followed by content-planning to generate the best SVO triple and surface realization to generate the final sentence.

Dataset

We used the English portion of the YouTube data collected by Chen et al. (2010), consisting of short videos each with

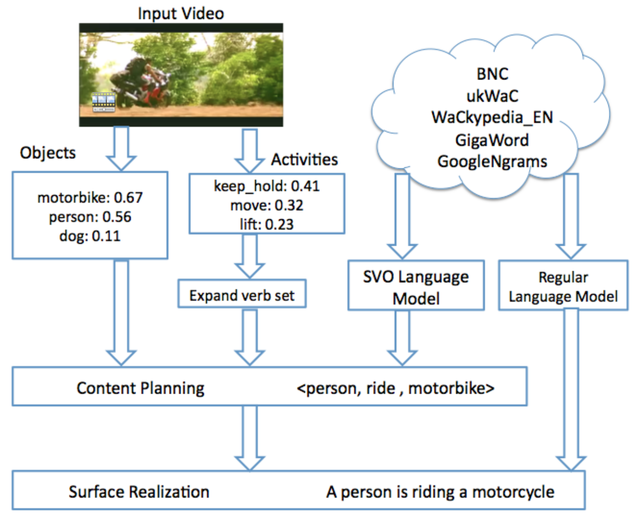

Figure 2: Summary of our approach

Figure 3: Activity clusters discovered by HAC

multiple natural-language descriptions. This data was previously used by Motwani and Mooney (2012), and like them, we ensured that the test data only contained videos in which we can potentially detect objects. We used Felzenswalb's (2008) object detector as it achieves the state-of-the-art performance on the PASCAL Visual Object Classes (VOC) Challenge. As such, we selected test videos whose subjects and objects belong to the 20 VOC object classes - *aeroplane, car, horse, sheep, bicycle, cat, sofa, bird, chair, motorbike, train, boat, cow, person, tv monitor, bottle, dining table, bus, dog, potted plant*. During this filtering, we also allow synonyms of these object names by including all words with a Lesk similarity (as implemented by Pedersen et al. (2004)) of at least 0.5.[1] Using this approach, we chose 235 potential test videos; the remaining 1,735 videos were reserved for training.

All the published activity recognition methods that work on datasets such as KTH (Schuldt, Laptev, and Caputo 2004), Drinking and Smoking (Laptev and Perez 2007) and

[1]Empirically, this method worked better than using WordNet synsets.

UCF50 (Reddy and Shah 2012) have a very limited recognition vocabulary of activity classes. Since we did not have explicit activity labels for our YouTube videos, we followed Motwani and Mooney (2012)'s approach to automatically discover activity clusters. We first parsed the training descriptions using Stanford's dependency parser (De Marneffe, MacCartney, and Manning 2006) to obtain the set of verbs describing each video. We then clustered these verbs using Hierarchical Agglomerative Clustering (HAC) using the *res* metric from WordNet::Similarity (Pedersen, Patwardhan, and Michelizzi 2004) to measure the distance between verbs. By manually cutting the resulting hierarchy at a desired level (ensuring that each cluster has at least 9 videos), we discovered the 58 activity clusters shown in Figure 3. We then filtered the training and test sets to ensure that all verbs belonged to these 58 activity clusters. The final data contains 185 test and 1,596 training videos.

Object Detection

We used Felzenszwalb's (2008) discriminatively-trained deformable parts models to detect the most likely objects in each video. Since these object detectors were designed for static images, each video was split into frames at one-second intervals. For each frame, we ran the object detectors and selected the maximum score assigned to each object in any of the frames. We converted the detection scores, $f(x)$, to estimated probabilities $p(x)$ using a sigmoid $p(x) = \frac{1}{1+e^{-f(x)}}$.

Activity Recognition

In order to get an initial probability distribution for activities detected in the videos, we used the motion descriptors developed by Laptev et al. (2008). Their approach extracts spatio-temporal interest points (STIPs) from which it computes HoG (Histograms of Oriented Gradients) and HoF (Histograms of Optical Flow) features over a 3-dimensional space-time volume. These descriptors are then randomly sampled and clustered to obtain a "bag of visual words," and each video is then represented as a histogram over these clusters. We experimented with different classifiers such as LIBSVM (Chang and Lin 2011) to train a final activity detector using these features. Since we achieved the best classification accuracy (still only 8.65%) using an SVM with the intersection kernel, we used this approach to obtain a probability distribution over the 58 activity clusters for each test video. We later experimented with Dense Trajectories (Wang et al. 2011) for activity recognition but there was only a minor improvement.

Text Mining

We improve these initial probability distributions over objects and activities by incorporating the likelihood of different activities occuring with particular subjects and objects using two different approaches. In the first approach, using the Stanford dependency parser, we parsed 4 different text corpora covering a wide variety of text: English Gigaword, British National Corpus (BNC), ukWac and WaCkypedia_EN. In order to obtain useful estimates, it is essential to collect text that approximates all of the written language in scale and distribution. The sizes of these corpora (after preprocessing) are shown in Table 1.

Corpora	Size of text
British National Corpus (BNC)	1.5GB
WaCkypedia_EN	2.6GB
ukWaC	5.5GB
Gigaword	26GB
GoogleNgrams	10^{12} words

Table 1: Corpora used to Mine SVO Triplets

Using the dependency parses for these corpora, we mined SVO triplets. Specifically, we looked for subject-verb relationships using *nsubj* dependencies and verb-object relationships using *dobj* and *prep_* dependencies. The *prep_* dependency ensures that we account for intransitive verbs with prepositional objects. Synonyms of subjects and objects and conjugations of verbs were reduced to their base forms (20 object classes and 58 activity clusters) while forming triplets. If a subject, verb or object not belonging to these base forms is encountered, it is ignored during triplet construction.

These triplets are then used to train a backoff language model with Kneser-Ney smoothing (Chen and Goodman 1999) for estimating the likelihood of an SVO triple. In this model, if we have not seen training data for a particular SVO trigram, we "back-off" to the Subject-Verb and Verb-Object bigrams to coherently estimate its probability. This results in a sophisticated statistical model for estimating triplet probabilities using the syntactic context in which the words have previously occurred. This allows us to effectively determine the real-world plausibility of any SVO using knowledge automatically mined from raw text. We call this the "SVO Language Model" approach (SVO LM).

In a second approach to estimating SVO probabilities, we used BerkeleyLM (Pauls and Klein 2011) to train an n-gram language model on the GoogleNgram corpus (Lin et al. 2012). This simple model does not consider synonyms, verb conjugations, or SVO dependencies but only looks at word sequences. Given an SVO triplet as an input sequence, it estimates its probability based on n-grams. We refer to this as the "Language Model" approach (LM).

Verb Expansion

As mentioned earlier, the top activity detections are expanded with their most similar verbs in order to generate a larger set of potential words for describing the action. We used the WUP metric from WordNet::Similarity to expand each activity cluster to include all verbs with a similarity of at least 0.5. For example, we expand the verb "move" with *go 1.0, walk 0.8, pass 0.8, follow 0.8, fly 0.8, fall 0.8, come 0.8, ride 0.8, run 0.67, chase 0.67, approach 0.67*, where the number is the WUP similarity.

Content Planning

To combine the vision detection and NLP scores and determine the best overall SVO, we use simple linear interpolation as shown in Equation 1. When computing the overall vision score, we make a conditional independence assumption and multiply the probabilities of the subject, activity and object. To account for expanded verbs, we additionally multiply by the WUP similarity between the original and expanded verbs. The NLP score is obtained from either the "SVO Language Model" or the "Language Model" approach, as previously described.

$$score = w_1 * vis\_score + w_2 * nlp\_score \tag{1}$$

$$\begin{aligned} vis\_score = P(S/vid) * P(V_{sim}/vid) \\ * Sim(V_{sim}, V_{orig}) * P(O/vid) \end{aligned} \tag{2}$$

After determining the top *n=5* object detections and top *k=10* verb detections for each video, we generate all possible SVO triplets from these nouns and verbs, including all potential verb expansions. Each resulting SVO is then scored using Equation 1, and the best is selected. We compare this approach to a "pure vision" baseline where the subject is the highest scored object detection (which empirically is more likely to be the subject than the object), the object is the second highest scored object detection, and the verb is the activity cluster with the highest detection probability.

Surface Realization

Finally, the subject, verb and object from the top-scoring SVO are used to produce a set of candidate sentences, which are then ranked using a language model. The text corpora in Table 1 are mined again to get the top three prepositions for every verb-object pair. We use a template-based approach in which each sentence is of the form:

"Determiner (A,The) - Subject - Verb (Present, Present Continuous) - Preposition (optional) - Determiner (A,The) - Object."

Using this template, a set of candidate sentences are generated and ranked using the BerkeleyLM language model trained on the GoogleNgram corpus. The top sentence is then used to describe the video. This surface realization technique is used for both the vision baseline triplet and our proposed triplet.

In addition to the one presented here, we tried alternative "pure vision" baselines, but they are not included since they performed worse. We tried a non-parametric approach similar to Ordonez, Kulkarni, and Berg (2011), which computes global similarity of the query to a large captioned dataset and returns the nearest neighbor's description. To compute the similarity we used an RBF-Chi^2 kernel over bag-of-words STIP features. However, as noted by Ordonez, Kulkarni, and Berg (2011), who used 1 million Flickr images, our dataset is likely not large enough to produce good matches. In an attempt to combine information from both object and activity recognition, we also tried combining object detections from 20 PASCAL object detectors (Felzenszwalb, McAllester, and Ramanan 2008) and from Object Bank (Li et al. 2010) using a multi-channel approach as proposed in (Zhang et al. 2007), with a RBF-Chi^2 kernel for the STIP features and a RBF-Correlation Distance kernel for object detections.

Method	Subject%	Verb%	Object%	All%
Vision Baseline	71.35	8.65	29.19	1.62
LM(VE)	71.35	8.11	10.81	0.00
SVO LM(NVE)	85.95	16.22	24.32	11.35
SVO LM(VE)	85.95	**36.76**	**33.51**	**23.78**

Table 2: SVO Triplet accuracy: Binary metric

Method	Subject%	Verb%	Object%	All%
Vision Baseline	87.76	40.20	61.18	63.05
LM(VE)	85.77	53.32	61.54	66.88
SVO LM(NVE)	94.90	63.54	69.39	75.94
SVO LM(VE)	94.90	**66.36**	**72.74**	**78.00**

Table 3: SVO Triplet accuracy: WUP metric

Experimental Results

Content Planning

We first evaluatated the ability of the system to identify the best SVO content. From the ~ 50 human descriptions available for each video, we identified the SVO for each description and then determined the ground-truth SVO for each of the 185 test videos using majority vote. These verbs were then mapped back to their 58 activity clusters. For the results presented in Tables 2 and 3, we assigned the vision score a weight of 0 ($w_1 = 0$) and the NLP score a weight of 1 ($w_2 = 1$) since these weights gave us the best performance for thresholds of 5 and 10 for the objects and activity detections respectively. Note that while the vision score is given a weight of zero, the vision detections still play a vital role in the determination of the final triplet since our model only considers the objects and activities with the highest vision detection scores.

To evaluate the accuracy of SVO identification, we used two metrics. The first is a binary metric that requires exactly matching the gold-standard subject, verb and object. Its results are shown in Table 2, where VE and NVE stand for "verb expansion" and "no verb expansion" respectively. However, the binary evaluation can be unduly harsh. If we incorrectly choose "bicycle" instead of a "motorbike" as the object, it should be considered better than choosing "dog." Similarly, predicting "chop" instead of "slice" is better than choosing "go". In order to account for such similarities, we also measure the WUP similarity between the predicted and correct items. For the examples above, the relevant scores are: *wup(motorbike,bicycle)=0.7826, wup(motorbike,dog)=0.1, wup(slice,chop)=0.8, wup(slice,go)=0.2857*. The results for the WUP metric are shown in Table 3.

Surface Realization

Figures 4 and 5 show examples of both good and bad sentences generated by our method compared to the vision baseline.

Automatic Metrics To automatically compare the sentences generated for the test videos to ground-truth human descriptions, we employed the BLEU and METEOR metrics

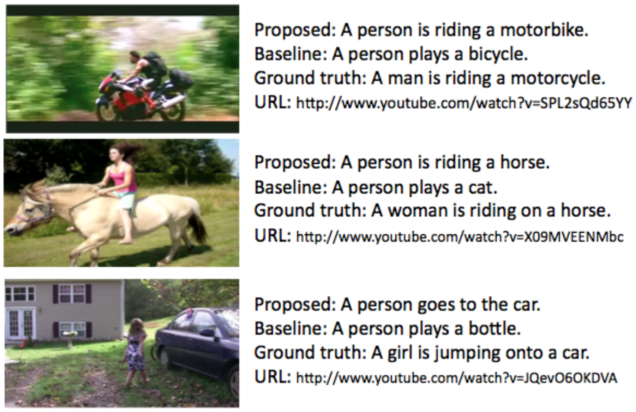

Figure 4: Examples where we outperform the baseline

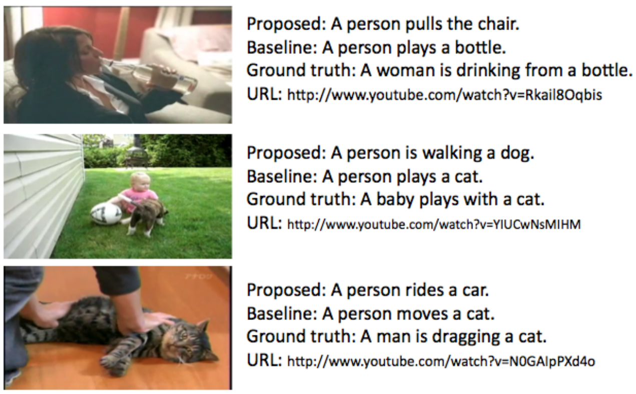

Figure 5: Examples where we underperform the baseline

used to evaluate machine-translation output. METEOR was designed to fix some of the problems with the more popular BLEU metric. They both measure the number of matching n-grams (for various values of n) between the automatic and human generated sentences. METEOR takes stemming and synonymy into consideration. We used the SVO Language Model (with verb expansion) approach since it gave us the best results for triplets. The results are given in Table 4.

Human Evaluation using Mechanical Turk Given the limitations of metrics like BLEU and METEOR, we also asked human judges to evaluate the quality of the sentences generated by our approach compared to those generated by the baseline system. For each of the 185 test videos, we asked 9 unique workers (with >95% HIT approval rate and who had worked on more than 1000 HITs) on Amazon Mechanical Turk to pick which sentence better described the video. We also gave them a "none of the above two sentences" option in case neither of the sentences were relevant to the video. Quality was controlled by also including in each HIT a gold-standard example generated from the human descriptions, and discarding judgements of workers who incorrectly answered this gold-standard item. Overall, when they expressed a preference, humans picked our descriptions to that of the baseline **61.04**% of the time. Out of the 84 videos where the majority of judges had a clear preference, they chose our descriptions **65.48%** of the time.

Method	BLEU score	METEOR score
Vision Baseline	0.37±0.05	0.25±0.08
SVO LM(VE)	0.45±0.05	0.36±0.27

Table 4: Automatic evaluation of sentence quality

Discussion

Overall, the results consistently show the advantage of utilizing text-mined knowledge to improve the selection of an SVO that best describes a video. Below we discuss various specific aspects of the results.

Vision Baseline: For the vision baseline, the subject accuracy is quite high compared to the object and activity accuracies. This is likely because the person detector has higher recall and confidence than the other object detectors. Since most test videos have a person as the subject, this works in favor of the vision baseline, as typically the top object detection is "person". Activity (verb) accuracy is quite low (8.65% binary accuracy). This is because there are 58 activity clusters, some with very little training data. Object accuracy is not as high as subject accuracy because the true object, while usually present in the top object detections, is not always the second-highest object detection. By allowing "partial credit", the WUP metric increases the verb and object accuracies to 40.2% and 61.18%, respectively.

Language Model(VE): The Language Model approach performs even worse than the vision baseline especially for object identification. This is because we consider the language model score directly for the SVO triplet without any object synonyms, verb conjugations and presence of determiners between the verb and object. For example, while the GoogleNgram corpus is likely to contain many instances of a sentence like "A person is walking with a dog", it will probably not contain many instances of "person walk dog", resulting in lower scores.

SVO Language Model(NVE): The SVO Language Model (without verb expansion) improves verb accuracy from 8.65% to 16.22%. For the WUP metric, we see an improvement in accuracy in all cases. This indicates that we are getting semantically closer to the right object compared to the object predicted by the vision baseline.

SVO Language Model(VE): When used with verb expansion, the SVO Language Model approach results in a dramatic improvement in verb accuracy, causing it to jump to 36.76%. The increase in WUP score for verbs is relatively minor between SVO Language Model(VE) and SVO Language Model(NVE). This is because even without verb expansion, semantically similar verbs are selected but not the one used in most human descriptions. So, the jump in verb accuracy for the binary metric is much more than the one for the WUP metric.

Importance of verb expansion: Verb expansion clearly improves activity accuracy. This idea could be extended to a scenario where the test set contains many activities for which we do not have any explicit training data. As such, we

Method	Subject%	Verb%	Object%	All%
Vision Baseline	71.35	8.65	29.19	1.62
Train Desc.	**85.95**	16.22	16.22	8.65
Gigaword	**85.95**	32.43	20.00	14.05
BNC	**85.95**	17.30	29.73	14.59
ukWaC	**85.95**	34.05	32.97	22.16
WaCkypedia_EN	**85.95**	35.14	**40.00**	**28.11**
All	**85.95**	**36.76**	33.51	23.78

Table 5: Effect of training corpus on SVO binary accuracy

Method	Subject%	Verb%	Object%	All%
Vision Baseline	87.76	40.20	61.18	63.05
Train Desc.	**94.95**	45.12	61.43	67.17
Gigaword	94.90	63.99	65.71	74.87
BNC	94.88	51.48	73.93	73.43
ukWaC	94.86	60.59	72.83	76.09
WaCkypedia_EN	94.90	62.52	**76.48**	77.97
All	94.90	**66.36**	72.74	**78.00**

Table 6: Effect of training corpus on SVO WUP accuracy

cannot train activity classifiers for these "missing" classes. However, we can train a "coarse" activity classifier using the training data that is available, get the top predictions from this coarse classifier and then refine them by using verb expansion. Thus, we can even detect and describe activities that were unseen at training time by using text-mined knowledge to determine the description of an activity that best fits the detected objects.

Effect of different training corpora: As mentioned earlier, we used a variety of textual corpora. Since they cover newswire articles, web pages, Wikipedia pages and neutral content, we compared their individual effect on the accuracy of triplet selection. The results of this ablation study are shown in Tables 5 and 6 for the binary and WUP metric respectively. We also show results for training the SVO model on the descriptions of the training videos. The WaCkypedia_EN corpus gives us the best overall results, probably because it covers a wide variety of topics, unlike Gigaword which is restricted to the news domain. Also, using our SVO Language Model approach on the triplets from the descriptions of the training videos is not sufficient. This is because of the relatively small size and narrow domain of the training descriptions in comparison to the other textual corpora.

Effect of changing the weight of the NLP score We experimented with different weights for the Vision and NLP scores (in Equation 1). These results can be seen in Figure 6 for the binary-metric evaluation. The WUP-metric evaluation graph is qualitatively similar. A general trend seems to be that the subject and activity accuracies increase with increasing weights of the NLP score. There is a significant improvement in verb accuracy as the NLP weight is increased towards 1. However, for objects we notice a slight increase in accuracy until the weight for the NLP component is 0.9 after which there is a slight dip. We hypothesize that this dip is caused by the loss of vision-based information about the objects which provide some guidance for the NLP system.

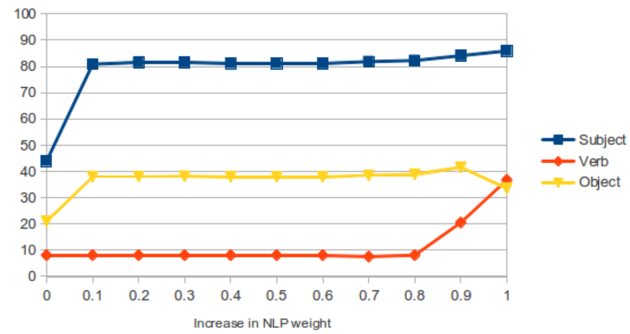

Figure 6: Effect of increasing NLP weights (Binary metric)

BLEU and METEOR results: From the results in Table 4, it is clear that the sentences generated by our approach outperform those generated by the vision baseline, using both the BLEU and METEOR evaluation metrics.

MTurk results: The Mechanical Turk results show that human judges generally prefer our system's sentences to those of the vision baseline. As previously seen, our method improves verbs far more than it improves subjects or objects. We hypothesize that the reason we do not achieve a similarly large jump in performance in the MTurk evaluation is because people seem to be more influenced by the object than the verb when both options are partially irrelevant. For example, in a video of a person riding his bike onto the top of a car, our proposed sentence was "A person is a riding a motorbike" while the vision sentence was "A person plays a car", and most workers selected the vision sentence.

Drawback of Using YouTube Videos: YouTube videos often depict unusual and "interesting" events, and these might not agree with the statistics on typical SVOs mined from text corpora. For instance, the last video in Figure 5 shows a person dragging a cat on the floor. Since sentences describing people moving or dragging cats around are not common in text corpora, our system actually down-weights the correct interpretation.

Conclusion

This paper has introduced a holistic data-driven approach for generating natural-language descriptions of short videos by identifying the best subject-verb-object triplet for describing realistic YouTube videos. By exploiting knowledge mined from large corpora to determine the likelihood of various SVO combinations, we improve the ability to select the best triplet for describing a video and generate descriptive sentences that are prefered by both automatic and human evaluation. From our experiments, we see that linguistic knowledge significantly improves activity detection, especially when training and test distributions are very different, one of the advantages of our approach. Generating more complex sentences with adjectives, adverbs, and multiple objects and multi-sentential descriptions of longer videos with multiple activities are areas for future research.

References

Barbu, A.; Bridge, A.; Burchill, Z.; Coroian, D.; Dickinson, S.; Fidler, S.; Michaux, A.; Mussman, S.; Narayanaswamy, S.; Salvi, D.; et al. 2012. Video in sentences out. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 102–12.

Chang, C., and Lin, C. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27.

Chen, S., and Goodman, J. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13(4):359–393.

Chen, D.; Dolan, W.; Raghavan, S.; Huynh, T.; Mooney, R.; Blythe, J.; Hobbs, J.; Domingos, P.; Kate, R.; Garrette, D.; et al. 2010. Collecting highly parallel data for paraphrase evaluation. *JAIR* 37:397–435.

De Marneffe, M.; MacCartney, B.; and Manning, C. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, 449–454.

Ding, D.; Metze, F.; Rawat, S.; Schulam, P.; Burger, S.; Younessian, E.; Bao, L.; Christel, M.; and Hauptmann, A. 2012. Beyond audio and video retrieval: towards multimedia summarization. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, 2. ACM.

Farhadi, A.; Hejrati, M.; Sadeghi, M.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. *Computer Vision–ECCV 2010* 15–29.

Felzenszwalb, P.; McAllester, D.; and Ramanan, D. 2008. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Khan, M., and Gotoh, Y. 2012. Describing video contents in natural language. *EACL 2012* 27.

Kojima, A.; Tamura, T.; and Fukunaga, K. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision* 50(2):171–184.

Kulkarni, G.; Premraj, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.; and Berg, T. 2011. Baby talk: Understanding and generating simple image descriptions. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1601–1608. IEEE.

Laptev, I., and Perez, P. 2007. Retrieving actions in movies. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 1–8. IEEE.

Laptev, I.; Marszalek, M.; Schmid, C.; and Rozenfeld, B. 2008. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Lee, M.; Hakeem, A.; Haering, N.; and Zhu, S. 2008. Save: A framework for semantic annotation of visual events. In *Computer Vision and Pattern Recognition Workshops, 2008*, 1–8. IEEE.

Li, L.; Su, H.; Xing, E.; and Fei-Fei, L. 2010. Object bank: A high-level image representation for scene classification and semantic feature sparsification. *Advances in Neural Information Processing Systems* 24.

Li, S.; Kulkarni, G.; Berg, T.; Berg, A.; and Choi, Y. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 220–228. Association for Computational Linguistics.

Lin, Y.; Michel, J.; Aiden, E.; Orwant, J.; Brockman, W.; and Petrov, S. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Motwani, T., and Mooney, R. 2012. Improving video activity recognition using object recognition and text mining. In *European Conference on Artificial Intelligence*. ECAI.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of NIPS*.

Packer, B.; Saenko, K.; and Koller, D. 2012. A combined pose, object, and feature model for action understanding. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 1378–1385. IEEE.

Pauls, A., and Klein, D. 2011. Faster and smaller n-gram language models. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 258–267.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, 38–41. Association for Computational Linguistics.

Reddy, K., and Shah, M. 2012. Recognizing 50 human action categories of web videos. *Machine Vision and Applications* 1–11.

Schuldt, C.; Laptev, I.; and Caputo, B. 2004. Recognizing human actions: A local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, 32–36. IEEE.

Wang, H.; Klaser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 3169–3176. IEEE.

Yang, Y.; Teo, C. L.; Daumé, III, H.; and Aloimonos, Y. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 444–454. Association for Computational Linguistics.

Yao, B., and Fei-Fei, L. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yao, B.; Yang, X.; Lin, L.; Lee, M.; and Zhu, S. 2010. I2t: Image parsing to text description. *Proceedings of the IEEE* 98(8):1485–1508.

Zhang, J.; Marszałek, M.; Lazebnik, S.; and Schmid, C. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision* 73(2):213–238.