A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching

Pradipto Das*, Chenliang Xu*, Richard F. Doell and Jason J. Corso
Computer Science and Engineering, SUNY at Buffalo
{pdas3,chenlian,rfdoell,jcorso}@buffalo.edu

Abstract

The problem of describing images through natural language has gained importance in the computer vision community. Solutions to image description have either focused on a top-down approach of generating language through combinations of object detections and language models or bottom-up propagation of keyword tags from training images to test images through probabilistic or nearest neighbor techniques. In contrast, describing videos with natural language is a less studied problem. In this paper, we combine ideas from the bottom-up and top-down approaches to image description and propose a method for video description that captures the most relevant contents of a video in a natural language description. We propose a hybrid system consisting of a low level multimodal latent topic model for initial keyword annotation, a middle level of concept detectors and a high level module to produce final lingual descriptions. We compare the results of our system to human descriptions in both short and long forms on two datasets, and demonstrate that final system output has greater agreement with the human descriptions than any single level.

1. Introduction

The problem of generating natural language descriptions of images and videos has been steadily gaining prominence in the computer vision community. A number of papers have been proposed to leverage latent topic models on low-level features [4, 6, 7, 22, 32], for example. The problem is important for three reasons: i) transducing visual data into textual data would permit well understood text-based indexing and retrieval mechanisms essentially *for free*; ii) fine grained object models and region labeling introduce a new level of semantic richness to multimedia retrieval techniques; and iii) grounding representations of visual data in natural language has great potential to overcome the inherent semantic ambiguity prominent in the data-driven high-level vision community (see [27] for a discussion of data-set bias and discussion on the different *meanings* common labels can have within and across data sets).

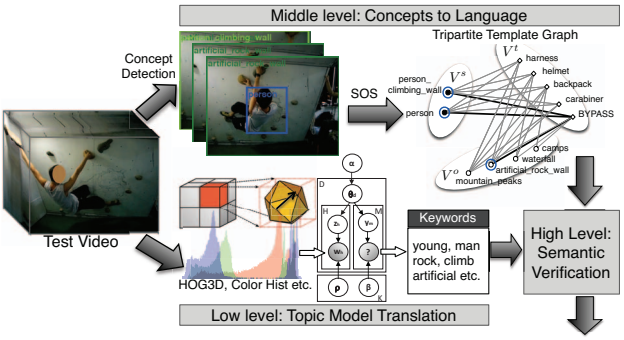


Figure 1: A framework of our hybrid system showing a video being processed through our pipeline and described by a few natural language sentences.

Fig. 1 shows our video to text system pipeline. To date, the most common approach to such lingual description of images has been to model the joint distribution over low-level image features and language, typically nouns. Early work on multimodal topic models by Blei et al. [4] and subsequent extensions [6, 7, 11, 22, 32] jointly model image features (predominantly SIFT and HOG derivatives) and language words as mixed memberships over latent topics with considerable success. Other non-parametric nearest-neighbor and label transfer methods, such as Makadia et al. [18] and TagProp [12], rely on large annotated sets to generate descriptions from similar samples. These methods have demonstrated a capability of lingual description on images at varying levels, but they have two main limitations. Being based on low-level features and/or similarity measures, first, it is not clear they can scale up as the richness of the semantic space increases. Second, the generated text has largely been in the form of word-lists without any semantic verification (see Sec. 2.3).

*Pradipto Das and Chenliang Xu contributed equally to this paper.

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.340

Alternatively, a second class of approaches to lingual description of images directly seeks a set of high-level concepts, typically objects but possibly others such as scene categories. Prominent among object detectors is the deformable parts model (DPM) [10] and related visual phrases [26] which have been successful in the task of "annotating" natural images. Despite being able to guarantee the semantic veracity of the generated lingual description, these methods have found limited use due to the overall complexity of object detection *in-the-wild* and its constituent limitations (i.e., noisy detection), and the challenge of enumerating all relevant world concepts and learning a detector for each.

In this work, we propose a hybrid model that takes the best characteristics of these two classes of methods. Namely, our model leverages the power of low-level joint distributions over video features and language by treating them as a set of lingual proposals which are subsequently filtered by a set of mid-level concept detectors. A test video is processed in three ways (see Fig. 1). First, in a bottom up fashion, low level video features predict keywords. We use multimodal latent topic models to find a proposal distribution over some training vocabulary of textual words [4, 7], then select the most probable keywords as potential subjects, objects and verbs through a natural language dependency grammar and part-of-speech tagging.

Second, in a top down fashion, we detect and stitch together a set of concepts, such as "artificial rock wall" and "person climbing wall" similar to [26], which are then converted to lingual descriptions through a tripartite graph template. Third, for high level semantic verification, we relate the predicted caption keywords with the detected concepts to produce a ranked set of well formed natural language sentences. Our semantic verification step is independent of any computer vision framework and works by measuring the number of inversions between two ranked lists of predicted keywords and detected concepts both being conditional on their respective learned topic multinomials.

Our method does not suffer from any lack of semantic verification as bottom-up models do, nor does it suffer from the tractability challenges of the top-down methods—it can rely on fewer well-trained concept detectors for verification allowing the correlation between different concepts to replace the need for a vast set of concept detectors.

Videos vs. Images Recent work in [9, 16, 34] is mainly focused on generating fluent descriptions of a single image—images not videos. Videos introduce an additional set of challenges such as temporal variation/articulation and dependencies. Most related work in vision has focused only on the activity classification side: example methods using topic models for activities are the hidden topic Markov model [33] and frame-by-frame Markov topic models [13], but these methods do not model language and visual topics jointly. A recent activity classification paper of relevance is the Action Bank method [25], which ties high-level actions to constituent low-level action detections, but it does not include any language generation framework.

The three most relevant works to ours are the Khan et al. [14], Barbu et al. [1] and Malkarnenkar et al. [19] systems. All of these methods extract high-level concepts, such as faces, humans, tables, etc., and generate language description by template filling; [19] additionally uses externally mined language data to help rank the best subject-verb-object triplet. The methods rely directly on all high-level concepts being enumerated (the second class of methods introduced above) and hence may be led astray by noisy detection and have a limited vocabulary, unlike our approach which not only uses the high-level concepts but augments them with a large corpus of lingual descriptions from the bottom-up. Furthermore, some have used datasets have simpler videos not *in-the-wild*.

We, in contrast, focus on descriptions of general videos (e.g., from YouTube) *directly* through bottom-up visual feature translations to text and top-down concept detections. We leverage both detailed object annotations and human lingual descriptions. Our proposed hybrid method shows more relevant content generation over simple keyword annotation of videos alone as observed using quantitative evaluation on two datasets—the TRECVID dataset [20] and a new in-house dataset consisting of cooking videos collected from YouTube with human lingual descriptions generated through MTurk (Sec. 3).

2. System Description

2.1. Low Level: Topic Model

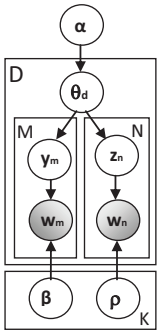


Figure 2: Low-level topic model.

Following [7], we adapt the GM-LDA model in [4] (dubbed MMLDA for MultiModalLDA in this paper) to handle a discrete visual feature space, e.g., we use HOG3D [15]. The original model in [4] is defined in the continuous space, which presents challenges for discrete features: it can become unstable during deterministic approximate optimization due to extreme values in high-dimensions and its inherent non-convexity [30]. We briefly explain the model and demonstrate how it is instantiated and differs from the original version in [4]. First, we use an asymmetric Dirichlet prior, $\boldsymbol{\alpha}$ for the document level topic proportions $\boldsymbol{\theta}_d$ following [31] unlike the symmetric one in [4]. In Fig. 2, D is the number of documents, each consisting of a video and a lingual description (the text is only available during training). The number of discrete visual words and lingual words per video document d are N and M. The parameters for corpus level topic multinomials over visual words are $\boldsymbol{\rho}_{1:K}$. The param-

eters for corpus level topic multinomials over textual words are $\boldsymbol{\beta}_{1:K}$—only the training instances of these parameters are used for keyword prediction. The indicator variables for choosing a topic are $\{z_{d,n}\}$ and $\{y_{d,m}\}$; $w_{d,m}$ is the text word at position m in video "document" d with vocabulary size V. Each $w_{d,n}$ is a visual feature from a bag-of-discrete-visual-words at position n with vocabulary size $corrV$ and each $w_{d,n}$ represents a visual word (e.g., HOG3D [15] index, transformed color histogram [28], etc.).

We use the mean field method of optimizing a lower-bound to the true likelihood of the data. A fully factorized q distribution with "free" variational parameters $\boldsymbol{\gamma}$, $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$ is imposed by: $q(\boldsymbol{\theta}, \mathbf{z}, \mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\lambda}) =$

$$\prod_{d=1}^{D} q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \left[\prod_{n=1}^{N_d} q(z_{d,n}|\boldsymbol{\phi}_{d,n}) \prod_{m=1}^{M_d} q(y_{d,m}|\boldsymbol{\lambda}_{d,m})\right]. \quad (1)$$

The optimal values of free variables and parameters are found by optimizing the lower bound on $\log p(\mathbf{w_M}, \mathbf{w_N}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho})$. The free multinomial parameters of the variational topic distributions ascribed to the corresponding data are $\boldsymbol{\phi}_d$s. The free parameters of the variational word-topic distribution are $\boldsymbol{\lambda}_d$s. The surrogate for the K-dimensional $\boldsymbol{\alpha}$ is $\boldsymbol{\gamma}_d$ which represents the expected number of observations per document in each topic. The free parameters are defined for every video document d. The optimal value expressions of the hidden variables in video document d for the MMLDA model are as follows:

$$\phi_{n,i} \propto \exp\left\{\psi(\gamma_i) + \log \rho_{i,w_{d,n}}\right\} \quad , \quad (2)$$

$$\lambda_{m,i} \propto \exp\left\{\psi(\gamma_i) + \log \beta_{i,w_{d,m}}\right\} \quad , \quad (3)$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N_d} \phi_{n,i} + \sum_{m=1}^{M_d} \lambda_{m,i} \quad , \quad (4)$$

where ψ is the digamma function. The expressions for the maximum likelihood of the topic parameters are:

$$\rho_{i,j} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{j=1}^{corrV} \phi_{d,n,i} \delta(w_{d,n}, j) \quad , \quad (5)$$

$$\beta_{i,j} = \sum_{d=1}^{D} \sum_{m=1}^{M_d} \sum_{j=1}^{V} \lambda_{d,m,i} \delta(w_{d,m}, j) \quad . \quad (6)$$

The asymmetric $\boldsymbol{\alpha}$ is optimized using the formulations given in [5], which incorporates Newton steps as search directions in gradient ascent.

A strongly constrained model, Corr-LDA, is also introduced in [4] that uses real valued visual features and shows promising image annotation performance. We have experimented with the model to use our discrete visual feature space (and name it Corr-MMLDA) but finally opt to not use it in our final experiments due to the following reasons.

K=200	n=1	n=5	n=10	n=15
MMLDA	0.03518	0.11204	0.18700	0.24117
Corr-MMLDA	0.03641	0.11063	0.18406	0.24840

Table 1: Average word prediction 1-gram recall for different topic models with 200 topics when the full corpus is used. The numbers are slightly lower for lower number of topics but are not statistically significant.

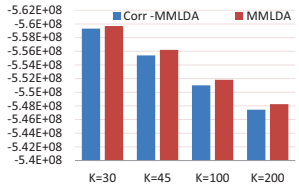


Figure 3: Prediction ELBOs from the two topic models for the videos in TRECVID dataset. Lower is better.

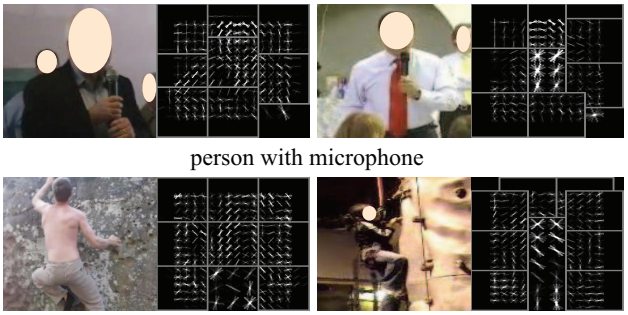
The correspondence between $w_{d,m}$ and $z_{d,n}$ necessitates checking for correspondence strengths over all possible dependencies between $w_{d,m}$ and $w_{d,n}$. This assumption is relaxed in the MMLDA model and removes the bottleneck in runtime efficiency for high dimensional video features without showing significant performance drain. Fig. 3 shows the held out log likelihoods or the Evidence Lower BOunds (ELBOs) on part of the TRECVID dataset (Sec. 3.1). The figures are obtained by topic modeling on the entire corpus of multimedia documents (video with corresponding lingual description). Using visual features, we predict the top n words as the description of the test videos. Table 1 shows the average 1-gram recall of predicted words (as in [14]). We observe that both models have approximately the same fit and word prediction power, and hence choose the MMLDA model since it is computationally less expensive.

2.2. Middle Level: Concepts to Language

The middle level is a top-down approach that detects concepts sparsely throughout the video, matches them over time, which we call *stitching*, and relates them to a tripartite template graph for generating language output.

2.2.1 Concept Detectors

Instead of using publicly available object detectors from datasets like the PASCAL VOC [8], or training independent object detectors for objects such as *microphone*, we build the concept object detectors like *microphone with upper body*, *group of people* etc., where multiple objects together form a single concept. A concept detector captures richer semantic information (from object, action and scene level) than object detectors, and usually reduces the visual complexity compared to individual objects, which requires less training examples for an accurate detector. These concept detectors are closely related to Sadeghi and Farhadi's visual phrases [26] but do not use any decoding process and



person with microphone

person climbing wall

Figure 4: Examples of DPM based concept detectors.

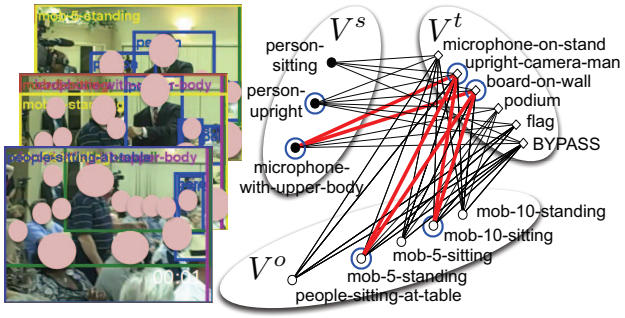


Figure 5: Lingual descriptions from tripartite template graphs consisting of concepts as vertices.

are applied on video.

We use the deformable parts model (DPM) [10] for the concept detectors, some examples of which are visualized in Fig. 4. The specific concepts we choose are based on the most frequently occurring object-groupings in the human descriptions from the training videos. We use the VATIC tool [29] to annotate the trajectories of concept detectors in training videos, which are also used in Sec. 2.2.3 for extracting concept relations.

2.2.2 Sparse Object Stitching (SOS)

Concept detectors act as a proxy to the trajectories being tracked in a video. However, tracking over detection is a challenging and open problem for videos *in-the-wild*. First, camera motion and the frame rate are unpredictable, rendering the existing tracking methods useless. Second, the scale of our dataset is huge (thousands of video hours), and we hence need a fast alternative. Our approach is called *sparse object stitching*; we sparsely obtain the concept detections in a video and then sequentially group frames based on commonly detected concepts.

For a given video, we run the set of concept detectors $\mathcal{L}$ on T sparsely distributed frames (e.g. 1 frame/sec) and denote the set of positive detections on each frame as $\mathcal{D}_i$. The algorithm tries to segment the video into a set of concept shots $\mathcal{S} = \{S_1, S_2, \ldots, S_Z\}$, where $\mathcal{S} = \cup\mathcal{D}_i$, and $Z \ll T$, so that each S_j can be independently described by some sparse detections similar in spirit to [14]. We start by uniformly splitting the video into K proposal shots $\{S'_1, S'_2, \ldots, S'_K\}$. Then we greedily traverse the proposed shots one by one considering neighboring shots S'_k and S'_{k+1}. If the Jaccard distance $J(S'_k, S'_{k+1}) = 1 - \frac{|S'_k \cap S'_{k+1}|}{|S'_k \cup S'_{k+1}|}$ is lower than a threshold σ (set as 0.5 using cross-validation), then we merge these two proposed shots into one shot and compare it with the next shot, otherwise shot S'_k is an independent shot. For each such concept shot, we match it to a tripartite template graph and translate it to language, as we describe next.

2.2.3 Tripartite Template Graph

We use a tripartite graph $\mathcal{G} = (V^s, V^t, V^o, E)$—$V^s$ for human subjects, V^t for tools, and V^o for objects—that takes the concept detections from each S_j and generates template-based language description. The vertex set $\mathcal{V} = V^s \cup V^t \cup V^o$ is identical to the set of concept detectors $\mathcal{L}$ in the domain at hand. Each concept detector is assigned to one of the three vertex sets (see Fig. 5). The set of paths $\mathcal{P} = \{(E_{\tau,\mu}, E_{\mu,\nu}) | \tau \in V^s, \mu \in V^t, \nu \in V^o\}$ is defined as all valid paths from V^s to V^o through V^t, and each forms a possible language output. However, we prune $\mathcal{P}$ so that it contains only those valid paths that were observed in the annotated training sequences. For a given domain, each such path, or triplet $\langle V^s, V^t, V^o \rangle$, instantiates a manually created template, such as "$\langle V^s \rangle$ *is cleaning* $\langle V^o \rangle$ *with* $\langle V^t \rangle$."

Language Output: Given the top confident concept detections $\mathcal{L}_c \subset \mathcal{L}$ in one concept shot S_j, we activate the set of paths $\mathcal{P}_c \subset \mathcal{P}$. A natural language sentence is output for paths containing a common subject using the template $\langle V^s, V^t, V^o \rangle$. For situations where $\mathcal{L}_c \cap V^t = \emptyset$, the consistency of the tripartite graph is maintained through a default "BYPASS" node in V^t (see Figs. 1 and 5). This node acts as a "backspace" production rule in the final lingual output thereby connecting the subject to an object effectively through a single edge. There is, similarly, a BYPASS node in V^o as well. In this paper, we generally do not consider the situation that $\mathcal{L}_c \cap V^s = \emptyset$, in which no human subject is present. Histogram counts are used for ranking the concept nodes for the lingual output.

Fig. 5 depicts a visual example of this process. The edges represent the action phrases or function words that stitch the concepts together cohesively. For example, consider the following structure: "([a person with microphone]) *is speaking to* ([a large group of sitting people] and [a small group of standing people]) *with* ([a camera man] and [board in the back])." Here the parentheses encapsulate a simple conjunctive production rule and the phrases inside the square brackets denote human subjects, tools or objects.

The edge labels in this case are "is speaking to" and "with" which are part of the template $\langle V^s, V^t, V^o \rangle$. In the figure, $\mathcal{L}_c$ is colored blue and edges in $\mathcal{P}_c$ with the common vertex "microphone-with-upper-body" are colored red. We delete repeated sentences in the final description.

2.3. High Level: Semantic Verification

The high level system joins the two earlier sets of lingual descriptions (from the low and middle levels) to enhance the set of sentences given from the middle level and at the same time to filter the sentences from the low level. Our method takes the predicted words from the low level and tags their part-of-speech (POS) with standard NLP tools. These are used to retrieve weighted nearest neighbors from the training descriptions, which are then ranked according to predictive importance, similar in spirit to how Farhadi et al. [9] select sentences. In contrast, we rank over semantically verified low level sentences, giving higher weight to shorter sentences and a fixed preference to middle level sentences.

We use the dependency grammar and part-of-speech (POS) models in the Stanford NLP Suite* to create annotated dictionaries based on word morphologies; the human descriptions provide the input. The predicted keywords from the low level topic models are labeled through these dictionaries. For more than two POS for the same morphology, we prefer verbs, but other variants can be retained as well without loss of generality. For the video in Fig. 5, we obtain the following labeled top 15 keywords: *"hall/OBJ town/NOUN meeting/VERB man/SUBJ-HUMAN speaks/VERB microphone/OBJ talking/VERB representative/SUBJ-HUMAN health/NOUN care/NOUN politician/SUBJ-HUMAN chairs/NOUN flags/OBJ people/OBJ crowd/OBJ."* The word annotation classes used are Subjects, Verbs, Objects, Nouns and "Other." Subjects which can be humans (SUBJ-HUMAN) are determined using WordNet synsets.

To obtain the final lingual description of a test video, the output from the middle level is used first. If there happen to be no detections, we rely only on the low-level generated sentences. For semantic verification, we train MMLDA on a vocabulary of training descriptions and training concept annotations available using VATIC. Then we compute the number of topic rank inversions for two ranked lists of the top P predictions and top C detections from a test video as:

$$L_{keywords} = \left\langle \left\{ k : \sum_{j=1}^{V} \sum_{m=1}^{P} p(w_m|\boldsymbol{\beta}_k)\delta(w_m, j) \right\}^{\uparrow} \right\rangle$$
$$L_{concepts} = \left\langle \left\{ k : \sum_{j=1}^{corrV} \sum_{n=1}^{C} p(w_n|\boldsymbol{\rho}_k)\delta(w_n, j) \right\}^{\uparrow} \right\rangle \quad . \quad (7)$$

*nlp.stanford.edu/software/corenlp.shtml

If the number of inversions is less than a threshold ($\leq \sqrt{P+C}$) then the keywords are semantically verified by the detected concept list.

Finally, we retrieve nearest neighbor sentences from the training descriptions by a ranking function. Each sentence s is ranked as: $r_s = bh(w_1 x_{s_1} + w_2 x_{s_2})$ where b is a boolean variable indicating that a sentence must have at least two of the labeled predictions, which are verified by the class of words to which the concept models belong. The boolean variable h indicates the presence of at least one human subject in the sentence. The variable indicating the total number of matches divided by the number of words in the sentence is x_{s_1}—this penalizes longer and irrelevant sentences. The sum of the weights of the predicted words from the topic model in the sentence is x_{s_2}—the latent topical strength is reflected here. Each of x_{s_1} and x_{s_2} is normalized over all matching sentences. The weights for sentence length penalty and topic strength respectively are w_1 and w_2 (set to be equal in our implementation).

3. Experimental Setup and Results

3.1. Datasets and Features

TRECVID MED12 dataset: The first dataset we use for generating lingual descriptions of real life videos is part of TRECVID Multimedia Event Detection (MED12) [20]. The training set has 25 event categories each containing about 200 videos of positive and related instances of the event descriptions. For choosing one topic model over another (Sec. 2.1) we use the positive videos and descriptions in the 25 training events and predict the words for the positive videos for the first five events in the Dev-T collection. The descriptions in the training set consist of short and very high level descriptions of the corresponding videos ranging from 2 to 42 words and averaging 10 words with stopwords. We use 68 concept models on this dataset.

A separate dataset released as part of the Multimedia Event Recounting (MER) task contains six test videos per event where the five events are selected from the 25 events for MED12. These five events are: 1) *Cleaning an appliance*; 2) *Renovating a home*; 3) *Rock Climbing*; 4) *Town hall meeting*; 5) *Working on a metal crafts project*. Since this MER12 test set cannot be publicly released for obtaining descriptions, we employ in-house annotators (blinded to our methodology) to write one description for each video.

In-house "YouCook" dataset on cooking videos: We have also collected a new dataset for this video description task, which we call YouCook. The dataset consists of 88 videos downloaded from YouTube, roughly uniformly split into six different cooking styles, such as baking and grilling. The videos all have a third-person viewpoint, take place in different kitchen environments, and frequently display dynamic camera changes. The training set consists of

49 videos with object annotations. The test set consists of 39 videos. The objects for YouCook are in the categories of utensils (31%), bowls (38%), foods and other; with 10 different object classes for utensils and bowls (we discard the other classes in this paper because of too few instances).

We use MTurk to obtain multiple human descriptions for each video. The annotators are shown an example video with a sample description focusing on the actions and objects therein. Participants in MTurk are instructed to watch a cooking video as many times as required to lingually describe the video in *at least* three sentences totaling a *minimum* of 15 words. We set our minimum due to the complex nature of the micro-actions in this dataset. The average number of words per summary is 67, the average number of words per sentence is 10 with stopwords and the average number of descriptions per video is eight. The recent data set [24] is also about cooking but it has a fixed scene and no object annotations.

Our new YouCook dataset, its annotations and descriptions, and the train/test splits are available at `http://www.cse.buffalo.edu/~jcorso/r/youcook`.

Low Level Features for Topic Model: We use three different types of low level video features: (1) HOG3D [15], (2) color histograms, and (3) transformed color histograms (TCH) [28]. HOG3D [15] describes local spatiotemporal gradients. We resize the video frames such that the largest dimension (height or width) is 160 pixels, and extract HOG3D features from a dense sampling of frames. We then use K-means clustering to create a 4000-word codebook for the MED12 data, and a 1000-word codebook for the YouCook data, due to sparsity of the dataset following [3]. Color histograms are computed using 512 RGB color bins. Further, they are computed over each frame and merged across the video. Due to large deviations in the extreme values, we use the histogram between the 15^{th} and 85^{th} percentiles averaged over the entire video. To account for poor resolution in some videos, we also use the TCH features [28] with a 4096 dimension codebook.

For a given description task, the event type is assumed known (specified manually or by some prior event detection output); we hence learn separate topic models for each event that vary based on the language vocabulary. However, the visual feature codebooks are not event specific. When learning each specific topic model, we use 5-fold cross validation to select the subset of best performing visual features. For example, on YouCook, we ultimately use HOG3D and color histograms, whereas on most of MED12 we use HOG3D and TCH (selected through cross-validation).

3.2. Quantitative Evaluation

We use the ROUGE [17] tool to evaluate the level of relevant content generated in our system output *video descriptions*. As used in [34], ROUGE is a standard for comparing text summarization systems that focuses on recall of relevant information coverage. ROUGE allows a perfect score of 1.0 in case of a perfect match given only *one* reference description. The BLEU [21] scorer is more precision oriented and is useful for comparing accuracy and fluency (usually using 4-grams) of the outputs of text translation systems as used in [2, 16] which is not our end task.

Quantitative evaluation itself is a challenge—in the UIUC PASCAL sentence dataset [23], five sentences are used *per image*. On the other hand we only allow at most five sentences *per video* per level – low or middle up to a maximum of ten. A human, on the other hand, can typically describe a video in just one sentence.

Table 2 shows the ROUGE-1 recall and precision scores obtained from the different outputs from our system for the MER12 test set. In Tables 2 and 3, "Low" is the sentence output from our low level topic models and NLP tools, "Middle" is the output from the middle level concepts, "High" is the semantically verified final output. We use the top 15 keywords with redundancy particularly retaining subjects like "man," "woman" etc. and verb morphologies (which otherwise stem to the same prefix) as proxies for ten-word training descriptions. All system descriptions are sentences, except the baseline [7], which is keywords.

From Table 2, it is clear that lingual descriptions from both the low and middle levels of our system cover more relevant information, albeit, at the cost of introducing additional words. Increasing the number of keywords improves recall but precision drops dramatically. The drop in precision for our final output is also due to increased length of the descriptions. However, the scores remain within the 95% confidence interval of that from the keywords for "Renovating home," "Town hall meeting" and "Metal crafts project" events. The "Rock climbing" event has very short descriptions as human descriptions and the "Cleaning an appliance" event is a very hard event both for DPM as well as MMLDA since multiple related concepts indicative of appliances in context appear in prediction and detection. From Table 2 we see the efficacy of the short lingual descriptions from the middle level in terms of precision while the final output of our system significantly outperforms relevant content coverage of the lingual descriptions from the other individual levels with regards to recall.

Table 3 shows ROUGE scores for both 1-gram and 2-gram comparisons. R1 means ROUGE-1-Recall and P1 means ROUGE-1-Precision. Similarly for R2 and P2. The length of all system summaries is truncated at 67 words based on the average human description length. The sentences from the low level are chosen based on the top 15 predictions only. For fair comparison on recall, the number of keywords ([7] columns in Table 3) is chosen to be 67. The numbers in bold are significant at 95% confidence over corresponding columns on the left. R2 is non-zero for key-

Events	Precision [7]	Low	Middle	High	Recall [7]	Low	Middle	High
Cleaning appliance	20.03	17.52	$11.69^{(*)}$	$10.68^{(*)}$	$19.16^{(-)}$	32.60	35.76	**48.15**
Renovating home	6.66	15.29	**12.55**	9.99	$7.31^{(-)}$	43.41	30.67	**49.52**
Rock climbing	24.45	$16.21^{(*)}$	24.52	$12.61^{(*)}$	44.09	59.22	46.23	**65.84**
Town hall meeting	17.35	14.41	**27.56**	13.36	$13.80^{(-)}$	28.66	45.55	**56.44**
Metal crafts project	16.73	18.12	**31.68**	15.63	$19.01^{(-)}$	41.87	25.87	**54.84**

Table 2: ROUGE-1 PR scores for the MER12 test set. A $(-)$ for the Recall-[7] column means significantly lower performance than the next 3 columns. The **bold** numbers *in the last column* is significantly better than the previous 3 columns in terms of recall. The **bold** numbers *in Precision-Middle column* are significantly better than those in Precision-[7] column. A $(*)$ in columns 3, 4 or 5 means significantly lower than Precision-[7]. A 95% confidence interval is used for significance testing.

[7]				High			
P2	P1	R2	R1	P2	P1	R2	R1
6E-4	15.47	6E-4	19.02	**5.04**	**24.82**	**6.81**	**34.2**

Table 3: ROUGE scores for our "YouCook" dataset.

words since some paired keywords are indeed phrases. Our method thus performs significantly well even when compared against longer descriptions. Our lingual descriptions built on top of concept labels and just a few keywords significantly outperform labeling with even *four times* as large a set of keywords. This can also tune language models to context since creating a sentence out of the predicted nouns and verbs does not increase recall based on unigrams.

3.3. Qualitative Examples

The first four rows in Fig. 6 show examples from the MER12 test set. The first one or two italicized sentences in each row are the result of the middle level output. The "health care reform" in the second row is a noise phrase that actually cannot be verified though our middle level but remains in the description due to our conservative ranking formula. Next we show one good and one bad example from our YouCook dataset. The human descriptions in the last two rows are shown for the purpose of illustrating their variance and yet their relevancy. The last cooking video has a low R1 score of 21% due to imprecise predictions and detections.

4. Conclusion

In this paper we combine the best aspects of top-down and bottom-up methods of producing lingual descriptions of videos *in-the-wild* that exploit the rich semantic space of both text and visual features. Our contribution is unique in that the class of concept detectors semantically verify low level predictions from the bottom up and leverage both sentence generation and selection that together outperforms output from the independent modules. Our future work will emphasize scalability in the semantic space to increase the generality of plausible lingual descriptions.

Acknowledgments We thank the Mori group at Simon Fraser University and Kitware Inc. for helpful discussions and feature extraction, and Philip Rosebrough, Cody Boppert, Yao Li, and David Molik for their work on data curation and annotation. This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF-11-1-0090), the DARPA Mind's Eye program (W911NF-10-2-0062), and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, DARPA, ARO, NSF or the U.S. Government.

References

[1] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. J. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. W. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. Video in sentences out. In *UAI*, 2012.

[2] A. Belz and E. Reiter. Comparing automatic and human evaluation of nlg systems. In *EACL*, 2006.

[3] P. Bilinski and F. Bremond. Evaluation of local descriptors for action recognition in videos. In *ICCV*, 2011.

[4] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR*, 2003.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[6] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 2007.

[7] P. Das, R. K. Srihari, and J. J. Corso. Translating related words to videos and back through latent topics. In *ACM WSDM*, 2013.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 2010.

[9] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: generating sentences from images. In *ECCV*, 2010.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models.

Cleaning an appliance

Keywords: refrigerator/OBJ cleans/VERB man/SUBJ-HUMAN clean/VERB blender/OBJ cleaning/VERB woman/SUBJ-HUMAN person/SUBJ-HUMAN stove/OBJ microwave/OBJ sponge/NOUN food/OBJ home/OBJ hose/OBJ oven/OBJ
Sentences from Our System 1. *A person is using dish towel and hand held brush or vacuum to clean panel with knobs and washing basin or sink.* 2. Man cleaning a refrigerator. 3. Man cleans his blender. 4. Woman cleans old food out of refrigerator. 5. Man cleans top of microwave with sponge.
Human Synopsis: Two standing persons clean a stove top with a vacuum clean with a hose.

Town hall meeting

Keywords: meeting/VERB town/NOUN hall/OBJ microphone/OBJ talking/VERB people/OBJ podium/OBJ speech/OBJ woman/SUBJ-HUMAN man/SUBJ-HUMAN chairs/NOUN clapping/VERB speaks/VERB questions/VERB giving/VERB
Sentences from Our System 1. *A person is speaking to a small group of sitting people and a small group of standing people with board in the back.* 2. A person is speaking to a small group of standing people with board in the back. 3. Man opens town hall meeting. 4. Woman speaks at town meeting. 5. Man gives speech on health care reform at a town hall meeting.
Human Synopsis: A man talks to a mob of sitting persons who clap at the end of his short speech.

Renovating home

Keywords: people/SUBJ-HUMAN, home/OBJ, group/OBJ, renovating/VERB, working/VERB, montage/OBJ, stop/VERB, motion/OBJ, appears/VERB, building/VERB, floor/OBJ, tiles/OBJ, floorboards/OTHER, man/SUBJ-HUMAN, laying/VERB
Sentences from Our System: 1. *A person is using power drill to renovate a house*. 2. *A crouching person is using power drill to renovate a house*. 3. A person is using trowel to renovate a house. 4. man lays out underlay for installing flooring. 5. A man lays a plywood floor in time lapsed video.
Human Synopsis: Time lapse video of people making a concrete porch with sanders, brooms, vacuums and other tools.

Metal crafts project

Keywords: metal/OBJ man/SUBJ-HUMAN bending/VERB hammer/VERB piece/OBJ tools/OBJ rods/OBJ hammering/VERB craft/VERB iron/OBJ workshop/OBJ holding/VERB works/VERB steel/OBJ bicycle/OBJ
Sentences from Our System 1. *A person is working with pliers.* 2. Man hammering metal. 3. Man bending metal in workshop. 4. Man works various pieces of metal. 5. A man works on a metal craft at a workshop.
Human Synopsis: A man is shaping a star with a hammer.

Cooking video: High ROUGE score

Keywords: bowl/OBJ pan/OBJ video/OBJ adds/VERB lady/OBJ pieces/OBJ ingredients/OBJ oil/OBJ glass/OBJ liquid/OBJ butter/SUBJ-HUMAN woman/SUBJ-HUMAN add/VERB stove/OBJ salt/OBJ
Sentences from Our System: 1. A *person is cooking with bowl and stovetop*. 2. In a pan add little butter. 3. She adds some oil and a piece of butter in the pan. 4. A woman holds up Bisquick flour and then adds several ingredients to a bowl. 5. A woman adds ingredients to a blender.
Human Synopsis1: A lady wearing red colored dress, blending (think butter) in a big sized bowl. Besides there is 2 small bowls containing white color powders. It may be maida flour and sugar. After she is mixing the both powders in that big bowl and blending together. **Human Synopsis2**: In this video, a woman first adds the ingredients from a plate to a large porcelain bowl. She then adds various other ingredients from various different bowls. She then mixes all the ingredients with a wooden spoon.

Cooking video: Low ROUGE score

Keywords: bowl/OBJ pan/OBJ video/OBJ adds/VERB ingredients/OBJ lady/OBJ woman/SUBJ-HUMAN add/VERB pieces/OBJ stove/OBJ oil/OBJ put/VERB added/VERB mixes/VERB glass/OBJ
Sentences from Our System: 1. A *person is cooking with pan and bowl*. 2. *A person is cooking with pan*. 2. A woman adds ingredients to a blender. 2. In this video, a woman adds a few ingredients in a glass bowl and mixes them well. 3. In this video, a woman first adds the ingredients from a plate to a large porcelain bowl 4. The woman is mixing some ingredients in a bowl. 5. the woman in the video has a large glass bowl.
Human Synopsis1: The woman is giving directions on how to cook bacon omelette. She shows the ingredients for cooking and was frying the bacon, scrambling the egg, melting the butter and garnishing it with onions and placed some cheese on top. The woman then placed the scrambled egg and bacon to cook and then placed it on a dish. **Human Synopsis2**: in this video the woman takes bacon, eggs, cheese , onion in different containers. On a pan she cooks the bacon on low flame. Side by side she beats the eggs in a bowl. she removes the cooked bacon on a plate. In the pan she fries onions and then adds the beaten eggs. She sprinkles grated cheese on the pan and cooks well. She then adds the fried bacon on the eggs in the pan and cook well. She transfers the cooked egg with bacon to as serving plate.

Figure 6: Qualitative results from MER12 and our "YouCook" dataset. Only the top 5 sentences from our system are shown.

TPAMI, 2010.

[11] Y. Feng and M. Lapata. Topic models for image annotation and text illustration. In *NAACL HLT*, 2010.

[12] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[13] T. M. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *ICCV*, 2009.

[14] M. U. G. Khan, L. Zhang, and Y. Gotoh. Towards coherent natural language description of video streams. In *ICCVW*, 2011.

[15] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[16] 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.

[17] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL HLT*, 2003.

[18] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008.

[19] G. Malkarnenkar, N. Krishnamoorthy, S. Guadarrama, K. Saenko, and R. Mooney. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, 2013.

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

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

[22] D. Putthividhya, H. T. Attias, and S. S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *CVPR*, 2010.

[23] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *NAACL HLT 2010*, 2010.

[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.

[25] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[26] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

[27] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[28] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 2010.

[29] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*.

[30] M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008.

[31] H. M. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why priors matter. In *NIPS*, 2009.

[32] C. Wang, D. M. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *CVPR*, 2009.

[33] J. Wanke, A. Ulges, C. H. Lampert, and T. M. Breuel. Topic models for semantics-preserving video compression. In *MIR*, 2010.

[34] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011.