# Youtube2Story: Unsupervised Joint-Representation of Instructional Videos

Anonymous ICCV submission

Paper ID ****

## Abstract

*The ABSTRACT*

## 1. Introduction

Leaning the instructions of a novel non-trivial task is both a challange and a necessity for both humans and autonomous systems. This necessity resulted in many community generated instruction collections [**?**, **?**] and expert curated recipe books[**?**, **?**]. However, this instructions are generally based on a language modality and explains a single way of performing the task although there are variety of ways. On the other hand, online video storage services are full of unstructured instructional videos[1] covering variety of ways, environment conditions and view angles. Although there have been many succesfull attempts in detecting activities from videos [**?**, **?**], structural representation of such a large and useful video collection is not possible. In this paper, we focus on joint semantic representation of YouTube videos as a response to a single query. We specficially study the unsupervised joint-detection of the activities from a collection of YouTube videos.

Understanding of the instructional videos, requires the careful processing of two complementary modaliities namely language and the vision. Luckily the target domain, YouTube videos, has unstructured subtitles as well. They are either generated by the content developer (5% of the time) or automatically generated by using the Automatic Speech Recognition (ASR) softwares. The main limitations of the existing activity detection literature for this problem is scalability and representation level. Existing approaches are mainly supervised and requires extensive training set which is not tractable in the scale of YouTube videos. Moreover, current activity detection research focuses on the low-level visual features. However, such videos in the wild have objects with completely different texture and shape characteristics from wide range of views. Instead, we focus on extracting high-level visual semantic representations and using salient words occuring among the videos.

We rely on the assumption that the videos collected as the response of a same instructional query, share similar activities performed by the similar objects. We start with the independent processing of the videos in order to create a large collection of visual object proposals and words. After the proposal generation, we jointly process the proposal collections and words to detect the visual objects and words which can be used to represent the unstructured information. Since we rely on high-level information instead of the low-level features, the resulting objects represent the semantic information instead of visual characteristic. By using the extracted objects, we compute the holistic representation of the multi-modal information in each frame.

Moving from frame-wise visual understading to activity understanding, requires the joint processing of all the videos with the temporal information. In order to exploit the temporal information, we model each video as a Hidden Markov Model using state space of activities. Since we assume that the videos share some of the activities and we have no supervision, we use a model based on *beta process mixture model*. Our model jointly learn the activities and detect them in the videos. Moreover, it does not require prior knowledge over the number of activities.

## 2. Related Work

**Video Summarization** Summarizing an input video as a sequence of keyframes (static) or as a sequence of video clips (dynamic) is useful for both multi-media search interfaces and retrieval purposes. Early works in the area are summarized in [27] and mostly focus on choosing key-frames for visualization. Idea of choosing key-frames is also extended by using the video tags by Hong et al[8] and using the spatio-temporal information by Gygli et al.[7].

Summarizing videos is crucial for ego-centric videos since the ego-centric videos are generally long in duration. There are many works which succesfully segment such videos in to a sequence of important shots [15, 16]; however, they mostly rely on specific features of ege-centric videos. Rui et al. [23] proposed another dynamic summarization method based on the excitement in the speech of the

[1] YouTube has 281.000 videos for *"How to tie a bow tie"*

reporter. Due to their domain specific designs, these algorithms are not applicable to the general instructional videos.

Same idea is also applied to the large image collecitons by recovering the temporal ordering and visual similarity of images [12]. This image colelctions are further used to choose important view points for video key-frame selection by Khosla et al.[10]. And further extended to video clip selection by Kim et al[11] and Potapov et al.[22]. Although they are different from our approach since they do not use any high-level semantic information or the language information, we experimentally compare our method with them.

**Understanding Multi-Modal Information:** Learning the relationship between the visual and language data is a crucial problem due to its immense multimedia applications. Early methods [1] in the area focus on learning a common multi-modal space in order to jointly represent language and vision. They are also extended to learning higher level relations between object segments and words [24]. Zitnick et al.[30, 29] used abstracted clip-arts to further understand spatial relations of objects and their language correspondences. Kong et al. [14] and Fidler et al. [5] both accomplished the same task by using the image captions only. Relations extracted from image-caption pairs, are further used to help semantic parsing [28] and activity recognition [20]. Recent works also focused on generating image captions automatically using the input image. These methods range from finding similar images and using their captions [21] to learning language modal conditioned on the image [13, 25, 4]. And the methods to learn language models vary from graphical models [4] to neural networks [25, 13, 9].

All afforementioned methods are using supervised labels either as strong image-word pairs or weak image-caption pairs. On the other hand, our method is fully unsupervised.

**Activity Detection/Recognition:**

**Recipe Understanding** Following the interest in community generated recipes in the web, there have been many attempts to automatically process recipes. Recent methods on natural language processing [18, 26] focus on semantic parsing of language recipes in order to extract actions and the objects in the form of predicates. Tenorth et al.[26] further process the predicates in order to form a complete logic plan. Mori et al.[19] also learns the relations of the actions in terms of a flow graph with the help of a supervision. The aforementioned approaches focus only on the language modality and they are not applicable to the videos. We have also seen recent advances [2, 3] in robotics which uses the parsed recipe in order to perform cooking tasks. They use supervised object detectors and report a succesfull autonomous cooking experiment. In addition to the language based approaches, Malmaud et al.[17] consider both language and vision modalities and propose a method to align an input video to a recipe. However, it can not extract the steps/actions automatically and requires a ground truth recipe to align. On the contrary, our method uses both visual and language modalities and extracts the actions while autonomously constructing the recipe. There is also an approach which generates multi-modal recipes from expert demonstrations [6]. However, it is developed only for the domain of *teaching user interfaces* and are not applicable to the videos.

## 3. Method

In this section, we explain the highlevel compenents of the method we develop to jointly represent multi-modal instructions. We explain the details of the each sub-system in the following sections. As shown in the Figure **??**, our proposed method consists of three major components; online query and filtering, frame-wise multi-modal representation and joint clustering to extract activities. **Query:** Our system starts with querying the YouTube with an *How to* question for top 100 videos. The text descriptions of the returned videos are represented as bag-of-words and clustered to eliminate outliers. **Framewise Representation:** In order to represent the frames of the returned videos, we process both the visual and language content of the videos. We extract object proposals and jointly cluster them in order to detect the salient objects of the recipe. For the language descriptions, we use the top salient words of the courpus generated as concatenation of the all subtitles. We represent each frame in terms of the resulting salient objects and words. **Unsupervised Activity Detection:** After describing each frame by using the salient objects and words, we apply a non-parametric Bayesian method in order to find temporally consistent clusters occuring over multiple videos. The resulting clusters are used to label activities of the test-videos and summarize them as a flow graph.

## 4. Semantic Multi-Modal Frame Representation

### 4.1. Learning Atoms to Represent Multi-Modal Data

We represent each frame in our video set as a distribution over set of language and visual atoms. Language atoms are the salient words detected by using a tf-idf like measure, and the visual atoms are found via clustering over object proposals which we extract from frames. We explain the details of finding the atoms and representing the frames in the subsequent sections.

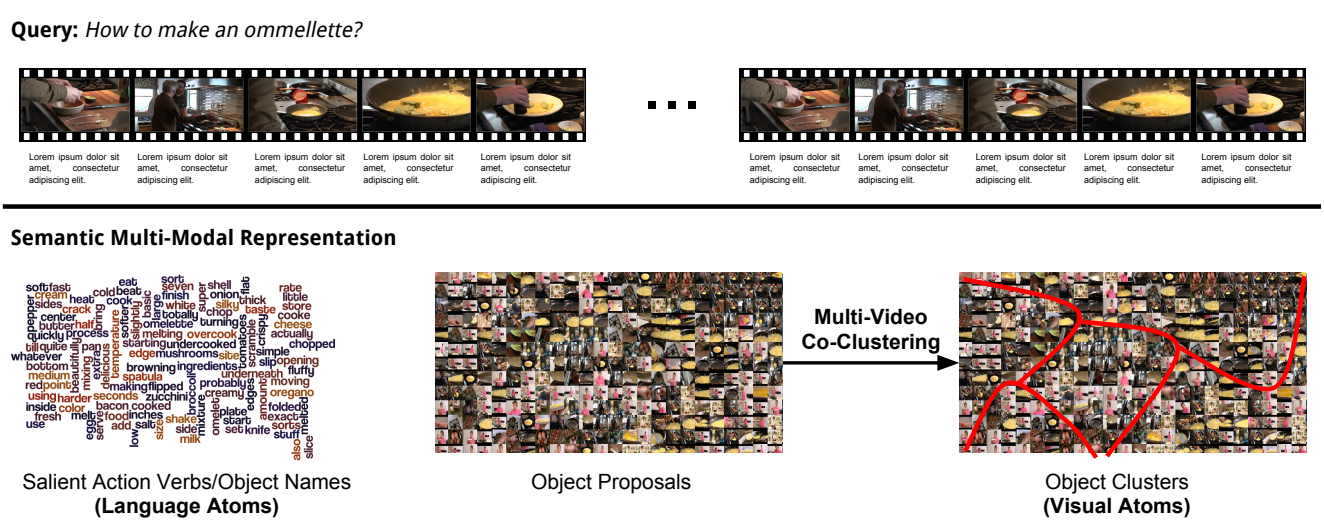



Figure 1: Components of our recipe understanding method. **Query:** We query the YouTube for top 100 *How To* videos and filter the outliers; **Framewise Representation:** We automatically extract object clusters and salient word in order to find multi-modal representation of each frame. **Unsupervised Activity Detection:** We jointly cluster videos in order to learn activities/steps related to the recipe.

#### 4.1.1 Learning Language Atoms

In order to detect salient words, we use tf-idf like measure. For each recipe, we concatanate all subtitles into a single term corpus. As a document corpus, we use all the words extracted from all recipes. Moreover, we compute the tf-idf as $tfidf(w, d, D) = f_{w,d} \times \log \frac{N}{n_w}$ where $w$ is the word, $d$ is the corpus of the recipe and $f_{w,d}$ is the frequency of the word in the recipe. Moreover, $N$ represents the total number of videos in all recipes and $n_w$ is the number of videos whose subtitle include the word $w$. After computing the tf-idf, we sort all words with their tf-idf values and choose top $K$ words as set of salient words.

#### 4.1.2 Learning Visual Atoms

In order to learn visual atoms, we initially generate object proposals from each frame of each videos.

**Generating Framewise Object Proposals**

**Joint Proposal Clustering to Detect Visual Objects**

### 4.2. Multi-Modal Representation via Learned Atoms

## 5. Unsupervised Activity Representation

Here we explain the generative model which we use in order to jointly learn the activities from videos. We start

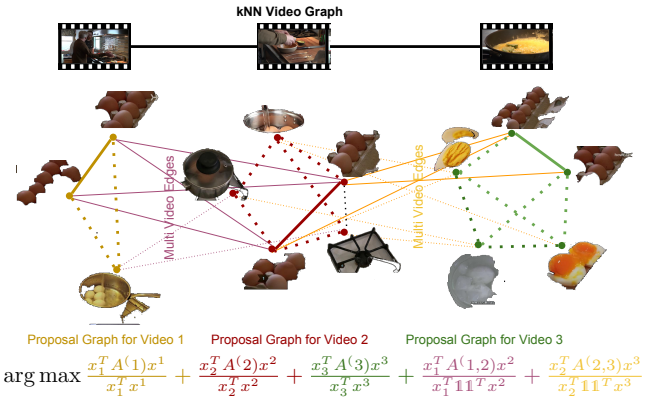



Figure 2: Visualization of the joint proposal clustering.

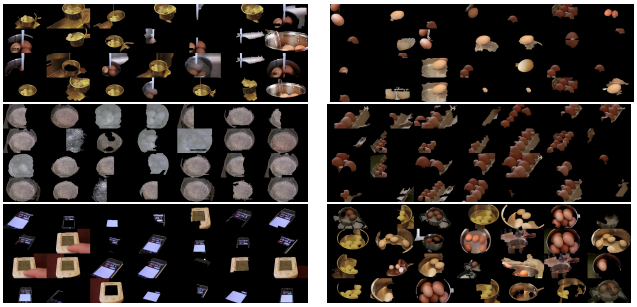


Figure 3: Randomly selected images of randomly selected clusters learned for *How to boil an egg?*

with explaining the basic notation we use in the model. We denote the $t^{th}$ frame of the $i^{th}$ video as $I_t^i$ and its subtitle as $L_t^i$. Moreover, we note the extracted frame representation as $y_t^i$. We model our algorithm based on activities and the note the activity of the $t^{th}$ frame of the $i^{th}$ video as $z_t^i$. Since our model is non-parametric, number of activities are not fixed and $z_t^i \in \mathcal{N}$.

**Frame:** Representation of each frame is computed as the occurence vector over the detected language and visual atoms. Formally, frame $y_t^i$ is represented as $y_t^i = [y_t^{i,l}, y_t^{i,v}]$ such that $k^t h$ entry of the $y_t^{i,l}$ is 1 if the frame as language atom $k$ and 0 otherwise. $y_t^{i,l}$ is also a binary vector similarly defined over visual atoms. We consider $y_t^i$ as the observed variable and consider the underlying activity label as the hidden state $z_t^i$. A sample state is visualized in the Figure 4.

**Activity:** We represent each activity as a Bernoulli distribution over the visual and language atoms. In other words, each frame's representation $y_t^i$ is sampled from its activity distribution as $y_t^i | z_t^i = k \sim Ber(\Theta_k)$. For the sake of clarity, we sample the $\Theta$ from its conjugate distribution *Beta distribution*.

In the following sections, we exlpain how these models

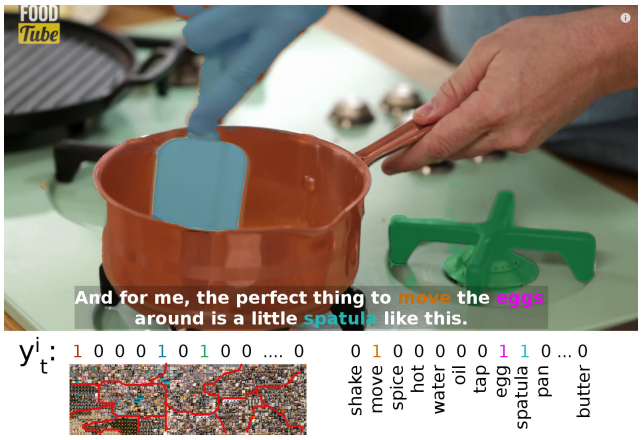



Figure 4: Visualziation of the representation of the Frame.

can be jointly learned and inferred by using the Beta Process Hidden Markov Models. s

### 5.1. Beta Process Hidden Markov Model

For joint understanding of the time-series information, Fox et al.[**?**] proposed the Beta Process Hidden Markov Models (BP-HMM) using the Indian Buffet Process[**?**] of time-series sequences over features. It assumes that there exist a set of features(activities in our case) which can explain the behaviour of all time-series data (all videos in our case). Each time-series data exhibits a subset of available features. This setup is similar to Hughes et al.[**?**]. However, we differ in the choices of the underlying distributions since we based our model on semantic multi-modal information.

In our model, each video $i$ chooses a set of activities through a activity vector $\mathbf{f^i}$ such that $f_k^i$ is 1 if $i^t h$ video has activity $k$, 0 otherwise. When the feature vectors of all videos in the courpus is concatanated, it becomes an activity matrix $\mathbf{F}$ such that $i^t h$ row of the $\mathbf{F}$ is the activity vector $\mathbf{f^i}$. Moreover, each feature $k$ also has an activity frequency $b_k$ and a distribution parameter $\Theta_k$. Distribution parameter $\Theta_k$ is the Bernoulli distribution as explain in the Section 5. Moreover, its base distribution ($B_0$) is *Beta random variable* since it is the conjugate of Bernoulli. Moreover, in this setting, the activity paremeters $\Theta_k$ and $b_k$ follow the *beta process* as;

$$B|B_0, \gamma, \beta \sim \text{BP}(\beta, \gamma B_o), B = \sum_{k=1}^{\infty} b_k \delta_{\Theta_k} \qquad (1)$$

where $B_0$ and the $b_k$ are determined by the underlying poisson process [**?**] and the feature vector is determined as independent Bernoulli draws as $f_k^i \sim Ber(b_k)$. After marginilizing over the $b_k$ and $\Theta_k$, this distribution is shown to be equivalent to Indian Buffet Process [**?**]. Where videos are customers and activities are dishes in the buffet. The

first video chooses a Poisson($\gamma$) unique dishes. The following video $i$ chooses previously sampled activity $k$ with probability $\frac{m_k}{i}$ proportional to number of videos ($m_k$) chosen the activity $k$, and it also chooses Poisson($\frac{\gamma}{i}$) new activities. Here, $\gamma$ controls the number of active features in each video and $\beta$ controls the likelihood of the features getting shared by multiple videos.

After each video chooses a subset of activities, we model the videos as an Hidden Markov Model (HMM) over the selected videos. Each frame has the hidden state activity id($z_t^i$) and we observe the binary representation $y_t^i$. Since we model each activity as a Bernoulli distribution, the emmition probabilities follow the Bernoulli distribution as $p(y_t^i|z_t^i) = Ber(\Theta_{z_t^i})$. Following the construction of the Fox et al.[**?**], we sample the transition probabilities from a normalized Gamma distribution. For each video $i$, we sample a Gamma random variable for the transition between activity $j$ and activity $k$ if both of the activities are included by the video (if $f_k^i$ and $f_j^i$ are both 1). After sampling these random variables, we normalize them to have proper transition probabilities. This procedure can be represented formally as

$$\eta_{j,k}^i \sim Gam(\alpha + \kappa\delta_{j,k}, 1), \quad \pi_j^i = \frac{\eta_j^i \circ f^i}{\sum_k \eta_{j,k}^i f_k^i} \qquad (2)$$

Where $\kappa$ is the persistence parameter promoting self state transitions to have more coherent temporal bounderies, $\circ$ is the element-wise product and $\pi_j^i$ is the transition probabilities in video $i$ from state $j$ to all states in the form of a vector.

This model is also presented as a graphical model in Figure 5

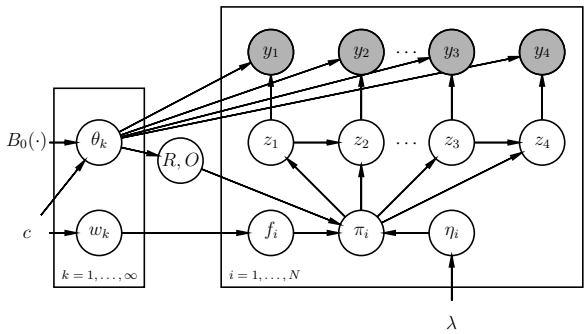



Figure 5: **Graphical model for BP-HMM:**Some explanation.

### 5.2. Gibbs sampling for BP-HMM

We employ Markov Chain Monte Carlo (MCMC) method for learning and inference of the BP-HMM. We base our algorithms on the MCMC procedure proposed by Fox et al.[**?**]. It marginilizes over blah and blah and sample blah and blah. For faster convergence, we also utilize a series of data driven samplers. Here we only discuss the proposed data driven samplers and move the details of the remainin samplers to the Supplementary Material.

**Data-Driven Sampler1**

**Data-Driven Sampler2**

## 6. Experiments

### 6.1. Dataset

We collected blah blah videos from YouTube and did blah blah... All numbers etc. These are recipes, 5 of them are evaluation set and labelled temporally and labels are matched.

### 6.2. Baselines

We compare against BP-HMM-O with local feauters and HMM with Semantic Features, HMM with local features, HMM with CNN feautes...

### 6.3. Qualitative Results

### 6.4. Accuracy over Activity Detection

### 6.5. Accuracy over Activity Learning

## 7. Discussions and Conclusions

Discuss which recipes worked and why. Discuss the importance of semantic representation, scaling features and multi-modality.

## References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003. 2

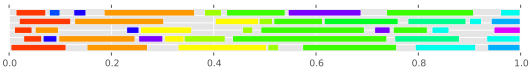
[2] M. Beetz, U. Klank, I. Kresse, A. Maldonado, L. Mosenlechner, D. Pangercic, T. Ruhr, and M. Tenorth. Robotic roommates making pancakes. In *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, pages 529–536. IEEE, 2011. 2

[3] M. Bollini, J. Barry, and D. Rus. Bakebot: Baking cookies with the pr2. In *The PR2 Workshop: Results, Challenges and Lessons Learned in Advancing Robots with a Common Platform, IROS*, 2011. 2

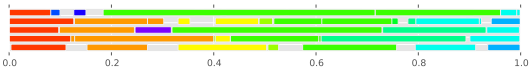
[4] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010. 2

Table 1: Notation of the Paper

| | | | | | |
|---|---|---|---|---|---|
| $y_t = [y_t^v, y_t^l]$ | feature representation of $t^{th}$ frame | $I_t$ | $t^{th}$ frame of the video | $x_{i,r}^p$ | 1 if $p^{th}$ cluster has $r^{th}$ proposal of $i^{th}$ video |
| $x^p$ | binary vector for $p^{th}$ cluster | $L_t$ | subtitle for $t^{th}$ frame | $O^{k,k'}$ | 1 if $\#(z_t = k, z_{t'}) = k') = 0 \forall \quad t \leq t'$ |
| $\Theta_k = [\Theta_k^v, \Theta_k^l]$ | emmition prob. of $k^{th}$ activity | $z_t$ | activity ID of frame $t$ | $f_i^k$ | 1 if $i^{th}$ video has $k^{th}$ activity $0 o.w.$ |
| $\eta_i^{k,k'}$ | $P(z_{t+1} = k' \| z_t = k)$ for $i^{th}$ vid | $\pi_i^{k,k'}$ | $\eta_i^{k,k'} \times f_i^k \times f_i^{k'}$ | | |




(a) Ground Truth Activity Labels




(b) Activity Labels extracted by Our Method

Crack the eggs one at a time into a bowl.

Remove the omelette onto a plate.

You can either use a fork or wire whisk to beat the eggs into a bowl.

Eggs cook quickly, so make sure the pan gets very hot first; the butter melt completely.

(c) Sample images and the automatically generated captions for some of the clusters.

[5] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1995–2002. IEEE, 2013. 2

[6] F. Grabler, M. Agrawala, W. Li, M. Dontcheva, and T. Igarashi. Generating photo manipulation tutorials by demonstration. *ACM Transactions on Graphics (TOG)*, 28(3):66, 2009. 2

[7] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *Computer Vision–ECCV 2014*, pages 505–520. Springer, 2014. 1

[8] R. Hong, J. Tang, H.-K. Tan, C.-W. Ngo, S. Yan, and T.-S. Chua. Beyond search: event-driven summarization for web videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 7(4):35, 2011. 1

[9] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *ArXiv e-prints*, Dec. 2014. 2

[10] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2698–2705. IEEE, 2013. 2

[11] G. Kim, L. Sigal, and E. P. Xing. Joint summarization of large-scale collections of web images and videos for story-line reconstruction. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 4225–4232. IEEE, 2014. 2

[12] G. Kim and E. P. Xing. Reconstructing storyline graphs for image recommendation from web community photos. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3882–3889. IEEE, 2014. 2

[13] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603, 2014. 2

[14] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3558–3565. IEEE, 2014. 2

[15] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, volume 2, page 6, 2012. 1

[16] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2714–2721. IEEE, 2013. 1

[17] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy. What's Cookin'? Interpreting Cooking Videos using Text, Speech and Vision. *ArXiv e-prints*, Mar. 2015. 2

[18] J. Malmaud, E. J. Wagner, N. Chang, and K. Murphy. Cooking with semantics. *ACL 2014*, page 33, 2014. 2

[19] S. Mori, H. Maeta, Y. Yamakata, and T. Sasada. Flow graph corpus from recipe texts. In *Proceedings of the Nineth International Conference on Language Resources and Evaluation*, pages 2370–2377, 2014. 2

[20] T. S. Motwani and R. J. Mooney. Improving video activity recognition using object recognition and text mining. In *ECAI*, pages 600–605, 2012. 2

[21] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011. 2

[22] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *Computer Vision–ECCV 2014*, pages 540–555. Springer, 2014. 2

[23] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the*

*eighth ACM international conference on Multimedia*, pages 105–115. ACM, 2000. 1

[24] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 966–973. IEEE, 2010. 2

[25] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. 2

[26] M. Tenorth, D. Nyga, and M. Beetz. Understanding and executing instructions for everyday manipulation tasks from the world wide web. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 1486–1491. IEEE, 2010. 2

[27] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1):3, 2007. 1

[28] H. Yu and J. M. Siskind. Grounded language learning from video described with sentences. In *ACL (1)*, pages 53–63, 2013. 2

[29] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3009–3016. IEEE, 2013. 2

[30] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1681–1688. IEEE, 2013. 2