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 [17] 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[5] and using the spatio-temporal information by Gygli et al.[4].

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 [9, 10]; however, they mostly rely on specific features of ege-centric videos. Rui et al. [15] 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 [8]. This image colelctions are further used to choose important view points for video key-frame selection by Khosla et al.[6]. And further extended to video clip selection by Kim et al[7] and Potapov et al.[14]. 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: Captioning Work, What are you talking about? Text-to-Image Coreference(urtasun), Bringing Semantics Into Focus Using Visual Abstraction, Learning the Visual Interpretation of Sentences Zitnick (parikh) clip art to understand spatial relations of objects and language, Improving Video Activity Recognition using Object Recognition and Text Mining Motwani uses captions and object detectors to learn activities automatically. Grounded Language Learning from Video Described with Sentences (yu) improve semantic parsing using object/activity detectors. Matching Words and Pictures, A Sentence is Worth a Thousand Pixels, Connecting Modalities: Semi-supervised Segmentation and Annotation of Images Using Unaligned Text Corpora, Grounded Compositional Semantics for Finding and Describing Images with Sentences, Multimodal Neural Language Models, Every Picture Tells a Story: Generating Sentences from Images, m2Text: Describing Images Using 1 Million Captioned Photographs

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 [12, 16] focus on semantic parsing of language recipes in order to extract actions and the objects in the form of predicates. Tenorth et al.[16] further process the predicates in order to form a complete logic plan. Mori et al.[13] 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 [1, 2] 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.[11] 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 [3]. However, it is developed only for the domain of *teaching user interfaces* and are not applicable to the videos.

3. Overview

Explanation of the Figure 3 with pointers to the Sections.

4. Semantic Multi-Modal Frame Representation

4.1. Learning Atoms to Represent Multi-Modal Data

4.1.1 Learning Language Atoms

4.1.2 Learning Visual Atoms

Generating Framewise Object Proposals

Joint Proposal Clustering to Detect Visual Objects

4.2. Multi-Modal Representation via Learned Atoms

5. Unsupervised Activity Representation

5.1. BP-HMM with Orderings (BP-HMM-O)

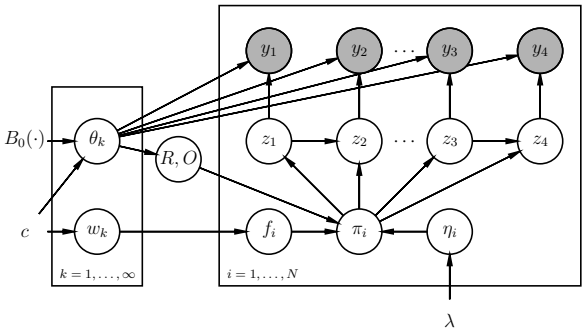
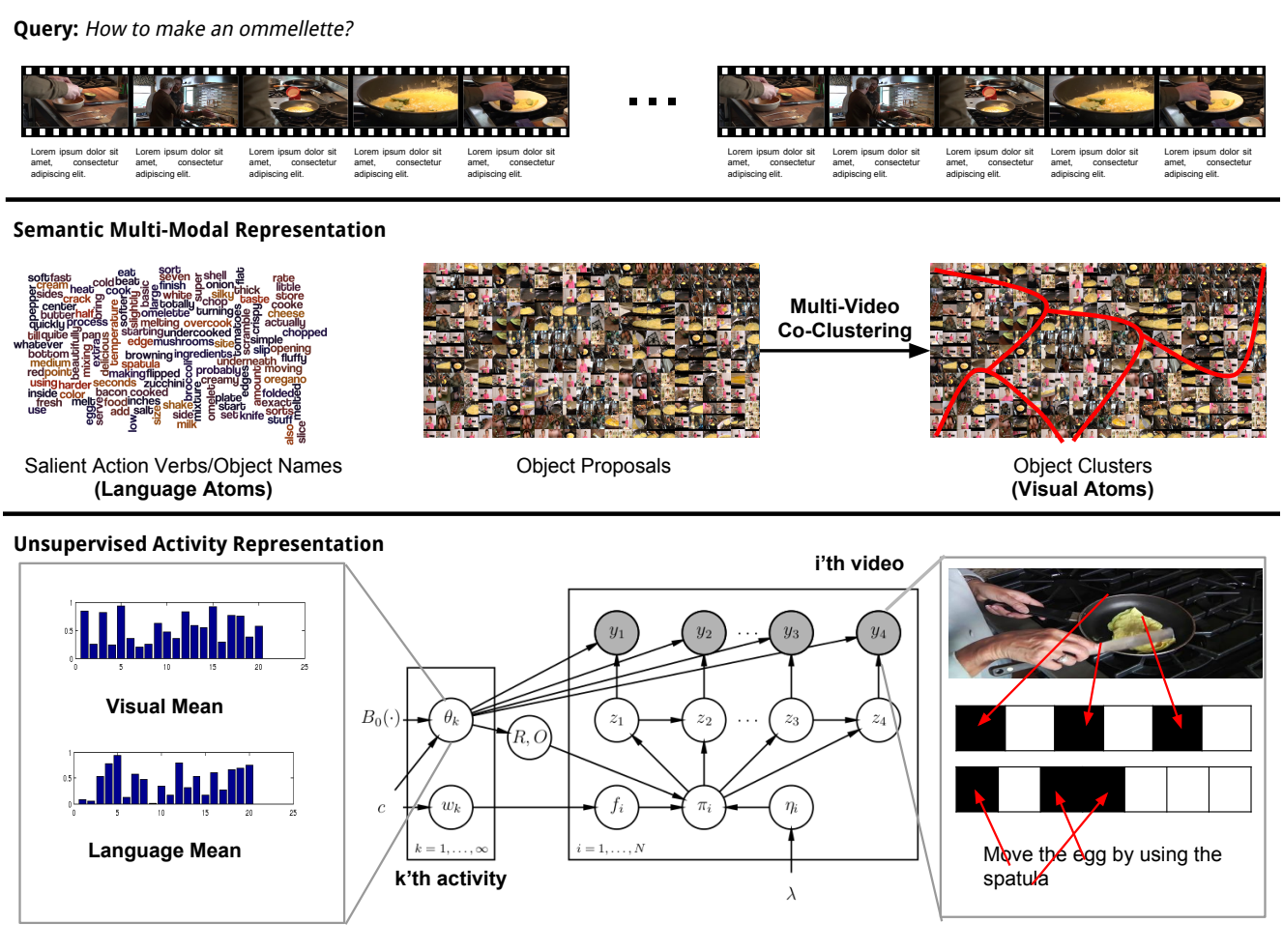


Figure 1. Graphical model for BP-HMM-O

5.2. Gibbs sampling for BP-HMM-O

Mostly deferred to supplementary, just say we ne need to sample a,b,c,etc. and we used MH for binary and Gibbs for continous/discrete ones.



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] 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

[2] 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

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 $0o.w.$
$\eta_i^{k,k'}$	$P(z_{t+1} = k' \| z_t = k)$ for i^{th} vid	$\pi_i^{k,k'}$	$\eta_i^{k,k'} \times f_i^k \times f_i^{k'}$		

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

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

[13] 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

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

[15] 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

[16] 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

[17] 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