

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

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 challenge and a necessity for both humans and autonomous systems. This necessity resulted in many community generated instruction collections [?, ?] and expert curated recipe books[?, ?]. However, these 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 videos¹ covering variety of ways, environment conditions and view angles. Although there have been many successful 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 specifically 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 modalities 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 us-

ing salient words occurring 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 understanding 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 keyframes 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 successfully segment such videos into a sequence of important shots [15, 16]; however, they mostly rely on specific features of ego-centric videos. Rui et al. [23] proposed another dynamic summarization method based on the excitement in the speech of the

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

108 reporter. Due to their domain specific designs, these algorithms are not applicable to the general instructional videos.
109

110 Same idea is also applied to the large image collections by recovering the temporal ordering and visual similarity
111 of images [12]. This image collections are further used to choose important view points for video key-frame selection
112 by Khosla et al.[10]. And further extended to video clip selection by Kim et al[11] and Potapov et al.[22]. Although
113 they are different from our approach since they do not use any high-level semantic information or the language information,
114 we experimentally compare our method with them.
115
116

117 **Understanding Multi-Modal Information:** Learning
118 the relationship between the visual and language data is
119 a crucial problem due to its immense multimedia applications.
120 Early methods [1] in the area focus on learning
121 a common multi-modal space in order to jointly represent
122 language and vision. They are also extended to learning
123 higher level relations between object segments and words
124 [24]. Zitnick et al.[30, 29] used abstracted clip-arts to further
125 understand spatial relations of objects and their language
126 correspondences. Kong et al. [14] and Fidler et
127 al. [5] both accomplished the same task by using the image
128 captions only. Relations extracted from image-caption
129 pairs, are further used to help semantic parsing [28] and activity
130 recognition [20]. Recent works also focused on generating
131 image captions automatically using the input image.
132 These methods range from finding similar images and using
133 their captions [21] to learning language modal conditioned
134 on the image [13, 25, 4]. And the methods to learn language
135 models vary from graphical models [4] to neural networks
136 [25, 13, 9].
137

138 All aforementioned methods are using supervised labels
139 either as strong image-word pairs or weak image-caption
140 pairs. On the other hand, our method is fully unsupervised.
141

142 Activity Detection/Recognition:

143 **Recipe Understanding** Following the interest in community
144 generated recipes in the web, there have been many attempts
145 to automatically process recipes. Recent methods on natural
146 language processing [18, 26] focus on semantic
147 parsing of language recipes in order to extract actions
148 and the objects in the form of predicates. Tenorth et al.[26]
149 further process the predicates in order to form a complete
150 logic plan. Mori et al.[19] also learns the relations of the
151 actions in terms of a flow graph with the help of a supervision.
152 The aforementioned approaches focus only on the language modality and they are not applicable to the videos.
153 We have also seen recent advances [2, 3] in robotics which uses the parsed recipe in order to perform cooking tasks.
154 They use supervised object detectors and report a successful
155 autonomous cooking experiment. In addition to the lan-
156 guage based approaches, Malmaud et al.[17] consider both
157 language and vision modalities and propose a method to align
158 an input video to a recipe. However, it can not extract the steps/actions automatically and requires a ground truth
159 recipe to align. On the contrary, our method uses both visual
160 and language modalities and extracts the actions while
161 autonomously constructing the recipe. There is also an approach which generates multi-modal recipes from expert
162 demonstrations [6]. However, it is developed only for the domain of *teaching user interfaces* and are not applicable
163 to the videos.
164

165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

3. Method

In this section, we explain the highlevel components 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 corpus 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 occurring 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.

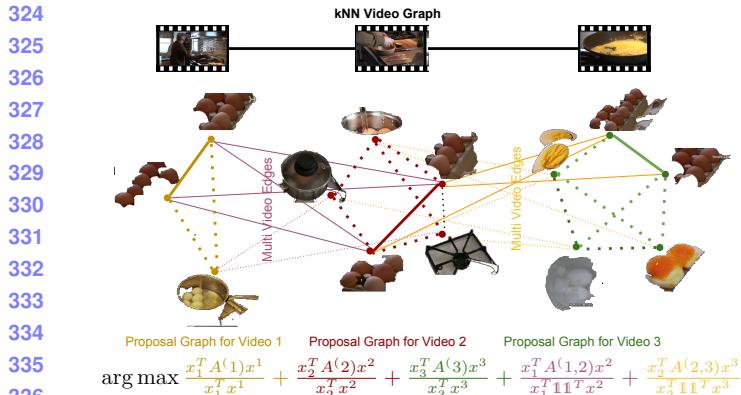
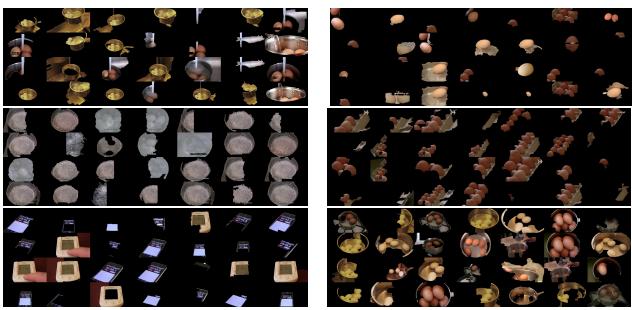


Figure 2: Visualization of the joint proposal clustering.

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 occurrence 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^{th} 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 Θ from its conjugate distribution *Beta distribution*.

In the following sections, we explain how these models

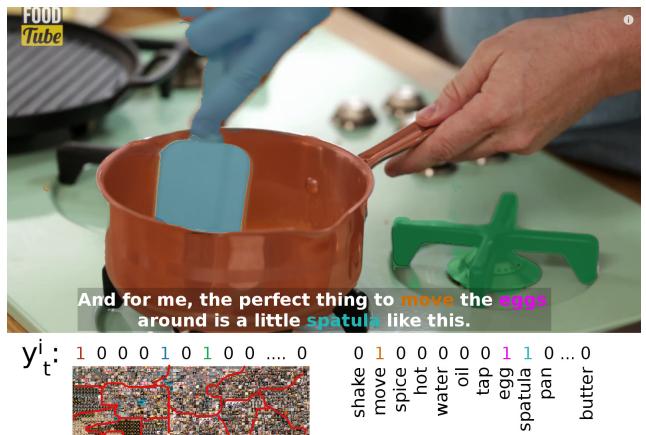


Figure 4: Visualziation of the representation of the Frame.

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 f^i such that f_k^i is 1 if i^{th} video has activity k , 0 otherwise. When the feature vectors of all videos in the corpus is concatenated, it becomes an activity matrix F such that i^{th} row of the F is the activity vector f^i . Moreover, each feature k also has an activity frequency b_k and a distribution parameter Θ_k . Distribution parameter Θ_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 Θ_k and b_k follow the *beta process* as;

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

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 Θ_k , this distribution is shown to be equivalent to Indian Buffet Process [?]. Where videos are customers and activities are dishes in the buffet. The

| |
|-----|
| 378 |
| 379 |
| 380 |
| 381 |
| 382 |
| 383 |
| 384 |
| 385 |
| 386 |
| 387 |
| 388 |
| 389 |
| 390 |
| 391 |
| 392 |
| 393 |
| 394 |
| 395 |
| 396 |
| 397 |
| 398 |
| 399 |
| 400 |
| 401 |
| 402 |
| 403 |
| 404 |
| 405 |
| 406 |
| 407 |
| 408 |
| 409 |
| 410 |
| 411 |
| 412 |
| 413 |
| 414 |
| 415 |
| 416 |
| 417 |
| 418 |
| 419 |
| 420 |
| 421 |
| 422 |
| 423 |
| 424 |
| 425 |
| 426 |
| 427 |
| 428 |
| 429 |
| 430 |
| 431 |

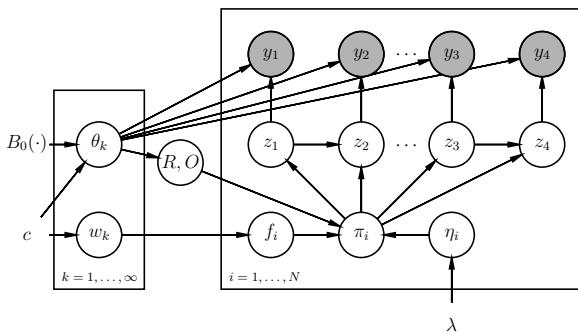
432 first video chooses a Poisson(γ) unique dishes. The fol-
 433 lowing video i chooses previously sampled activity k with
 434 probability $\frac{m_k}{i}$ proportional to number of videos (m_k) cho-
 435 sen the activity k , and it also chooses Poisson($\frac{\gamma}{i}$) new activi-
 436 ties. Here, γ controls the number of active features in each
 437 video and β controls the likelihood of the features getting
 438 shared by multiple videos.
 439

440 After each video chooses a subset of activities, we model
 441 the videos as an Hidden Markov Model (HMM) over the
 442 selected videos. Each frame has the hidden state activity
 443 id(z_t^i) and we observe the binary representation y_t^i . Since
 444 we model each activity as a Bernoulli distribution, the
 445 emmission probabilities follow the Bernoulli distribution as
 446 $p(y_t^i|z_t^i) = Ber(\Theta_{z_t^i})$. Following the construction of the
 447 Fox et al.[?], we sample the transition probabilities from a
 448 normalized Gamma distribution. For each video i , we sam-
 449 ple a Gamma random variable for the transition between ac-
 450 tivity j and activity k if both of the activities are included by
 451 the video (if f_k^i and f_j^i are both 1). After sampling these ran-
 452 dom variables, we normalize them to have proper transition
 453 probabilities. This procedure can be represented formally
 454 as

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

455 Where κ is the persistence parameter promoting self state
 456 transitions to have more coherent temporal boundries, \circ is
 457 the element-wise product and π_j^i is the transition probabili-
 458 ties in video i from state j to all states in the form of a
 459 vector.
 460

461 This model is also presented as a graphical model in Figure 5
 462



463 Figure 5: **Graphical model for BP-HMM:**Some explana-
 464 tion.
 465

466 5.2. Gibbs sampling for BP-HMM

467 We employ Markov Chain Monte Carlo (MCMC)
 468 method for learning and inference of the BP-HMM. We
 469

470 base our algorithms on the MCMC procedure proposed by
 471 Fox et al.[?]. It marginilizes over blah and blah and sam-
 472 ple blah and blah. For faster convergence, we also utilize
 473 a series of data driven samplers. Here we only discuss the
 474 proposed data driven samplers and move the details of the
 475 remainin samplers to the Supplementary Material.
 476

477 Data-Driven Sampler1

478 Data-Driven Sampler2

479 6. Experiments

480 6.1. Dataset

481 We collected blah blah videos from YouTube and did
 482 blah blah... All numbers etc. These are recipes, 5 of them
 483 are evaluation set and labelled temporally and labels are
 484 matched.
 485

486 6.2. Baselines

487 We compare against BP-HMM-O with local feauters and
 488 HMM with Semantic Features, HMM with local features,
 489 HMM with CNN feautes...
 490

491 6.3. Qualitative Results

492 6.4. Accuracy over Activity Detection

493 6.5. Accuracy over Activity Learning

494 7. Discussions and Conclusions

495 Discuss which recipes worked and why. Discuss the im-
 496 portance of semantic representation, scaling features and
 497 multi-modality.
 498

499 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 $\#(z_t = k, z_{t'} = k') = 0 \forall t \leq t'$ |
| $\Theta_k = [\Theta_k^v, \Theta_k^l]$ | 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 | $\pi_i^{k,k'}$ | $\eta_i^{k,k'} \times f_i^k \times f_i^{k'}$ | | |

- 540
- 541
- 542
- 543
- 544
- 545
- 546
- 547
- 548
- 549
- 550
- 551
- 552
- 553
- 554
- 555
- 556
- 557
- 558
- 559
- 560
- 561
- 562
- 563
- 564
- 565
- 566
- 567
- 568
- 569
- 570
- 571
- 572
- 573
- 574
- 575
- 576
- 577
- 578
- 579
- 580
- 581
- 582
- 583
- 584
- 585
- 586
- 587
- 588
- 589
- 590
- 591
- 592
- 593
- [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 storyline 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 Cook-

- 594
- 595
- 596
- 597
- 598
- 599
- 600
- 601
- 602
- 603
- 604
- 605
- 606
- 607
- 608
- 609
- 610
- 611
- 612
- 613
- 614
- 615
- 616
- 617
- 618
- 619
- 620
- 621
- 622
- 623
- 624
- 625
- 626
- 627
- 628
- 629
- 630
- 631
- 632
- 633
- 634
- 635
- 636
- 637
- 638
- 639
- 640
- 641
- 642
- 643
- 644
- 645
- 646
- 647
- ing 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 Ninth 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 Vi-*

| | | |
|-----|-----------------------------------------------------------|-----|
| 648 | sion (ICCV), 2013 IEEE International Conference on, pages | 702 |
| 649 | 1681–1688. IEEE, 2013. 2 | 703 |
| 650 | | 704 |
| 651 | | 705 |
| 652 | | 706 |
| 653 | | 707 |
| 654 | | 708 |
| 655 | | 709 |
| 656 | | 710 |
| 657 | | 711 |
| 658 | | 712 |
| 659 | | 713 |
| 660 | | 714 |
| 661 | | 715 |
| 662 | | 716 |
| 663 | | 717 |
| 664 | | 718 |
| 665 | | 719 |
| 666 | | 720 |
| 667 | | 721 |
| 668 | | 722 |
| 669 | | 723 |
| 670 | | 724 |
| 671 | | 725 |
| 672 | | 726 |
| 673 | | 727 |
| 674 | | 728 |
| 675 | | 729 |
| 676 | | 730 |
| 677 | | 731 |
| 678 | | 732 |
| 679 | | 733 |
| 680 | | 734 |
| 681 | | 735 |
| 682 | | 736 |
| 683 | | 737 |
| 684 | | 738 |
| 685 | | 739 |
| 686 | | 740 |
| 687 | | 741 |
| 688 | | 742 |
| 689 | | 743 |
| 690 | | 744 |
| 691 | | 745 |
| 692 | | 746 |
| 693 | | 747 |
| 694 | | 748 |
| 695 | | 749 |
| 696 | | 750 |
| 697 | | 751 |
| 698 | | 752 |
| 699 | | 753 |
| 700 | | 754 |
| 701 | | 755 |