

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

127 All aforementioned methods are using supervised labels either as strong image-word pairs or weak image-caption
 128 pairs. On the other hand, our method is fully unsupervised.
 129

130 Activity Detection/Recognition:

131 **Recipe Understanding** Following the interest in community generated recipes in the web, there have been many attempts to automatically process recipes. Recent methods
 132 on natural language processing [18, 26] focus on semantic parsing of language recipes in order to extract actions
 133 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 successful autonomous cooking experiment. In addition to the lan-

guage 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.
 162
 163
 164
 165
 166
 167
 168
 169
 170
 171
 172
 173

174 3. Method

175 Explanation of the Figure 3 with pointers to the Sections.
 176

177 4. Semantic Multi-Modal Frame Representation

178 4.1. Learning Atoms to Represent Multi-Modal Data

179 4.1.1 Learning Language Atoms

180 4.1.2 Learning Visual Atoms

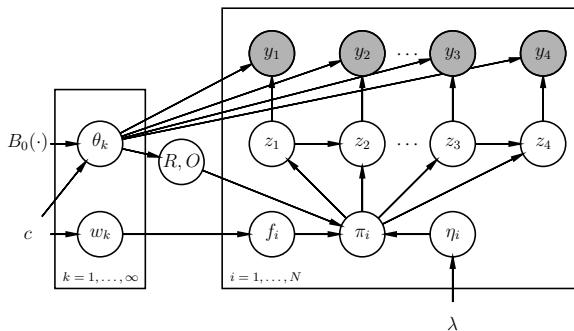
181 Generating Framewise Object Proposals

182 Joint Proposal Clustering to Detect Visual Objects

183 4.2. Multi-Modal Representation via Learned Atoms

184 5. Unsupervised Activity Representation

185 5.1. BP-HMM with Orderings (BP-HMM-O)



186 Figure 1. Graphical model for BP-HMM-O

216

Query: How to make an ommellette?

270

217

271

218

272

219

273

220

274

221

275

222

276

223 Semantic Multi-Modal Representation

277

224

278

225

279

226

280

227

281

228

282

229

283

230

284

231

285

232

286



Lorem ipsum dolor sit amet, consectetur adipiscing elit.

soft fast eat sort seven to shell had rabe
besides heat cold beat q white silk thick score
center center edge top edge
bulldog lighting so totally burning cook
quaker popcorn so melting overcooked
tall quiet pan starting undercooked
whatever in the edge mushrooms star
body medium browning ingredients
medium ring making flipped 3 probably 5 oregano
redpointing m seconds zucchini
Inside color melt foodchines
fresh add salsa shrimp
use side milk
also slice
slice



Multi-Video Co-Clustering

Salient Action Verbs/Object Names
(Language Atoms)

Object Proposals

Object Clusters
(Visual Atoms)**233 Unsupervised Activity Representation**

287

234

288

235

289

236

290

237

291

238

292

239

293

240

294

241

295

242

296

243

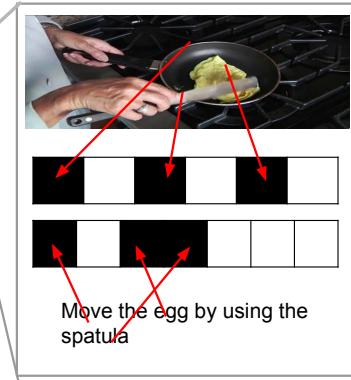
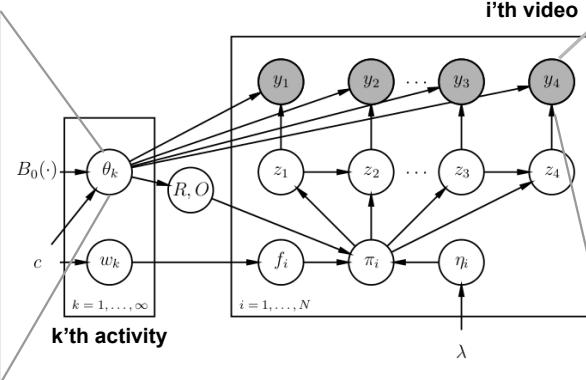
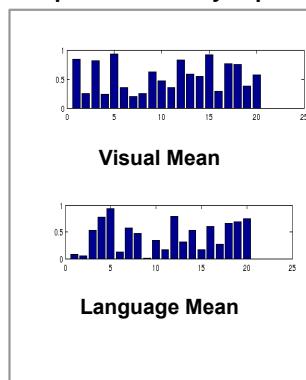
297

244

298

245

299

**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.

303

304

246

305

247

306

248

307

249

308

250

309

251

310

252

311

253

312

254

313

255

314

256

315

257

316

258

317

259

318

260

319

261

320

262

321

263

322

264

323

265

324

266

325

267

326

268

327

269

328

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

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

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

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

- [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
- [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. Rabovich, 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 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

432	[28] H. Yu and J. M. Siskind. Grounded language learning from	486
433	video described with sentences. In <i>ACL (1)</i> , pages 53–63,	487
434	2013. 2	488
435	[29] C. L. Zitnick and D. Parikh. Bringing semantics into fo-	489
436	cus using visual abstraction. In <i>Computer Vision and Pat-</i>	490
437	<i>tern Recognition (CVPR), 2013 IEEE Conference on</i> , pages	491
438	3009–3016. IEEE, 2013. 2	492
439	[30] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learn-	493
440	ing the visual interpretation of sentences. In <i>Computer Vi-</i>	494
441	<i>sion (ICCV), 2013 IEEE International Conference on</i> , pages	495
442	1681–1688. IEEE, 2013. 2	496
443		497
444		498
445		499
446		500
447		501
448		502
449		503
450		504
451		505
452		506
453		507
454		508
455		509
456		510
457		511
458		512
459		513
460		514
461		515
462		516
463		517
464		518
465		519
466		520
467		521
468		522
469		523
470		524
471		525
472		526
473		527
474		528
475		529
476		530
477		531
478		532
479		533
480		534
481		535
482		536
483		537
484		538
485		539