YouCookII Dataset

Luowei Zhou
Robotics Institute
University of Michigan
luozhou@umich.edu

Chenliang Xu
Department of CS
University of Rochester
chenliang.xu@rochester.edu

Jason J. Corso
Department of EECS
University of Michigan
jjcorso@umich.edu

Abstract

Learning from instructional video is a promising direction that may help ground the vision and language problem. To move toward this goal, we collect a large-scale cooking video dataset, called YouCookII, with 2000 videos downloaded from YouTube. All the videos are untrimmed, under unconstrained environment and in third person viewpoint. They represent a more challenging visual problem than existing instructional datasets. The annotations of the videos include the temporal boundaries for procedure steps of each video the corresponding English descriptions for each step. All the frame-wise features and annotations are available for download on the dataset webpage: http://youcook2.eecs.umich.edu.

1 Introduction

YouCookII contains 2000 long untrimmed videos from 89 cooking recipes. **The procedure steps for each video are annotated with temporal boundaries and described by imperative English sentences** (see example in Fig. 1). The videos were downloaded from YouTube and are all in the third-person viewpoint. All the videos are unconstrained and can be preformed by individual persons at their houses with unfixed cameras. YouCookII contains rich recipe types and various cooking styles from all over the world.

We compare YouCookII with commonly-used instructional video datasets, namely, YouCook [3], MPII [6], 50Salads [7], Breakfast [4] and Coffee [1] in Tab. 1. Some important features that distinguish YouCookII from existing datasets of instructional videos are summarized as follows:

- Large-scale cooking dataset with 2000 annotated videos.
- Unconstrained videos, can be preformed by individual persons at their houses with unfixed cameras.
- Long untrimmed videos, up to 10 minutes.
- Extremely rich recipe types and various cooking styles from all over the world.
- Procedure steps temporally localized and described by English sentences.

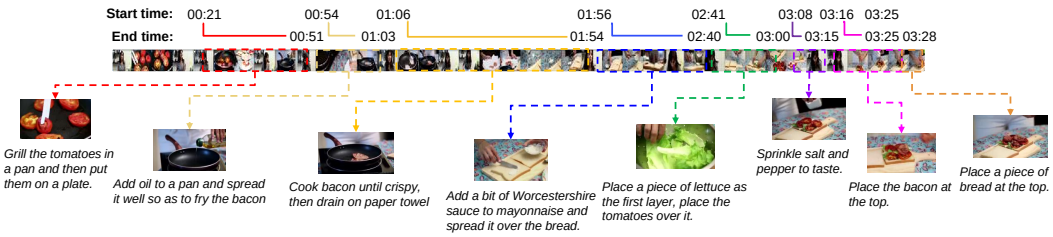


Figure 1: An example annotation on making a BLT sandwich. Each procedure step has time boundaries annotated and is described by a English sentence. Video from YouTube with ID: 4eWzsx1vAi8.

Table 1: Comparisons of instructional video datasets. UnCons. stands for Unconstrained Scene and Proc. Ann. is short for Procedure Annotation.

Name	Duration	UnCons.	Proc. Ann.
YouCook	140 m	Yes	No
MPII	490 m	No	No
50Salads	320 m	No	No
Coffee	120 m	Yes	No
Breakfast	67 h	Yes	No
YouCookII	**176h**	**Yes**	**Yes**

The dataset is organized as follows in the all-in-one file.

YouCookII/annotations: contains the single annotation file youcookii-annotations-trainval.json for training and validation sets.

YouCookII/splits: contains 6 files. train_list.txt, val_list.txt and test_list.txt are lists of videos for each splits. train_duration_totalframe.txt, val_duration_totalframe.txt and test_duration_totalframe.txt store the duration and total frame for each video.

YouCookII/features/feat_csv: frame-wise resnet-34 feature in .csv format for training, validation and testing sets. See Sec. 5 for more details.

YouCookII/features/feat_dat: frame-wise resnet-34 feature in binary format (.dat) for training, validation and testing sets. Can only be read by Lua Torch for fast access.

YouCookII/scripts: contains the script for downloading YouCookII videos.

YouCookII/label_foodtype.csv: mappings of recipe ID to recipe names.

YouCookII/youcookii_readme.pdf: same as this document.

Please cite the following paper if you find the dataset useful:

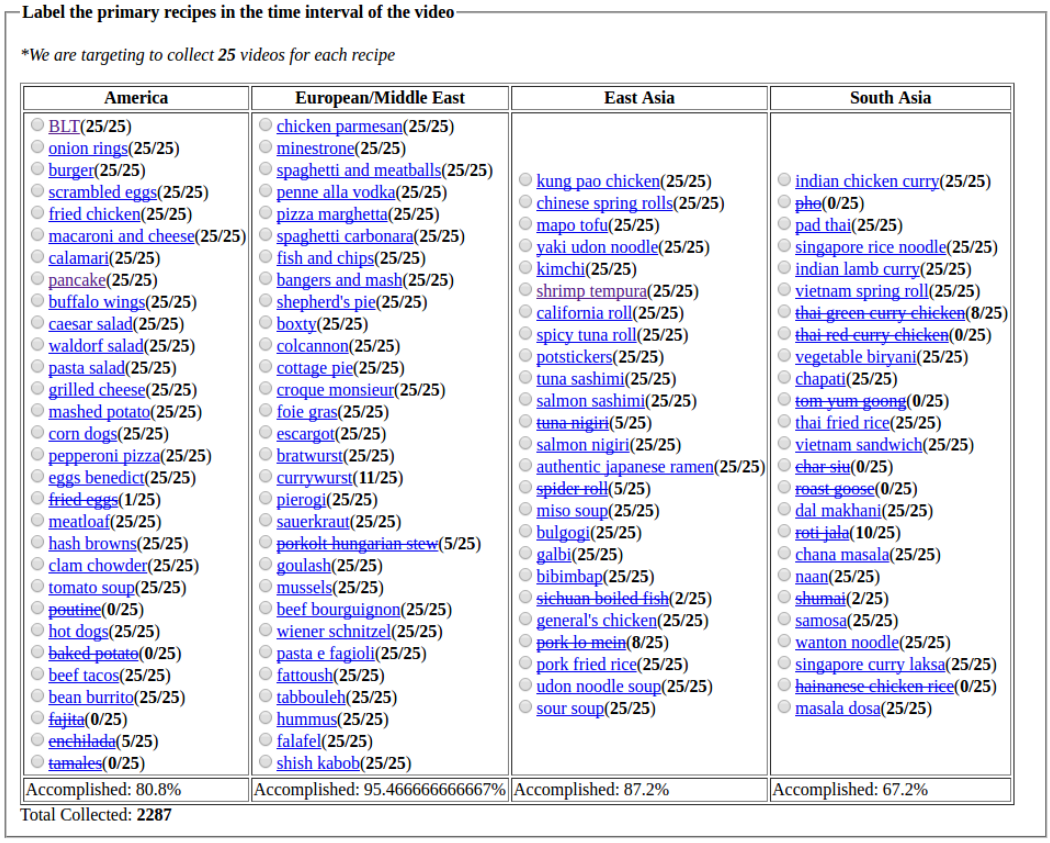
```
@article{zhou2017procnets,
  title={Towards Automatic Learning of Procedures from Web Instructional Videos},
  author={Zhou, Luowei and Xu, Chenliang and Corso, Jason J},
  journal={arXiv preprint arXiv:1703.09788},
  year={2017}
}
```

In the follow sections, we describe some details of YouCookII dataset in four aspects: 1) Data acquisition, 2) Annotations, 3) Dataset statistics, and 4) Precomputed feature.

2 Data acquisition

We allow workers to collect video data through a web interface. The interface can record various information for a given video, which is listed as follows.

- Video URL.
- Number of cooks.
- Sound options, include English Speech, Native Language Speech, English Speech and Background Music, and Native Language Speech and Background Music.
- Time length estimation.
- Time interval of the cooking process. Start time and end time.
- Cooking style. Choose one from 11 types. Broiling/Baking, Steaming, Grilling, Roasting, Stewing, Frying, Raw/Cold-prep, Slow Cooking, Boiling, Kneading and Deep-frying.



Label the primary recipes in the time interval of the video

We are targeting to collect **25 videos for each recipe*

America	European/Middle East	East Asia	South Asia
BLT(25/25) onion rings(25/25) burger(25/25) scrambled eggs(25/25) fried chicken(25/25) macaroni and cheese(25/25) calamari(25/25) pancake(25/25) buffalo wings(25/25) caesar salad(25/25) waldorf salad(25/25) pasta salad(25/25) grilled cheese(25/25) mashed potato(25/25) corn dogs(25/25) pepperoni pizza(25/25) eggs benedict(25/25) ~~fried eggs~~(1/25) meatloaf(25/25) hash browns(25/25) clam chowder(25/25) tomato soup(25/25) ~~poutine~~(0/25) hot dogs(25/25) ~~baked potato~~(0/25) beef tacos(25/25) bean burrito(25/25) ~~fajita~~(0/25) ~~enchilada~~(5/25) ~~tamales~~(0/25)	chicken parmesan(25/25) minestrone(25/25) spaghetti and meatballs(25/25) penne alla vodka(25/25) pizza margherita(25/25) spaghetti carbonara(25/25) fish and chips(25/25) bangers and mash(25/25) shepherd's pie(25/25) boxty(25/25) colcannon(25/25) cottage pie(25/25) croque monsieur(25/25) foie gras(25/25) escargot(25/25) bratwurst(25/25) currywurst(11/25) pierogi(25/25) sauerkraut(25/25) ~~porkolt hungarian stew~~(5/25) goulash(25/25) mussels(25/25) beef bourguignon(25/25) wiener schnitzel(25/25) pasta e fagioli(25/25) fattoush(25/25) tabbouleh(25/25) hummus(25/25) falafel(25/25) shish kabob(25/25)	kung pao chicken(25/25) chinese spring rolls(25/25) mapo tofu(25/25) yaki udon noodle(25/25) kimchi(25/25) shrimp tempura(25/25) california roll(25/25) spicy tuna roll(25/25) potstickers(25/25) tuna sashimi(25/25) salmon sashimi(25/25) ~~tuna nigiri~~(5/25) salmon nigiri(25/25) authentic japanese ramen(25/25) ~~spider roll~~(5/25) miso soup(25/25) bulgogi(25/25) galbi(25/25) bibimbap(25/25) ~~sichuan boiled fish~~(2/25) general's chicken(25/25) ~~pork lo mein~~(8/25) pork fried rice(25/25) udon noodle soup(25/25) sour soup(25/25)	indian chicken curry(25/25) ~~pho~~(0/25) pad thai(25/25) singapore rice noodle(25/25) indian lamb curry(25/25) vietnam spring roll(25/25) ~~thai green curry chicken~~(8/25) ~~thai red curry chicken~~(0/25) vegetable biryani(25/25) chapati(25/25) ~~tom yum goong~~(0/25) thai fried rice(25/25) vietnam sandwich(25/25) ~~char siu~~(0/25) ~~roast goose~~(0/25) dal makhani(25/25) ~~roti jala~~(10/25) chana masala(25/25) naan(25/25) ~~shumai~~(2/25) samosa(25/25) wanton noodle(25/25) singapore curry laksa(25/25) ~~hainanese chicken rice~~(0/25) masala dosa(25/25)
Accomplished: 80.8%	Accomplished: 95.466666666667%	Accomplished: 87.2%	Accomplished: 67.2%

Total Collected: **2287**

Figure 2: Cooking video acquisition status by recipe type. The goal for each recipe is to acquire 25 unique videos. For some recipes, existing video tutorials are rare. We end up picking 89 recipes closest to the goal to construct the YouCookII dataset. The rest might appear in the future version of the dataset.

- Recipe type. 110 recipes in total. 89 of them are used to construct YouCookII.

Recipe types are specified as the keywords to retrieve videos from YouTube. For each recipe type, we collect at most 25 videos. Each video is within 10 mins and should be recorded by camera devices but not slideshows. All the videos are in third-person viewpoint and are made under various kitchen environments. The videos collected should have high resolutions and be recently uploaded.

YouCookII consists of 89 recipes from four major cuisine locales, i.e., America, European/Middle East, East Asia and South Asia. The full recipe list is shown in Fig. 2. After removing videos that are no longer available online, we collect a total of 2000 unique YouTube videos with 23 videos per recipe in average. We randomly split the videos belonging to each recipe into 67%:23%:10% as the training, validation and testing sets. This leads to 1333 videos for training, 457 videos for validation and 210 videos for testing.

3 Annotations

We acquire structured recipe descriptions for each video. Recipe steps are temporally annotated with the starting time and ending time. Each step is described by a human annotator with a English sentence. Each recipe contains 3 to 16 steps, where each step is described by a sentence in imperative form, such as *grill the tomatoes in a pan*.

The annotators have access to audio and subtitles but are required to organize and summarize the descriptions in their own way. Also, the annotators should not be biased by the user-uploaded recipe

Figure 3: Top 100 actions/objects in the YouCookII recipe descriptions. Generated from TagCrowd.

descriptions, mostly of which are in casual forms. The full requirement of the YouCookII annotation task is shown below.

- Each clip (recipe step) is less than two minutes long.
- Each recipe includes 5 to 16 steps; each step should be described with one sentence. Ordering among the steps counts.
- Generally, each sentence should have less than 20 words, use proper grammar and punctuation, and be in imperative form.
- The starting time and the ending time for each step are recorded in the form HH:MM:SS.
- For some videos, the user-uploaded recipe is present in the text box. For the annotation task, these recipes should be discarded and be not read or used. Only use the video/audio/subtitle to generate the recipe description.

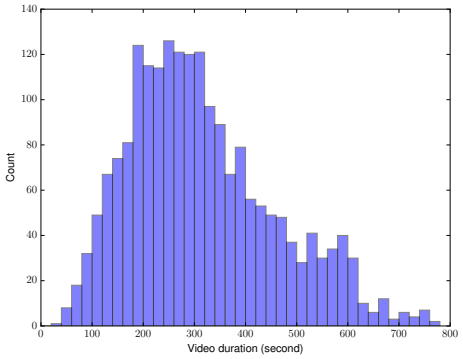
An example of the annotation is shown in Tab. 2. Due to the complexity of fine-grained recipe annotation, we hire well-trained native English speakers as the annotators instead of crowdsourcing. As indicated in prior work [2], people generally agree with boundaries of salient events in video and hence we collect one annotation per video. To reflect the human consensus on how a procedure should be segmented, we annotate each video with two annotators, one for the major effort and the other one for verification.
So far, fine-grained recognition from video streams is rather challenging, especially for some ingredients and utensils in cooking videos. We might further annotate the dataset by spatial-temporally segmenting cooking tools, ingredients, related objects and actions in the future update of YouCookII.

4 Dataset Statistics.

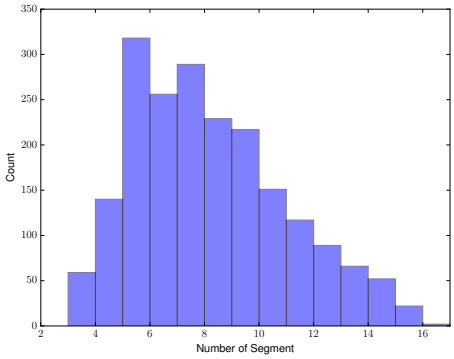
The distribution of video duration is shown in Fig. 4(a). The total video length is 175.6 hours with an average duration of 5.27 min per video. The distribution of number of segments per video is shown in Fig. 4(b) and the overall mean and standard deviation are 7.7 and 2.8. The mean and standard deviation of the number of procedure segments for each recipe are shown

Table 2: An example of the annotated recipes. ID: 4eWzsx1vAi8.

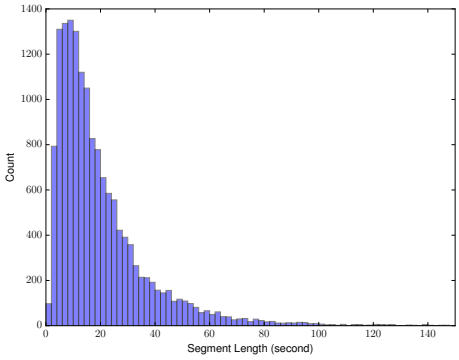
ID	Start Time	End Time	Description
1	00:00:21	00:00:51	Grill the tomatoes in a pan and then put them in a plate
2	00:00:54	00:01:03	Add oil to a pan and spread it well so as to fry the bacon
3	00:01:06	00:01:54	Cook bacon until crispy, then drain on paper towel
4	00:01:56	00:02:40	Add a bit of Worcestershire sauce to mayonnaise and spread it over the bread
5	00:02:41	00:03:00	Place a piece of lettuce as the first layer, place the tomatoes over it
6	00:03:08	00:03:15	Sprinkle salt and pepper to taste
7	00:03:16	00:03:25	Place the bacon at the top
8	00:03:25	00:03:28	Place a piece of bread at the top



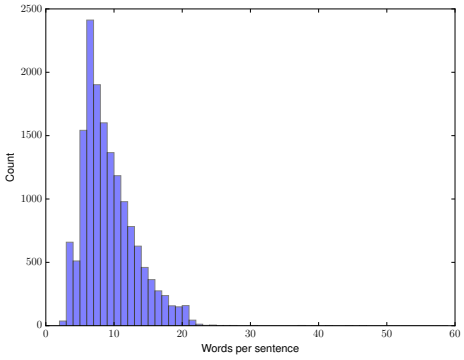
(a) Distribution of video duration.



(b) Distribution of number of segments per video.



(c) Distribution of segment duration.



(d) Distribution of number of words per sentence.

Figure 4: Dataset statistics.

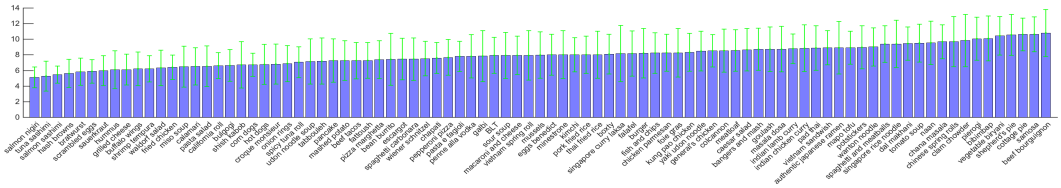


Figure 5: Mean and standard deviation of number of procedure segments for each recipe.

in Fig. 5. The distribution of segment durations is shown in Fig. 4(c) with mean and standard deviation of 19.6s and 18.2s, respectively. The longest segment lasts 264s and the shortest one lasts 1s.

For the recipe descriptions, the total vocabulary is around 2600 words and the top 100 frequent actions/objects are shown in Fig. 3. The distribution of number of words per sentence is shown in Fig. 4(d) with mean and standard deviation of 8.8 words and 3.9 words.

Other facts include: 1) the majority of the recipes are operated by a single person, 2) 56% of the video sound is in English speech and the rest is in native language.

5 Precomputed Feature.

We first explain how we sample frames from each video. Suppose the total frame for the video is T, number of sampled frames is F, then the sampling interval is $I = \lceil \frac{T}{F} \rceil$ frames starting from frame 0.

In our case, $F = 500$ for all the videos so the average sampling rate is 1.58 fps. During training, we temporally augment the data by sampling each video at most $R = 10$ times starting from frame $\max(\lfloor \frac{I}{R} \rfloor, 1) \times r$, where $r = 0, 1, ..., R-1$. Information on video duration and total frame can be downloaded from the dataset website.

We provide frame-wise ResNet-34 feature [1] pretrained on ImageNet dataset on image classification task and fine-tuned on image captioning tasks [8, 9] on MSCOCO dataset [5]. The feature vector is the activation output before the last fully-connected layer. The feature files are stored in .csv format. Note that for fast access, we also provide binary feature files (.dat) that can be read by Lua Torch. Other features, such as two stream features (RGB + optical flows), will be provided upon request.

References

[1] J.-B. Alayrac, P. Bojanowski, N. Agrawal, I. Laptev, J. Sivic, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[2] C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, and K. A. Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721, 2017.

[3] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641, 2013.

[4] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 780–787, 2014.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[6] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201. IEEE, 2012.

[7] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738. ACM, 2013.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[9] L. Zhou, C. Xu, P. Koch, and J. J. Corso. Image caption generation with text-conditional semantic attention. *arXiv preprint arXiv:1606.04621*, 2016.

[1] `https://github.com/facebook/fb.resnet.torch`