A Survey of Procedural Video Datasets

Hui Li Tan, Keng-Teck Ma, Hongyuan Zhu, Mark Rice, Joo-Hwee Lim, Cheston Tan
Institute for Infocomm Research, A*STAR
{hltan, makt, zhuh, mdrice, joohwee, cheston-tan}@i2r.a-star.edu.sg

Abstract

Motivated by building knowledge bases of procedural knowledge for video understanding, we present the first survey of procedural video datasets. This survey covers 15 procedural video datasets, sub-categorized into instructional and non-instructional video datasets. The goal of the survey is to examine the current state of procedural video datasets, as well as to discuss the future of such datasets, suggesting possible steps to bring this area to the next level.

1. Introduction

Human knowledge can be divided into declarative and procedural knowledge. Declarative knowledge refers to facts or information while procedural knowledge refers knowledge of how to perform or operate. There have been various works at building knowledge bases encompassing declarative knowledge, such as Cyc [11], WordNet [20], ImageNet [6], NEIL [3], etc. On the other hand, constructing knowledge bases encompassing procedural knowledge remains little discussed. Nonetheless, the inclusion of procedural knowledge will greatly enrich existing knowledge bases, and are valuable for helping human and robots learn and execute new tasks [16] [2]. In this paper, we explore valuable video dataset resources for procedural knowledge that are readily available in the research community.

Procedural videos, videos containing structured information on how a task should be completed, are important resources for procedural knowledge. These videos typically depict series of actions performed in some constrained but non-unique order to achieve some intended outcomes. Examples include videos on cooking, assembly, repair, crafts, beauty tutorials, academic tutorials, etc.

Advantageously, instructional videos are rich in procedural knowledge, as they offer explicit guidance on the procedures. Made with the intention to teach on performing certain task, they are typically well paced with clear section demarcations, and have consistent viewpoints with minimal occlusion and shake/jitter. Auxiliary information in the form of audio or text is typically available. However, these videos require significant effort to create, involving careful set-up or post-processing to create or align the auxiliary information. In addition, there are also non-instructional videos with rich procedural knowledge. As these videos are not meant to be didactic, they only offer implicit guidance on the procedures. Auxiliary information in the form of audio and text may also not be available.

In this paper, we systematically survey all known procedural video datasets, including both instructional and non-instructional video datasets. Through the survey, we seek to understand the trends and gaps in existing datasets, as well as gain insights into the future of such datasets.

2. Datasets Covered

This survey covers 15 procedural video datasets, including six instructional and 10 non-instructional video datasets. The instructional video datasets are: YouCook [4], What's Cookin' [12], YouCookII [21], What's Cookin' with reference resolution [8], "5 tasks" [1], and Arduino Assembly [9]. The non-instructional video datasets are: TUM Kitchen [19], CMU Multi-Modal Activity (CMU-MMAC) [5], Actions for Cooking Eggs (ACE) [17], MPII cooking activities [14], 50Salads [18], Human Manipulation Action [13], Breakfast Actions [10], MPII cooking 2 [15], Ikea Furniture Assembly (Ikea FA) [7], and Arduino Assembly [9]. The Arduino Assembly dataset falls under both categories as it comprises instructional videos to teach the subjects, as well as, videos of subjects performing the tasks after watching the instructional videos.

3. Analysis and Discussion

The datasets will be characterized and discussed along the following aspects:

- Modalities covered by dataset (e.g., video, depth information, motion capture, inertial measurement sensor data, audio, text, etc.)
- Scale of dataset (e.g., size of dataset, length of videos, number of videos, etc.)
- Type of task being performed (e.g., food preparation, mechanical tasks, etc.)

- Type of environment (from laboratory settings, to real-world surveillance environments, and to in the wild settings crawled from the internet)
- Human subject characteristics (number of subjects, single or multiple subjects, novice or expert, mode of data collection, etc.)
- Variety of objects (e.g., ingredients, tools, etc.)
- Problem (fine-grain and composite activity recognition to procedure segmentation, “state-action-state” discovery, visual linguistic reference resolution, etc.)
- Type of ground-truth labels (temporal granularity and spatial granularity of ground-truth labels)

These datasets will then be analysed as follows:

- What is the current variety of modalities, and how have multiple modalities been leveraged together?
- What is the current variety of tasks covered, and are some important task types missing?
- How diverse are the datasets, and is there any obvious bias, e.g., in the atomic actions covered, or in the way in which subjects perform the tasks?
- How are the datasets used and how do they fit into the bigger picture of building knowledge bases with procedural knowledge?
- Overall, what is a possible roadmap for the evolution of such datasets, and what are the potential next steps?

4. Concluding Remarks

There is a need for growth in the scale and variety of the procedural video datasets. For instance, the majority of the datasets are on food preparation, with some recent exploration towards mechanical tasks. Moreover, most of these datasets involve single subjects. Datasets on procedures involving other task types and multiple subjects would help in understanding other scenarios involving complex interactions between subjects and objects. Towards automatically building large-scale knowledge bases with procedural knowledge, there are various challenging problems, such as visual linguistic reference resolution and unsupervised learning from procedural videos.

References

[1] J.-B. Alayrac, P. Bojanowski, N. Agrawal, I. Laptev, J. Sivic, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *IEEE CVPR*, 2016. 1

[2] M. Beetz, M. Tenorth, and J. Winkler. Open-EASE. In *IEEE ICRA*, pages 1983–1990, May 2015. 1

[3] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. In *IEEE ICCV*, pages 1409–1416, Dec 2013. 1

[4] 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 *IEEE CVPR*, pages 2634–2641, Jun 2013. 1

[5] F. de la Torre, J. K. Hodgins, J. Montano, and S. Valcarcel. Detailed human data acquisition of kitchen activities: the CMU-multimodal activity database (CMU-MMAC). In *CHI Workshop*, 2009. 1

[6] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, Jun 2009. 1

[7] T. Han, J. Wang, A. Cherian, and S. Gould. Human action forecasting by learning task grammars. *CoRR*, abs/1709.06391, 2017. 1

[8] D. A. Huang, J. J. Lim, L. Fei-Fei, and J. C. Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. In *IEEE CVPR*, pages 1032–1041, 2017. 1

[9] W. L. Koh, J. Kaliappan, M. Rice, K. T. Ma, H. H. Tay, and W. P. Tan. Preliminary investigation of augmented intelligence for remote assistance using a wearable display. In *IEEE TENCON*, pages 2093–2098, Nov 2017. 1

[10] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE CVPR*, pages 780–787, Jun 2014. 1

[11] D. B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, Nov. 1995. 1

[12] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. *CoRR*, abs/1503.01558, 2015. 1

[13] A. Pieropan, G. Salvi, K. Pauwels, and H. Kjellstrm. Audio-visual classification and detection of human manipulation actions. In *IEEE IROS*, pages 3045–3052, Sep 2014. 1

[14] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *IEEE CVPR*, pages 1194–1201, Jun 2012. 1

[15] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *IJCV*, pages 1–28, 2015. 1

[16] A. Saxena, A. Jain, O. Sener, A. Jami, D. K. Misra, and H. S. Koppula. Robobrain: Large-scale knowledge engine for robots. *CoRR*, abs/1412.0691, 2014. 1

[17] A. Shimada, K. Kondo, D. Deguchi, G. Morin, and H. Stern. Kitchen scene context based gesture recognition: A contest in ICPR2012. In *IEEE ICPR*, pages 168–185, 2013. 1

[18] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM Ubicomp*, 2013. 1

[19] M. Tenorth, J. Bandouch, and M. Beetz. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE ICCV Workshops*, pages 1089–1096, Sep 2009. 1

[20] P. University. About wordnet. 2010. 1

[21] L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. *arXiv preprint arXiv: 1703.09788*. 1