Supplementary: My View is the Best View: Procedure Learning from Egocentric Videos

Siddhant Bansal[1], Chetan Arora[2], and C.V. Jawahar[1]

[1] Center for Visual Information Technology, IIIT, Hyderabad
[2] Indian Institute of Technology, Delhi
siddhant.bansal@research.iiit.ac.in

1 Appendix

This document contains additional details to support the main paper.

Fig. 1. Issues with standard datasets for procedure learning. Existing datasets [1, 6, 10, 12, 15–17] majorly consist of third-person videos. They contain issues like occlusion and atypical camera locations that make them ill-suited for procedure learning. Additionally, the datasets rely on noisy videos from YouTube [6, 12, 15, 17]. In contrast, we propose to use egocentric videos that overcome the issues posed by third-person videos. To this end, we create the EgoProceL dataset.

1.1 Outline

Figure 1 highlights issues with standard third-person datasets, motivating us to use egocentric videos for procedure learning. In Section 2, we discuss the annotation protocols, task-level details, and datasets excluded while creating the EgoProceL dataset. In Section 3, we highlight multiple use-cases for our work. In Section 4.1, we provide additional ablation results on EgoProceL. To facilitate reproducing the results reported in the main paper and supplementary, Section 4.2 lists the hyper-parameters used for CnC. Furthermore, we release the EgoProceL dataset and code for the work on project's webpage[3].

[3] Link 1: http://cvit.iiit.ac.in/research/projects/cvit-projects/egoprocel; Mirror link 2: https://sid2697.github.io/

2 EgoProceL

This section contains additional details on the proposed EgoProceL dataset.

2.1 Annotation Protocols followed for EgoProceL

CMU-MMAC [3], EPIC-Tents [8], MECCANO [13], PC Assembly, PC Disassembly: A list of key-steps required to perform the task was created upon viewing the videos. Two annotators were asked to identify the key-steps in the videos and temporally mark the start and end locations. Once an annotator added temporal segments to the videos, the other annotator verified them. We use the ELAN software [5] to annotate the videos.

EGTEA Gaze+ [11]: We used the recipes provided by the dataset curators to create the key-step's list for each task. The dataset offers dense activity annotations for all the videos. We created a one-to-many mapping between the key-steps and the provided annotations; this accelerated the annotations process. The mapping generated was used to create key-step annotations for all videos. Three people further watched the videos and verified the annotations generated.

To accelerate future research, we release the EgoProceL dataset on the project web page[3].

2.2 Task-level details of EgoProceL

In Table 1, we share the statistics for each of the 16 tasks in the EgoProceL dataset. Let N be the number of videos, K be the number of key-steps for a task, u_n be the number of unique key-steps and g_n be the number of annotated key-steps for n^{th} video. Following [6], we calculate the following:

Foreground Ratio: It is the ratio of total duration of the key-steps to the total duration of the video. This helps to understand the amount of background actions a task has. The foreground ratio is inversely proportional to the amount of background. It is calculated as:

$$F = \frac{\sum_{n=1}^{N} \frac{t_k^n}{t_v^n}}{N} \tag{1}$$

Here, t_k^n and t_v^n are the key-step duration and video duration for n^{th} video, respectively. The range of F is between 0 and 1.

From Table 1, we can see that the tasks have significant variance in the foreground ratio. Conversely, tasks like "PC Assembly" and "Tent Assembly" have a high foreground ratio, suggesting fewer background actions. On the other hand, tasks like preparing "Bacon and Eggs" and "Turkey Sanwich" have low foreground ratios, suggesting more background actions.

Table 1. Statistics of the EgoProceL across different tasks. The high range of the foreground ratio and repeated steps highlights the complexity of the tasks involved in EgoProceL

Task	Videos Count	Key-steps Count	Foreground Ratio	Missing Key-steps	Repeated Key-steps
PC Assembly	14	9	**0.79**	0.02	0.65
PC Disassembly	15	9	0.72	**0.00**	0.60
Toy Bike Assembly	20	**17**	0.50	0.06	0.32
Tent Assembly	29	12	0.63	0.14	0.73
Bacon and Eggs	16	11	0.15	0.22	0.51
Cheese Burger	10	10	0.22	0.22	0.65
Continental Breakfast	12	10	0.23	0.20	0.36
Greek Salad	10	4	0.25	0.18	0.77
Pasta Salad	19	8	0.25	0.19	**0.86**
Hot Dog Pizza	6	8	0.31	0.13	0.62
Turkey Sandwich	13	6	0.21	0.01	0.52
Brownie	**34**	9	0.44	0.19	0.26
Eggs	33	8	0.26	0.05	0.26
Pepperoni Pizza	33	5	0.53	**0.00**	0.26
Salad	**34**	9	0.32	0.30	0.14
Sandwich	31	4	0.25	0.03	0.37

Missing Key-steps: This measure captures the count of missed key-steps in each video. It is defined as:

$$M = 1 - \frac{\sum_{n=1}^{N} u_n}{KN} \tag{2}$$

The range of M is between 0 and 1. It helps understand if a task can be done even if we miss some steps. For example, in Table 1, "Salad" has the highest missing key-steps ratio suggesting that salad can be made if we miss multiple key-steps. This makes sense, as one can miss adding mayonnaise to the salad but still create an edible salad. On the other hand, tasks like "PC Disassembly" and "Pepperoni Pizza" can not afford to miss key-steps as the task won't be complete. So, for such tasks, we see a missing key-step ratio of 0.

Repeated Key-steps: This measure captures the repetitions of key-steps across the videos. It is defined as:

$$R = 1 - \frac{\sum_{n=1}^{N} u_n}{\sum_{n=1}^{N} g_n} \tag{3}$$

The range of R is between 0 and 1. Higher values of R indicate repetitions of key-steps across videos. From Table 1, we can see preparing "Pasta Salad" has the highest repeated key-steps and preparing "salad" has the lowest. Methods that do not consider repetitions of the key steps, will not perform well for such tasks. As CnC takes repetitions of the key steps into consideration, it performs well.

2.3 Datasets not included in EgoProceL

As mentioned in the main paper, we followed a set of criteria to select videos from existing datasets for including in EgoProceL. Here we discuss two potential datasets which we could not use for EgoProceL.

The Charades-Ego dataset [14], consisting of paired egocentric and third-person videos, is essential for activity recognition. However, it is not practical for procedure learning. The subjects do not perform a series of key-steps to achieve a goal; instead, they perform activities like pouring a drink in a cup and having it. Additionally, the average duration of the videos is 31.2 seconds compared to 13 minutes in EgoProceL, suggesting the briefness of the tasks acted out.

The EPIC-Kitchens dataset [2], consisting of 100 hours of kitchen recordings, comes quite close to our requirements. However, due to the unscripted nature of the dataset (which sets it apart from [11]), it becomes unsuitable. As for procedure learning, we need videos of the same tasks performed multiple times.

3 Applications

Learning a procedure by observing multiple videos of the same task opens up a range of possible applications.

Monitoring procedures: Consider a system trained to know the key-steps for performing a task; if a new person does the same task again, the system will identify if the person misses a step or does a step differently.

Guidance systems: A system trained to know the key-steps for performing a task can identify the current step and show the next possible step for performing the task.

Automated systems: The proposed framework benefits by enabling automated robotic systems to autonomously learn the key-steps for performing the task by observing the task being performed. Once the automated system learns the key-steps, the next time, it can do the task without any human assistance.

4 Additional Experimental Details

4.1 Ablation Results

This section contains ablation results on parts of EgoProceL. Table 2 contains the results obtained upon replacing the TC3I loss with TCC [4], LAV [7], and a combination of LAV and TCC [7]. Additionally, Table 3 shows the results obtained upon using various values of K. Finally, Table 4 shows the results obtained after considering different combination of losses along with HC and SS for [3, 11].

Consistent with the results obtained in the main paper, in Table 2, we observe highest results when using the proposed TC3I loss. This is because TC3I accounts for the loss of temporal coherency by TCC [4] with the help of C-IDM loss [7]. Additionally, the TC3I loss focuses on correspondences at the frame level as compared to global alignment employed by LAV [7].

Table 2. Effectiveness of the TC3I loss. Results after replacing TC3I loss in CnC with TCC, LAV, and a combination of LAV and TCC. For the majority of the cases, the proposed TC3I loss outperforms all the losses as it focuses on the frame-level correspondences and adds temporal coherency by adopting the C-IDM loss

Experiment	MECCANO [13]			EPIC-Tent [8]		
	Precision	F-Score	IoU	Precision	F-Score	IoU
TCC+PCM	15.1	17.9	**8.7**	14.2	14.9	7.8
LAV+TCC+PCM	13.4	15.6	7.3	16.0	16.7	**8.5**
LAV+PCM	14.6	17.4	7.1	15.2	15.8	8.3
TC3I+PCM (CnC)	**15.5**	**18.1**	7.8	**17.1**	**17.2**	8.3

Experiment	PC Assembly			PC Disassembly		
	Precision	F-Score	IoU	Precision	F-Score	IoU
TCC+PCM	19.9	21.7	11.6	22.0	23.4	12.2
LAV+TCC+PCM	21.6	21.1	10.8	21.0	24.3	12.3
LAV+PCM	21.5	22.7	11.7	26.4	26.5	12.9
TC3I+PCM (CnC)	**25.0**	**25.1**	**12.8**	**28.4**	**27.0**	**14.8**

Consistent with our observations in the main paper, in Table 3, we achieve the highest scores when $K = 7$. Additionally, for most cases, CnC results in the highest scores for all the values of K.

Table 3. Selecting the value of K. Numbers in **bold** are highest in the respective row and underlined numbers are highest in the respective column

Experiment	MECCANO [13]				EPIC-Tents [8]			
	K=7	K=10	K=12	K=15	K=7	K=10	K=12	K=15
Random	**13.4**	10.1	8.8	7.4	**14.1**	10.6	9.1	8.3
TC3I+HC	**16.6**	14.0	11.4	10.8	**15.4**	12.1	10.6	9.9
TC3I+SS	**16.3**	12.6	12.2	10.7	**15.9**	11.9	10.7	10.4
CnC	**18.1**	15.2	13.5	11.9	**17.2**	11.1	12.1	9.46

Experiment	PC Assembly				PC Disassembly			
	K=7	K=10	K=12	K=15	K=7	K=10	K=12	K=15
Random	**15.1**	11.0	10.4	9.2	**15.3**	11.8	10.7	9.6
TC3I+HC	**21.7**	17.3	20.7	19.2	**24.9**	18.3	18.0	20.7
TC3I+SS	**24.7**	18.1	18.1	19.7	**23.6**	19.7	21.0	20.7
CnC	**25.1**	18.7	20.7	19.0	**27.0**	26.5	24.5	23.6

Table 4 shows the results after using various losses with HC, SS, and PCM for procedure learning [3, 11]. Nearly all the experiments using PCM achieve the

Table 4. Effectivenss of PCM. Results after replacing PCM with HC and SS with different losses

Experiment	CMU-MMAC [3]				EGTEA Gaze+ [11]			
	Precision	Recall	F-Score	IoU	Precision	Recall	F-Score	IoU
TCC+HC	17.06	19.47	18.08	8.55	16.78	20.00	18.25	8.33
TCC+SS	17.34	19.71	18.31	8.66	16.96	20.29	18.48	8.18
TCC+PCM	**18.46**	**21.45**	**19.71**	**9.46**	**17.46**	**22.71**	**19.74**	**8.81**
LAV+TCC+HC	17.37	18.40	17.76	8.61	**16.59**	19.72	18.02	7.35
LAV+TCC+SS	17.46	17.94	17.57	8.53	16.16	20.05	17.90	7.39
LAV+TCC+PCM	**18.80**	**21.11**	**19.71**	**9.03**	16.44	**21.40**	**18.60**	**7.45**
LAV+HC	18.44	19.78	19.07	8.66	16.59	18.18	17.35	7.87
LAV+SS	17.82	18.99	18.36	8.53	16.08	18.13	17.04	7.87
LAV+PCM	**20.62**	**21.95**	**21.11**	**9.40**	**17.42**	**21.17**	**19.12**	**8.02**
TC3I+HC	18.47	20.27	19.15	8.98	18.74	23.70	20.82	7.93
TC3I+SS	18.53	21.13	19.66	8.86	17.71	24.09	20.36	7.94
CnC	**21.62**	**24.38**	**22.72**	**11.08**	**19.58**	**24.68**	**21.72**	**9.51**

highest scores for other losses. Additionally, we achieve the highest scores with CnC. Due to the characteristics of TC3I loss and PCM, the results are consistent with our previous observations.

4.2 Hyper-parameters

Table 5 lists the hyper-parametes used for CnC.

Table 5. Hyper-parameter settings for CnC.

Hyper-parameter	Value
No. of key-steps (K)	7
No. of sampled frames	32
Batch Size	5
Learning Rate	10^{-4}
Weight Decay	10^{-5}
Window size (σ)	300
Margin (ζ)	2.0
Regularization parameter (ξ)	1.0
No. of context frames (c)	2
Context stride	15
Embedding Dimension	128
Optimizer	Adam [9]

References

1. Alayrac, J.B., Bojanowski, P., Agrawal, N., Laptev, I., Sivic, J., Lacoste-Julien, S.: Unsupervised learning from Narrated Instruction Videos. In: Computer Vision and Pattern Recognition (CVPR) (2016) 1
2. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In: European Conference on Computer Vision (ECCV) (2018) 4
3. De La Torre, F., Hodgins, J., Bargteil, A., Martin, X., Macey, J., Collado, A., Beltran, P.: Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. In: Robotics Institute (2008) 2, 4, 5, 6
4. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal Cycle-Consistency Learning. In: Computer Vision and Pattern Recognition (CVPR) (2019) 4
5. ELAN (Version 6.0) [Computer software]. (2020).: Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan, https://archive.mpi.nl/tla/elan 2
6. Elhamifar, E., Naing, Z.: Unsupervised Procedure Learning via Joint Dynamic Summarization. In: International Conference on Computer Vision (ICCV) (2019) 1, 2
7. Haresh, S., Kumar, S., Coskun, H., Syed, S.N., Konin, A., Zia, Z., Tran, Q.H.: Learning by Aligning Videos in Time. In: Computer Vision and Pattern Recognition (CVPR) (2021) 4
8. Jang, Y., Sullivan, B., Ludwig, C., Gilchrist, I., Damen, D., Mayol-Cuevas, W.: EPIC-Tent: An Egocentric Video Dataset for Camping Tent Assembly. In: International Conference on Computer Vision (ICCV) Workshops (2019) 2, 5
9. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations, (ICLR) (2015) 6
10. Kuehne, H., Arslan, A.B., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Computer Vision and Pattern Recognition (CVPR) (2016) 1
11. Li, Y., Liu, M., Rehg, J.M.: In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video. In: European Conference on Computer Vision (ECCV) (2018) 2, 4, 5, 6
12. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: International Conference on Computer Vision (ICCV) (2019) 1
13. Ragusa, F., Furnari, A., Livatino, S., Farinella, G.M.: The MECCANO Dataset: Understanding Human-Object Interactions From Egocentric Videos in an Industrial-Like Domain. In: Winter Conference on Applications of Computer Vision (WACV). pp. 1569–1578 (2021) 2, 5
14. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and Observer: Joint Modeling of First and Third-Person Videos. In: Computer Vision and Pattern Recognition (CVPR) (2018) 4
15. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: COIN: A Large-Scale Dataset for Comprehensive Instructional Video Analysis. In: Computer Vision and Pattern Recognition (CVPR) (2019) 1

16. Zhou, L., Xu, C., Corso, J.J.: Towards Automatic Learning of Procedures From Web Instructional Videos. In: AAAI Conference on Artificial Intelligence (2018) 1
17. Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: Computer Vision and Pattern Recognition (CVPR) (2019) 1