NurViD: A Large Expert-Level Video Database for Nursing Procedure Activity Understanding

Ming Hu[1,2,3]*, Lin Wang[1,2,4]*, Siyuan Yan[1,2]*, Don Ma[1]*, Qingli Ren[5], Peng Xia[1,2,3], Wei Feng[1,2,3], Peibo Duan[3], Lie Ju[1,2,3], Zongyuan Ge

[1]AIM Lab, Faculty of IT, Monash University

[2]Airdoc-Monash Research, Airdoc

[3]Faculty of Engineering, Monash University

[4]College of Intelligent Systems Science and Engineering, Harbin Engineering University

[5]Nursing College, Shanxi Medical University

Motivation

Monash Medical AI

NEURAL INFORMATION PROCESSING SYSTEMS

The application of deep learning to nursing procedure activity understanding has the potential to greatly enhance the quality and safety of nurse-patient interactions. By utilizing the technique, we can facilitate training and education, improve quality control, and enable operational compliance monitoring. However, the development of automatic recognition systems in this field is currently hindered by the scarcity of appropriately labeled datasets. The existing video datasets pose several limitations:

1) **Small-scale:** these datasets are small-scale in size to support comprehensive investigations of nursing activity.

2) **Lack of diversity and professionalism:** they primarily focus on single procedures, lacking expert-level annotations for various nursing procedures and action steps.

3) **Lack of localization annotation:** they lack temporally localized annotations, which prevents the effective localization of targeted actions within longer video sequences.

Motivation

To mitigate these limitations, we proposed NurViD, a large-scale video benchmark for nursing procedure activity understanding. Compared to existing datasets, NurViD incorporates characteristics from the following aspects:

1) **Diverse procedure and action:** It also contains 1,538 videos depicting 51 nursing procedure categories, covering the majority of common procedures, along with 177 action steps, providing much more comprehensive coverage, compared to previous datasets that primarily focus on single procedures with limited action steps.

2) **Real-world clinic settings:** Videos in NurViD were captured from over ten real clinical environments according to our statistics, including hospitals, clinics, and nursing homes.

3) **Expert-level annotations:**NurViD was labeled by professionals with high expertise 59 and knowledge in nursing.

Dataset Construction

1) Procedure and Action Definition

2) Online Video Crawling

3) Localization Annotation and Quality Control

Dataset Construction

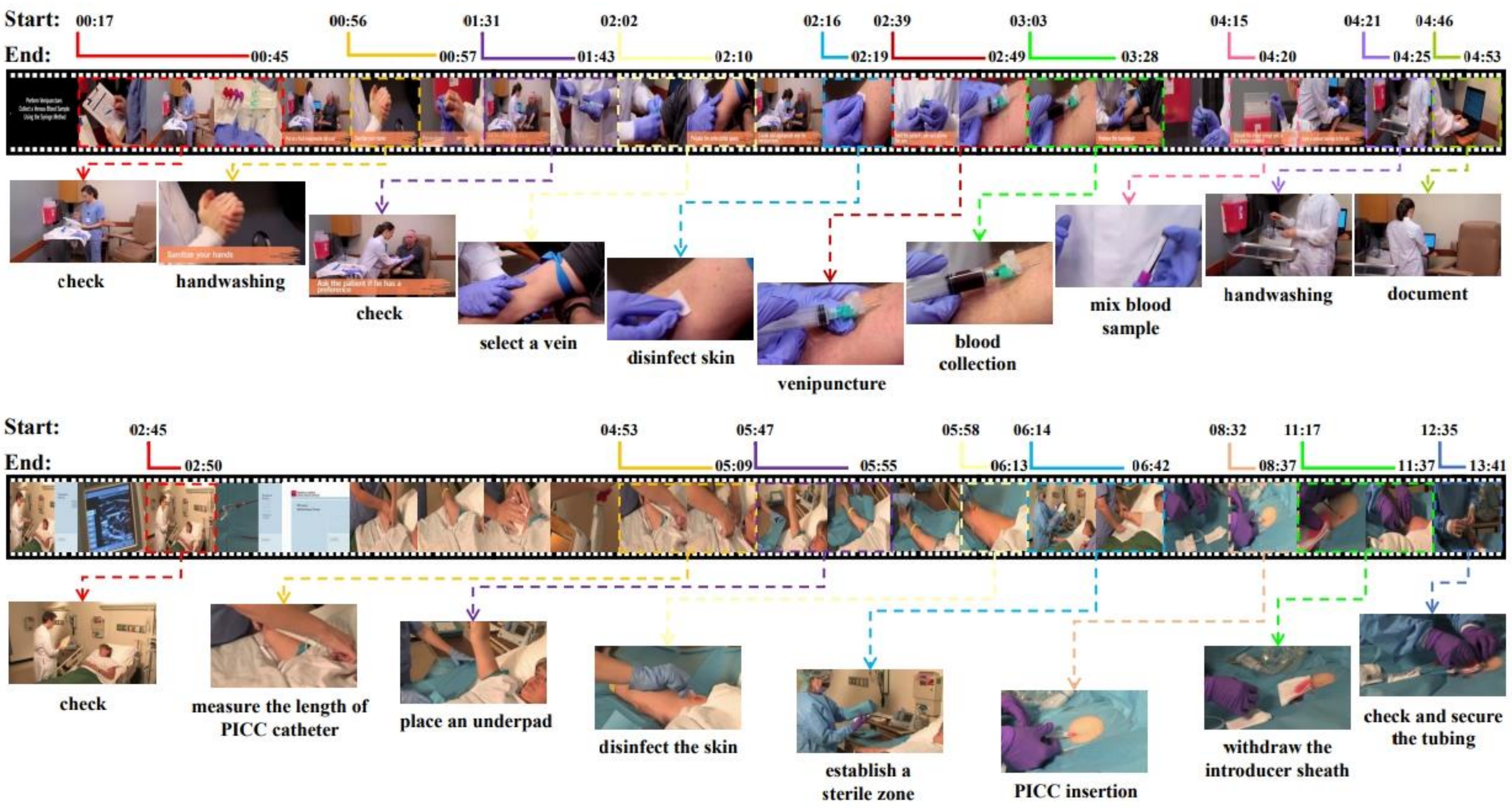


Figure 1: The examples for the annotated target action boundaries for *Intravenous Blood Sampling* and *Modified Seldinger Technique with Ultrasound for PICC Placement* procedures. The frames marked in colored boxes denote the annotated temporal boundaries for the target action steps.

NurViD Statistics

Datasets	Dataset Properties								Tasks		
	Publicly Available?	Expert-Level Annotations?	Real-World Settings?	No. of Videos	No. of Segments	No. of Procedures	No. of Actions	Total Duration	Procedure Recognition	Action Recognition	Action Detection
Llorca et al. [15]	✕	✓	✕	8	-	1	7	4.2min	✕	✓	✕
Ameling et al. [7]	✕	-	✕	24	-	1	6	-	✕	✓	✕
Zhong et al. [48]	✕	-	-	200	1,400	1	7	-	✕	✓	✕
Wang et al. [42]	✕	✕	✓	280	2,760	1	8	-	✕	✓	✕
Kaggle [4]	✓	✕	✕	292	3,504	1	12	23.3h	✕	✓	✕
Lulla et al. [26]	✓	-	✓	3,185	6,689	1	7	38.9h	✕	✓	✕
NurViD (Ours)	✓	✓	✓	1,538	5,608	51	177	144.4h	✓	✓	✓

Table 1: The comparison among existing nursing procedure activity video datasets. Compared to other datasets, NurViD annotates the procedures and actions by following expert-level standards, focuses on more comprehensive coverage of various nursing procedure categories, collects a large number of videos, totaling 144 hours, and also enables action detection tasks.

NurViD Statistics

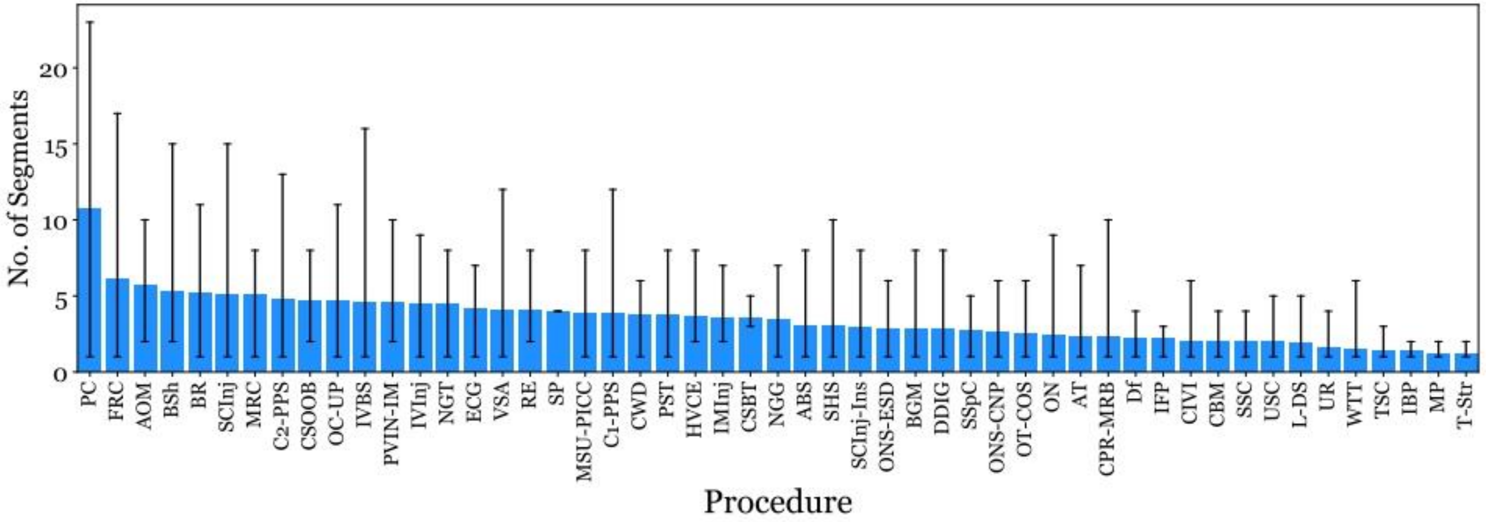
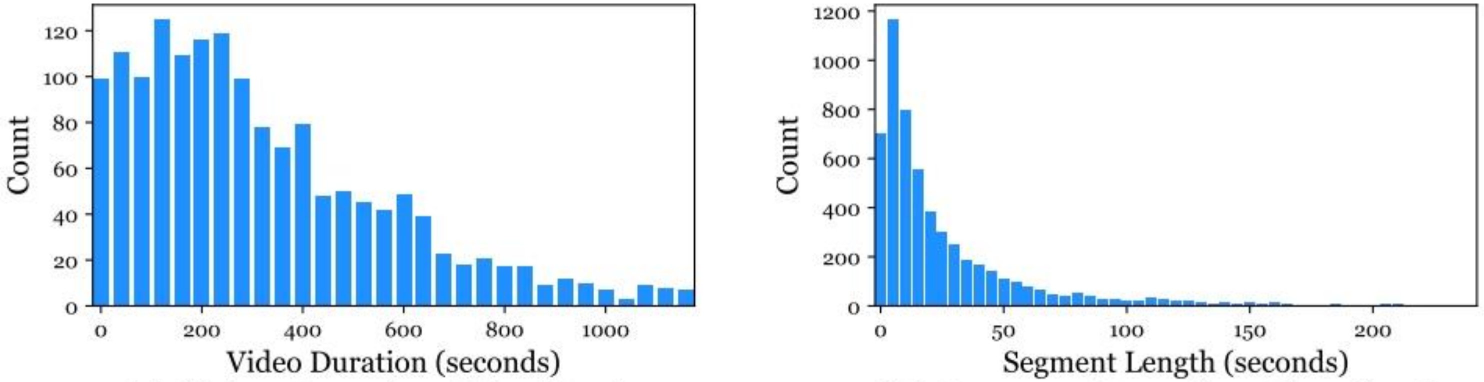


Figure 2: The average, maximum, and minimum number of action segments for each procedure



(a) Video Duration Distribution **(b) Segment Duration Distribution**

Figure 3: NurViD dataset duration statistics.

Experiments

Procedure Classification on Untrimmed Videos

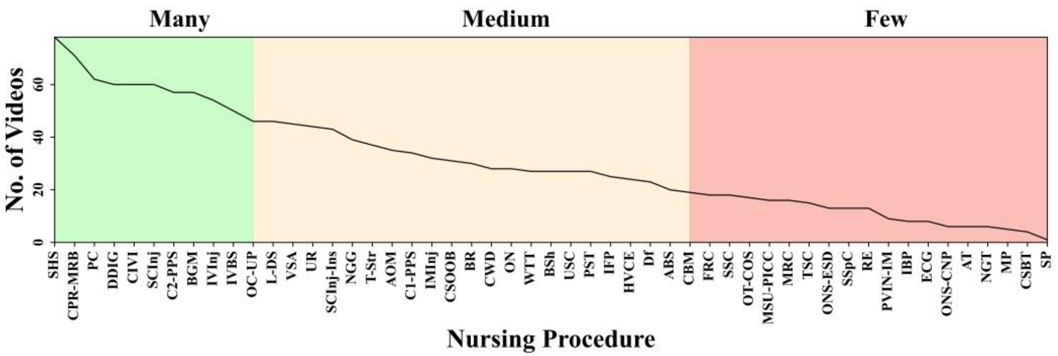


Figure 7: *The number of untrimmed videos per each procedure.* The procedures with the frequency ≥ 50 are grouped into *many*. The procedures with the frequency < 50 and ≥ 20 are grouped into *medium*. The procedures with the frequency < 20 are grouped into *few*. We rank the procedures based on their frequency.

Baselines	Procedure Classification			
	Many 10	Medium 22	Few 18	All 50
SlowFast [14]	9.9	7.5	0.1	7.4
C3D [38]	10.7	5.1	1.8	7.7
I3D [9]	9.9	9.0	2.8	8.7
SlowFast*	19.9	10.2	5.0	13.5
C3D*	**21.5**	11.3	**5.8**	**14.8**
I3D*	19.8	**12.5**	5.6	13.1

Table 2: Per-class Top-1 accuracy for procedure prediction on untrimmed videos. The best performance for each split has been highlighted in **bold**.

Experiments

Monash Medical AI

NEURAL INFORMATION PROCESSING SYSTEMS

Procedure and Action Classification on Trimmed Videos

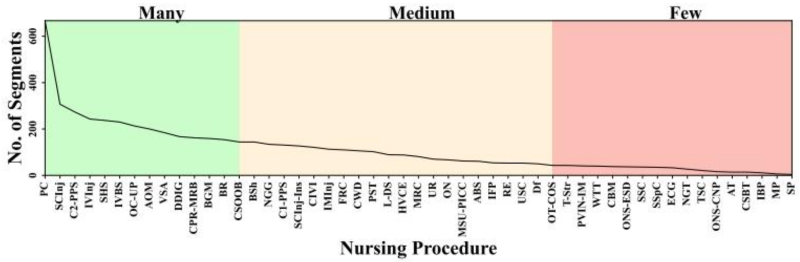


Figure 8: *The number of trimmed videos per each procedure.* The procedures with the frequency ≥ 150 are grouped into *many*. The procedures with the frequency < 150 and ≥ 45 are grouped into *medium*. The procedures with the frequency < 45 are grouped into *few*. We rank the procedures based on their frequency.

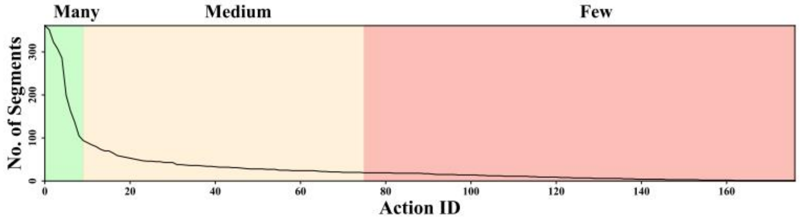


Figure 9: *The number of trimmed videos per each action.* The actions with the frequency ≥ 100 are grouped into *many*. The actions with the frequency < 100 and ≥ 20 are grouped into *medium*. The actions with the frequency < 20 are grouped into *few*. We rank the actions based on their frequency.

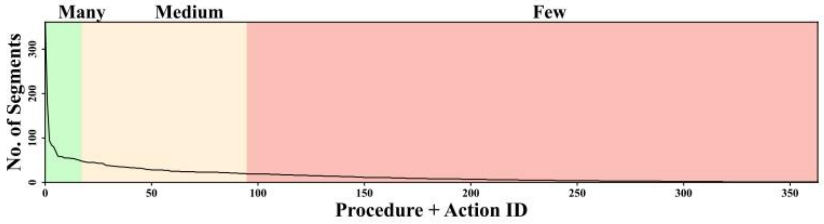


Figure 10: *The number of trimmed videos per each composition of procedure category and action.* The compositions with the frequency ≥ 50 are grouped into *many*. The compositions with the frequency < 50 and ≥ 20 are grouped into *medium*. The compositions with the frequency < 20 are grouped into *few*. We rank the compositions based on their frequency.

Baselines	Procedure Classification				Action Classification				Joint Classification			
	Many 13	Medium 21	Few 17	All 51	Many 9	Medium 66	Few 87	All 162	Many 17	Medium 78	Few 224	All 302
SlowFast [14]	68.9	50.0	33.0	63.0	25.7	10.2	3.2	17.1	12.5	7.2	3.3	7.5
C3D [38]	70.1	48.8	33.0	63.9	22.9	9.3	2.9	15.9	13.8	7.3	3.5	7.7
I3D [9]	67.6	49.9	32.9	62.9	26.3	9.8	4.1	17.9	12.7	7.9	4.0	7.9
SlowFast*	71.2	**61.8**	39.0	68.9	29.8	**15.5**	7.9	21.1	21.2	9.4	5.6	12.8
C3D*	**73.2**	60.0	39.6	**71.2**	28.1	14.6	7.3	**22.8**	**21.8**	**10.8**	**5.6**	**13.1**
I3D*	70.7	60.4	**40.9**	70.0	**31.3**	14.8	**8.2**	21.5	19.5	9.9	4.7	12.5

Table 3: Per-class Top-1 accuracy (%) for the procedure, action, and their joint prediction on trimmed videos. * denotes the initialization from the model pre-trained on Kinetics 400 [23]. The best performance for each split has been highlighted in **bold**.

Experiments

Monash Medical AI

NEURAL INFORMATION PROCESSING SYSTEMS

Action Detection on Untrimmed Videos

Baselines	mAP (%) 0.5	0.6	0.7	0.8	0.9	Avg.
TriDet [33]	30.3	26.7	24.3	20.1	10.7	20.8
TAGS [28]	31.4	26.5	22.6	19.2	11.5	22.4
ActionFormer [46]	**32.9**	**29.6**	**25.8**	**20.8**	**12.7**	**23.9**

Table 4: The results of action detection. We report mAP at the IoU thresholds of [0.5:0.1:0.9]. The average mAP is calculated by averaging the mAP scores across various tIoU thresholds.

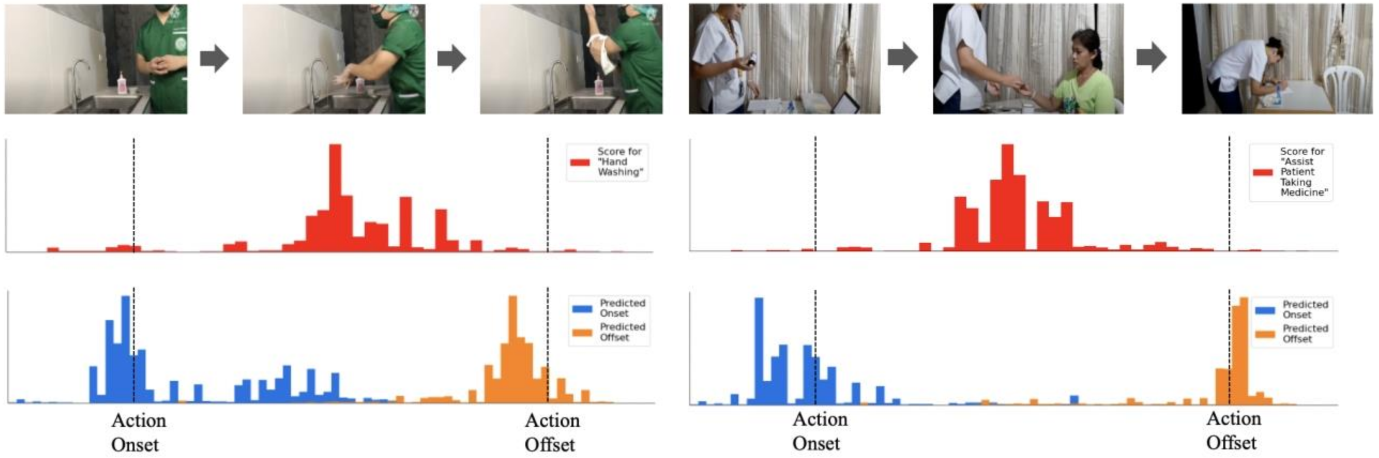


Figure 4: Visualization of action detection results. From top to bottom: (1) input video frames; (2) action scores at each time step; (3) histogram of action onsets and offsets computed by weighting the regression outputs using action scores.

Limitations

1) Intended/Foreseeable Uses

2) Potential Privacy

3) Employment Risks

4) Contestability/Explainability Issues

5) Potential Regional Biases

6) Comprehensiveness of Nursing Procedures and Actions