IndustReal: a dataset for **procedure step recognition** handling execution errors in egocentric videos in an industrial-like setting

Tim J. Schoonbeek[1], Tim Houben[1], Hans Onvlee[2], Peter H.N. de With[1], Fons van der Sommen[1]

[1]Department of Electrical Engineering, Eindhoven University of Technology [2]ASML Research, Veldhoven

Motivation

Increasing interest in industrial action recognition, but:

- No measure of completeness for each action
- No measure of correctness for each action
- Procedure knowledge not explicitly leveraged
- 3D geometry data rarely used, but frequently available

Task definition

New task proposed: **procedure step recognition (PSR)**

Goal. Extract an estimate of all procedure steps correctly performed by a person up to time t.

For some computational model $\mathcal{F}$, PSR is defined as

$$\hat{y}_t = \mathcal{F}(X_t, P)$$

$\hat{y}_t$: recognized procedure execution

X_t : sensory inputs (e.g. video, depth, audio) up to time t

P : descriptive set of the procedural actions

For evaluation criteria, we propose **procedure order similarity** (POS), based on Damerau-Levenshtein distance:

$$POS = 1 - \min\left(\frac{\mathrm{wDamLev}(y, \hat{y})}{|y|}, 1\right),$$

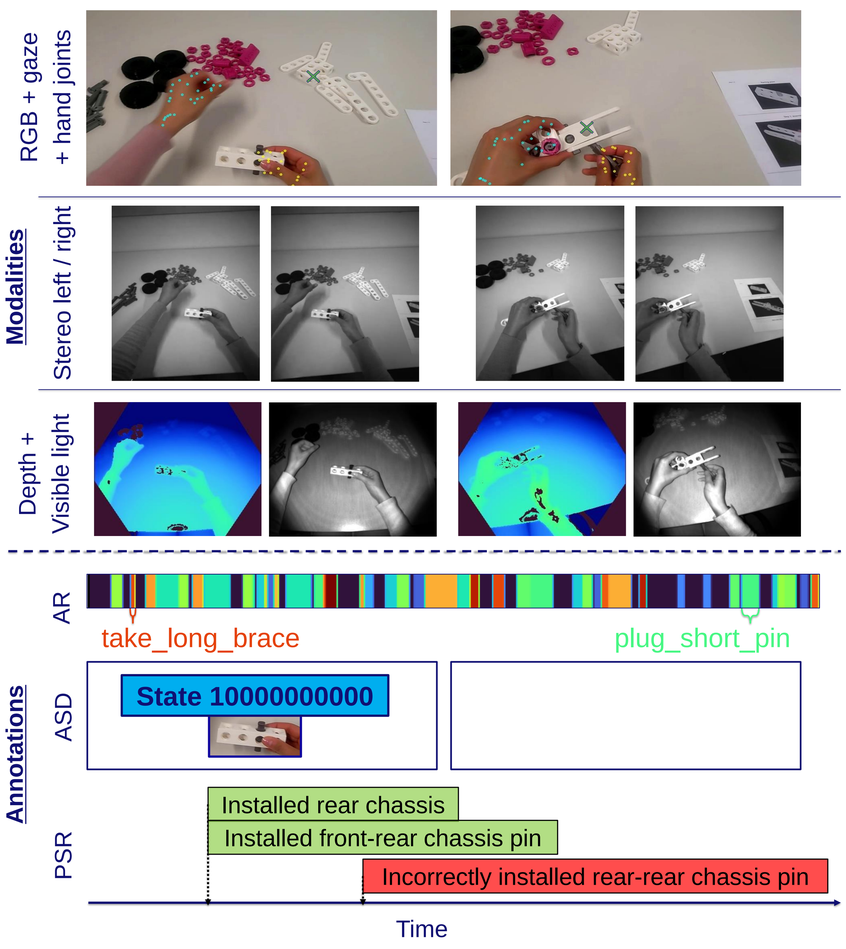
as well as the F_1-score and average delay τ [seconds]

Our proposed **procedure step recognition** task focuses on recognizing *correctly completed* procedural actions, which is key to creating value for industrial applications

We publish the new **IndustReal dataset**:

- ✓ 6 hours of RGB, stereo, depth, gaze, pose tracking data
- ✓ Includes various execution and procedural errors
- ✓ Accompanying CAD data for all parts
- ✓ Benchmarks for procedure step recognition, action recognition, and assembly state detection

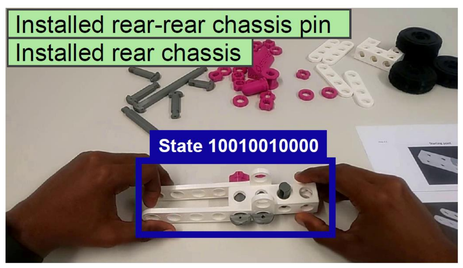
IndustReal dataset

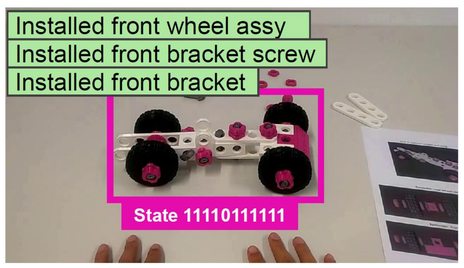


PSR baseline benchmark

	All recordings			Recordings with errors		
	POS	F_1	τ [s]	POS	F_1	τ [s]
B1	0.570	0.779	**14.9**	0.480	0.698	**14.4**
B1-S	0.014	0.206	36.9	0.000	0.174	48.4
B2	0.731	0.860	22.3	0.636	0.784	20.2
B2-S	0.240	0.573	44.4	0.107	0.516	60.5
B3	**0.797**	**0.883**	22.4	**0.731**	**0.816**	20.4
B3-S	0.597	0.734	49.5	0.571	0.731	71.4

B1 : directly translates assembly state detections to procedure steps,
B2 : accumulates the confidence for each detection up to a threshold *T*
B3 : same as *B2*, but limits the possible step completions to those expected in the correct execution of the given procedure with *P*
S : assembly state detections based entirely on synthetic training data





Two fail cases: model does not recognize the execution errors.

Can you?

Other benchmarks

Action recognition

Model	Modalities	Top-1 acc. [%]	Top-5 acc. [%]
SlowFast [11]	RGB	60.39	85.21
MViTv2 [22]	RGB	65.25	87.93
SlowFast [11]	RGB, VL, stereo	62.34	85.97
MViTv2 [22]	RGB, VL, stereo	**66.45**	**88.43**

Assembly state recognition

Pre-trained	Fine-tuned	mAP (b-boxed)	mAP (entire videos)
COCO	Synthetic	0.573	0.341
COCO	IndustReal	0.753	0.553
Synthetic	IndustReal	0.779	0.575
COCO	IndustReal + synthetic	**0.838**	**0.641**

Project page

Let's connect!