Poster 1213

IPN Hand: A Video Dataset and Benchmark for Real-Time Continuous Hand Gesture Recognition

Gibran Benitez-Garcia[1], Jesus Olivares-Mercado[2], Gabriel Sanchez-Perez[2] and Keiji Yanai[1]

[1]The University of Electro-Communications, Tokyo, Japan
[2]Instituto Politecnico Nacional, Mexico City, Mexico

Continuous Hand Gesture Recognition

Problem:

- Recognize dynamic hand gestures with **zero lag**.
- Currently available datasets only include **segmented gestures clips** (one per video).

Objective:

- Introduce a **new dataset** able to evaluate real-time continuous hand gesture recognition with sufficient size for deep learning models.

Contributions

- ***A new dataset named IPN Hand, which includes:***
  - ***Multiple gestures*** *per video with temporal labels (200 videos in total).*
  - *Sufficient* ***number of samples*** *(4,218 instances) from different subjects (50 in total).*
  - ***Dynamic and static hand gestures*** *(13 in total) for controlling touchless screens.*
  - *Natural movements of the hand as* ***non-gesture segments****.*
  - *Real-world scenario with* ***different backgrounds*** *(28 in total) .*
  - ***Publicly available*** *including RGB frames, optical flow maps, and hand masks.*

- ***An alternative input for multimodal real-time hand gesture recognition:***
  - *Real-time* ***hand semantic segmentation*** *to obtain hand masks.*
  - *Evaluation of* ***3D-CNNs*** *with multimodal inputs:* ***RGB-Seg****, and* ***RGB-flow****.*

Challenges of the dataset

- ***Continuous gestures without transition states:***

Challenges of the dataset

- ***Natural behaviors of users' hands as non-gesture states:***

Challenges of the dataset

- ***Intra-class variability of gestures' duration:***

Example gesture: “Double click with one finger”

Challenges of the dataset

- ***Different real-world backgrounds:***

Dataset gestures (*static*)

Class 1: Pointing with one finger

Point-1f: 1010 instances

Class 2: Pointing with two fingers

Point-2f: 1007 instances

Dataset gestures (*dynamic*)

Class 3: Click with one finger

Click-1f: 200 instances

Class 4: Click with two fingers

Click-2f: 200 instances

Dataset gestures (*dynamic*)

Class 5: Throw up

Th-up: 200 instances

Class 6: Throw down

Th-down: 200 instances

Dataset gestures (*dynamic*)

Class 7: Throw left

Th-left: 200 instances

Class 8: Throw right

Th-right: 200 instances

Dataset gestures (*dynamic*)

Class 9: Open twice

Open-2: 200 instances

Class 10: Double click with one finger

2click-1: 200 instances

Dataset gestures (*dynamic*)

Class 11: Double click with two fingers

2click-2: 200 instances

Class 12: Zoom in

Zoom-in: 200 instances

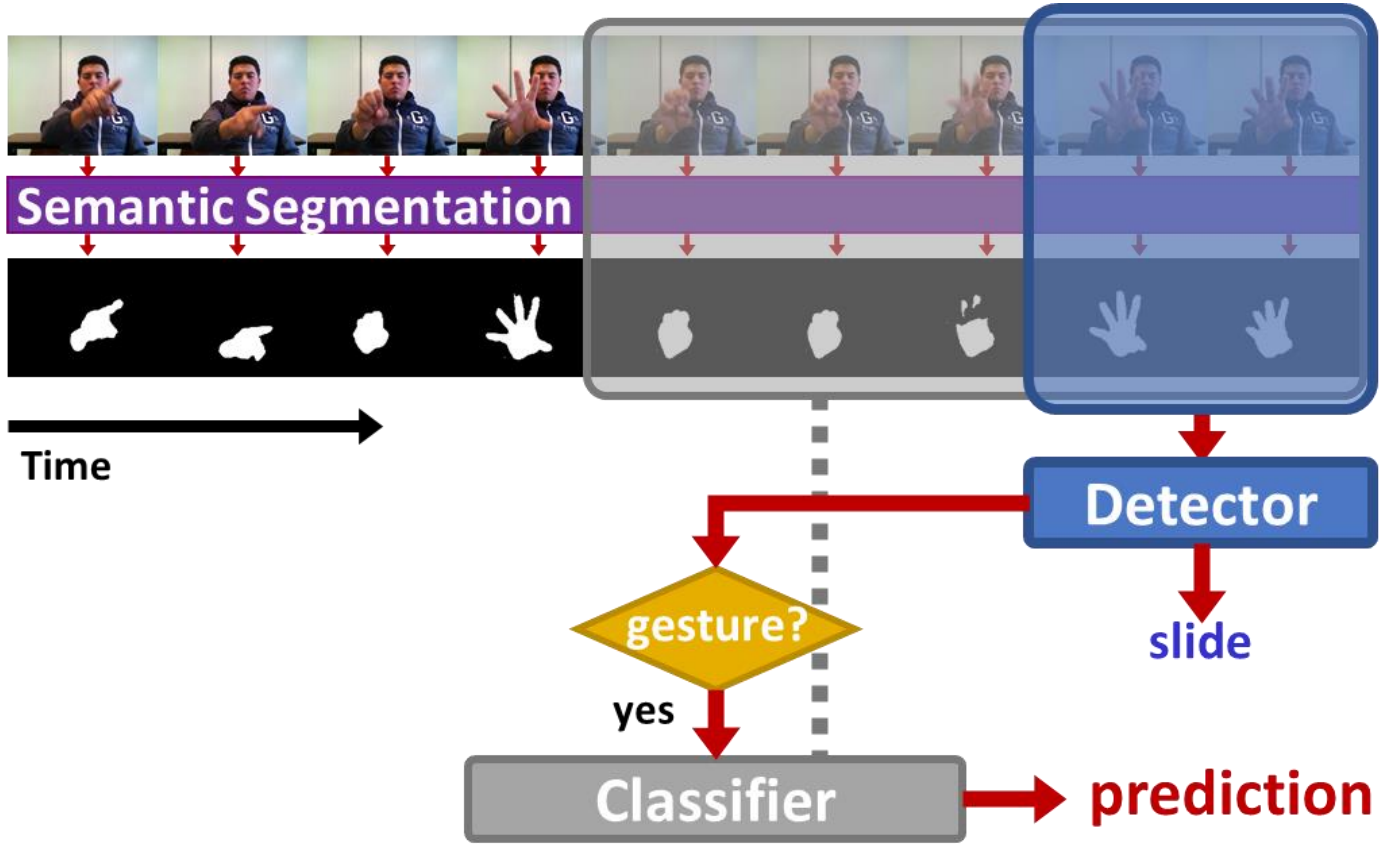
Dataset gestures (*dynamic*)

Class 13: Zoom out

Zoom-out: 200 instances

Benchmark evaluation

- ***Continuous hand gesture recognition:***
  - *Based on a **two hierarchical 3D-CNNs** approach* [1].



[1] Kopuklu, O, et.al. *Online Dynamic Hand Gesture Recognition Including Efficiency Analysis*. T-BIOM 2020

Evaluation results

- ***Continuous hand gesture recognition:***
  - ***Levenshtein accuracy*** *is used as evaluation metric for continuous recognition.*

Model	Modality	Results	Model size			Inference time		
			detector	classifier	total	detector	classifier	total
ResNeXt-101	RGB	25.34	**6.83 MB**	**363 MB**	**370 MB**	**2.9 ms**	**27.7 ms**	**30.1 ms**
ResNeXt-101	RGB-Flow	**42.47**	12.4 MB	363 MB	375 MB	11.1 ms	28.9 ms	53.7 ms
ResNeXt-101	RGB-seg	39.01	22.7 MB	363 MB	386 MB	24.8 ms	28.9 ms	39.9 ms
Resnet-50	RGB	19.78	**6.83 MB**	**353 MB**	**360 MB**	**2.9 ms**	**17.8 ms**	**20.4 ms**
Resnet-50	RGB-Flow	**39.47**	12.4 MB	353 MB	365 MB	11.1 ms	18.2 ms	43.1 ms
Resnet-50	RGB-seg	33.27	22.7 MB	353 MB	376 MB	24.8 ms	18.2 ms	29.2 ms

Evaluation results

- ***Isolated hand gesture recognition:***
  - *The best accuracy is obtained by **RGB-flow with 86.32%** (13 classes only).*

	P1	P2	C1	C2	T-u	T-d	T-l	T-r	O2	2c1	2c2	Z-i	Z-o
Point-1f	92	6	0	0	1	0	0	1	0	0	0	0	0
Point-2f	4	95	0	0	0	0	0	1	0	0	0	0	0
Click-1f	4	4	73	0	0	2	0	0	0	**13**	0	2	2
Click-2f	0	6	0	**63**	0	0	0	4	0	2	**21**	0	4
Th-up	2	2	0	0	85	2	0	0	10	0	0	0	0
Th-down	4	0	0	0	2	92	0	0	2	0	0	0	0
Th-left	0	0	0	0	0	0	94	2	2	0	0	0	2
Th-right	2	0	0	0	0	2	0	96	0	0	0	0	0
Open-2	0	0	0	0	2	0	0	0	87	0	0	**12**	0
2click-1f	6	0	**31**	0	0	0	0	0	0	**60**	2	2	0
2click-2f	0	4	0	**33**	0	0	0	0	0	2	**60**	2	0
Zoom-in	0	0	0	0	0	0	2	2	0	0	0	90	6
Zoom-o	0	4	2	4	0	0	0	4	0	0	0	**8**	79

Conclusion

- ***Conclusions:***
  - *Our new dataset is able to **evaluate hand gesture recognition** for isolated and continuous benchmarks.*
  - *Due to non-gesture segments and specific challenges, the SOTA 3D-CNNs models only **achieve 42.47% accuracy** on the continuous benchmark.*
  - *This demonstrates that our dataset will help to **push advances in this field**.*
- ***Available at:***
  - https://github.com/GibranBenitez/IPN-hand

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Thank You

Code & Models