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

Gibran Benitez-Garcia¹, Jesus Olivares-Mercado², Gabriel Sanchez-Perez² and Keiji Yanai¹

¹The University of Electro-Communications, Tokyo, Japan
²Instituto 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:**

Class: Pointing with one finger
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 \textit{(static)}

\textit{Class 1: Pointing with one finger}
\textbf{Point-1f}: 1010 instances

\textit{Class 2: Pointing with two fingers}
\textbf{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 \((\textit{dynamic})\)

\textbf{Class 11: Double click with two fingers}
\textbf{2click-2:} 200 instances

\textbf{Class 12: Zoom in}
\textbf{Zoom-in:} 200 instances
Dataset gestures \textit{(dynamic)}

\textbf{Class 13: Zoom out}
\textbf{Zoom-out:} 200 instances
Benchmark evaluation

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

---

Evaluation results

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

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

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>C1</th>
<th>C2</th>
<th>T-u</th>
<th>T-d</th>
<th>T-l</th>
<th>T-r</th>
<th>O2</th>
<th>2c1</th>
<th>2c2</th>
<th>Z-i</th>
<th>Z-o</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Point-1f</strong></td>
<td>92</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Point-2f</strong></td>
<td>4</td>
<td>95</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Click-1f</strong></td>
<td>4</td>
<td>4</td>
<td>73</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Click-2f</strong></td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>63</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>21</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>Th-up</strong></td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>85</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Th-down</strong></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>92</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Th-left</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>94</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Th-right</strong></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Open-2</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>2click-1f</strong></td>
<td>6</td>
<td>0</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>2click-2f</strong></td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>60</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Zoom-in</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><strong>Zoom-o</strong></td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>79</td>
<td></td>
</tr>
</tbody>
</table>
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](https://github.com/GibranBenitez/IPN-hand)
Thank You

Code & Models