Recognition of isolated complex mono- and bi-manual 3D hand gestures using discrete IOHMM

Agnès JUST(*) , Sébastien MARCEL(*) & Olivier BERNIER(**)

(*) IDIAP
Rue du Simplon, 4
CH-1920 MARTIGNY

(**) France Telecom Research & Development
FR-22300 LANNION
Outline

1. Introduction/Motivation
2. Hidden Markov Models (HMMs)
3. Input Output Hidden Markov Models (IOHMMs)
4. Description of the Database
5. Preprocessing
6. Experimental Protocol
7. Results
8. Conclusions and future work
Motivation

Related work

Applications
- Recognition of Sign Languages
- Human-Computer Interaction

Limitations
- Various techniques applied to many applications
- No real evaluation of the algorithms

Motivation:
- Comparison of ML algorithm on the same database
- Dynamic gestures $\iff$ sequential ML algorithms (HMM, IOHMM)
- Segmented gestures $\iff$ classification problem
HMMs

- $N$ states, non-observable
- Transition probabilities between these states
  \[ P(q_t = i | q_{t-1} = j), \forall i, j = 1, \ldots, N \]
- Emission probabilities to model the observations
  \[ P(y_t | q_t = i), \forall i = 1, \ldots, N \]
- Training using the EM algorithm
- One HMM per class
Extension of the HMM: map an input sequence to an output sequence

Two conditional distributions:
- Transition probabilities: \( P(q_t = i | q_{t-1}, x_{t-1}) \)
- Emission probabilities: \( P(y_t | q_t, x_t) \)

Trained by the EM algorithm
Continuous IOHMMs
- Modeling of conditional distributions using Neural Networks
- Difficult to train (time, convergence)

Discrete IOHMMs
- Modeling of conditional distributions using look-up tables
- Quantization of the data sequences required
- Easy to train, fast
- Theoretically, better discrimination than with HMMs
Outline

1. Introduction/Motivation
2. Hidden Markov Models (HMMs)
3. Input Output Hidden Markov Models (IOHMMs)
4. Description of the Database
5. Preprocessing
6. Experimental Protocol
7. Results
8. Conclusions and future work
16 gestures: mono- and bi-manual gestures

Examples:

- Stop/No
- Raise/Hello
- Directions
- Pointing
- Bi-manual
The Database 2/2

- Raw Data: Stereo video recordings
- 20 persons × 5 sessions × 10 shots
- Features: 3D trajectories of the hands, head and torso using stereo blob tracking

---

O. Bernier and D. Collobert, “Head and Hands 3D Tracking in Real Time by the EM algorithm” Proceeding of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001
Preprocessing

- **Normalization step:**
  - Maximum arm spread
  - Range of the 3D coordinates between $-0.5$ and $0.5$
  - Keep only the 3D hand trajectories
  - Compute the $\Delta$ for each coordinate and each hand $\Rightarrow$ 12 features $(x_l, y_l, z_l, x_r, y_r, z_r, \Delta x_l, \Delta y_l, \Delta z_l, \Delta x_r, \Delta y_r, \Delta z_r)$

- **Quantization step:**
  - K-means for each gesture class (75 clusters)
  - 16 K-means models merged into a single one
  - Quantization of each frame of each sequence: one discrete value = index of the nearest cluster
Protocol

- Database: divided into 3 subsets (person independent)
- Minimum, average and maximum number of frames for the different subsets

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Evaluation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimum number of frames</td>
<td>12</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>average number of frames</td>
<td>25</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td>maximum number of frames</td>
<td>64</td>
<td>71</td>
<td>89</td>
</tr>
<tr>
<td># sequences</td>
<td>4000</td>
<td>4000</td>
<td>8000</td>
</tr>
</tbody>
</table>
Experiments

- Choice of the hyper-parameters: number of K-means and number of states
- cross-validation on the training set to select \( K \)
- selection and validation of the IOHMM models on the training and evaluation set respectively
- retrain a model with the selected parameters
- results on the test set

Classification: \( \arg\max_c P(y_1^T = c | x_1^T) \)
Results

Classification rate on the test set:

<table>
<thead>
<tr>
<th></th>
<th>stop</th>
<th>no</th>
<th>raise</th>
<th>hello</th>
<th>direction</th>
<th>bi-manual gestures</th>
<th>pointing</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOHMM</td>
<td>59.8%</td>
<td>91.2%</td>
<td>54.8%</td>
<td>82.2%</td>
<td>65.8%</td>
<td>97.3%</td>
<td>68.9%</td>
</tr>
<tr>
<td>HMM</td>
<td>54.8%</td>
<td>81%</td>
<td>26.4%</td>
<td>87.2%</td>
<td>60.4%</td>
<td>98.3%</td>
<td>67.4%</td>
</tr>
</tbody>
</table>

Comparison between HMMs and IOHMMs:

<table>
<thead>
<tr>
<th></th>
<th>error rate</th>
<th>#parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 N</td>
<td>31%</td>
<td>80600</td>
</tr>
<tr>
<td>15 N</td>
<td>30%</td>
<td>123000</td>
</tr>
<tr>
<td>IOHMM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 N</td>
<td>31%</td>
<td>68400</td>
</tr>
<tr>
<td>5 N</td>
<td>26%</td>
<td>126000</td>
</tr>
</tbody>
</table>
Conclusions and Future Work

- Interesting results
  - Bi-manual gestures well classified
  - Still many misclassifications between mono-manual gestures
    - certainly due to the quantization

- Future work
  - Apply other features to increase the classification power of IOHMMs
  - Work on unsegmented gestures
  - Use Continuous IOHMM
## Results

<table>
<thead>
<tr>
<th></th>
<th>stop</th>
<th>no</th>
<th>raise</th>
<th>hello</th>
<th>left</th>
<th>right</th>
<th>up</th>
<th>down</th>
<th>front</th>
<th>back</th>
<th>swim</th>
<th>fly</th>
<th>clap</th>
</tr>
</thead>
<tbody>
<tr>
<td>stop</td>
<td>59.8</td>
<td>0.8</td>
<td>1.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.4</td>
<td>0.2</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>no</td>
<td>20.6</td>
<td>91.2</td>
<td>2.8</td>
<td>8.8</td>
<td>0.6</td>
<td>3.4</td>
<td>2.4</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>raise</td>
<td>3.8</td>
<td>0</td>
<td>54.8</td>
<td>8.8</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>hello</td>
<td>2.2</td>
<td>1.4</td>
<td>39.6</td>
<td>82.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>left</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>87.8</td>
<td>10</td>
<td>0.4</td>
<td>0.8</td>
<td>0.6</td>
<td>1.4</td>
<td>0</td>
<td>0</td>
<td>3.8</td>
</tr>
<tr>
<td>right</td>
<td>0</td>
<td>5.4</td>
<td>0</td>
<td>0</td>
<td>4.2</td>
<td>77.2</td>
<td>0.2</td>
<td>0.6</td>
<td>1.6</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>up</td>
<td>2.6</td>
<td>0.8</td>
<td>0.4</td>
<td>0</td>
<td>0.6</td>
<td>0.6</td>
<td>43</td>
<td>19.2</td>
<td>6</td>
<td>7.4</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>down</td>
<td>7.6</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>3.6</td>
<td>3.8</td>
<td>38.8</td>
<td>71.4</td>
<td>27.4</td>
<td>17.4</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
</tr>
<tr>
<td>front</td>
<td>0.4</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td>0.4</td>
<td>1.8</td>
<td>0.4</td>
<td>1.4</td>
<td>56.6</td>
<td>2.8</td>
<td>0</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>back</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
<td>5.6</td>
<td>2.4</td>
<td>3.2</td>
<td>59</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>swim</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td>0.4</td>
<td>2</td>
<td>0.8</td>
<td>1.2</td>
<td>1</td>
<td>0.6</td>
<td>99</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>fly</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>clap</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.4</td>
<td>0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.4</td>
<td>1</td>
<td>0</td>
<td>93.8</td>
</tr>
<tr>
<td>pt left</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
<td>0.4</td>
<td>0</td>
<td>0.8</td>
<td>1.2</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>71.8</td>
</tr>
<tr>
<td>pt front</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.4</td>
<td>0</td>
<td>0.4</td>
<td>1</td>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20.2</td>
</tr>
<tr>
<td>pt right</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>1.2</td>
<td>0.6</td>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Mean classification rate = 74%, IOHMM with 5 states