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

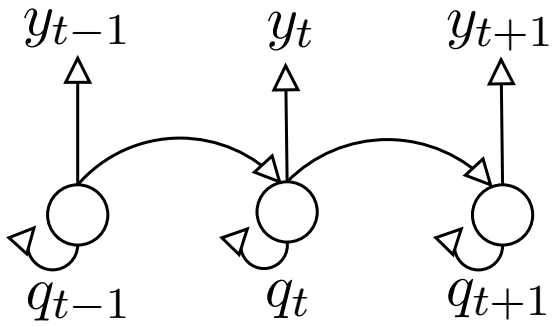
Outline

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

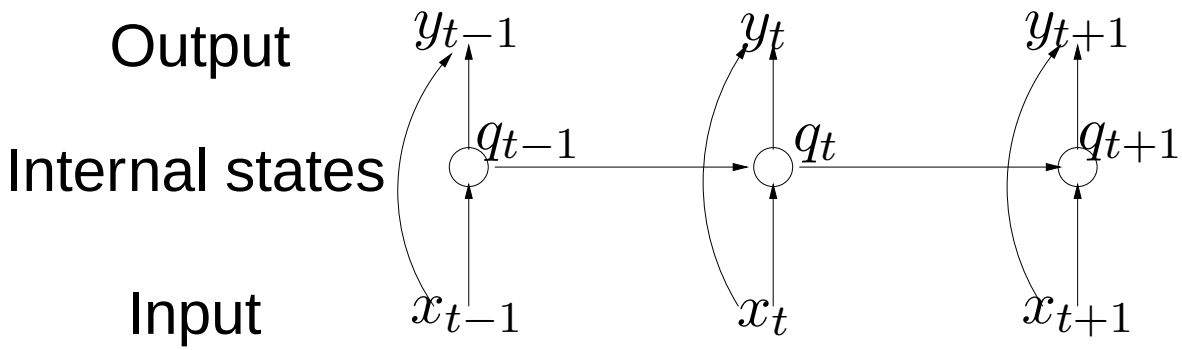
- Related work
  - Applications
    - Recognition of Sign Languages
    - Human-Computer Interaction
  - Limitations
    - Various techniques applied to many applications
    - No real evaluation of the algorithms
- Motivation:
  - Comparison of ML algorithm on the same database
  - Dynamic gestures $\Longrightarrow$ sequential ML algorithms (HMM, IOHMM)
  - Segmented gestures $\Longrightarrow$ 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

IOHMMs 1/2



- 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

IOHMMs 2/2

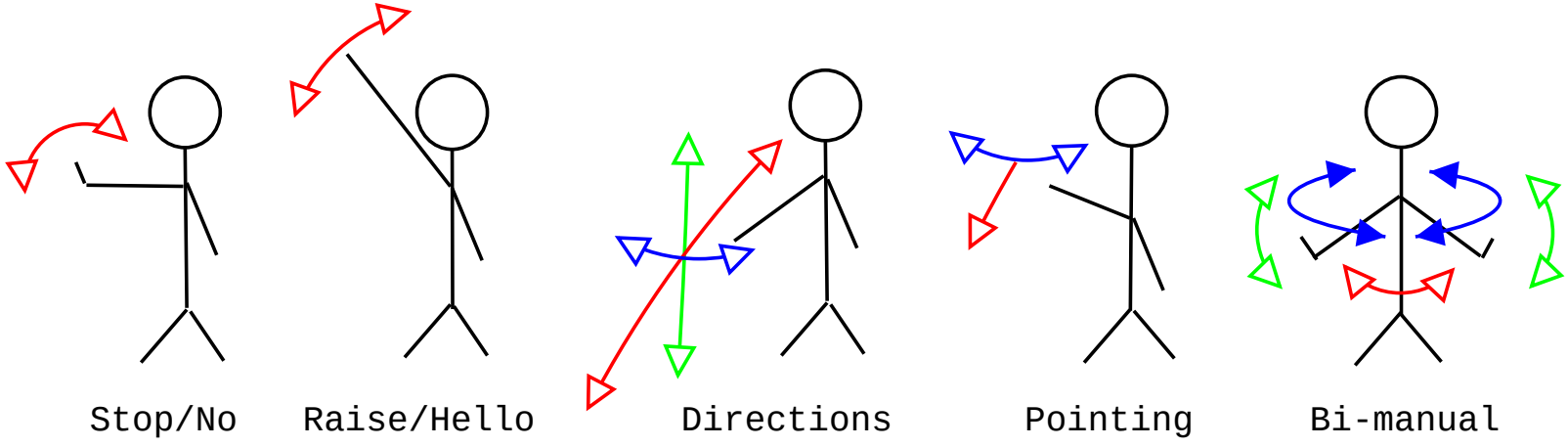
- Continuous IOHMMs
  - Modeling of conditional distributions using Neural Networks
  - Difficult to train (time, convergence)
- Discrete IOHMMs
  - Modeling of conditional distributions using look-up tables
  - Quantization of the data sequences required
  - Easy to train, fast
- Theoretically, better 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

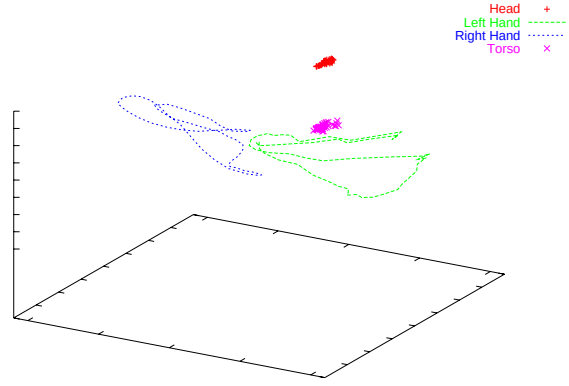
The Database 1/2

- 16 gestures: mono- and bi-manual gestures
- Examples:



The Database 2/2

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



[a] 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 Δ for each coordinate and each hand $\Longrightarrow$ 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

	Training set	Evaluation set	Test set
minimum number of frames	12	6	10
average number of frames	25	24	28
maximum number of frames	64	71	89
♯ sequences	4000	4000	8000

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: $argmax_c P(y_1^T = c|x_1^T)$

Results

- Classification rate on the test set:

	stop	no	raise	hello	direction	bi-manual gestures	pointing
IOHMM	59.8%	91.2%	54.8%	82.2%	65.8%	97.3%	68.9%
HMM	54.8%	81%	26.4%	87.2%	60.4%	98.3%	67.4%

- Comparison between HMMs and IOHMMs:

			error rate	♯parameters
HMM	10 N	20 G	31%	80600
HMM	15 N	20 G	30%	123000
IOHMM	3 N	1200 K	31%	68400
IOHMM	5 N	1200 K	26%	126000

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

	stop	no	raise	hello	left	right	up	down	front	back	*swim*	*fly*	*clap*	pt left	pt front	pt right
stop	**59.8**	0.8	1.6	0	0	0	6.4	0.2	0	4	0	0	0	0	0	0
no	20.6	**91.2**	2.8	8.8	0.6	3.4	2.4	0	1	6	0	0	0	2	0	0.2
raise	3.8	0	**54.8**	8.8	0	0	0.2	0	0	0	0	0	0	0.4	3.6	5.2
hello	2.2	1.4	39.6	**82.2**	0	0	0	0	0	0	0	0	0	0.2	1.6	4
left	0	0.2	0	0	**87.8**	10	0.4	0.8	0.6	1.4	0	0	0	3.8	0.4	0
right	0	5.4	0	0	4.2	**77.2**	0.2	0.6	1.6	0.2	0	0	0	0	0	5.6
up	2.6	0.8	0.4	0	0.6	0.6	**43**	19.2	6	7.4	0	0	0	0	1.4	2
down	7.6	0.2	0.2	0.2	3.6	3.8	38.8	**71.4**	27.4	17.4	0	0	0	0.6	3	9.6
front	0.4	0	0.2	0	0.4	1.8	0.4	1.4	**56.6**	2.8	0	0	0	0.4	7.6	9
back	3	0	0	0	0	0.6	5.6	2.4	3.2	**59**	0	0	0	0	0	0
swim	0	0	0.2	0	0.4	2	0.8	1.2	1	0.6	**99**	1	6.2	0.2	2	0.2
fly	0	0	0	0	0	0	0	0	0	0	0	**99**	0	0	0	0
clap	0	0	0	0	1.4	0	0.2	0.4	0.6	0.4	1	0	**93.8**	0.4	0	0
pt left	0	0	0	0	0.6	0.4	0	0.8	1.2	0.8	0	0	0	**71.8**	3	0.4
pt front	0	0	0	0	0.4	0	0.4	1	0.4	0	0	0	0	20.2	**73.8**	2.6
pt right	0	0	0.2	0	0	0.2	1.2	0.6	0.4	0	0	0	0	0	3.6	**61.2**

mean classification rate = 74%, IOHMM with 5 states