Fast Gesture Recognition with Multiple Stream Discrete HMMs on 3D Skeletons

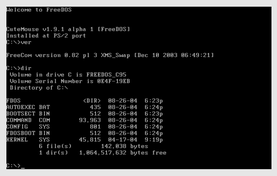
Guido Borghi, Roberto Vezzani, Rita Cucchiara
{name.surname}@unimore.it
University of Modena and Reggio Emilia, Italy

Introduction

Human Computer Interaction (HCI) is the discipline that studies models and techniques for the interaction between people and computers.

Historical evolution:

- **Command Line Interface** (CLI, '70s)

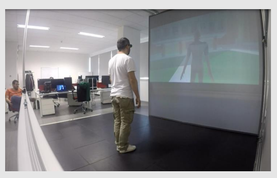


  - Quick
  - Mnemonic

- **Graphical User Interface** (GUI, '80s)

  - User friendly
  - New devices
  - New metaphors

- **Natural User Interface** (NUI, today)



  - Intuitive
  - Invisible
  - New low cost technologies

NUIs have recently got prestige thanks to **new low cost technologies**.
In NUI, systems must be able to **automatically segment and classify actions** in continuos action/gesture streams.

Dynamic body gestures for explicit Human Computer Interaction

1. **Dynamic**:
   The target posture requires a **movement**; thus, we neglect static potures (*e.g.* sitting, reading a book…);
2. **Body**:
   The target posture is potentially performed using the **whole body**; thus, we discard too local gestures such as finger movements or facial expressions;
3. **Gestures**:
   A gesture is a **well-defined** and **time-limited** body movement; continuous actions such as running, walking are not considered;
4. **Explicit HCI**:
   We focus on gestures provided by a user wich has spontaneously **decided to interact** with the system; thus, the gesture recognition subsumes a corresponding reaction or feedback at the end of each gesture.

Objectives

1. Overtake the **static sliding window** approach
2. No **prior strong hypothesis** (action duration, typology…)
3. Automatically detect the **beginning** and the **end** of actions
4. **Reliability**: avoid unwanted interactions

Setup

1. Depth sensor (Kinect, Kinect One, Intel R200…)

2. Hidden Markov Model

The classification of an observation sequence O is carried out selecting the model λ^* whose *likelihood* is highest. If the classes are a-priori equally likely, this solution is optimal also in a **Bayesian** sense:

$$\lambda^* = arg \min_{1 \le c \le C} [P(O|\lambda^c)]$$

Proposed Method

1. Multiple Stream Discrete HMM

Instead of a Gaussian Mixture Model, we adopted a **set of weighted distributions** to model **discrete observations** for each HMM's hidden state; so, we can write:

Streams — Weighting term

$$b_j(o_t) = \prod_{d=1}^{D} (h_j^d(o_t^d))^{\alpha_d}$$

Histogram — Discrete emission probability

2. Skeleton Features

Nine features are extracted for each selected body joint K_i. Given the sequence of 3D position of the joint $K_i^t = (x_i^t, y_i^t, z_i^t)$, we define $o_t(i) = \{o_t^1(i), \dots, o_t^9(i)\}$ as:

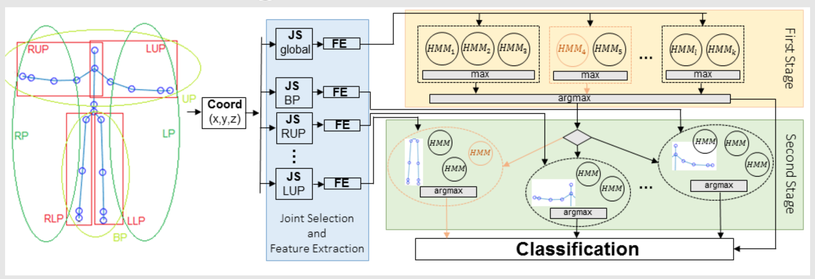
$$o_t^1(i) = x_i^t - x_k^t, \quad o_t^2(i) = y_i^t - y_k^t, \quad o_t^3(i) = z_i^t - z_k^t$$
$$o_t^4(i) = x_i^t - x_i^{(t-1)}, o_t^5(i) = y_i^t - y_i^{(t-1)}, o_t^6(i) = z_i^t - z_i^{(t-1)}$$
$$o_t^7(i) = x_i^t - 2x_i^{(t-1)} + x_i^{(t-2)}, o_t^8(i) = y_i^t - 2y_i^{(t-1)} + y_i^{(t-2)}$$
$$o_t^9(i) = z_i^t - 2z_i^{(t-1)} + z_i^{(t-2)}$$

Where:

- $\{o^1, \dots, o^3\}$: **offset** with respect to a reference joint
- $\{o^4, \dots, o^6\}$: **velocity** component
- $\{o^7, \dots, o^9\}$: **acceleration** component

Limited dependencies: **fast** and **parallel** computation is allowed.

3. Double-Stage Classification



- **First-stage**: *left-right* discrete HMMs, detect which body part is involved in performing gesture; they activate a specific set of second-stage HMMs.
- **Second-stage**: *ergodic* discrete HMMs; specialized for a particular body part, they provide the final classification.

For each classification, the *forward* algorithm for HMMs with the three well known *initialization*, *induction* and *termination* equations can be applied:

$$\alpha_1(j) = \pi_i b_i(o_1), 1 \le i \le N$$
$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1})$$
$$P(O|\lambda) = \sum_{j=1}^{N} \alpha_T(j)$$

4. Online Temporal Segmentation

4a. Gesture Beginning Detection

The beginning of a gesture is detected by analizing the first hidden state of each HMM; a **voting mechanism** is exploited, each HMM votes in this way:

$$\Phi(HMM_k) = \begin{cases} 0, & \alpha_t(1) = \left[\sum_{i=1}^{N} \alpha_t(i)\alpha_{i1}\right] b_1(o_t) \ge th \\ 1, & \alpha_t(1) = \left[\sum_{i=1}^{N} \alpha_t(i)\alpha_{i1}\right] b_1(o_t) < th \end{cases}$$

4b. Gesture End Detection

If a gesture is currently performed, a **probability distribution analysis** of the **last** state detects the gesture end:

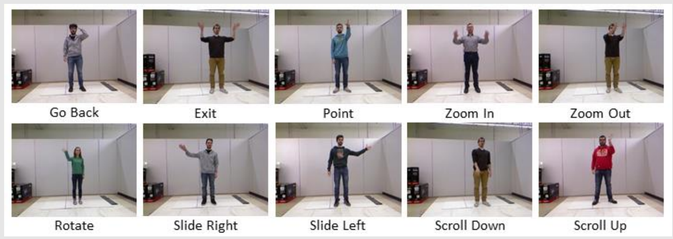
$$\alpha_t(N) = \left[\sum_{i=1}^{N} \alpha_t(i)\alpha_{iN}\right] b_j(o_t) \ge th$$

4c. Reliability Check: to filter out **false candidates**

$$\overline{S} \subseteq S, \overline{S} = \{s_j | \vee t \in [t_s, t_e], \alpha_t(j) \ge th\}, S_{N-1} \in \overline{S}_{\wedge} \#\overline{S} \ge {}^2/_3 N$$

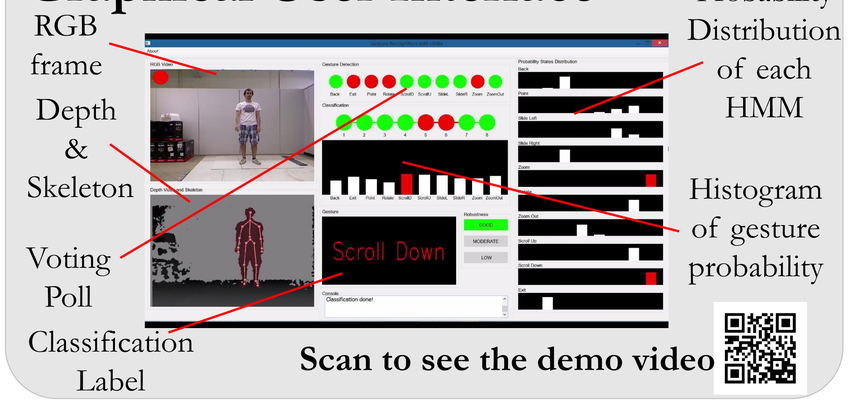
Datasets

1. MSRAction3D
2. UTKinect-Action
3. MSRDailyActivity3D
4. **Kinteract Dataset:** *our* dataset HCI oriented



Classes from **Kinteract** Dataset

Graphical User Interface



Scan to see the demo video

Results

1. Classification

Methods	Accuracy		
	1/3	2/3	cross
HOJ3D	0.962	0.971	0.789
HMM + DBM	-	-	0.820
EigenJoints	0.958	0.977	0.823
HMM + GMM (our implementation)	0.861	0.929	0.825
Actionlet Ensemble (skeleton data)	-	-	0.882
Skeletal Quads	-	-	0.898
LDS	-	-	0.900
Cov3DJ	-	-	0.905
➡ Our	0.943	0.983	0.905
FSFJ3D	-	-	0.909
KPLS	-	-	0.923

MSRAction3D Dataset

Methods	Accuracy
DSTIP+DCSF	0.858
FSFJ3D (skeleton data)	0.879
SNV	0.889
➡ Our	0.895
HOJ3D	0.909

UTKinect-Action Dataset

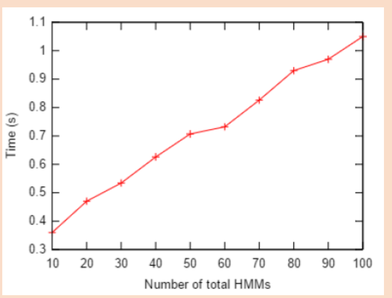
Methods	Accuracy
DTW	0.540
Moving Pose	0.738
HON4D	0.800
➡ Our	0.833
Actionlet Ensemble	0.845

MSRDailyAction Dataset

2. Online Temporal Segmentation

Dataset	Detection Rate	Recognition Rate
MSRAction3D	0.782	0.871
Kinteract	0.892	0.943

2. Speed Performance



Framework runs at about **80 fps**. The total number of HMM of each stage corresponds to the potential number of recognizable gesture classes.
Total time for feature extraction, stream weight evaluation and classification: $4.4x10^{-2}s$ (single action)

References and on-line material

- Rabiner, Lawrence, and B. Juang. "An introduction to hidden Markov models." IEEE assp magazine 3.1 (1986)
- Shotton, Jamie, et al. "Real-time human pose recognition in parts from single depth images." Communications of the ACM 56.1 (2013)

Project page **GitHub page**