HIERARCHICAL ARM-HAND GESTURES MODELLING AND RECOGNITION

Gori[1] I., Fanello[1] S. R., Demiris[2] Y.

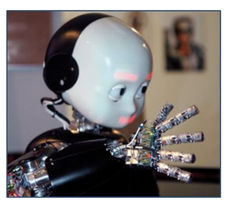
1. Italian Institute of Technology - Department of Robotics, Brain and Cognitive Sciences
2. Imperial College London - Department of Electrical and Electronic Engineering

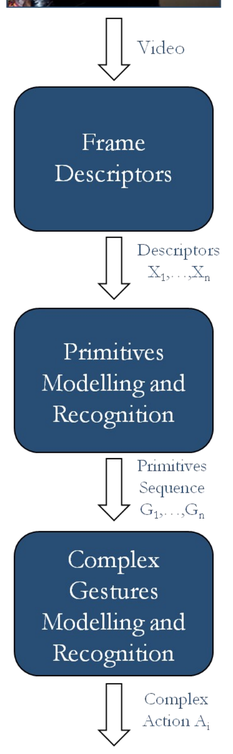
Abstract

We present an original, reliable and real-time gesture recognition system. Our system exploits motion information without prior knowledge of the presence of humans in the scene. We adopted a hierarchical approach; in particular we define a set of ***action primitives*** *that can be combined in order to obtain* ***complex gestures****. Accordingly, our system consists of three levels: the first one is based on features extraction, the second one regards the modelling and the recognition of simple actions, and the third one has been conceived for the modelling and the recognition of complex gestures.*

Introduction

We face the problem of modelling arm-hand behaviours from a robot perspective; the framework we will show has been designed to improve the iCub [1] robot's perception capabilities. The entire system has been built with the purpose of recognizing arm -hand gestures and it is based on the so called *Visual hypothesis* [2], which asserts that the understanding of an action derives from the analysis of primitive elements composing the action. Hence, we define a set of **action primitives**, and we describe every complex gesture as a combination of simple gestures. The proposed approach shows improved performance with respect to classical statistical approaches for structured temporal patterns like Hidden Markov Models (HMM); indeed we model separately the observations (i.e. action primitives) and the structure of the whole action. Furthermore, we employ our gesture recognition system in order to make the robot coarsely imitate the observed action; indeed with our system the iCub robot can imitate a complex action as a sequence of embedded action primitives.





Frame Descriptors

- Detection of the subject of interest (i.e. demonstrator) via disparity map: we employ Hirschmuller's algorithm [3]. In order to update the cameras relative position even when the robot moves its eyes, we exploit both information embedded in the kinematic model of the robot, and the Horn's relative orientation algorithm [4], relying on visual cues i.e. SIFT, SURF.

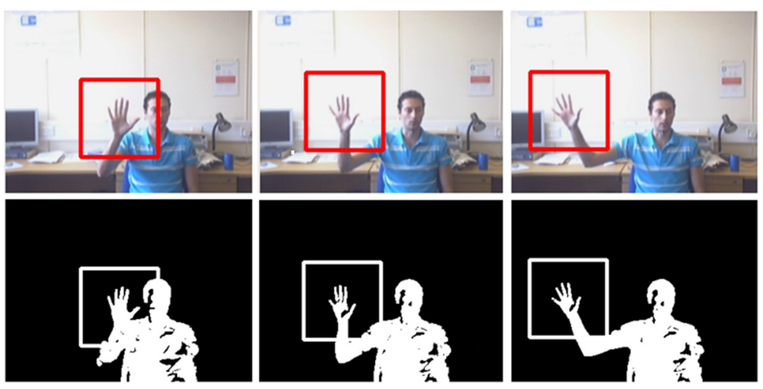
- Let [U(x,y,t),V(x,y,t)] the optical flow vector for each pixel (x,y) at frame t in the ROI, frame descriptors based on *2k* principal directions are defined as follows:

$$dir(x,y,t) = \frac{\pi}{2k}\left(\left\lfloor \frac{k}{\pi} arctan\left(\frac{V(x,y,t)}{U(x,y,t)}\right)\right\rfloor + \left\lfloor \frac{k}{\pi} arctan\left(\frac{V(x,y,t)}{U(x,y,t)}\right)\right\rfloor\right)$$

Then, we build a histogram using the *2k* principal directions of the plane as bins, so that we obtain a vector $X(t) \in R^{2k}$.

Primitives Modelling

Given a set of descriptors $X_1,\ldots,X_n$ each action primitive has been modelled as a Mixture of Gaussians. In order to estimate the number of clusters, the means, the covariance matrix and the mixing coefficients we employ self-tuning Spectral Clustering algorithm [5].



Primitives Recognition

Given a set of n observations the classification of an action primitive A arises naturally:

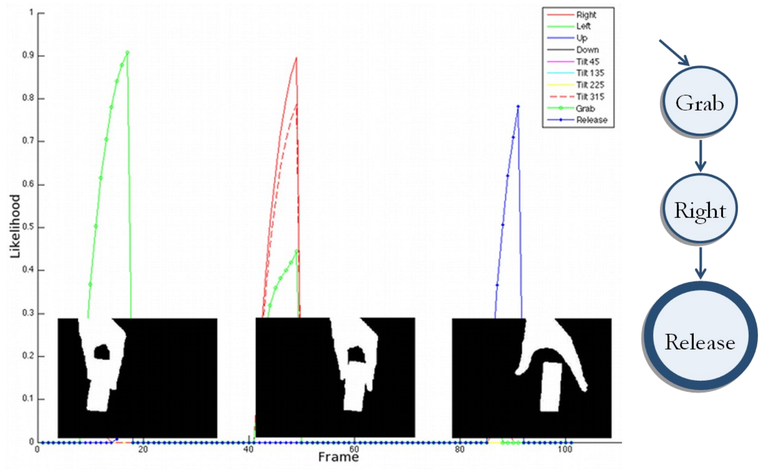
$$A = arg\max_j \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_{kj} N(x_i|\mu_{kj}, \Sigma_{kj})$$

In order to recognize a sequence of action primitives we developed an online recognition algorithm based on the analysis of the first derivative of the likelihoods:

```
max = 0
for i = 0 to numActions do
    compute likelihood(t)(i)
    if likelihood(t)(i) > max then
        max = likelihood(t)
        index = i
    end if
end for
if likelihood(t)(oldIndex) < likelihood(t − 1)(oldIndex) then
    if likelihood(t − 1)(oldIndex) > threshold & startRec == 1 then
        recognize gesture primitive Action(oldIndex)
        startRec = 0
    end if
end if
if likelihood(t)(index) < threshold then
    startRec = 1
end if
if index ≠ oldIndex & likelihood(t)(index) > likelihood(t − 1)(index) then
    startRec = 1
end if
```

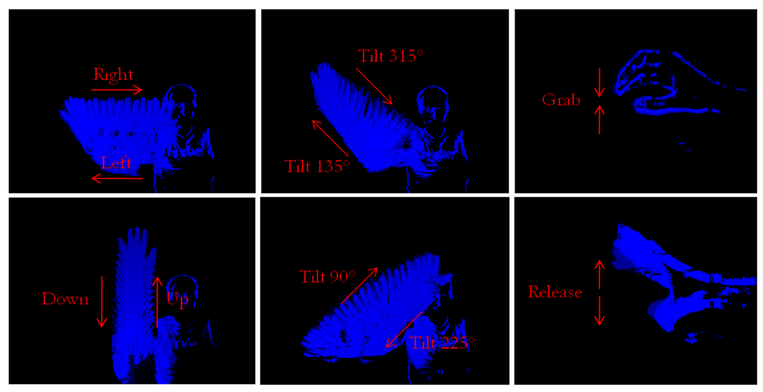
Complex Gestures

Each action primitive is conceived as a symbol of an alphabet, and each complex action, under opportune conditions, as a regular language. Since one complex gesture can be described by many different sequences, we employ Non-Deterministic Finite State Automata (NDFSA), which can manage this variability and they are built via induction. The learning phase is based on demonstration.



Experiments & Results

The employed data-set of action primitives is composed of 10 simple gestures, which are based on arm-hand movements and they are performed 10 times by 5 different actors. The confusion matrix showed below is about the performances



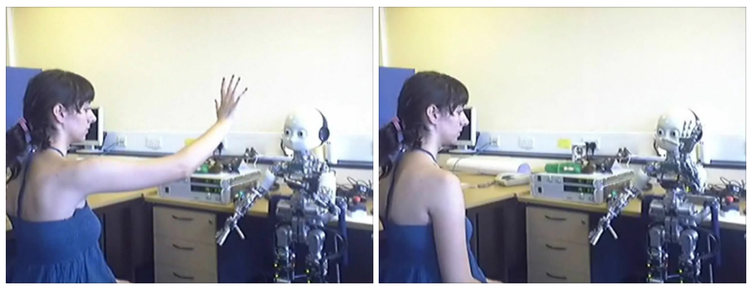
analysis on action primitives. Complex actions are modeled by fixed NDFAs, hence the recognition phase is just a language acceptance problem, without any kind of ambiguity. If the action is not present in the data-set, the complex gesture recognition classifies the action as unknown. The accuracy, which has been obtained via k-fold Cross-Validation procedure over 500 samples, is around 98%.

	Right	Left	Up	Down	T45°	T135°	T225°	T315°	Grab	Rel
Right	0.98							0.02		
Left		0.97				0.03				
Up			1.0							
Down				1.0						
T45°	0.02		0.03		0.95					
T135°		0.03				0.97				
T225°		0.03					0.97			
T315°				0.02				0.98		
Grab									1.0	
Rel										1.0

Conclusion

The presented work can be considered as an original contribute in the field of Gestures Recognition.

- The employment of the disparity map allows a reliable segmentation that can be used even when the robot moves.
- The optical flow generalizes across colour and pose.
- The features we chose are invariant to distance and position.
- Finite State Machines are real-time, fast and they allow the recognition of actions that do not belong to the dataset.



References

[1] G. Metta, G. Sandini, D. Vernon, L. Natale, F. Nori, *The iCub humanoid robot: an open platform for research in embodied cognition*. In Proc. of PerMIS, 2008.

[2] G. Rizzolatti, V. Gallese, *Neurophysiological mechanisms underlying the understanding and imitation of action*. Nature Reviews, 2001.

[3] H. Hirschmuller, *Stereo vision in structured environments by consistent semi-global matching* , CVPR, 2006.

[4] B.K.P. Horn, *Relative Orientations* , IJCV 1990.

[5] L. Zelnik-Manor, P. Perona, M. Fernandes, *Self-tuning spectral clustering* , In Advances in Neural Information Processing Systems, 2004.

This work has been partially supported by the EU project EFAA (FP7-270490)