Learning dictionaries of kinematic primitives for action classification

Alessia Vignolo[1], Nicoletta Noceti[2], Alessandra Sciutti[1], Francesca Odone[2], Giulio Sandini[3]

[1]CONTACT Unit - Isitituo Italiano di Tecnologia, Genova

[2]MaLGa – Machine Learning Genoa center, DIBRIS, Università di Genova

[3]RBCS Unit - Isitituo Italiano di Tecnologia, Genova

UniGe | MaLGa

Motivations, context and goals

Decomposing a movement

- Among the earliest processing stages of human development is the ability to precisely localize in space and time an action and its sub-parts

- Our work focuses on this ability and explores the concept of visual motion primitives, a limited number of actions sub-components that allow to describe and reconstruct a wide range of actions

Goals

- The aim of this research is to assess the use of a simple representation based on kinematics motion primitives as a backbone of a general approach to action understanding

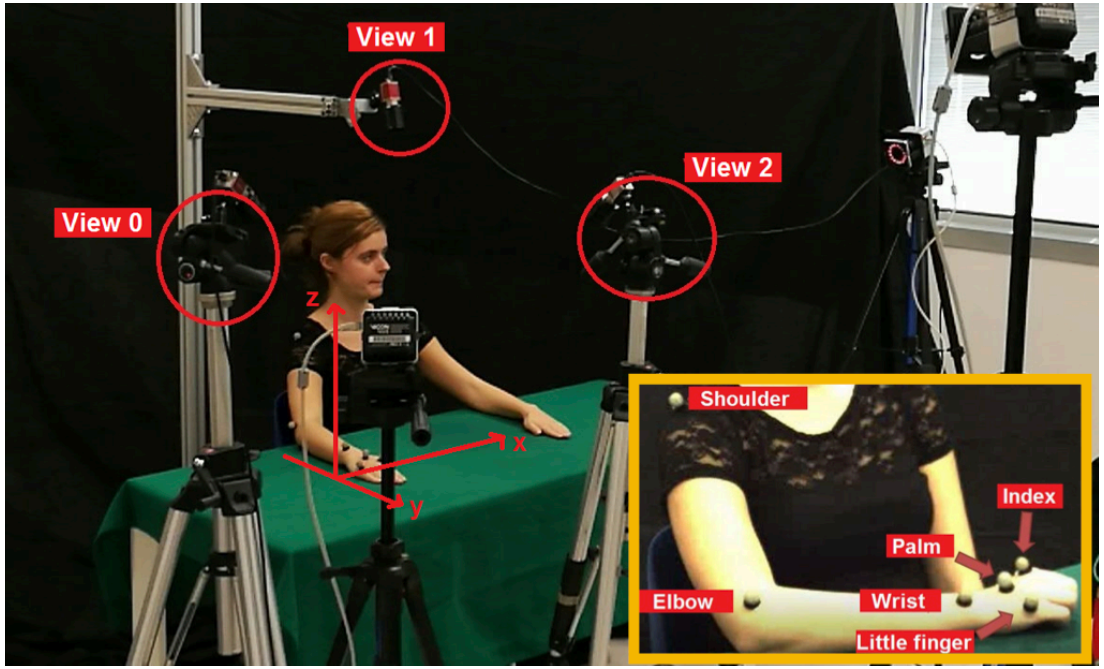
- To the purpose we combine a simple motion segmentation with dictionary learning and sparse coding to derive a motion representation to be used in classification scenarios

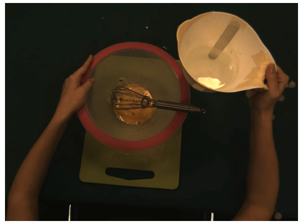
UniGe | MaLGa

Our approach

The MoCA dataset

A bi-modal (videos and motion capture) dataset



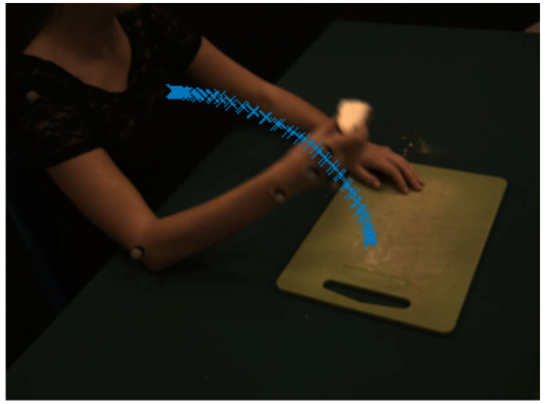


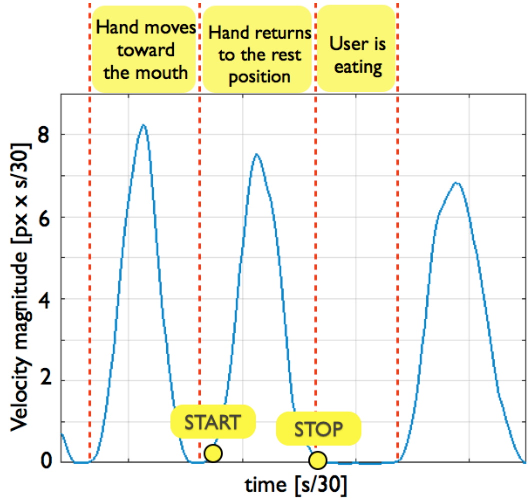
1-*Grating the carrot*, 2-*Cutting the bread*, 3-*Cleaning a dish*, 4-*Eating*, 5-*Beating eggs*, 6-*Squeezing the lemon*, 7-*Cutting with a mezzaluna*, 8-*Mixing*, 9-*Open the bottle*, 10-*Turning the omelette*, 11-*Pestling*, 12-*Pouring water*, 13-*Reaching an object*, 14-*Rolling the dough*, 15-*Washing the salad*, 16-*Salting*, 17-*Spreading cheese on a bread*, 18- *Cleaning the table*, 19-*Transporting an object*

E. Nicora, G. Goyal, N. Noceti, A. Vignolo, A. Sciutti, F. Odone, "The MoCA dataset, kinematic and multi-view visual streams of fine-grained cooking actions. Scientific Data, to appear

Motion segmentation

We represent a video as a sequence of velocities derived from **optical flow magnitudes** (Vignolo et al., 2017) collapsed in a single point and we segment the sequence detecting **dynamic instants** (Rea et al., 2019)



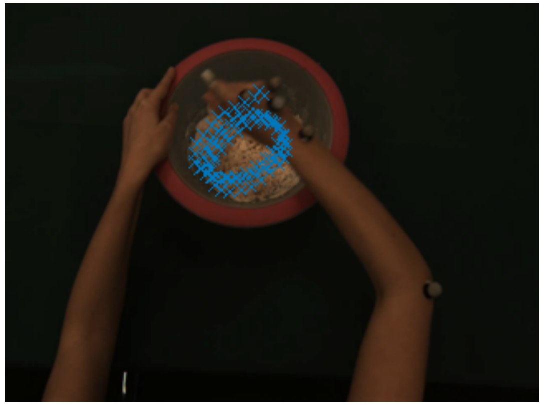


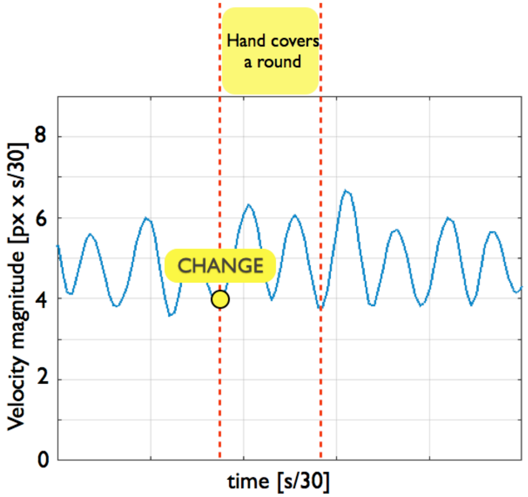
A. Vignolo, N. Noceti, F. Rea, A. Sciutti, F. Odone, and G. Sandini. Detecting biological motion for human-robot interaction: A link between perception and action. Frontiers in Robotics and AI, 2017

F.Rea,A.Vignolo,A.Sciutti,andN.Noceti.Humanmotionunderstand- ing for selecting action timing in collaborative human-robot interaction. Frontiers in Robotics and AI, 2019

Motion segmentation

We represent a video as a sequence of velocities derived from **optical flow magnitudes** (Vignolo et al., 2017) collapsed in a single point and we segment the sequence detecting **dynamic instants** (Rea et al., 2019)

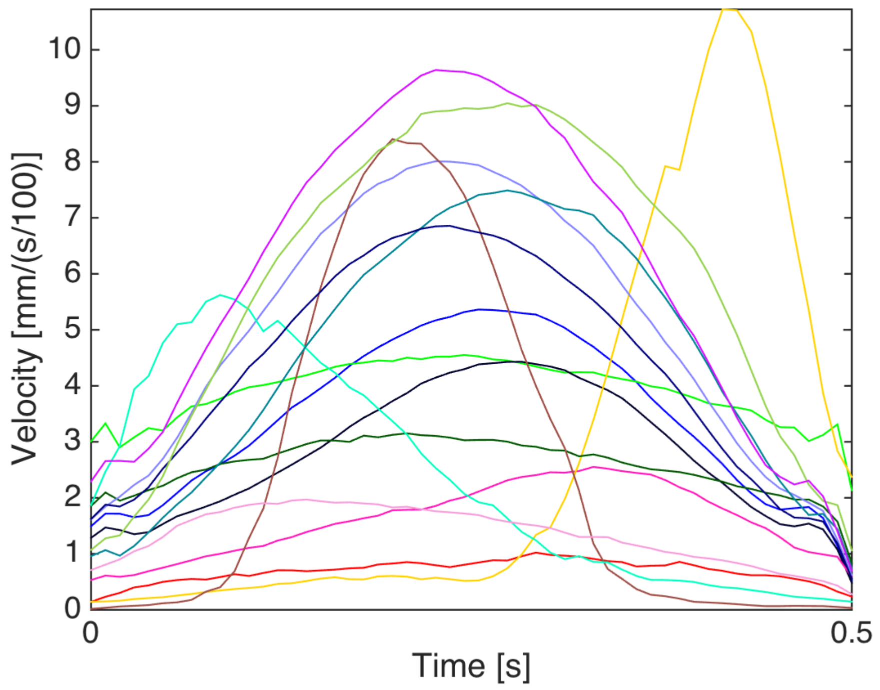




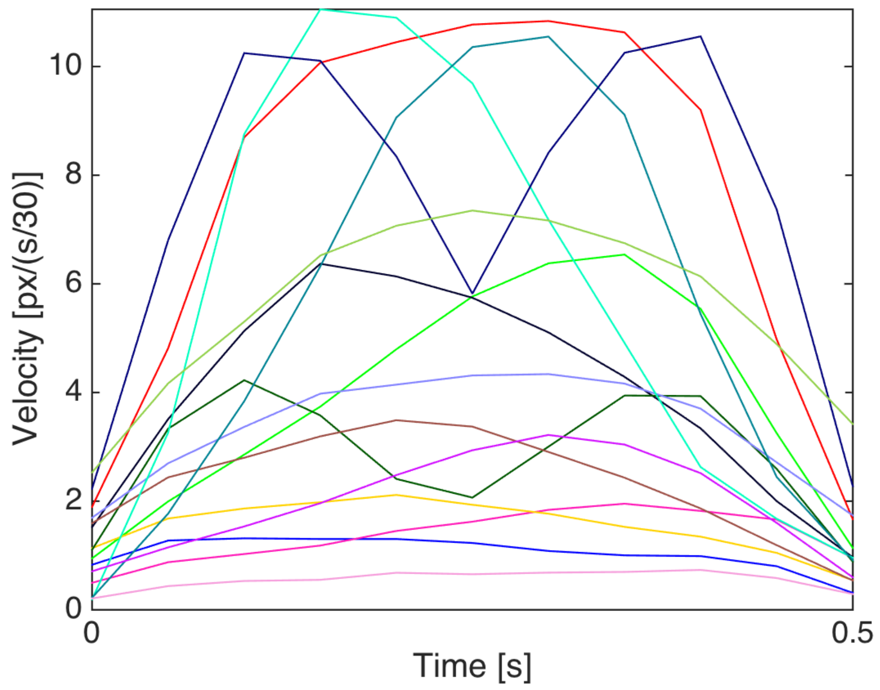
A. Vignolo, N. Noceti, F. Rea, A. Sciutti, F. Odone, and G. Sandini. Detecting biological motion for human-robot interaction: A link between perception and action. Frontiers in Robotics and AI, 2017

F.Rea,A.Vignolo,A.Sciutti,andN.Noceti.Humanmotionunderstand- ing for selecting action timing in collaborative human-robot interaction. Frontiers in Robotics and AI, 2019

The sub-movements



Motion capture



Videos (frontal view)

Representation and classification

A dictionary of visual motion primitives is derived from the segmented sub-movements. They are represented using sparse coding, and the obtained codes are classified using a simple *Regularized Least Squares*, to focus on the descriptive power of the representations

- Dictionary of visual motion primitives learnt with K-means

$$\min_{\mathbf{D},\mathbf{U}} \|\mathbf{X} - \mathbf{D}\mathbf{U}\|_F^2 \ . \ \text{s.t.} \quad \text{Card}(\mathbf{u}_i) = 1, |\mathbf{u}_i| = 1, \\ \mathbf{u}_i \geq 0, \forall i = 1, \ldots, T$$

- Representations derived as sparse codes

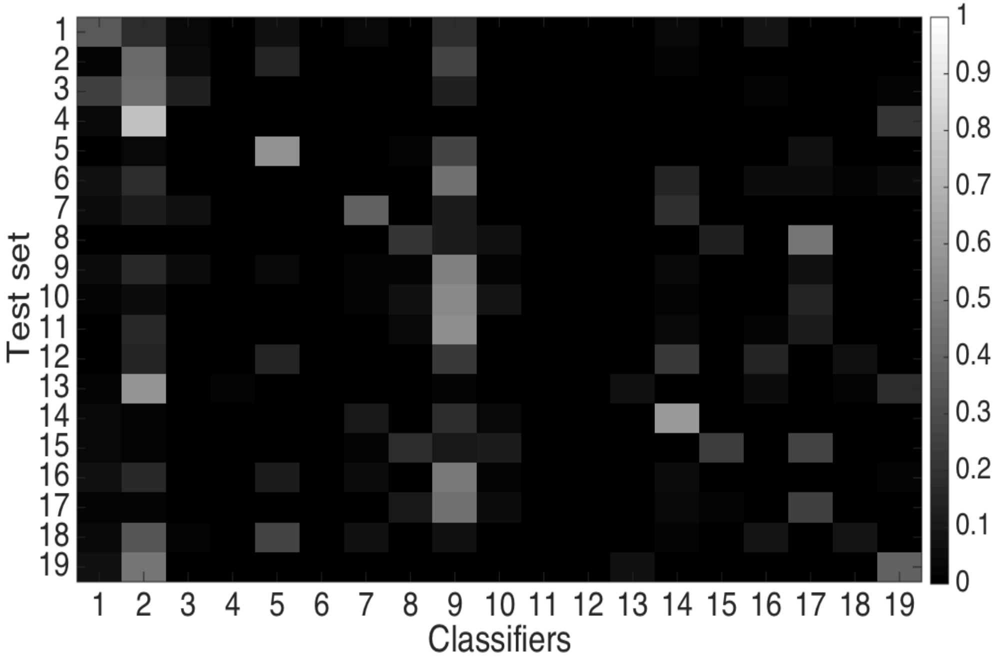
$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \|\mathbf{x} - \mathbf{D}\mathbf{u}\|^2 + \lambda\|\mathbf{u}\|_1$$

UniGe | MaLGa

Experiments

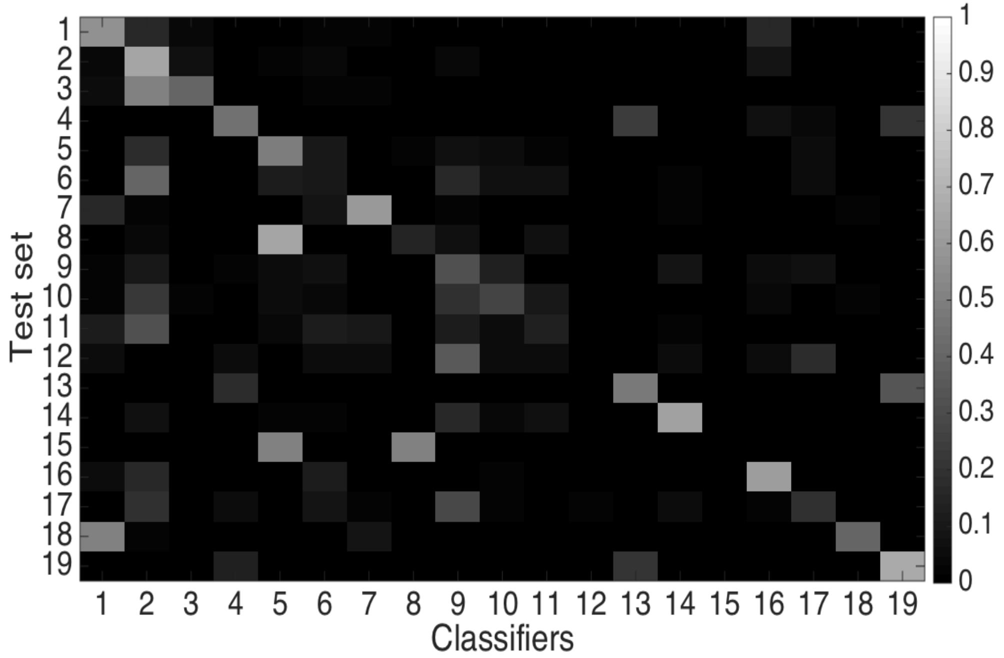
Classification of one sub-movement

Avg. acc.: 0.24; Over. Acc.: 0.29



Videos (frontal view)

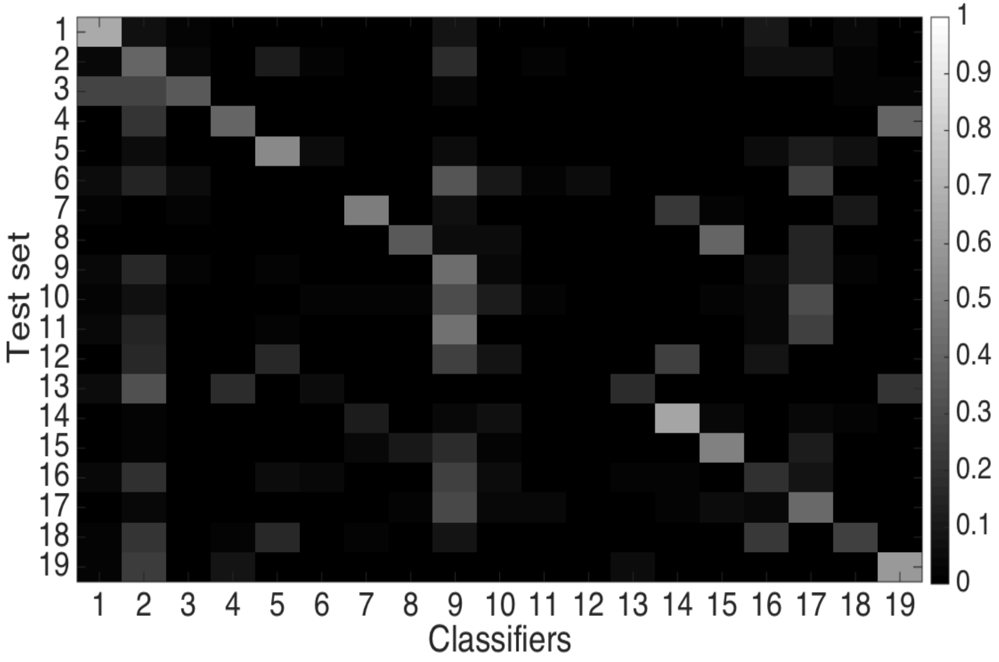
Avg. acc.: 0.37; Over. Acc.: 0.48



Motion capture

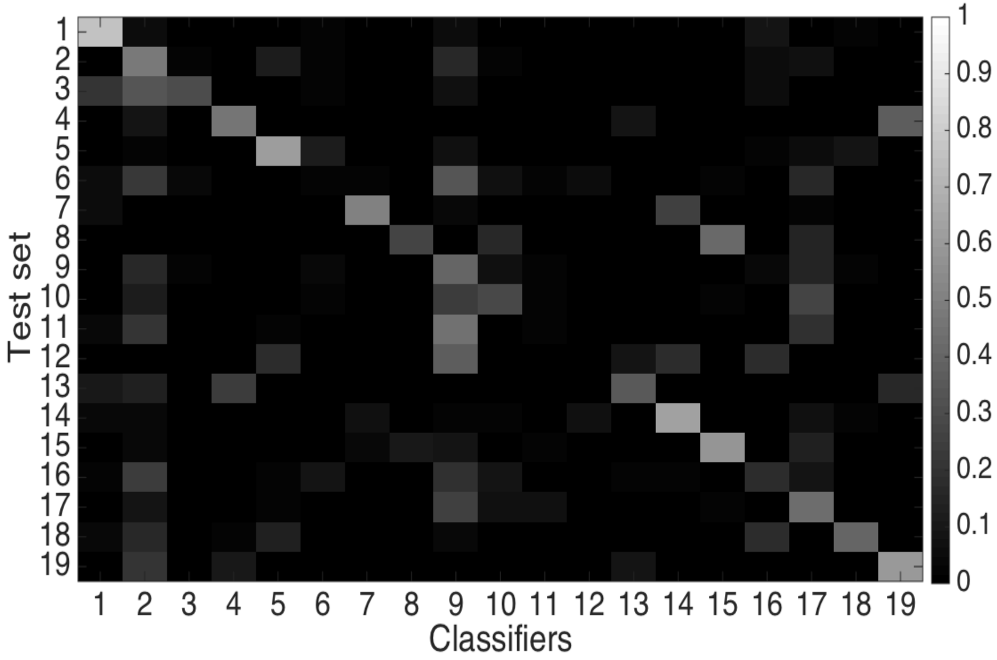
Classification of more sub-movements

Avg. acc.: 0.34; Over. Acc.: 0.37



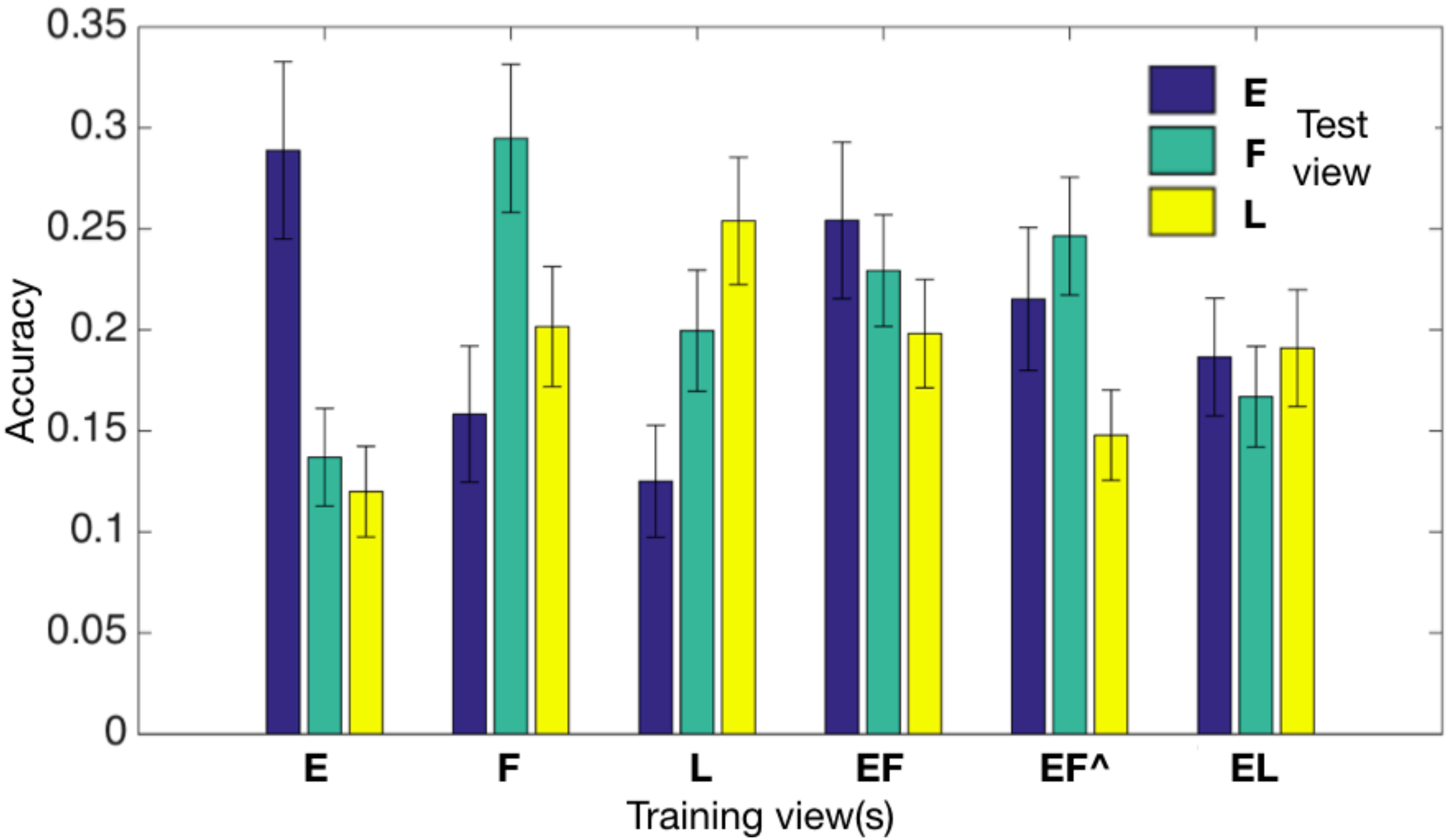
2 sub-movements

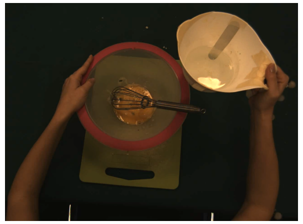
Avg. acc.: 0.38; Over. Acc.: 0.41



3 sub-movements

Classification across views





UniGe

MaLGa