Developmental Learning for Object Perception

Natalia Lyubova, David Filliat

Developmental Learning for Object Perception

Natalia Lyubova, David Filliat

Flowers team, ENSTA ParisTech, INRIA Bordeaux Sud-Ouest, France
natalia.lyubova@ensta.fr, david.filliat@ensta.fr

The goal of this work is to design a visual system for a humanoid robot. Taking inspiration from child's perception and following the principles of developmental robotics [1], the robot should detect and learn objects from interactions with people and own experiments.

The computer vision community provides a large amount of object recognition approaches which are mostly based on prior knowledge, either in the form of algorithm choices or image databases. In contrast, our algorithm is designed by analogy with the human perceptual learning in terms of input data and learning organisation. It does not require image databases nor face/skin detectors, and allows to be robust to changes in the environment, background, lighting conditions and camera motion. We acquire all knowledge iteratively from low-level features. Our approach differs from Deep Belief Networks and Deep Energy Models [2] in that it does not use training and testing sets, but characterizes each object by a hierarchy of shared and complementary features (see Fig. 1) built up incrementally, during interactions with objects.

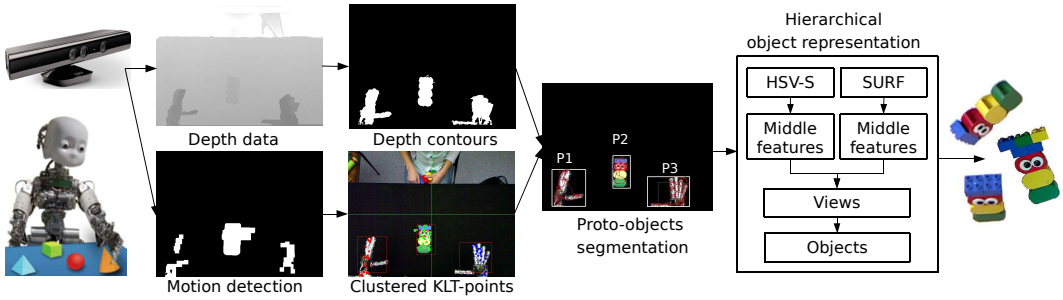

Fig. 1: Main processing steps of the object learning.

Our object detection algorithm is based on visual attention. In our scenario, people in front of the robot interact with objects, encouraging the robot to focus on them. We assume that the robot is attracted by motion; therefore we compute a saliency map based on the optical flow. The depth data obtained from a Kinect camera by the RGB Demo software[1], is used to filter the visual input, considering the constraints of the robot's working area. Then we segment the visual space into units of attention, so-called proto-objects [3], by analyzing motion behavior of key-points. We perform KLT (Kanade-Lucas-Tomasi) tracking and cluster detected points according to their speed and distance within the visual field. The groups of coherent points give an idea about possible objects, and their boundaries are refined with the contours extracted from the depth information.

The robot should be able to deal with various kinds of objects. SURF descriptor is a good solution for areas with a large amount of details, but its isolated key-points can not describe well homogeneous areas. So, we developed an additional descriptor using the Superpixels algorithm [4] based on the watershed segmentation on LoG (Laplacian of Gaussian) with local extrema as seeds, and dividing images into regularly shaped regions that follow image edges and allow to characterize uniformly colored areas. Each

[1]Software Kinect RGB Demo v0.6.1 available at http://nicolas.burrus.name

superpixel is described by the HSV model (hue, saturation and value). Therefore, our combined descriptor increases the completeness of the encoded information, and it is robust to object texture level, illumination, object rotations and scale variations.

The appearance of objects is encoded in a hierarchical way; we start from simple local features, group them to more complex, describing larger regions, and use them to describe the appearence of views and finally group multiple views to represent an object. Extracted SURF and HSV descriptors are first quantized to visual words and stored in incremental vocabularies. To avoid rapid dictionary growth, we implemented a short- and long-term memory in order to keep only features observed over consecutive frames. The next middle layer groups the closest SURF points and superpixels into pairs, incorporating local object geometry as the relative position of colors and key-points. Mid-features encode object views through the Bag of Visual Words approach with incremental dictionaries [5].

Our view recognition algorithm is based on TF-IDF (Term-frequency - Inverse-document frequency) of mid-features. This statistical measure evaluates the importance of mid-features with respect to segmented views. A voting method is used to compute the likelihood of a current view being an already known view. These learned views describe objects from a single perspective. To construct a multi-view model, we track the object during manipulations and associate detected views with the same label.

Object manipulations introduce additional difficulties in image processing. When the human or the robot's hand holds an object, they compose a single moving blob, and multiple occlusions divide an object into parts. In order to overcome these problems, we use a two-stage recognition, that identifies the presence of multiple objects.

Our perception system is implemented on the iCub robot, which detects objects in the visual space and characterizes them using a hierarchy of complementary features, their relative position and their occurrence statistics. Ten objects were presented in the experiment; each of them was manipulated by a person for 1-2 minutes. Once the vocabulary reached a sufficient amount of knowledge, the robot was able to reliably recognize human hands and most of objects. In future work, this system will be used as the basis for object categorization through robot experiments and affordance learning.

References

[1] J. Weng, J. McClelland, A. Pentland, O. Sporns and E. Thelen. Autonomous mental development by robots and animals. In *Science*, 291(5504): 599–600, 2001.

[2] J. Ngiam, Z. Chen, P. Koh and A. Y. Ng. Learning deep energy models. In *Proc. International Conference on Machine Learning*, 2011.

[3] Z. W. Pylyshyn. Visual indexes, preconceptual objects, and situated vision. In *Cognition*, 80: 127–158, Elsevier, 2001.

[4] B. Micusik and J. Kosecka. Semantic segmentation of street scenes by superpixel co-occurrence and 3D geometry. In *Proc. IEEE Workshop on Video-Oriented Object and Event Classification*, 625–632, 2009.

[5] D. Filliat. A visual bag of words method for interactive qualitative localization and mapping. In *Proc. IEEE ICRA*, 3921–3926, 2007.