

# Object-based Model of the Visual Attention for Imitation

Francesco Orabona

*francesco.orabona@unige.it*  
*University of Genova, Italy*

To imitate movements based on visual observation we need to recognize objects, bodies, hands, and finally, actions. How to do this? How to do this *efficiently*? Consider for example a person: he does not have a holistic view of the world, his vision system is not able to see the whole field of view in which a task is performed but focuses only on a limited region of the visual scene. This leads us to use an active vision system that has only a limited field of view of the scene, but, driven by an attention system, it is able to select and fixate “salient” regions of space. How does visual attention work to perform efficient selectivity?

There are two traditional assumptions in the literature attempting to account for this. On the one hand the space-based attention theory holds that attention is allocated to a region of space, with processing of everything within this spatial window of attention like a spotlight, internal eye or zoom-lens. On the other hand, object-based attention theory argues that attention is actually directed to an object or a group of objects to process any properties of selected object(s) rather than regions of space.

Moreover attention can be directed in two ways. One approach uses bottom-up information including basic features such as color, orientation, motion, depth, conjunctions of features, etc. A feature or stimulus catches the attention of the system if it differs from its immediate surrounding in some dimensions and the surround is reasonably homogeneous in those dimensions. However the bottom up salience cannot always capture attention if it is focused or directed elsewhere in advance. For this reason it is necessary to recognize the importance of how attention is also controlled by top-down information relevant to a particular task/behavior.

In literature a number of attentional models that use the first hypothesis have been proposed; most of them are derived from Treisman’s Feature Integration Theory, which employs separate low-level feature maps that are combined together by a spatial attention window operating on a master map or saliency map.

We propose an object-based model of the visual attention system that integrates bottom-up and top-down cues; in particular top-down information works as a prime for certain regions in the visual search task. The visual scene is divided in blobs of different colors; objects are represented as collections of blobs and their relative position. In this way the task of the attention system is more strongly coupled with the object recognition system. The model of a target object is created statistically through different views of the object, which are collected during manipulation (thus linking together vision and action). A histogram of the number of times a particular blob is seen is used to estimate the probability that the blob belongs to the object.

A first use of the system is to create a visual model of the hand of the robot (a special object in the environment) that can be used to distinguish it from the rest of the environment. When an object is found after visual search, a possible figure-ground segmentation is attempted, using the information gathered on the object during the exploration phase.