

How intimately may Visual Attention and Object Recognition Share Resources?

Vidhya Navalpakkam
navalpak@usc.edu

Laurent Itti
itti@usc.edu

University of Southern California, Computer Science Department, Los Angeles, CA 90089 - USA

In most modelling efforts, guiding attention to the salient visual targets in a scene and recognizing those targets have typically been treated as two independent problems. Hence, most models use different sets of low level features or resources to solve the above problems. However, our brain imposes severe limitations on the available resources because we have only one visual cortex, and hence, have one low level visual processing system. Moreover, nature generally exhibits a policy of sharing resources. Hence, we are motivated to investigate whether bottom-up attention and object recognition may share resources. In particular, we ask if the specific low level cues computed for bottom-up attention may be reused for recognition. We also investigate whether the basic processes involved in bottom-up attention, object detection and recognition may be similar, and, whether a common representation may be used for object detection and recognition. Our investigations demonstrate that with minimal enhancement of the attentional system, we can achieve detection and recognition of a variety of simple objects in natural, complex scenes.

In our model, attention is deployed according to the bottom-up salience model [1]. During the initial learning stage of our enhanced model, we learn the representations of the target objects by first choosing a few locations around the most salient location in the target. Then, for each chosen location, we extract the normalized center-surround features at multiple scales called a “view”. Specifically, a 42-component feature vector represents a view (six center-surround scale pairs, for four orientation, two color opponent, and one intensity feature types).

One of the key abilities of our brain is the capacity to integrate or pool across different sources of information that leads to robustness in the presence of noise, occlusion and other sources of ambiguity. Inspired by this feature, we attempt to combine the different views contained within the object to form a more stable, general representation of an instance of the object. We repeat this process by combining the various instances to form a general representation of the object and so on to generate an object heirarchy [2].

When we want to detect a specific target object in any scene, we use the previously learned representation to bias the competition or combination of different feature maps to form the saliency map, so that the relevant features characteristic of the target may be promoted. A feature is considered to be relevant and reliable if its mean feature value is high, and its feature variance is low. In order to promote the target in all the feature channels, each channel promotes itself proportionally to the maximum feature weight of its subchannels. For instance, if the target has a high value of redness at some scale, then the weight of the red channel increases, and so does the weight of the color channel. Hence, those channels that are irrelevant for this target are weighted down or not considered while contributing to the salience (e.g., for detecting a red object, the orientation of edges is irrelevant; so the orientation channel’s weights are decreased so as to promote only color). The weighted feature maps are then combined to form conspicuity maps that are in turn combined in a weighted manner to form the salience map. In the salience map thus formed by biasing the competition of all feature maps, all scene locations whose local features are the same as the target’s relevant features become more salient.

To recognize the objects in any scene, we use our attention model to fixate on any one location in the object and extract the normalized center-surround feature vector from that location or fixation. We attempt to recognize the entity at the current fixation by matching the extracted feature vector with those already learned. In doing so, we progressively find the best match by first matching the feature vector with the coarse representations denoting the object categories and then matching with finer representations denoting the individual object or instance or view. We use the maximum likelihood estimate to find

the match between any learned representation and the current fixation, i.e., find the object o that maximises the chances of occurrence of the fixation f .

The object recognition model that we investigated is simple, and shares its resources intimately with the attentional system, and uses minimal extra hardware (only to compute the maximum likelihood estimate) to achieve object recognition. The gradual matching from coarse representations like object categories to finer representations like the specific object or instance or view allows us to terminate the search at the appropriate level of representation, depending on our task requirements. Further, by pruning the subtrees (in the object hierarchy) that do not match, we can speed the search for the best match.

To test the performance of our enhanced model, we ran our model on databases of training images that ranged from artificial images of simple geometrical objects like squares, circles at different orientations and sizes to natural images of more complex objects like coke cans, traffic signs, and handicap signs in diverse backgrounds. The model learnt the features of the different training objects, and we organized the information in a hierarchical manner with objects composed of instances that are in turn composed of views.

Our model efficiently detected the target even in scenes with poor resolution or significant noise or clutter and complex backgrounds. Its performance in new scenes was further improved due to its ability to generalize. In most of nearly 300 test images, our enhanced model detected the targets in half the number of fixations taken by the bottom-up salience model.

To test the ability of our model to recognize arbitrary fixations, we ran it on the same test images or scenes mentioned previously. At each fixation, the model extracted the normalized center-surround feature vector, and found the learned object representation that gave the best match. Despite the simplicity of our current approach (simple in that it attempts to recognize the scene entity at the fixation by just considering the feature vector at one location, i.e., the current fixation), it was successful in recognizing a wide variety of simple objects that constituted our database. There were few false negatives and false positives. The good performance of our lightweight model suggests that the human visual system may indeed be sharing resources extensively, and, attention and object recognition may be very intimately related.

References

- [1] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, May 2000.
- [2] V. Navalpakkam and L. Itti. Sharing resources: Buy attention, get recognition. In *Proc. International Workshop on Attention and Performance in Computer Vision (WAPCV'03)*, Graz, Austria, in-press.