A neural network model of visual attention and group classification, and its performance in a visual search task

Hayden Walles, Anthony Robins and Alistair Knott

Dept of Computer Science, University of Otago, New Zealand

Abstract. Humans can attend to and categorise objects individually, but also as groups. We present a computational model of how visual attention is allocated to single objects and groups of objects, and how single objects and groups are classified. We illustrate the model with a novel account of the role of stimulus similarity in visual search tasks, as identified by Duncan and Humphreys [1].

1 Introduction

Humans can represent objects individually, but also collectively, as groups. We can attend to and categorise individual objects, but we can also attend to several objects as a group—and if the objects are all of the same type, we can classify them collectively as being of that type. In fact there is evidence that the visual object classification system is relatively insensitive to the number of items in a group. In monkeys, Nieder and Miller [2] showed that neurons in the inferotemporal (IT) cortex are sensitive to the type of objects in a group but relatively insensitive to their cardinality, while neurons in the intraparietal sulcus show the opposite pattern. A similar sensitivity to type but not number has been found in imaging studies of human IT, using a habituation paradigm where either the type of objects in a group or the size of the group was selectively changed (e.g. [3]). In previous work [4] we coined the term **cardinality blindness** to describe this phenomenon. We showed that a classifier called a **convolutional neural network** (CNN) shows cardinality blindness, and argued that this property also characterises the object classifier in the IT cortex of humans and other primates. Our classifier assigns the same class ('X') to a single visually presented X shape and to a homogeneous group of X shapes. However, when it is presented with a group of objects with different shapes (a heterogeneous group), it typically refuses to make a classification at all.

If the classifier in IT is cardinality blind, this may be expected to have consequences for the design of the attentional system that selects spatial regions to be classified [5, 6]. For one thing, attention should be able to deliver *homogeneous groups* to the classifier as well as single objects, so that the objects in these groups can be classified in parallel. There should also be a system that acts in parallel with object classification, to compute the number information which is not provided by the classifier. In this paper we present a computational model of visual attention and object classification, in which the attentional system selects individuals and groups for the classifier. We also describe the performance of this model in a visual search task.

2 The model of visual attention and object classification

The structure of our model of visual attention and classification is outlined in Figure 1a. The **attentional subsystem** (dorsal pathway) determines the salient regions on the

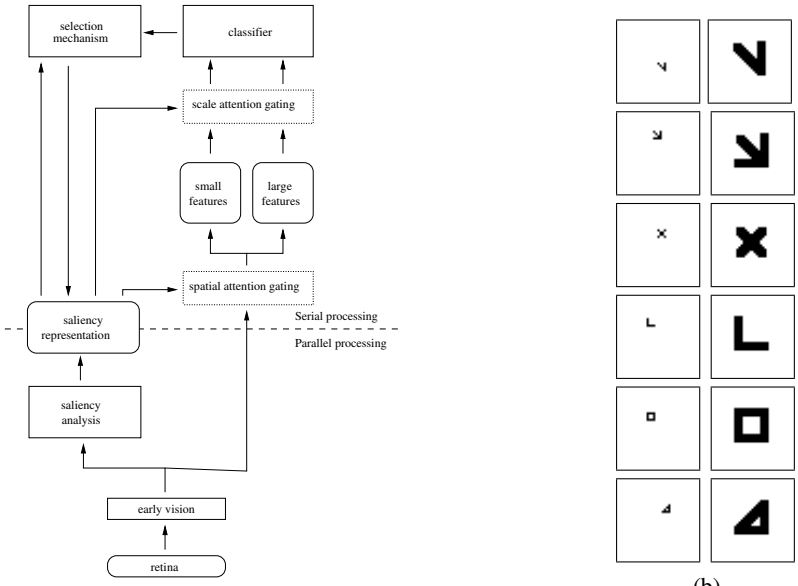

Fig. 1: (a) Our model of the visual attention system (left-hand pathway) and the object classification system (right-hand pathway). (b) The shapes used to train the model.

retina, and activates these regions one at a time. The **classification subsystem** (ventral pathway) categorises the retinal stimulus in the currently activated region; its output changes as different regions are selected [7].

In the attentional subsystem in our model, the saliency of a region is determined by two factors: one is local contrast (how different it is from the surrounding region), the other is homogeneity (how similar its texture elements are). Salient regions can contain isolated visual features which contrast from their surroundings, but also regions containing repeated visual features. Computations of saliency are performed at multiple spatial frequencies, so salient regions containing isolated visual features can be of different sizes. Salient regions containing repeated visual features (i.e. homogeneous textures) can also be of different sizes.

There are several existing computational models of saliency that detect salient regions of different sizes (e.g. [8]), and numerous models of texture identification that detect regions containing repeated visual features (e.g. [9]). There are also many existing computational models of classification that allow objects of different sizes to be classified, by taking as input primitive visual features at a range of different scales (e.g. [10]). The main innovations in our model are in how the saliency mechanism interacts with the classifier. Firstly, in our system, classification is influenced not only by the location of the currently selected salient region, but also by its size. Our classifier can work with primitive features of several different scales as input, but at any given point the scale it uses, called the **classification scale**, is selected by the attentional system. By default, the classification scale is a function of the size of the currently selected salient region,

so that large regions are classified using correspondingly large features, and small regions with correspondingly small ones. Our model is novel in proposing that the scale of the salient region selected by the attentional system determines a default scale for the classifier to use. Secondly, we envisage that the selected classification scale can be changed *without changing the spatial region to be classified*, so that the classifier can reanalyse the currently selected region at a different spatial frequency. In our model, this attentional operation is crucial for the classification of homogeneous groups, and for an account of the difference between single objects and plural groups. We suggest that in order to classify a group of objects occupying a given salient region, the observer must attentionally select a new classification scale which is smaller than the scale established by default. In this account, the distinction between singular and plural can be read from the current classification scale measured *in relation to* the default classification scale for the currently attended region. It is well known that observers can selectively attend to the global or local features of visual stimuli (e.g. [11]), and there is good evidence that this attention involves selective activation of particular spatial frequency channels (e.g. [12]). It has recently been found that the spatial frequencies associated with local and global features of an object are defined in relative not absolute terms ([13]). Our model makes use of this notion of relative classification scale to support an account of group classification and of the distinction between singular and plural in the visual system.

2.1 The classification subsystem

The visual classification subsystem is modelled by a convolutional neural network (CNN) previously described [4]. The classifier takes, as input, retinotopic maps of simple oriented visual features at two different spatial frequencies, or scales: one of these scales is selected by the attentional system. The classifier was trained with six shapes at each spatial frequency (see Figure 1b). The classifier has seven output units: six of these provide localist encodings of the six shape categories and the seventh encodes the verdict 'unknown category'. The units have activations ranging from zero to one. We define the classifier's decision to be the strongest output over 0.5. If no unit's activation exceeds 0.5 the classifier's decision is assumed to be 'unknown category'. In summary, the classifier provides two pieces of information: first, whether classification is possible and, if so, what that classification is.

The classifier exhibits two types of invariance which have been observed in IT [14] and are generally acknowledged to be crucial for a model of vision [10], namely location (or translation) invariance and scale invariance. Location invariance is a result of the architecture of the CNN, which intersperses feature combination layers with layers that abstract over space [4]. Scale invariance depends on the input having been prefiltered for the desired frequency: the small shapes must be classified with the high-frequency visual features, and the large ones with the low-frequency features. Importantly for the current paper, the classifier is also blind to the cardinality of homogeneous groups of small shapes: its accuracy varies from 95% for a single shape to 97% for a homogeneous group of five shapes. Interestingly, these results show a redundancy gain effect similar to that found in humans: the classifier's performance improves the more instances of a type it classifies.

2.2 The attentional subsystem

As shown in Figure 1a, the attentional subsystem can be divided into two interacting stages: a preattentive, or parallel, stage and an attentive, or serial, stage.

The preattentive stage includes an operation called saliency analysis. The job of saliency analysis is to analyse the local contrast and texture homogeneity of the input in parallel. These are used to implement the Gestalt grouping properties of proximity and similarity respectively. The result of this is a saliency representation, or saliency map [15]. This representation is the point of communication between saliency analysis and the selection mechanism. The selection mechanism uses the saliency representation to decide how best to deploy attention. Once processing of attended stimuli is complete the representation is updated and then used to redeploy attention.

The saliency representation is also used to gate the input to the classifier. Input is gated in two different ways. It is gated by location, which is a well-known idea [6, 16]). And it is also independently gated by scale, which is a new idea in our model. The initial scale selected by the attentional system is the **default classification scale** for the selected region. In order to recognise a figure within a region, the primitive visual features which the classifier uses must be of an appropriate spatial scale—not too large and not too small (see Sowden and Schyns [17]). If they are too large, they cannot be combined to represent a complex shape within the region. And if they are too small, then their combinations are not guaranteed to represent the global form of the figure occupying the region. A novel idea in our model is that a selected region can first be classified at the default classification scale, and then subsequently at a finer classification scale. If the classifier returns a result in this second case, it is identifying the type of objects in a homogeneous *group* occupying the selected region. In the remainder of this section we will provide more details about the attentional subsystem and its interaction with the classifier. Full technical details are given in [18].

3 Performance of the system in a visual search task

In this section we describe two experiments investigating the behaviour of our complete system in the domain of visual search There are well-known similarity and grouping effects in search, which our model may be able to explain.

In a visual search task, a subject searches for a target stimulus in a field of distractors. The search time is a function of the number of distractors, but also on the visual properties of the target and distractor stimuli. The earliest visual search experiments reported a discrete difference between 'parallel search', in which search time is independent of the number of distractors, and 'serial search', where search time is linearly proportional to the number of distractors (Treisman and Gelade [5]). In the original model explaining this finding, feature integration theory (FIT), parallel search is possible if there is a single 'visual feature' that the target possesses and the distractors do not, allowing it to 'pop out' of the field of distractors; if the target is distinguished from the distractors by a specific conjunction of visual features, items in the visual field must be attended serially, to allow their features to be integrated.

Later experiments uncovered more complex patterns of visual search performance. Treisman [19] found that perceptual grouping affects search because subjects serially

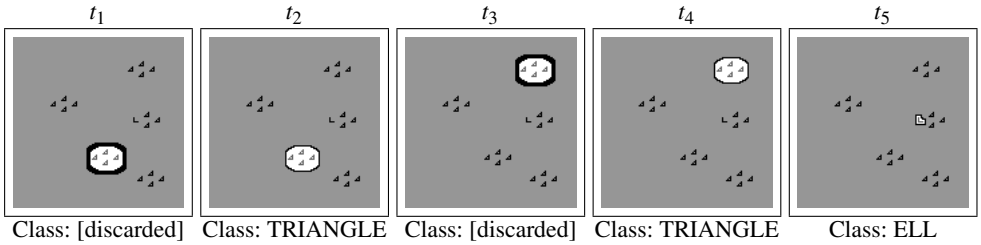

Fig. 2: An example sequence of operations during simple search. At t_1 the input is presented and at subsequent time steps attention is directed as shown until the target (ell) is found. Thick borders around a region indicate attention to the low spatial frequency, thin borders attention to the high spatial frequency.

scan groups of items where possible, not just individual items. Treisman and Gormican's group scanning theory [20] drew this finding into the FIT model. In group scanning theory, when parallel search fails because the target cannot be discriminated from the distractors, attention is used to limit the spatial scope of the parallel search to a region where parallel search by feature discrimination *can* work. Parallel search then continues inside the attended area.

Duncan and Humphreys [1] presented results that challenged the basic assumption of a simple dichotomy between parallel and serial search. They gave subjects search tasks where the similarity between targets and distractors, and the similarity between distractors (i.e. the homogeneity of the set of distractors) were varied continuously. They found that increasing the degree of similarity between target and distractors progressively increases the slope of the search graph, and that increasing the similarity between distractors has the opposite effect. In Duncan and Humphreys' stimulus similarity theory (SST), pop-out and item-by-item serial search are opposite ends of a continuum of search processes, rather than discrete alternatives.

Our model of visual attention and classification is able to identify homogeneous groups and classify their elements in single operations; it therefore has some interest as a model of visual search. In this section we examine its performance on search tasks where target-distractor similarity and distractor-distractor similarity are varied, as in the experiment of Duncan and Humphreys.

To test the search performance of our model, we created four different search tasks, defined by varying two independent binary parameters based on those used by Duncan and Humphreys: target-distractor similarity (with values 't-d similar' and 'td-different') and distractor-distractor similarity (with values 'd-d similar' and 'd-d different'). Details of these tasks are given in [18]. We presented displays of each type to the model, and recorded how many serial attentional steps were taken for it to find the target. Figure 2 shows the steps taken by the system during a td-different/dd-similar search. We found that different search tasks have different slopes. Our simulation reproduces Duncan and Humphreys' main experimental results: when targets are dissimilar to distractors but distractors are similar to one another the search slope is close to flat, and when targets are similar to distractors the slopes are highest. Details of these findings, and a comparison with other computational models of visual search, are given in [18].

References

1. Duncan, J., Humphreys, G.W.: Visual search and stimulus similarity. Psychological Review **91**(3) (1989) 433–458
2. Nieder, A., Miller, E.K.: A parieto-frontal network for visual numerical information in the monkey. Proceedings of the National Acadamey of Sciences **101**(19) (2004) 7457–7462
3. Izard, V., Dehaene-Lambertz, G., Dehaene, S.: Distinct cerebral pathways for object identity and number in human infants. PLoS Biology **6**(2) (2008) 275–285
4. Walles, H., Knott, A., Robins, A.: A model of cardinality blindness in inferotemporal cortex. Biological Cybernetics **98**(5) (2008) 427–437
5. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology **12** (1980) 97–136
6. Moran, J., Desimone, R.: Selective attention gates visual processing in the extrastriate cortex. Science **229**(4715) (1985) 782–784
7. Zhang, Y., Meyers, E., Bichot, N., Serre, T., Poggio, T., Desimone, R.: Object decoding with attention in inferior temporal cortex. Proceedings of the National Academy of Sciences of the USA **108**(21) (2011) 8850–8855
8. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision **45**(2) (2001) 83–105
9. Kadir, T., Hobson, P., Brady, M.: From salient features to scene description. In: Workshop on Image Analysis for Multimedia Interactive Services. (2005)
10. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience **2**(11) (1999) 1019–1025
11. Fink, G., Halligan, P., Marshall, J., Frith, C., Frackowiak, R., Dolan, R.: Where in the brain does visual attention select the forest and the trees. Nature **382** (1996) 626–628
12. Flevaris, A., Bentin, S., Robertson, L.: Local or global? attentional selection of spatial frequencies binds shapes to hierarchical levels. Psychological Science **21**(3) (2010) 424–431
13. Flevaris, A., Bentin, S., Robertson, L.: Attention to hierarchical level influences attentional selection of spatial scale. Journal of Experimental Psychology: Human Perception and Performance **37**(1) (2011) 12–22
14. Logothetis, N.K., Sheinberg, D.L.: Visual object recognition. Annual Review of Neuroscience **19** (1996) 577–621
15. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research **40** (2000) 1489–1506
16. Moore, T., Armstrong, K.M.: Selective gating of visual signals by microstimulation of frontal cortex. Nature **421** (2003) 370–373
17. Sowden, P., Schyns, P.: Channel surfing in the visual brain. Trends in Cognitive Sciences **10**(12) (2006) 538–545
18. Walles, H., Robins, A., Knott, A.: A neural network model of visual attention and object classification: technical details. Technical Report OUCS-2013-09, Dept of Computer Science, University of Otago (2013)
19. Treisman, A.: Perceptual grouping and attention in visual search for features and for objects. Journal of Experimental Psychology: HPP **8**(2) (1982) 194–214
20. Treisman, A., Gormican, S.: Feature analysis in early vision: Evidence from search asymmetries. Psychological Review **95**(1) (1988) 15–48