A simple recurrent network model of active vision for object recognition

Timothy T. Rogers and Maarten van Casteren MRC Cognition and Brain Sciences Unit, Cambridge, UK.

tim.rogers@mrc-cbu.cam.ac.uk maarten.van-casteren@mrc-cbu.cam.ac.uk

Introduction

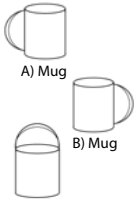
To derive meaning from a visually complex scene, the observer must direct her gaze to relevant objects, identify the objects and establish their spatial relationships to one another, and integrate this information across fixations to form a coherent representation of the scene. Questions about how components of scenes are selected for attention, or how successive fixations give rise to a stable percept, have been fundamental to theories of visual perception since Helmholtz. The same questions are seldom raised in the context of single object recognition; yet object recognition also requires the observer to identify the important components of an object, establish their spatial relationships to one another, and integrate this information into a stable percept. We describe a simple recurrent network model in which direction of gaze critically supports the ability to bind object parts together into a recognisable whole. The model simultaneously learns to a) direct its gaze to informative regions of a visual scene and b) integrate inputs over fixations to recognise objects in the scene. We show how the model addresses key phenomena in object recognition, including generalisation across spatial location, representation of "what" and "where," and configural object recognition.

The role of space in recognition

Some aspects of scene recognition are sensitive to spatial information, whereas others are not.

For instance, to interpret the scenes to the left, we must be able to recognise the dog and the cat regardless of their orientations or spatial locations in the scene...

...but to determine whether the dog is chasing the cat, we must be sensitive both to the orientations of the individual animals, and to their spatial positions relative to one another.



Single object recognition may also exhibit varying sensitivity to spatial information.

For example, each of the objects to the left is composed of the same two parts, set in different spatial configurations.

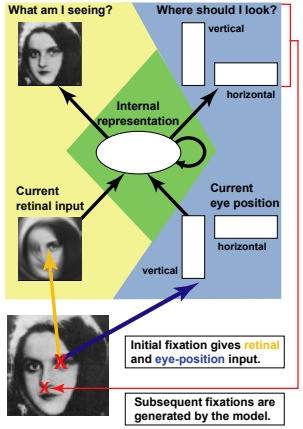
To recognise that the mugs are the same kind of thing but that the bucket is different, we must understand that differences in the arrangement of parts for A and B are irrelevant, whereas a third arrangement of parts in C is significant.

From Yarbus (1967)

In scene perception, it is clear that eye movements can be used to establish the spatial relationships among objects.

Perhaps eye-movements also play a role in establishing the spatial relationships among parts, and in learning which spatial arrangements are critical for recognition and which not.

Concept



Initial fixation gives first retinal and eye position inputs.

Retinal input is faithful in the centre and degraded in periphery.

Model output provides "best guess" about what the object is...

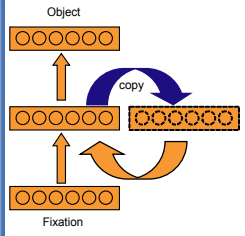
...and where it should move it's eyes next.

Outputs in the "where should I look" layer are used to determine the next fixation point.

New retinal inputs and eye position inputs are given to the network, based on the new fixation point.

The model takes another "best guess" as to what the object is, and where it should look next.

Implementation

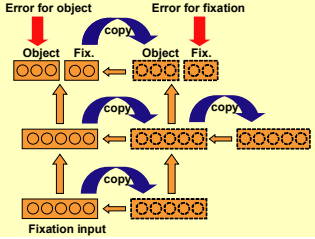


Standard backprop enables us to associate 1 input with 1 output

If we want to integrate the information from several fixations we need a recurrent network

Here the hidden layer has a recurrent connection, enabling it to integrate several inputs over time.

Still, we would like the network to also generate the fixations...



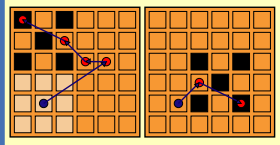
We only know the effect of a new fixation after it has been processed. So, how are we going to give training feedback on the fixation output? The solution is to inject the error produced by the current input back into the fixation output units, but one timestep earlier.

Error for a generated fixation is calculated by comparing the network error on the object-identity units before and after the fixation. A good fixation should reduce the error. To keep this differential error measure within reasonable values we use this transformation, which limits the error to the interval (-1,1):

Error (fixation) = Tanh (Error(object after fixation) - Error(object before fixation))

Initially the network selects locations at random, but as it begins to learn the mapping from "fixation" to "object identity," some fixations are reinforced and others are punished. Eventually the model finds efficient scan paths...

A simple example

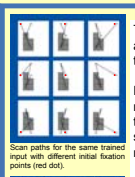


We first trained the model to recognise 9 simple patterns, each appearing in 16 possible locations. The Figure shows scan paths for the trained model on two test trials. Pale blocks indicate the extent of the model's "visual field."

Given the same starting location, the model generates a different sequence of fixations for different objects or locations.

It also appears to search the environment in a reasonable way. In the left-most panel, intially seeing nothing in its visual field, the model first looks to a location that would detect any object in the right side of the environment. Failing to find anything there, it moves its gaze to the left, "checking" the top left quadrant. Upon finding the object, the model identifies it in 2 further fixations. In the right panel, the object falls within the network's field of view at first fixation, and the model recognises it in 2 subsequent fixations.

Generalization across location



Scan paths for the same trained input with different initial fixation points (red dot).

Density of fixations across novel trials for same item in different locations. The model locates the item horizontally, then moves its gaze up and down.

The network shows a natural inclination to generalise across spatial location. To illustrate this, we trained the model to recognise 26 letters of the alphabet.

Each letter could appear in one of 16 locations, making a total of 416 possible inputs, but during training, each letter appeared in only one randomly-selected location (or 7% of the full corpus). The model was then tested on the full corpus.

For both novel and trained inputs, the model responded correctly 99% of the time, and took 3-4 fixations on average to identify the letter. Thus the network shows near-perfect learning, and perfect generalization.

The reason is that, once the network has found the object in the visual field, it can "line up" its retina with the stimulus so that a given letter always produces the same pattern of input across the fovea, regardless of its location in the environment.

Reporting "what" or "where"

The network is not insensitive to spatial information, however. To show this, we trained the model to report either the identity of a letter, or its spatial location relative to a visual reference point.

On each trial, a reference point (X in figures below) and one of 5 different letters (A-E) both appeared in the environment, each occupying one of 4 possible locations.

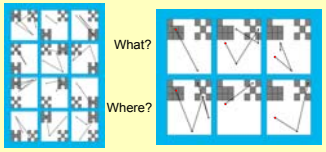
Two input units were added to instruct the model to report either the identity of the letter, or its position relative to the reference point.

The trained model was able to correctly report a) the identity and location of familiar letters appearing in novel (i.e. untrained) locations, and b) the location of completely unfamiliar letters (left Figure).

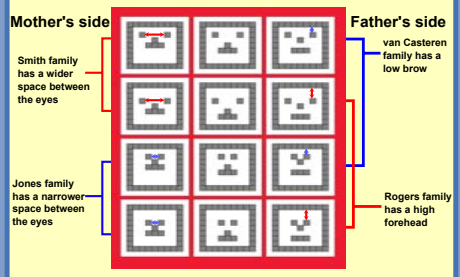
The model learned to generate different scan paths for the same input, depending on whether it was reporting identity or location (right Figure).

Left. Scan-path for trained model queried for the location of a novel letter (H) relative to the reference point (X). Despite never having seen an H, the model produces the correct response in an average of 5-6 fixations.

Right. Scan path for same input when model is queried for identity or location of letter b. Gaze is more likely to encompass reference point when location is relevant.



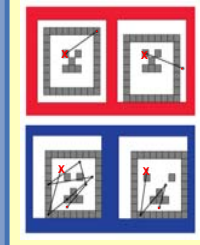
Configural object recognition



In this recognition problem, all faces have the same "features" (2 eyes, nose, mouth), but in different spatial arrangements.

To determine the parentage of a given face, the observer must take into account a) the spatial arrangement of the features themselves (mother's side), and b) the spatial location of the features within the reference frame of the face boundary (father's side).

The model was trained with 8 of the 12 faces above, with each face appearing in one of 9 possible locations. It was taught to report both the mother's family and the father's family for all 8 faces in every possible location. It was then tested with the 4 new faces shown below.



For the first test pair, the features appear in the same absolute location, but the location of the face boundary differs. The model correctly concludes that the left face is a Smith-Rogers and the right is a Smith-van Casteren, taking only 2 fixations, which encompass the informative feature and the reference frame.

In the second test pair, the face boundary is the same, but the position of the eyes differs. In an average of 5-6 fixations, the model correctly concludes that the left face is a Jones-van Casteren whereas the right is a Smith-van Casteren. In both cases, the scan path terminates at the diagnostic position.

Conclusions

In visual scene perception, it is clear that knowledge about spatial locations and object identity must converge to yield a Gestalt representation of the scene. Perhaps because object recognition is relatively robust to changes in location or viewpoint, spatial information is not usually considered critical to single object recognition. Yet classic work by Marr and Nishihara (1978) and Biederman (1986) suggests that object recognition depends on establishing the spatial relations among geometric primitives. A theory of visual perception must explain how people capitalise on spatial information when it is useful, and ignore this information when it is not--both for object recognition and scene perception. Our model provides a simple framework for thinking about how visual and spatial information are integrated over time to support object recognition and scene perception, in a manner that is sensitive to spatial information only when that information is relevant. The framework shows promise for addressing certain key phenomena in the study of visual recognition, but it remains for further empirical work to determine whether it has any basis in reality.