PhD Thesis

Fintan S. Nagle

June 8, 2014

Supervisors: Alan Johnston and Peter McOwan
n words as counted by the TeXcount script at
`http://app.uio.no/ifi/texcount/index.html`.

Acknowledgements
...

Please note
Throughout this document, error bars show one standard error of the mean.

Contents

0.1 Introduction

0.2 Thesis organisation

This document is organised into the following sections:

0.3 Literature review

0.4 Terms

0.5 The overall goal

0.6 Modularity

The property of modularity is the possibility to divide a system into multiple components!

0.7 Forms of neural coding

- **Single neuron activation**. The firing of a single neuron can convey binary information.
- **Single spike frequency** can code a real-valued quantity.
- **Spike frequency across multiple neurons** can code relative information between two real-valued quantities.
- **Connection patterns** between neurons (the existence of a connection, or its strength) can code complex information, but this information cannot be extracted without activating the neurons and monitoring the outputs.

0.8 Static and dynamic faces are processed differently

The first evidence of a difference in the perception of expression between static and dynamic faces was found in 1991[1].

0.9 Identity vs. expression

There is a substantial body of evidence that identity (information which is invariant within individuals) and expression (information which is invariant across perceived emotional states) are processed differently. On the high level, identity judgement and expression judgement have been observed to be doubly dissociated in prosopagnosics[2]. However, this observation may not allow us to generalise deductions to the normally-functional population, as prosopagnosics may have developed alternative recognition strategies such as non-holistic feature recognition (as is used to recognise classes of objects for which we do not possess a specialised representation or processing system).

On a slightly lower level, judgement reaction times differ depending on whether expression or identity is being judged; when judging identity, familiar faces are matched faster, but familiarity confers no advantage when judging expression[3]. This could imply that the computation of identity is intrinsically more complex or that other neural actions such as memory retrieval of biographical data are triggered.

On the lowest level, it is possible to find individual neurons which are receptive to either identity or expression[4]. Multidimensional scaling methods on their spike train data allow stimuli to be classified in either identity or expression space solely by neural response.

However, the location in one test subject of a small number of individual neurons which correlate with a particular condition provides no information about the algorithmics of face processing; it simply demonstrates that the brain can judge identity and expression at some level (which is intuitively obvious) and that this information can be coded by neural activation as opposed to connection patterning or higher-level codes such as spike train phase.

0.10 Correlates between the two decouplings

It is tempting to connect the identity-expression dichotomy with the static-dynamic dichotomy, as dynamic faces have constant identity but changing expression. This would be erroneous, as static faces can vary in both expression and identity.

0.11 Object perception

0.12 Visual perception as dimensionality reduction

Visual perception creates percepts from visual input. Photons arrive on the retina and induce signals in the optic nerve, which then pass to the LGN, dorsal and ventral visual pathways, and eventually effect conscious perception (such as when we perceive a face) or motor control (such as when we press a button to indicate that we have seen a face).

The number of photons arriving per unit time is so high that they cannot all be losslessly recorded, as shown by the reduced information capacity of the optic nerve[5] compared to the retina, so information is compressed before dispatch. Motion representations are a simple form of compression; rather than recording the positions of a dot at each time-step (1,2,3,...,99,100), we can simply record its initial position (1) and speed (1 unit per second). Averaging is another simple compressor, as is nonlinear activation of cone cells (which require several afferent photons to change their membrane potential).

The bandwidth of the optic nerve is also smaller than that of incoming light signals, and this is dealt with by retinal adaptation.

These forms of compression can all be seen as transfer functions from low- to higher-level representations. The ultimate low-level representation of visual input is to record every photon arriving on the retina, but as this is impractical, optic nerve representations are compressed.

The process continues as we move further away from the retina and into the early visual system. Colour perception is another compression strategy, allowing any combination of wavelengths to be described by three coordinates in colour spaces like RGB, HSV or LAB.

Compression is evident in Marr's theory of vision, as in [6]:

"A representation is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this. And I shall call the result of using a representation to describe a given entity a description of the entity in that representation."

Provided that Marr's "result" contains less information than his "entities or types of information," representation is precisely a process of compression.

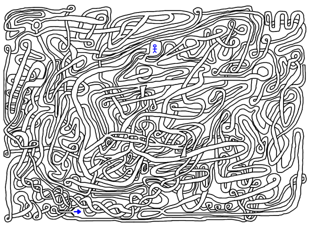
Different types of representation record and miss different types of information. For example, neurons in V5 are sensitive to motion but not to colour; neurons in FFA are sensitive to faces but not houses. Many early visual centres are retinotopically mapped, keeping account of the position on the retina of a stimulus. This information, however, is not always useful: maintaining a map requires separate channels for each part of the visual field, even those which are not of interest.

In a retinotopically-mapped representation, it is easy to compare objects in the same position by simply subtracting or Pearson-correlating the corresponding areas. However, if the same object is present in the top-left of map A and the bottom-right of map B, comparing the two maps will not detect object identity.

In reality, object recognition is position-invariant: we are able to track an object as it moves around on the retina, and also to compare objects in different positions in the map. How position-invariant representations are built is a key problem in understanding the brain[7].

Position-invariance is an important quality of object percepts, but is it sufficient to say that something is perceived as an object? Object perception is often associated with two other properties: spatiotemporal extent, and a canonical coordinate frame.

Spatiotemporal extent Most objects have spatial extent: defined borders. Prominent work on segmentation[8, 9, 10, 11, 12] and figure-ground separation[13, 14, 15] (which is simply segmentation from the background) underlines how important this process is to vision. Segmentation is task-specific: although we *can* segment this object:



we do not automatically perform that complex computation (and cannot without scanning the fovea over the image, as it is too detailed) unless engaged in a task which requires it, like copying the object.

When reasoning about dynamic stimuli, some objects have temporal extent: they appear at a certain moment, then disappear. Temporal extent is often task-sensitive. For example, a changing traffic light can be seen either as three permanent “light” objects which change colour, or as “red-light”, “yellow-light” and “green-light” objects which appear when they illuminate and disappear when they darken.

Canonical coordinate frames Translation is not the only transformation under which objects are invariant: they also preserve their identity under 2D or 3D rotation and scaling. Marr pointed out that many objects are more easily recognised from certain points of view and inferred that they possess principal axes[16]. This property allows different objects of the same class to be compared; their principal axes can be aligned, allowing features to be registered.

In summary, objects are percepts which admit position invariance and have defined extent. However, all visual input is not segmented into objects. One of the main stimulus classes which we do not segment consists of textures.

0.12.1 Textures

There is much work on texture perception[], individuation[] and classification[]

Textures differ from objects

0.13 Representations

It is important to note that there is never a unique representation of a visual stimulus, and it makes no sense to speak of ”the” representation of a face or a house. Representations of a scene include:

Retinal photon trace (similar to a digital camera image) Optic nerve representation Neural recordings from V1 Neural recordings from the FFA A verbal description of a scene A written description of the scene

In terms of size, representations range from the very small (a recording from a single face-sensitive neuron can be taken to represent the presence or absence of a face in its receptive field) to the very large (such as Gallant’s reconstruction of visual input from multivoxel MR imaging[17]).

The Marrean view sees representations as processes. Like other processes, such as functions, they can be composed so that information flows through them sequentially. Visual information flows from the retina to a single face-sensitive neuron as follows:

Retina - optic nerve - visual centres - FFA - neuron of interest

0.13.1 Levels

Representations can be seen to operate on different levels. We say that we move ”up” from a low-level representation (the retina, or image space) to a high-level representation (FFA neurons, or face space). ”Top-down control” indicates that cognitive representations accessible to consciousness are influencing low-level representations like motor neuron activity.

This up-and-down metaphor is very imprecise, despite being very common in the literature (over 369,000 results for the search ”top-down control vision” on the Google Scholar literature search engine). It can generally be interpreted in two ways.

1. Top-bottom as distance from consciousness This view sees representations as being organised according to their interaction with consciousness. Qualia, intentions and percepts are the most high-level representations, as they are consciously accessible. Early visual system representations are seen as lower-level as they can be hidden from consciousness by processes like masking, crowding and adaptation. We refer to this as the **awareness scale**.

2. Top-bottom as representational information This view sees representations as being organised according to their information content, or entropy. Consider our two alternative codes for the 1D positions of a moving point, R1:(1,2 3,4,...,99,100) and R2:(start=1, speed=1). Although they describe the same thing, R1 contains 50 times more information than R2 (100 numbers compared to 50). We refer to this as the **information scale**.

These two metaphors describe completely different things, yet are mixed under the monikers ”top-down” and ”bottom-up.” It is necessary to be very clear about which one we mean.

0.13.2 Operations on representations

Matching. Two representations can be compared for identity. This usually happens on two representations of equal information level

Chapter 1

Modelling

1.1 Forgetting functions

Chapter 2

Image domain analysis

Chapter 3

Experiments on fire alone

3.1 Recording and processing of stimuli

A continuous 45-minute recording was acquired from a hearth fire using a Sony INS camera recording at 50 Hz. The scene was lit by a mixture of natural and artificial light and CCD gain was set to zero. Video was saved directly to the compressed AVCHD format at an initial resolution of 1024 by 768.

Before presentation, stimuli were cropped to 564 by 641 pixels, removing the background and most of the fireplace. Individual frames were decompressed and saved as bitmaps.

3.2 Experimental set-up

Experiments were coded in MATLAB using Psychtoolbox. Video was displayed by loading bitmaps into video memory and manually displaying each one to the screen. This allowed precise control of frame rate.

Stimuli were displayed at 50 Hz on an INS monitor with a refresh rate of 100 Hz and a resolution of INS. The active video area subtended a visual angle of 14; subjects used a chin-rest at a distance of 57 cm from the screen and were asked not to deviate their head angle from the vertical. Subjects were not requested to fixate, and the experiment took place in a darkened room.

All monitors used during these experiments were identically calibrated using a Cambridge Research Systems ColorCal or ColorCal MKII.

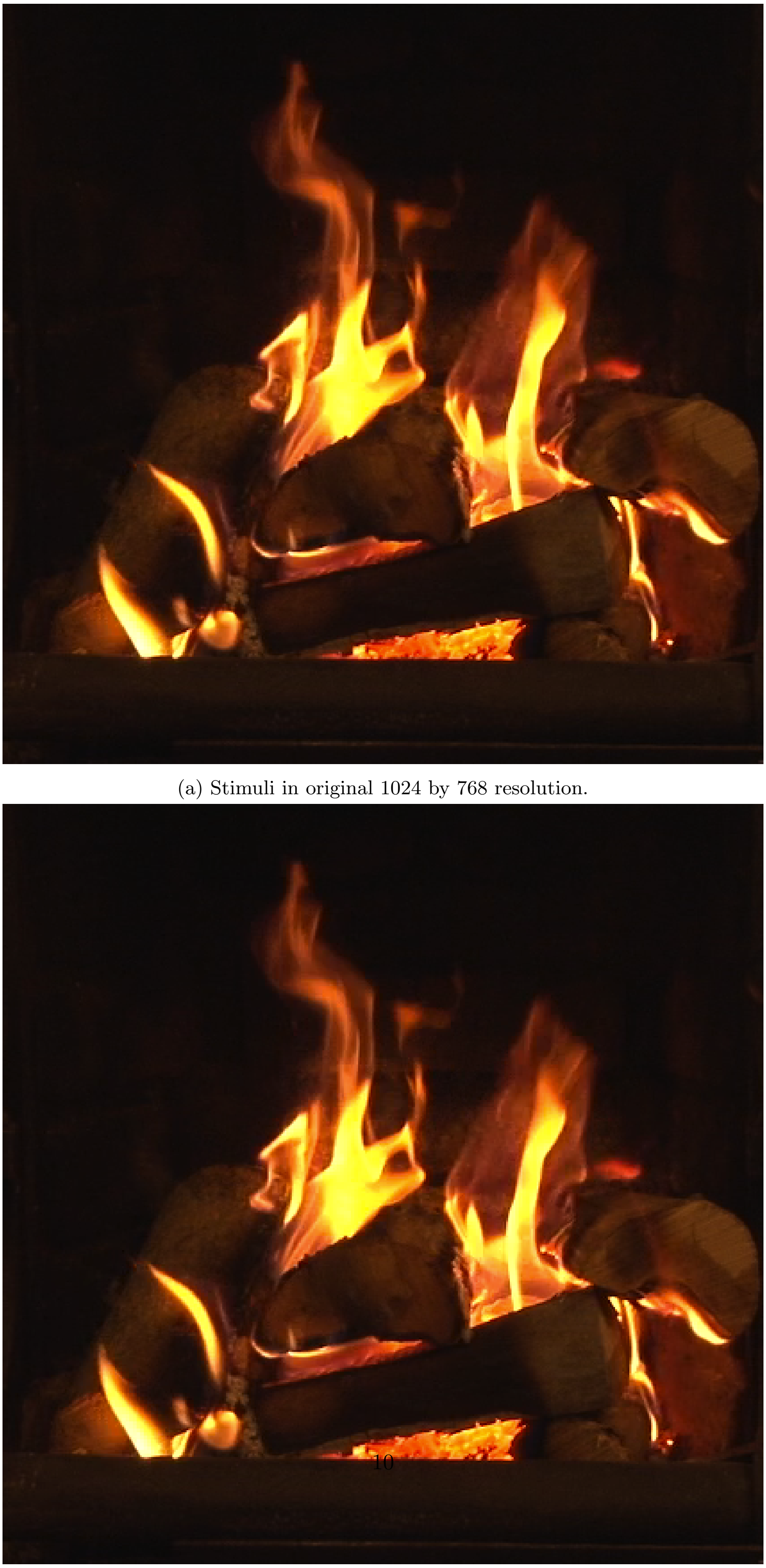
3.3 Experiment 1: search-ratio

3.3.1 Methodology

Stimuli A 1000-frame corpus of consecutive fire images was used.

Subjects 12 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Trial structure In each trial, a sample was presented first, followed by two tests. Subjects indicated which test they thought corresponded to the sample using the left arrow (first sample) and right arrow (second sample) keys.



(a) Stimuli in original 1024 by 768 resolution.

(b) Stimuli cropped to 564 by 641 pixels, as presented to subjects.

Figure 3.1: Stimuli used in our visual search experiments.

Factors Sample length (sL) was one of (10 25 50) frames, equivalently (0.2 0.5 1) seconds. Ratio of sample to test was one of (1.2 1.4 1.6 1.8 2).

This gave the following sample lengths: 10-frame sample: 12 14 16 18 20 frames, 0.24, 0.28, 0.32, 0.36, 0.4 seconds 25-frame sample: 30 35 40 45 50 frames, 0.6,0.7,0.8, 0.9, 1 seconds 50-frame sample: 60 70 80 90 100 frames, 1.2,1.4,1.6,1.8, 2 seconds

There were 3*5 = 15 conditions.

Block structure 25 training trials were presented first.

Sample length was varied across blocks. Target length was varied within blocks.

We presented 3 blocks, one corresponding to each target length, in random order. Subjects took a short break between blocks.

We used a total of 600 trials (40 trials per condition).

3.3.2 Results

Sample length We observed the following mean accuracies:

Sample length	**Mean accuracy**
10 frames (0.2 s)	0.728
25 frames (0.5 s)	0.704
50 frames (1 s)	0.687

Paired-sample *t*-tests revealed a significant accuracy drop between the 0.2 s samples and the 1 s samples (*p*¡0.05) but not between any other pairs of levels. Subjects are more capable of matching longer samples.

Test/sample ratio The ratio by which the test was longer than the sample (which rises as the search space increases) had a significant effect on subject performance.

Test/sample ratio	**Mean accuracy**
1.2	0.746
1.4	0.724
1.6	0.704
1.8	0.696
2	0.664

Paired-sample *t*-tests revealed significant differences between ratio=1.2 and each of the other levels; between ratio=1.8 and ratio=2 (p <0.05); and between ratio=1.4 and ratio=5 (p <0.01).

Learning rate We measured the subjects' learning rate by arranging the correct/incorrect responses in the order in which they were presented during the experimental run, blocking them into sequential groups of 20, and calculating the mean accuracy of each group. As shown in Fig. 3.4a

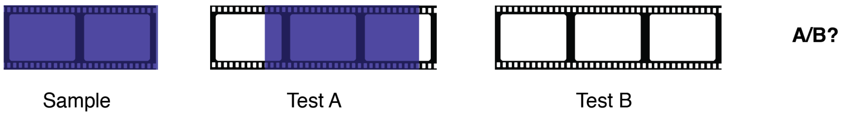
3.3.3 Discussion

A two-factor repeated-measures ANOVA shows a highly significant effect of test/sample ratio (p <0.0001) but not of sample length ((p =0.203) or of the ratio/sample length interaction (p =0.503). Neither is there an effect of absolute test length: see Fig. **??**. The three sample length conditions ranged over three sets of test lengths which did not overlap at all; but, when re-expressed in terms of ratios, the effects on accuracy were consistent.

It is therefore the ratio between test and sample which determines the difficulty of the task[1], not the absolute lengths of either the target or the test.

	Experiment 1
Design	2AFC delayed match-to-sample (sample clip followed by two test clips)
Stimuli	1000-frame corpus
Factors	sample length (10, 25, 50) frames or (0.2, 0.5, 1)seconds. sample/test ratio (1.2 1.4 1.6 1.8 2).
Block design	Sample length varied across blocks Test length varied within blocks 15 conditions 40 trials per condition 600 trials 25 training trials

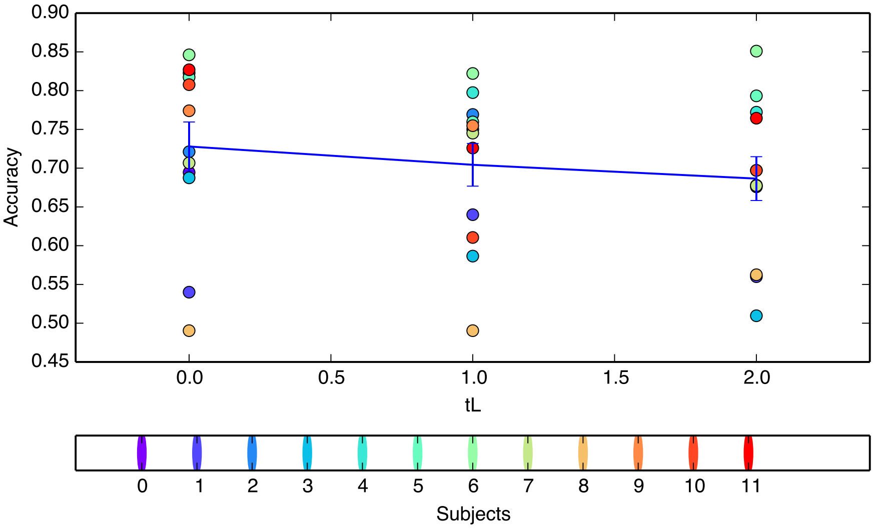
(a) Design summary.



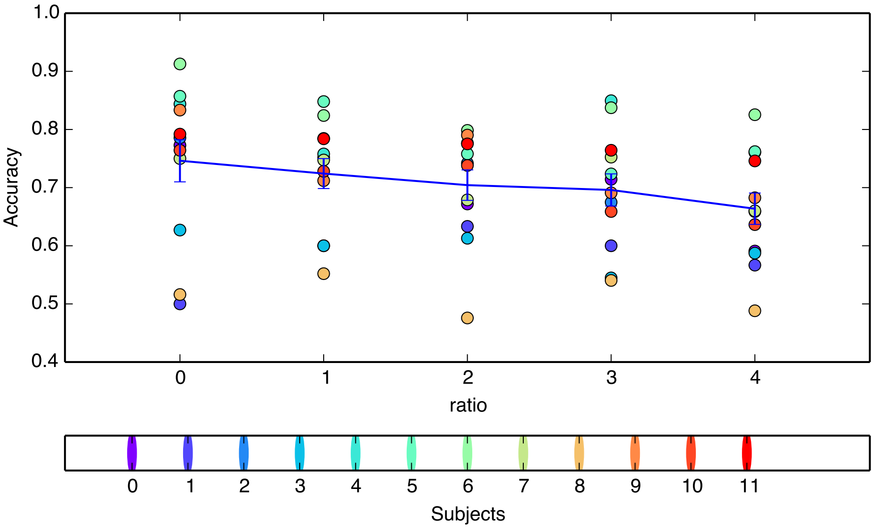
(b) A short sample was followed by two longer tests, one of which contained the sample.

Figure 3.2: Experiment 1: design summary and trial structure.

[1] We use difficulty here as a proxy for accuracy, not to indicate the perceived "hardness" of the task, which we did not measure.

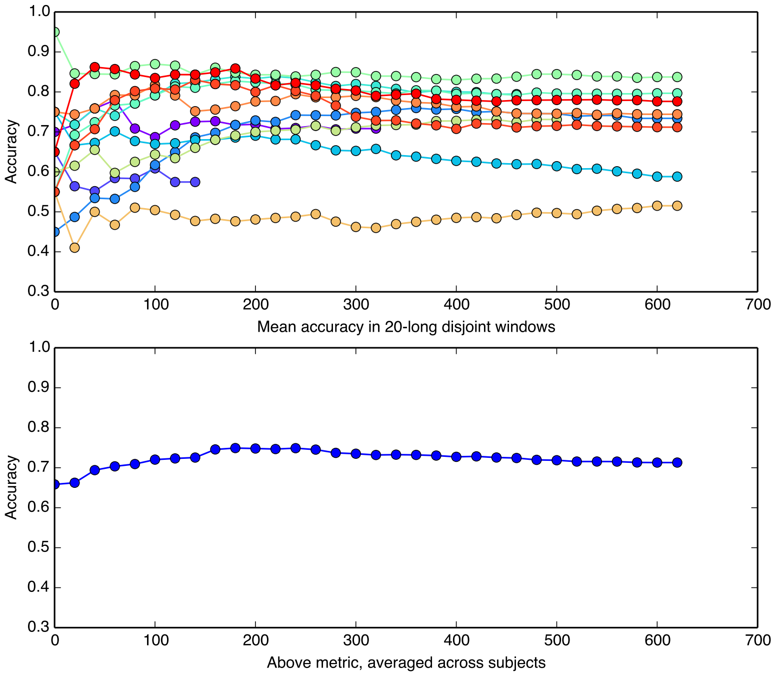


(a) Accuracy against sample length. Longer samples are more easily matched.



(b) Accuracy against test/sample ratio. Longer tests render the sample harder to find.

Figure 3.3: Experiment 1: the effects on accuracy of varying sample length, and test/sample ratio.



(a) Learning: accuracy vs. trial number, blocked into groups of 20.

Figure 3.4: Experiment 1: the effects on accuracy of varying sample length, and test/sample ratio. As throughout this document, error bars are 1 SEM.

Chapter 4

Unsorted

4.1 Characterising fire

4.2 Evolutionary value

From an evolutionary point of view, mastery of fire was key to human development. Being able to control fire allowed early humans to cook food, defend themselves from predators and survive in cold, challenging environments. Fire was the first of a long line of technologies which release stored energy from fuel and turn it to human purposes; the earliest archaeological evidence of fire use dates back 1.8 million years, with frequent use found from 100,000 years ago[18]. Even before this, hominids regularly encountered flame in the form of bushfires, although these were perceived as a threat, not a controllable, exploitable entity.

The evolving human visual system has therefore been exposed to a large amount of flamelike stimuli in the last 1.8 million years. These stimuli have often appeared in dangerous or life-threatening contexts, either posing a threat or aiding survival. In sufficiently extreme situations, such as extreme cold or heavy predation, those early humans who could successfully control fire had an increased chance of survival.

It is therefore natural to enquire whether the human visual system has become adapted in any way to the perception of flamelike stimuli. Does the visual system employ any specific representations or specialised models when attending to fire, does it use the same general-purpose systems employed when observing a novel moving stimulus?

This question recalls the ongoing debate concerning the specialisation of face perception. We find increased activation of the fusiform face area and inferior temporal sulcus while viewing faces[19]; this can be explained either by innate specialisation or learned proficiency. In the same way, observation of fire may recruit neurons and systems which respond preferentially to, and perform better on, flame stimuli. On the other hand, observing fire may stimulate the same neural populations as observing other moving stimuli.

4.3 High-level and low-level representations

The goal of neuroscience is to impose structure and explanatory power on the neural systems present in the brain. This task is accomplished using descriptions in several different domains of representation:

	Experiment 1
Design	2AFC delayed match-to-sample (sample clip followed by two test clips)
Stimuli	1000-frame corpus
Factors	sample length (10, 25, 50) frames or (0.2, 0.5, 1)seconds. sample/test ratio (1.2 1.4 1.6 1.8 2).
Block design	Sample length varied across blocks Test length varied within blocks 15 conditions 40 trials per condition 600 trials 25 training trials

Table 4.1: default

Bibliography

[1] G.W. Humphreys, N. Donnelly, and M.J. Riddoch. Expression is computed separately from facial identity, and it is computed separately for moving and static faces: Neuropsychological evidence. *Neuropsychologia*, 31(2):173–181, 1993.

[2] J. Archer, DC Hay, and AW Young. Movement, face processing and schizophrenia: evidence of a differential deficit in expression analysis. *British Journal of Clinical Psychology*, 33(4):517–528, 1994.

[3] V. Bruce and A. Young. Understanding face recognition. *British journal of psychology*, 1986.

[4] M.E. Hasselmo, E.T. Rolls, and G.C. Baylis. The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural brain research*, 32(3):203–218, 1989.

[5] J Gerard Wolff. Computing, cognition and information compression. *AI Communications*, 6(2):107–127, 1993.

[6] David Marr and Herbert Keith Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140):269–294, 1978.

[7] Simon M Stringer and Edmund T Rolls. Position invariant recognition in the visual system with cluttered environments. *Neural Networks*, 13(3):305–315, 2000.

[8] Daniel C Kiper, Karl R Gegenfurtner, and J Anthony Movshon. Cortical oscillatory responses do not affect visual segmentation. *Vision research*, 36(4):539–544, 1996.

[9] Stephen Grossberg and Ennio Mingolla. The role of illusory contours in visual segmentation. In *The perception of illusory contours*, pages 116–125. Springer, 1987.

[10] Janette Atkinson and Oliver Braddick. Visual segmentation of oriented textures by infants. *Behavioural Brain Research*, 49(1):123–131, 1992.

[11] Edward M Riseman and Michael A Arbib. Computational techniques in the visual segmentation of static scenes. *Computer Graphics and Image Processing*, 6(3):221–276, 1977.

[12] Zhaoping Li. Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. *Network: computation in neural systems*, 10(2):187–212, 1999.

[13] Frank Kelly and Stephen Grossberg. Neural dynamics of 3-d surface perception: Figure-ground separation and lightness perception. *Perception & Psychophysics*, 62(8):1596–1618, 2000.

[14] Stephen Grossberg. Figure-ground separation. *CAS/CNS Technical Report Series*, (062), 1993.

[15] Stephen Grossberg. 3-d vision and figure-ground separation by visual cortex. *Perception & psychophysics*, 55(1):48–121, 1994.

[16] David Marr. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, 1982.

[17] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

[18] David MJS Bowman, Jennifer K Balch, Paulo Artaxo, William J Bond, Jean M Carlson, Mark A Cochrane, Carla M DAntonio, Ruth S DeFries, John C Doyle, Sandy P Harrison, et al. Fire in the earth system. *science*, 324(5926):481–484, 2009.

[19] Truett Allison, Aina Puce, and Gregory McCarthy. Social perception from visual cues: role of the sts region. *Trends in cognitive sciences*, 4(7):267–278, 2000.

Appendices here