PhD Thesis

Fintan S. Nagle

June 19, 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.
When discussing accuracy, we use both the $[0, 1]$ interval and the $[0\%, 100\%]$ interval.
Thus a drop in accuracy of 0.03 is a drop of 3 percentage points.

Contents

Chapter 1

Introduction

1.1 Thesis organisation

This document is organised into the following sections:

Chapter 2

Literature review

2.1 Terms

2.2 The overall goal

2.3 Modularity

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

2.4 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.

2.5 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[**?**].

2.6 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[**?**]. 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[?]. 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[?]. 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.

2.7 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.

2.8 Object perception

2.9 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[?] 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 [?]:

”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.

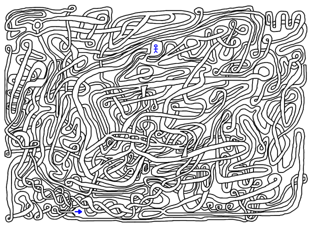
Different types of representation record and miss different types of information. For example, neurons in V5 are sensitive to motion but not to colour; neurons in FFA are sensitive to faces but not houses. Many early visual centres are retinotopically mapped, keeping account of the position on the retina 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[**?**].

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[**?**, **?**, **?**, **?**, **?**, **?**] and figure-ground separation[**?**, **?**, **?**] (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[**?**]. 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.

2.9.1 Textures

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

Textures differ from objects

2.10 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[**?**]).

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

2.10.1 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.

2.10.2 Operations on representations

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

Chapter 3

Image domain analysis

Chapter 4

Experiments on fire alone

4.1 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.

4.2 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.

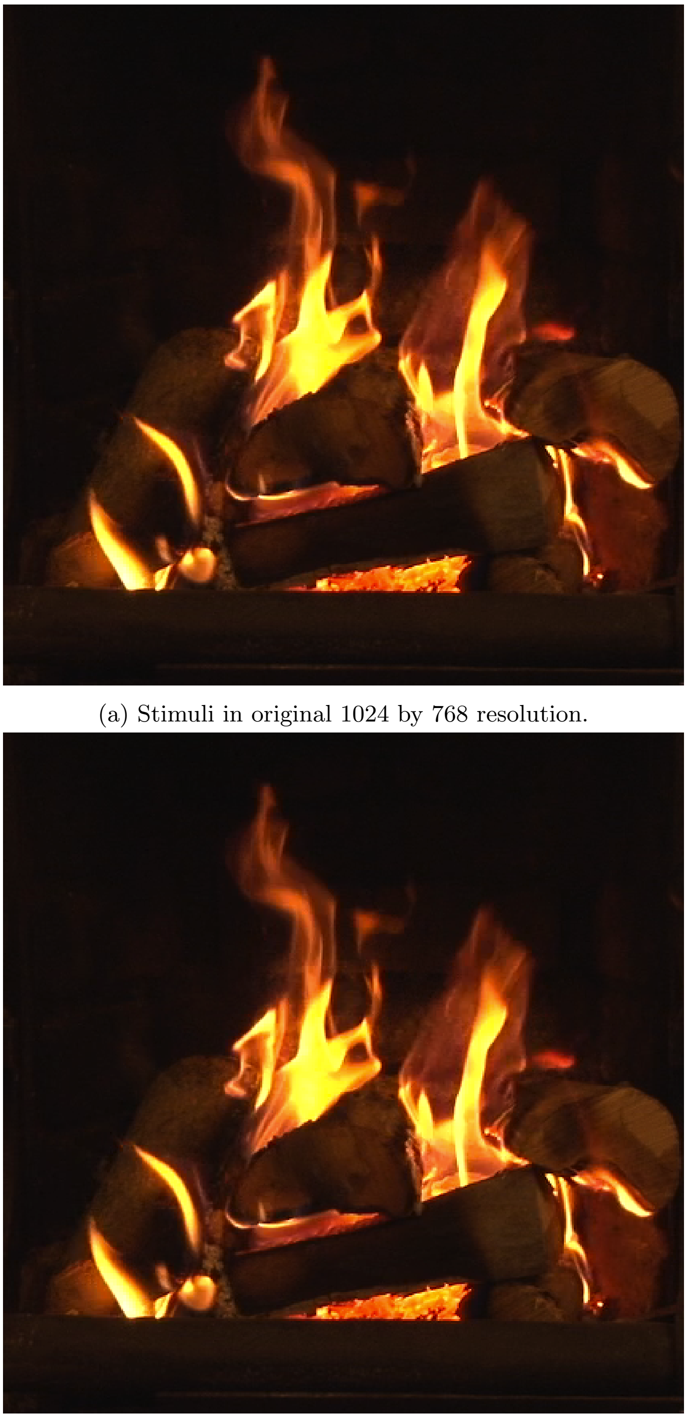
All monitors used during these experiments were identically calibrated using a Cambridge Research Systems ColorCal or ColorCal MKII.

4.3 Choice of subjects

So that we could study the fire perception of the general population, rather than that of trained vision scientists, we recruited from an email list operated by University College London. Most subjects were degree or Masters students aged between 20 and 25. There was an approximately equal gender balance.

During each experiment, subjects were screened for accuracy during the training phase, which was very easy (average accuracy 80% or higher). We followed the policy of rejecting subjects if they could not perform about 70% during the training phase. However, this rarely occurred; only two subjects were rejected during the entire course of experiments described in this thesis.

We therefore consider that our subjects form a representative sample of the general population.



(a) Stimuli in original 1024 by 768 resolution.

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

Figure 4.1: Stimuli used in our visual search experiments.

4.4 Motivating subjects

4.5 Experiment 1: search-ratio

4.5.1 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.

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).

4.5.2 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

Unfolding across sample lengths and ratios, accuracies are the following:

Sample length	ratio=1.2	ratio=1.4	ratio=1.6	ratio=1.8	ratio=2
0.2 s	0.766	0.728	0.767	0.696	0.682
0.5 s	0.759	0.731	0.691	0.681	0.658
1 s	0.711	0.714	0.654	0.709	0.648

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. 4.4a

4.5.3 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.

Subjects' mean accuracy peaks at about trial number 150, showing that this task is learnt quickly.

Across subjects, there are large individual differences in accuracy.

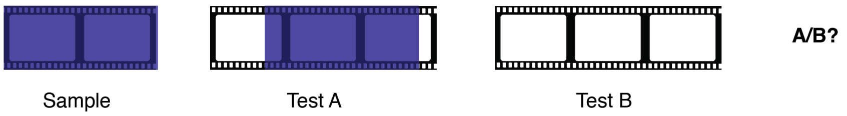
Even at maximum accuracy, subjects only perform at 74%: this is a relatively hard task, even with sub-1-second clips.

This experiment only characterised samples between 10 anf 50 frames (0.2 and 1 seconds). To investigate subjects' performance on shorter clips and smaller amounts of image data, we conducted a similar experiment using samples between 1 and 12 frames (0.02 and 0.24 seconds).

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

	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 480 trials 25 training trials
Subjects	n

(a) Design summary.



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

Figure 4.2: Experiment 1: design summary and trial structure.

4.6 Experiment 2: search-shortsamples

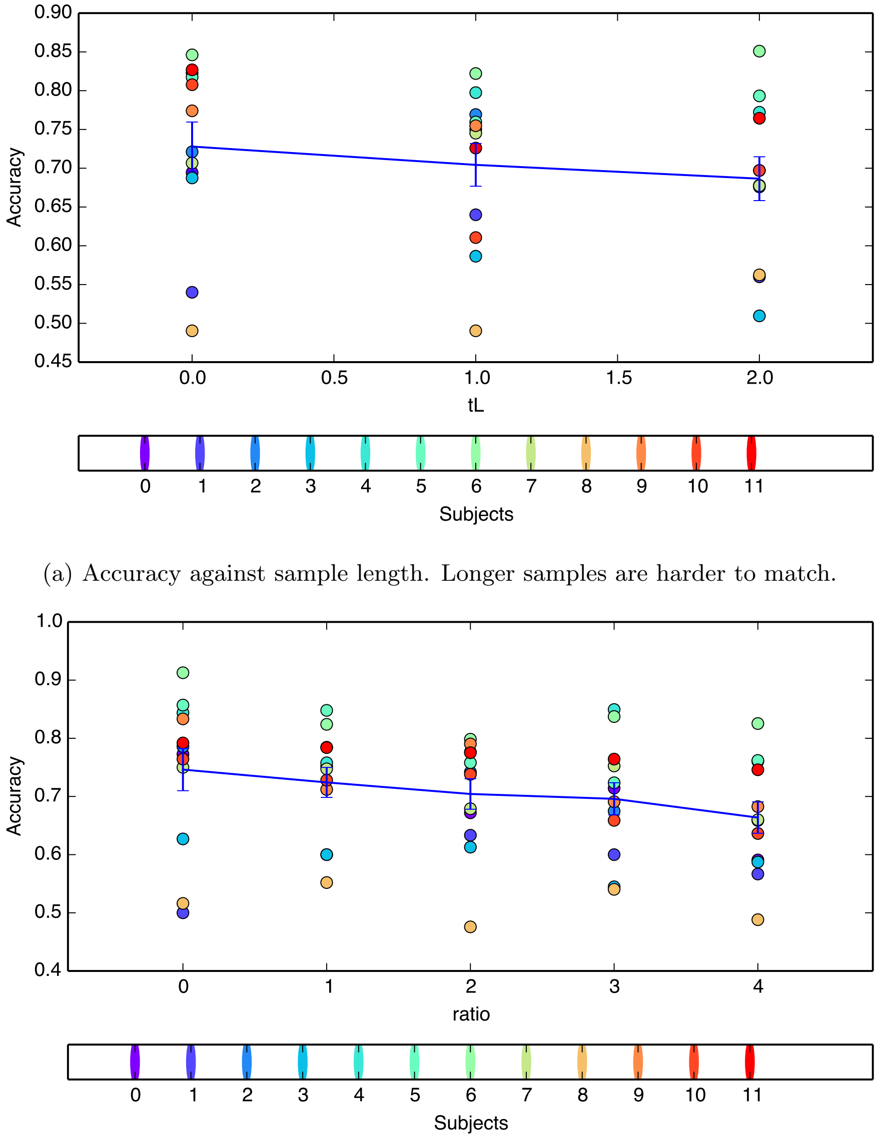
Experiment 1 showed us that the visual system can effectively discriminate the complex patterns of motion and form found in fire, but cannot effectively search for them in longer video sequences. However, it only looked at test clips between 0.2 and 1 seconds in length.

In order to look at the discriminability and searchability of shorter clips, we performed Experiment 2. This study used an identical 2AFC delayed-match-to-sample technique, but with shorter video clips lasting between 0.02 and 0.24 seconds (1 to 12 frames at 50Hz).

4.6.1 Methodology

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

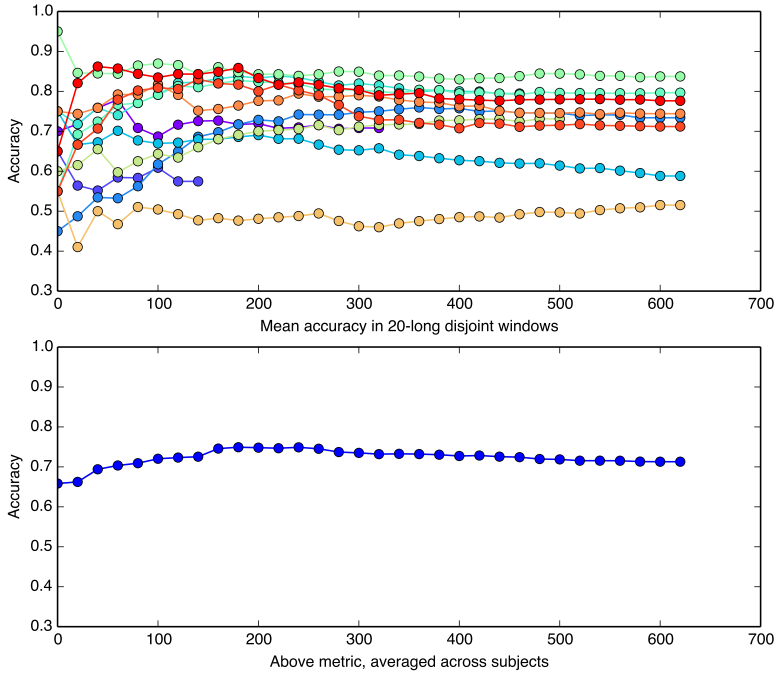
Subjects 12 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.



(a) Accuracy against sample length. Longer samples are harder to match.

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

Figure 4.3: 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 4.4: Experiment 1: during approximately the first 150 trials, subjects learn to perform this task.

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.

Factors Sample length (sL) was one of (1, 3, 6, 12) frames, equivalently (0.02, 0.06, 0.12, 0.24) seconds.

Test length was one of (15,20,40) frames, equivalently (0.3,0.4,0.8) seconds.

There were 3*4 = 12 conditions. We presented a total of 480 trials (40 trials per condition).

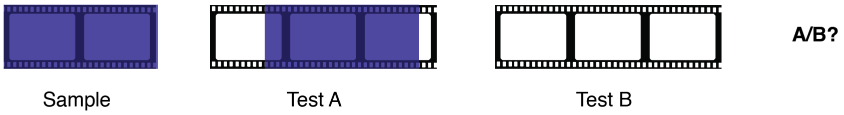
Block structure 24 training trials were presented first.

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

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

	Experiment 2
Design	2AFC delayed match-to-sample (sample clip followed by two test clips)
Stimuli	1000-frame corpus
Factors	sample length: (0.02, 0.06, 0.12, 0.24) seconds or (1, 3, 6, 12) frames. sample/test ratio: (0.3,0.4,0.8) seconds or (15,20,40) frames.
Block design	Test length varied across blocks Sample length varied within blocks 12 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.

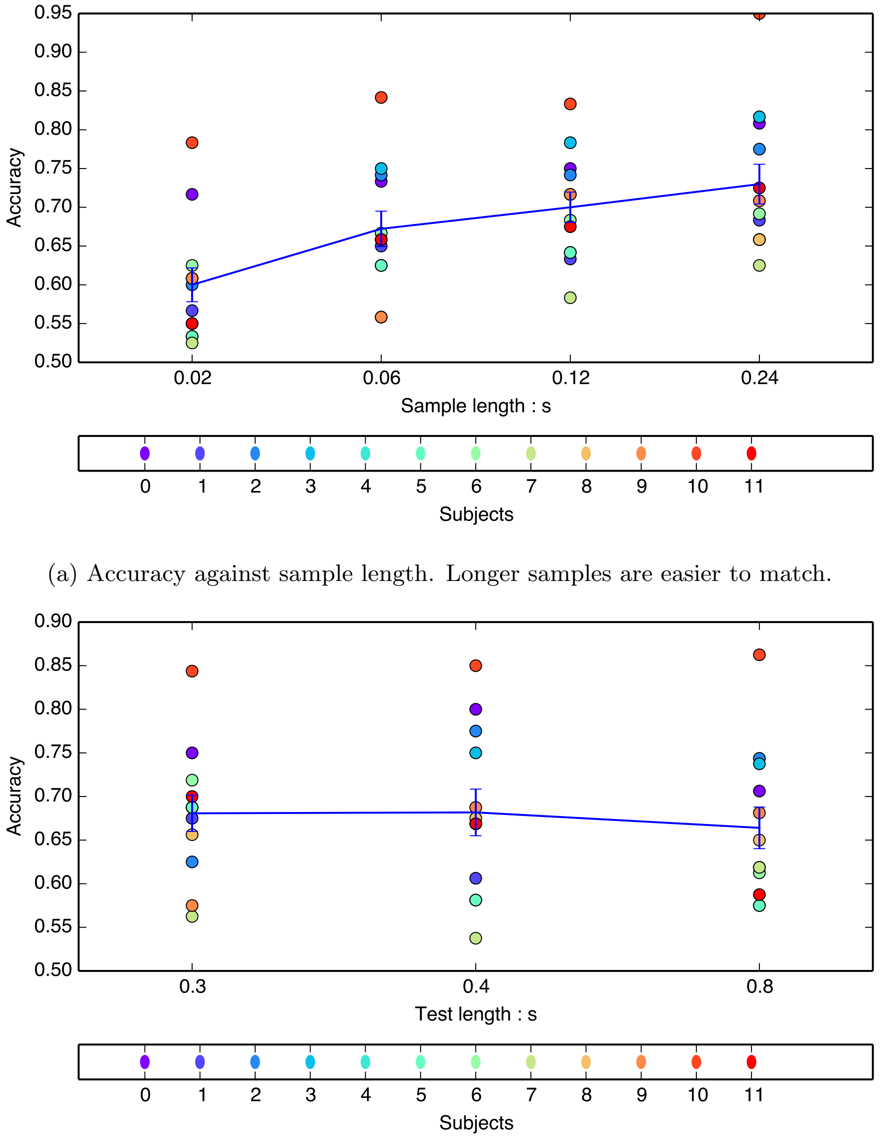
Figure 4.5: Experiment 2: design summary and trial structure.

4.6.2 Results

Sample length Accuracy increased significantly with the length of the sample, and *t*-tests showed that it was always above chance across different sample lengths. Two-way repeated-measures ANOVA revealed a significant effect of sample length (p <0.00001), but not of test length (p =0.652), and no interaction between the two (p <0.395). Mean accuracies by target length were:

Learning As shown in Fig. 4.11b, there is no discernible effect of learning during this task: subjects' accuracy did not improve during the first 100 trials, as in Experiment 1. It is tempting to explain this result by the fact that Experiment 1 kept sample lengths constant within blocks, allowing subjects to tune their mental set to a particular sample length. However, Experiment 1 used 3 blocks of 200 trials each, and we note no drop in accuracy after the end of the first or second block. We thus conclude that the lack of learning is related to the shorter sample lengths.

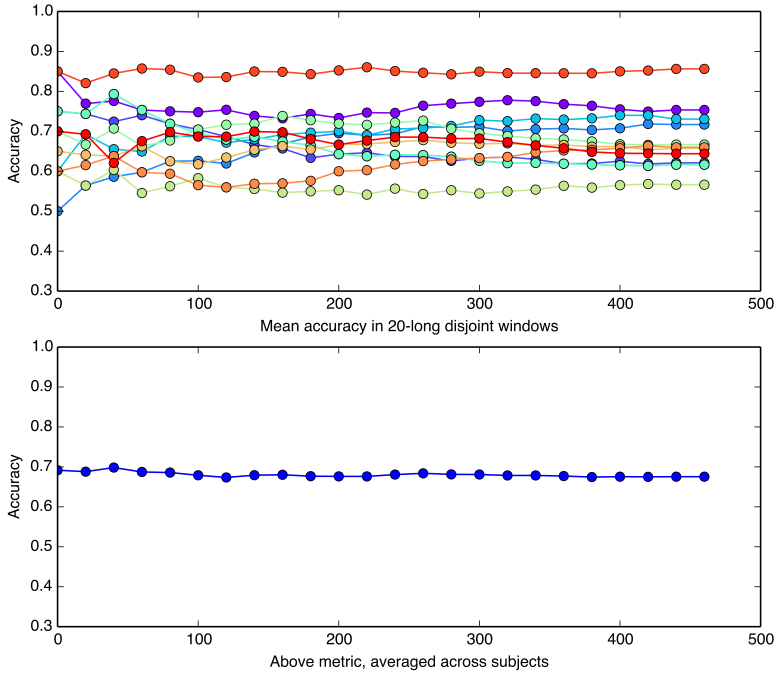
Sample length	Mean accuracy
0.02 s	0.6
0.06 s	0.67
0.12 s	0.70
0.24 s	0.73



(a) Accuracy against sample length. Longer samples are easier to match.

(b) Accuracy against test length. Search for samples under 0.24 seconds is not affected by test length.

Figure 4.6: Experiment 2: For samples under 0.24 seconds, search is more effective with longer samples; however, test length makes no difference up to 0.8 seconds.



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

Figure 4.7: Experiment 2: for short samples under 0.24 seconds, there is no discernible effect of learning.

4.6.3 Discussion

4.7 Experiment 3: Visual search in more detail

4.7.1 Methodology

Stimuli A 10,000-frame corpus of consecutive fire images was used.

Subjects 11 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Trial structure Yes-no delayed match-to-sample with altered sample. In each trial, a sample was presented first, followed by a single test. Subjects indicated whether they thought the sample corresponded to the test (up arrow) or now (down arrow).

Samples were all one second (50 frames).

Test clips consisted of the sample, surrounded by a pre-clip and a post-clip, which could both be of length zero.

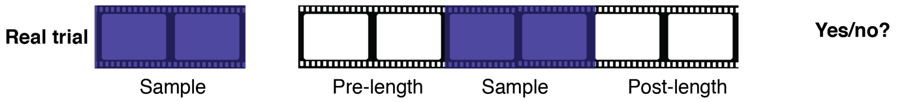
Factors The lengths of the pre-clip and the post-clip (which we term prelength and postlength) were picked from (0, 25,50,100) frames or (0, 0.5,1,2) seconds. Each factor thus had four levels, leading to 16 conditions.

Block structure We used 30 training trials with both samples and tests of length one second, and no pre-or post-clips.

The experiment was divided into 10 blocks; subjects took a break between each block. There were 40 trials per block (400 trials total). There were 16 conditions (4*4 factors) and 25 trials per condition.

	Experiment 3
Design	2AFC delayed match-to-sample, with filtered or inverted sample
Stimuli	10000-frame corpus 1 second sample (50 frames) variable-length test
Factors	prelength: (0, 25,50,100) frames or (0, 0.5,1,2) seconds postlength: (0, 25,50,100) frames or (0, 0.5,1,2) seconds
Block design	prelength and postlength varied within blocks 10 identical blocks 25 trials per condition 40 trials per block 400 trials
Training	30 training trials
Subjects	11

(a) Design summary.



(b) An altered (filtered or inverted) sample was followed by two untouched tests, one of which contained the sample.

Figure 4.8: Experiment 3: design summary and trial structure.

4.7.2 Results

Prelength and postlength Prelength and postlength could only be defined for the true trials (tests in which the subject was present), not for the foil trials (tests in which the subject was absent).

Over the true trials, a two-way repeated-measures ANOVA revealed no significant effect of prelength or postlength.

Total distractor time and test presence The pattern of accuracy in function of total time was related to whether the test was present (true trials) or absent (foil trials).

For true trials, accuracy first decreased and then increased as distractor length went up. For foil trials, accuracy smoothly decreased as distractor length went up.

In other words, as distractor length increased:

- Hits stayed high (with a slight down then up trend), and misses stayed low (with a slight up then down trend)
- Correct rejects declined, and false positives increased

As search space increased, the source of errors is mainly due to false positives, not misses. Accurate judgements were mainly due to hits, not correct rejects. Observers were good at finding a present target, but tended to confuse an absent target for a present one.

Beginning trials and end trials We define beginning trials as true trials where the sample was present at the beginning of the test, and end trials as those where the sample was present at the end of the test. In the set of beginning trials, the trial counts by level of postlength were constant; similarly, in the set of end trials, the trial counts by level of prelength were constant. This provided a useful contrast to the analysis by totaltime, where there were neccesarily vastly different numbers of trials at each level.

Signal detection analysis As this was a yes/no experiment, we analysed the accuracy pattern using the framework of signal detection theory[**?**].

First we calculated the collective d' across all subjects for each level of total distractor time. The d' curve followed the same shape as subject accuracy: first descending, then increasing as total time increased.

ROC analysis We plotted each subject in ROC space, the 2-D space defined by hit rate over false positive rate[**?**].

4.7.3 Discussion

Observers were good at finding a present target, but tended to confuse an absent target for a present one. This allows us to rule out a range of search systems: those where the choice of representation was independent of the features observed during the test.

4.8 Experiment 4: the importance of edges

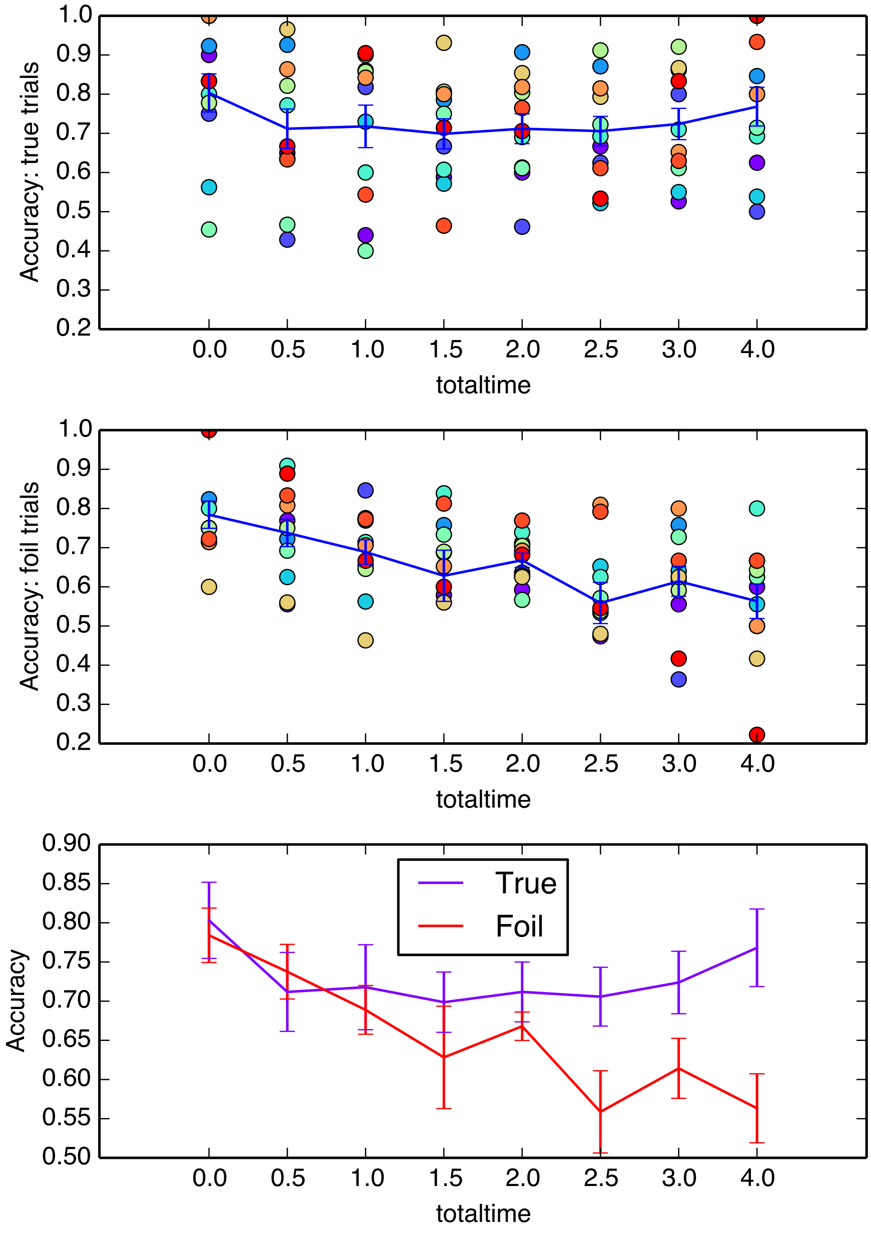
4.8.1 Methodology

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

Subjects 15 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Trial structure 2AFC Delayed match-to-sample with altered sample. In each trial, a sample was presented first, followed by two tests. A manipulation was applied to the sample; the tests were unchanged. Subjects indicated which test they thought corresponded to the sample using the left arrow (first sample) and right arrow (second sample) keys.

Sample length (sL) was one 10 frames (0.2 seconds) Test length was 15 frames (0.3 seconds)



(a) Accuracy in function of the five manipulations.

Figure 4.9: Experiment 3: Detection was above chance under all manipulations, but was too low to discern a contrast between the effects.

Factors In half the trials, the sample was manipulated using a Sobel edge filter. The test clips were left unchanged.

Each frame was convolved with the 3x3 filter

0.1250 0.2500 0.1250 0 0 0 -0.1250 -0.2500 -0.1250

to give an estimate of the image derivative. All pixels above a certain threshold value in the derivative image were returned as edges. The implementation used was MATLAB's edge() function.

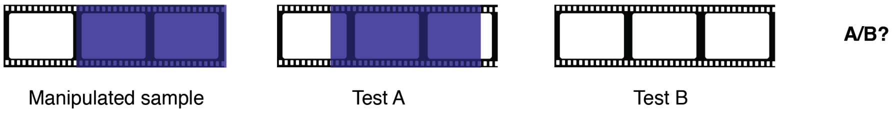
Block structure Firstly, we presented 30 training trials with static samples and tests (displayed for 0.2 and 0.3 seconds respectively), half of which used the edge-filtered sample.

Next, we presented 15 training trials with dynamic samples and tests and the same clip lengths, but with samples and tests unaltered.

There were 2 block types; we presented each block 7 times (in random order), giving a total of 14 blocks. We presented 40 trials per block, giving a total of 560 trials.

	Experiment 4
Design	2AFC delayed match-to-sample (sample clip followed by two test clips)
Stimuli	1000-frame corpus
Factors	sample edge-filtered or untouched
Block design	14 blocks, half edge-filtered 280 trials per condition 560 trials 45 training trials
Subjects	15

(a) Design summary.



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

Figure 4.10: Experiment 4: design summary and trial structure.

4.8.2 Results

Edge filtering Edge-filtering the sample induced a 4 percentage point drop in accuracy compared to the normal condition:

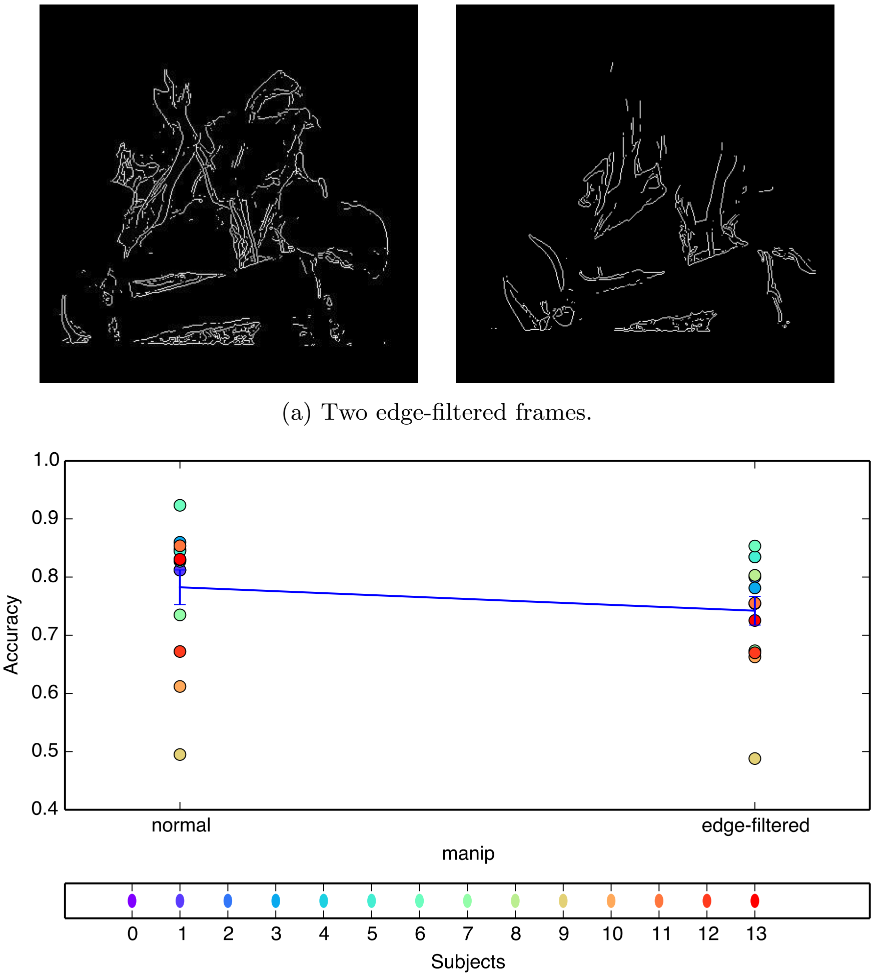
Sample	**Mean accuracy**
Normal	0.782
Edge-filtered	0.742

This difference was significant (paired-samples t-test, p¡0.005).

Order Presenting the sample as the first test (as opposed to the second test) induced an accuracy drop of 4.5 percentage points:

Test order	Mean accuracy
True test first	0.785
Foil test first	0.740

However a paired-samples *t*-test did not find this difference significant (p=0.22).



(a) Two edge-filtered frames.

(b) Edge-filtering the samples induced a 4 percentage point accuracy drop.

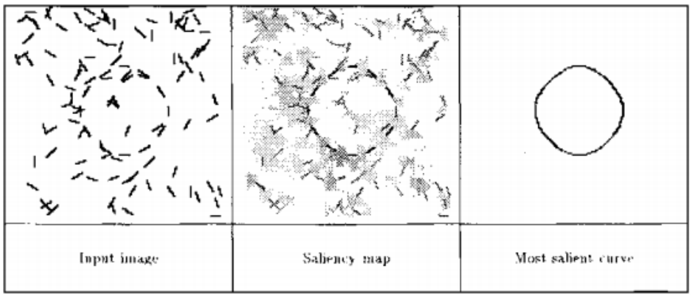
Figure 4.11: Experiment 4: detecting a sample based on its edges alone.

4.8.3 Discussion

Edges are very important in the representation of fire. We are able to match a sample clip from which all of the texture and fine contrast information has been removed, with an unaltered test, with only a small impairment from normal performance. This shows that the visual system is able to extract very useful information from flame edges alone, and employ it effectively for comparison.

This observation allows us to reject the hypothesis that fire matching is done based on the global average luminance signal, which is not preserved under edge filtering.

Edges are very important in the generation of salience maps[**?**], which guide attention to features useful for discrimination. The generation of salience percepts is not local, as evidenced by its tendency to highlight features which form global shapes:



*The circle in the first image is detectable both by your visual system and by the Shashua and Ullman model described by this figure from [**?**].*

However, the edges in this experiment were generated purely locally, by convolving each frame with the matrix

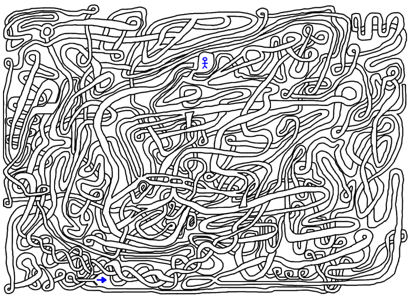
$$M = \begin{bmatrix} 0.125 & 0.25 & 0.125 \\ 0 & 0 & 0 \\ -0.125 & -0.25 & -0.125 \end{bmatrix}$$

which is only 3-by-3 and thus very local. Edge detection has often been used as an input to saliency map generators, which are more global; one small object by itself in a large image is very salient, but when surrounded by dozens of similar objects, salience is quickly lost.

Salience-map calculation is related to grouping, but is quite different. We define grouping as the process by which local features are bound together into global features, as, for example, the local edge elements in the image just above are unified into the percept of a circle. It is often assumed that grouping, like salience, can be implemented by a map: a representation composed of local pixel-like elements in which each element is tagged with the identity of the object it belongs to.

There are several problems with this idea:

- Maps make sense for scalar quantities like luminance and salience, but object identity is not scalar: how do we code *teapot* or *face* in a scalar manner?
- Maps do not make sense for hierarchical representations: one pixel of a human body can belong to *finger*, *hand*, *arm*, *upper body* and *John.* There is no room in a map to code all of this information.
- Object segmentation can be a very complex computation and is not always performed unless required. For example, in the following image, we can decide whether two points belong to the same object if asked:



but this computation is not performed in an automatic manner.

4.9 Experiment 5: colour, inversion and reversal

4.9.1 Methodology

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

Subjects 10 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Trial structure Delayed match-to-sample with altered sample. In each trial, a sample was presented first, followed by two tests. A manipulation was applied to the sample; the tests were unchanged. Subjects indicated which test they thought corresponded to the sample using the left arrow (first sample) and right arrow (second sample) keys.

Sample length (sL) was 50 frames (1 second) Test length was 60 frames (1.2 seconds)

Factors We varied the manipulation applied to the sample clips: none, luminance-inverted,colour-inverted, backwards, or spatially inverted.

Block structure Firstly, we presented 24 training trials

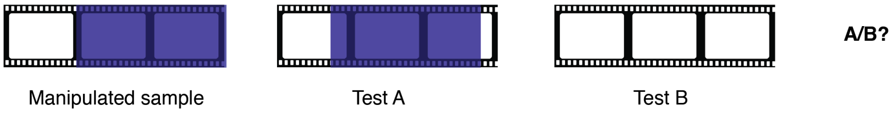
Next, we presented 30 training trials with dynamic samples and tests and the same clip lengths, but with samples and tests unaltered.

We presented 5 blocks (corresponding to each manipulation) in random order.

We used 80 trials per block, a total of 400 trials.

	Experiment 3
Design	2AFC delayed match-to-sample, with filtered or inverted sample
Stimuli	1000-frame corpus 1 second sample (50 frames) 1.2 second test (60 frames)
Factors	Manipulation: a) none b) colour inversion c) luminance inversion d) temporal inversion e) spatial inversion
Block design	Manipulation varied across blocks 5 blocks in random order 80 trials per condition 400 trials
Training	25 training trials
Subjects	10

(a) Design summary.



(b) An altered (filtered or inverted) sample was followed by two untouched tests, one of which contained the sample.

Figure 4.12: Experiment 4: design summary and trial structure.

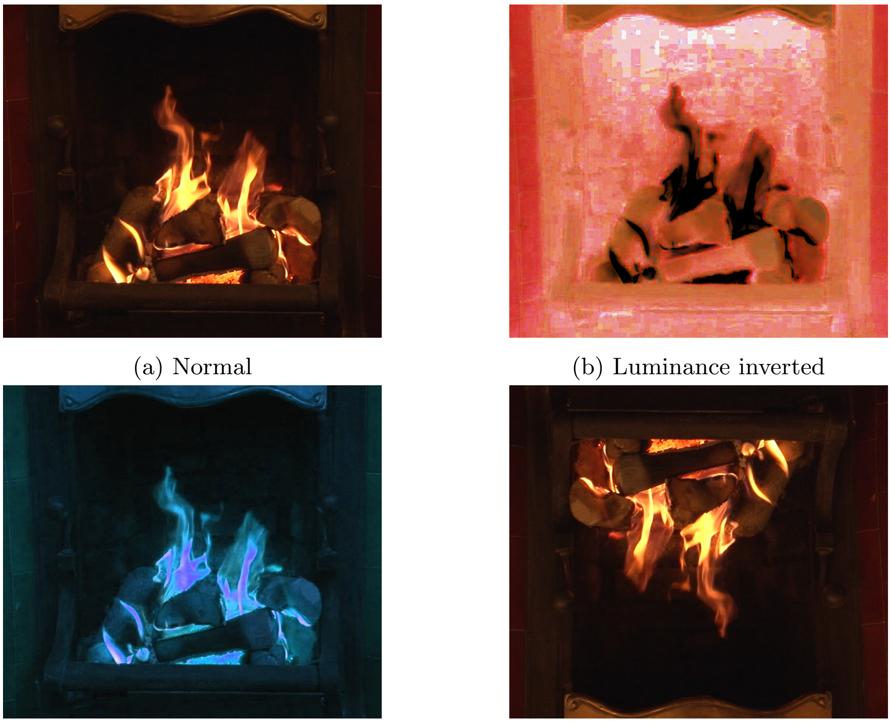
4.9.2 Results

Accuracy by manipulation We observed the following mean accuracies:

Sample manipulation	**Mean accuracy**
None	0.60
Negative	0.59
Chromatic	0.61
Reversed	0.58
Inverted	0.59

Single-sample t-tests showed that subjects were more accurate than chance (p¡0.05) under each one of these conditions. However, paired-samples t-tests showed no difference in means between any pair of conditions (p¿0.2 in each case).

Learning



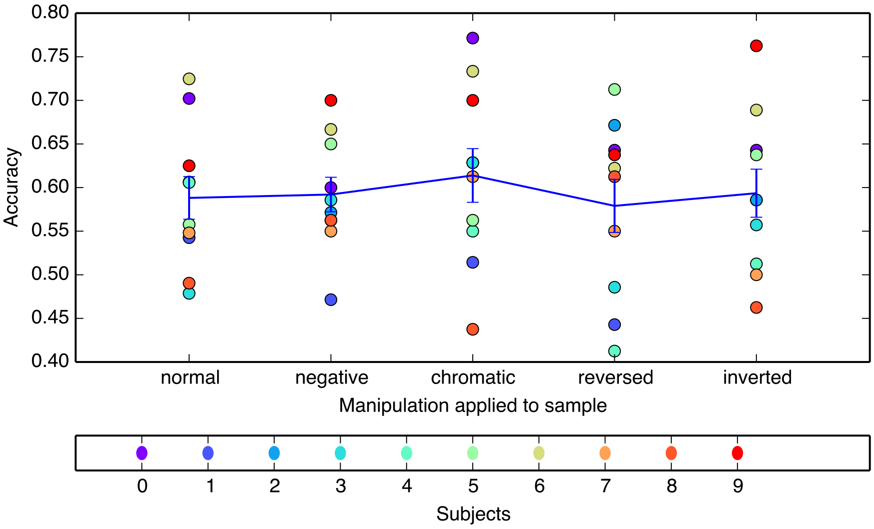
(a) Normal

(b) Luminance inverted

(c) Hue inverted

(d) Spatially inverted

Figure 4.13: Experiment 4: Examples of the manipulations we used, applied to one frame. Reversal is not shown.



(a) Accuracy in function of the five manipulations.

Figure 4.14: Experiment 3: Detection was above chance under all manipulations, but was too low to discern a contrast between the effects.

4.9.3 Discussion

4.10 Experiment 6: colour, inversion and reversal

4.10.1 Methodology

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

Subjects 8 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Trial structure Delayed match-to-sample with altered sample. In each trial, a sample was presented first, followed by two tests. A manipulation was applied to the sample; the tests were unchanged. Subjects indicated which test they thought corresponded to the sample using the left arrow (first sample) and right arrow (second sample) keys.

Sample length (sL) was one 10 frames (0.2 seconds) Test length was 15 frames (0.3 seconds)

Factors We varied the manipulation applied to the sample clips: none, colour-inverted, backwards, or spatially inverted.

Colour inversion was done by expressing each image in HSV space and rotating each pixel by 180 degrees about the hue axis.

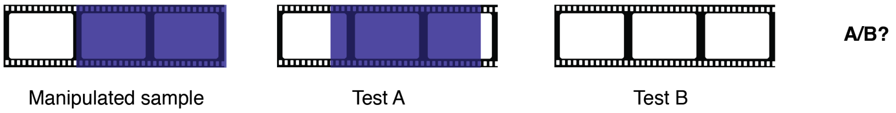
Block structure Firstly, we presented 30 training trials with static samples and tests (displayed for 0.2 and 0.3 seconds respectively) with the four manipulations applied to the sample.

Next, we presented 30 training trials with dynamic samples and tests and the same clip lengths, but with samples and tests unaltered.

We used 4 block types (corresponding to each manipulation) with 4 repetitions of each block (16 blocks total) in random order.

	Experiment 3
Design	2AFC delayed match-to-sample, with filtered or inverted sample
Stimuli	1000-frame corpus 1 second sample (50 frames) 1.2 second test (60 frames)
Factors	Manipulation: a) none b) colour inversion c) temporal inversion d) spatial inversion
Block design	Manipulation varied across blocks 4 block types 16 blocks in random order
Training	60 training trials
Subjects	8

(a) Design summary.



(b) An altered (filtered or inverted) sample was followed by two untouched tests, one of which contained the sample.

Figure 4.15: Experiment 6: design summary and trial structure.

4.10.2 Results

Accuracy by manipulation The following table shows the mean accuracies for each manipulation, as well as the paired-sample *t*-test *p*-value between that condition and the no-manipulation trials.

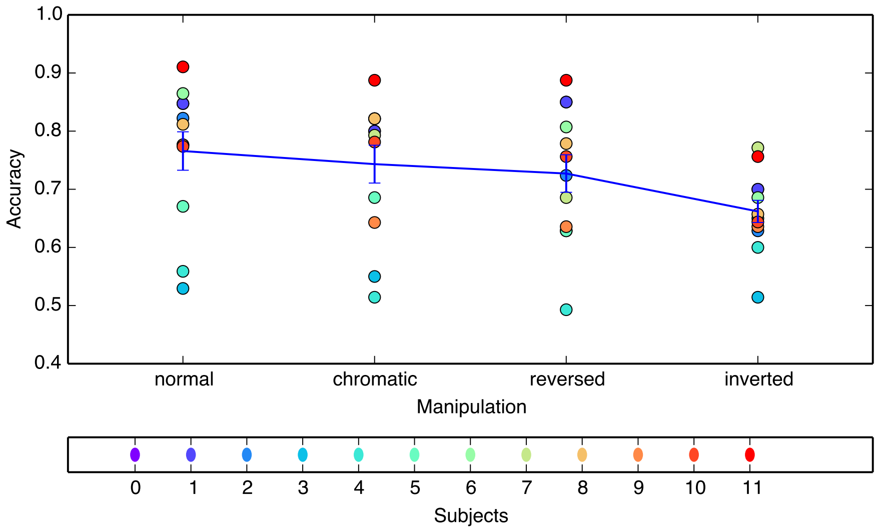
Sample manipulation	**Mean accuracy**	*p*-value from normal
None	0.766	N/A
Chromatic	0.743	0.11
Reversed	0.727	0.05
Inverted	0.662	0.00

We also note a significant difference between reversion and inversion (p=0.03, paired-samples *t*-test).

All of the conditions gave accuracy greater than chance ($p < 0.001$, single-sample *t*-tests).

4.10.3 Discussion

In terms of design, no-manipulation condition of this experiment exactly replicates the (0.2 second sample, 0.24 second test) condition of experiment 1 (p. 13). Results were



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

Figure 4.16: Experiment 6: effects of each manipulation.

also comparable: accuracy was exactly 0.766 in both cases. As completely different subjects were used, this is useful evidence that we used a large enough sample size.

Colour While hue reversal drops the mean by 2 percentage points, a paired-samples *t*-test shows a low probability of difference from the mean (p=0.11). Observers do not require the correct colour in order to match fire samples.

Reversal Reversing a video clip alters many of its simple motion properties (such as direction of motion) but not certain higher-level motion properties, such as speed. It also does not alter the position of salient motion features: if for example a salient curling flame is tracked in the upper left of the frame, its position will not have changed in the reversed stimulus. Provided that it is just as salient when played in reverse (which small flames usually are, due to their luminance), it will be easily detectable.

Here reversion is associated with a 3.9 percentage point drop in accuracy.

Inversion Inversion alters the local processing of motion features found in fire clips: it transforms upwards motion into downwards motion and leftwards into righwards. It does not, however, alter any of the temporal properties of the clip, either locally or globally.

Inverting a video clip alters the spatial location of all the features contained therein. The signature of a fire clip is not merely a bag of features; each feature is linked to its location in space ("a flare in the the upper left of the frame"). This information is disrupted by inversion.

Here inversion is associated with a 10 percentage point drop in accuracy.

Comparison The accuracy drop under inversion (10 p.p.) is much larger than that under reversal (3.9 p.p). For clips of length 0.2 s, then, spatial properties are much more important than temporal properties. Subjects are helped much more by knowing where features are in the image, than when they occur.

Subjects definitely have *some* temporal resolution during the 0.2 second period: persistence of vision blurs details together within a window of 0.04 s[**?**], a much shorter duration than the clips shown here. However, the information derived from the spatial location of features is much more important than temporal information.

Chapter 5

Temporal search for faces

5.1 Face stimuli

We also used several different portrait video clips of facial motion, recorded at 50fps. We show the frame counts here:

50fps

x3_Free 2538 xs_HCB 830 x3_TGDoY 1064 x3_TTLs 1332 x4_Free 2208 x4_HCB 814 x4_TGODoY 1002 x4_TTLS 1158 x5_Free 1830 x5_HCB 798 x5_TGODoY 859 x5_TTLS 1318 x8_Free 2937 x8_HCB 737 x8_TGODoY 958 x8_TTLS 1306

These clips consisted of either free facial motion (the subject was instructed to speak freely) or video of the subject singing along with nursery rhymes.

5.2 Experiment n: visual search for faces with short tests

5.2.1 Methodology

5.2.2 Results

5.2.3 Discussion

Chapter 6

Temporal search for faces and fire

6.1 Experiment n: Loading the perceptual system with faces and fire

6.1.1 Methodology

Stimuli A 10,000-frame corpus of consecutive fire images was used.

Subjects 8 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Trial structure Yes-no dual delayed match-to-sample. In each trial, we presented two tests (one fire clip and one face clip), followed by two samples(one fire clip and one face clip). After each test, subjects indicated whether they thought it matched the corresponding test (up arrow) or not (down arrow).

We thus derived two accuracy measures: fire accuracy and face accuracy.

Sample lengths were all 100 frames (2 s) and test lengths were all 120 frames (2.4 s).

Factors We varied the order in which samples were shown (fire first or face first). We always showed the fire test before the face test.

Block structure No training trials were used.

Chapter 7

Modelling

7.1 Forgetting functions

Chapter 8

Discussion

8.1 Hierarchies

Are the top-down/bottom-up continuum and the automatic/deliberate continuum actually the same thing?

Chapter 9

Unsorted

9.1 Characterising fire

9.2 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[?]. 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[?]; 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.

9.3 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 9.1: 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*, volume 36, 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*, 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 Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*, volume 1974. Wiley New York, 1966.

[19] Peter A Flach. The geometry of roc space: understanding machine learning metrics through roc isometrics. In *ICML*, pages 194–201, 2003.

[20] Tao D Alter and Ronen Basri. Extracting salient curves from images: An analysis of the saliency network. *International Journal of Computer Vision*, 27(1):51–69, 1998.

[21] FW Edridge-Green. Persistence of vision. *Nature*, 155:178, 1945.

[22] 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.

[23] 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