Detection of dynamic form in faces and fire

Fintan S. Nagle

September 23, 2015

A thesis submitted to the Faculty of Brain Sciences
of University College London
for the degree of Doctor of Philosophy

I, Fintan S. Nagle, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed,

Supervisors: Alan Johnston and Peter W. McOwan

50,435 words as counted by the TeXcount script at `http://app.uio.no/ifi/texcount/index.html`.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my primary supervisor, Alan Johnston, for a great deal of incisive and patient supervision and the opportunity to write this thesis.

I would also like to thank my secondary supervisor, Peter W. McOwan, for his excellent support and advice.

I am very grateful to John Raymond Towler for many fascinating discussions.

Finally, I would like to thank my parents for making all of this possible.

Please note

Throughout this document, error bars show $\pm$ 1 standard error of the mean.

Abstract

Moving natural scenes pose a challenge to the human visual system, containing diverse objects, clutter, and backgrounds. Well-known models of object recognition do not fully explain natural scene perception, ignoring segmentation or the recognition of dynamic objects. In this thesis, we use a familiar natural stimulus, moving flames, to evaluate the human visual system's ability to match and search for complex examples of dynamic form.

What can analysis in the image domain tell us about dynamic flame? Using image statistics, Fourier analysis and motion evaluation algorithms, we analysed a high-resolution dataset typical of moving flame. We characterise it as a motion-rich stimulus with an exponential power spectrum and few long-range spatial or temporal correlations.

Are observers able to effectively encode and recognise dynamic flame stimuli? What visual features play an important role in matching? To investigate, we set observers matching tasks using clips from the same dataset. Colour changes do not affect matching on short clips, but inversion and reversal do. We show that dynamic edges are a key component of flame representations.

Can observers search well for flame stimuli? Can they detect targets (short flame clips) in equally-sized longer clips? Using temporal search tasks, we show that observers' accuracy drops quickly as the search space grows; there is no pop-out. Accuracy is not so strongly affected by a blank ISI, however, showing that search difficulties, rather than representational decay, are to blame.

In conclusion, we find that the human visual system is capable of matching the complex motion patterns of dynamic flame, but finds search much harder. We find no evidence of category orientation specialisation. Combining several experimental results, we suggest that the representation of dynamic flame is neither snapshot-based nor dedicated and high-level, but relies on the encoding of sparse, local spatiotemporal features.

Contents

1 Literature review **11**

1.1 Theories of static object recognition 11

1.2 Recognising dynamic stimuli 29

1.3 Our experimental approach 39

2 General methods **42**

2.1 Recording and processing of stimuli 42

2.2 Choice of subjects 44

2.3 Experimental set-up 44

2.4 The task: delayed match-to-sample 46

3 Image domain analysis **50**

3.1 Image statistics 51

3.2 Fourier transforms 58

3.3 Motion 75

3.4 How can we encode fire? 86

3.5 General discussion 89

4 The features of dynamic flame **93**

4.1 Experiment 4.1: Feature manipulation on long clips 97

4.2 Experiment 4.2: Feature manipulation on short clips 100

4.3 Experiment 4.3: Edge filtering 104

4.4 Experiment 4.4: Can observers detect backwards playback? 108

4.5 Experiment 4.5: Motion direction percepts in dynamic flame 112

4.6 Learning 117

4.7 General discussion 118

5 Visual search for dynamic flames **123**

5.1 Experiment 5.1: Matching dynamic flame samples 132

5.2 Experiment 5.2: Matching shorter flame samples 136

5.3 Experiment 5.3: Visual search in more detail 139

5.4 Experiment 5.4: Position dependence 144

5.5 Experiment 5.5: Memory 147

5.6 General discussion 154

6 The decision process in face matching and flame matching **159**

6.1 Pilot study: Delayed match-to-sample on faces 161

6.2 Experiment 6.1: Parallel loading with faces and fire 163

6.3 Do observers show an inversion effect for dynamic flame? 168

6.4 Experiment 6.2: Looking for a pure inversion effect in dynamic flame . 171

6.5 Learning 174

6.6 Intertrial dependence 174

6.7 General discussion 179

7 General discussion and conclusions **185**

7.1 Summary of results 185

7.2 How is dynamic flame represented? 189

7.3 How is dynamic flame matched? 190

7.4 Contribution to knowledge 192

7.5 Aesthetics 193

7.6 Further work 193

7.7 Conclusions 194

Bibliography **196**

Appendices **208**

A Video data on CD **208**

List of Experiments

No.	Protocol	Investigates
4.1	Matching, 2AFC	Feature manipulations
4.2	Matching, 2AFC	Feature manipulations
4.3	Matching, 2AFC	Matching edge-filtered and normal clips
4.4	Playback direction detection	Flame category representations
4.5	Motion direction adjustment	Motion direction percepts
5.1	Temporal search, 2AFC	Long clips
5.2	Temporal search, 2AFC	Short clips
5.3	Temporal search, Yes/No	Pre- and post-lengths
5.4	Temporal search, Yes/No	Search: position dependence
5.5	Matching	Matching and memory
Pilot	Matching, 2AFC	Dynamic face matching
6.1	Matching, Yes/No	Parallel face and fire matching
6.2	Matching, 2AFC	Flame inversion effect

List of Figures

1.1 Hierarchical models of object recognition. 14
1.2 The challenges of visual search in a continuous search space. 26
1.3 An illustration of the implausibility of automatic total segmentation. . . 27

2.1 Fire dataset. 43
2.2 Face dataset. 45
2.3 Trial structure in temporal search 48

3.1 Four randomly chosen images from our dynamic flame dataset. Frame rate is 50 Hz. 52
3.2 Four sequential images from our dynamic flame dataset. 53
3.3 Image similarity measures over time. 55
3.4 Mean flame images. 56
3.5 Variance images of the flame dataset. 57
3.6 Spectrum of the mean luminance signal. 60
3.7 The 1D brightness signal spectrum in semi-log-y space. 61
3.8 Power spectra from 3 individual frames. 62
3.9 The mean of 5000 individual-frame power spectra. 63
3.10 Power spectra of three individual frames. 64
3.11 The power of each pixel at specific frequencies. 66
3.12 An individual pixel spectrum. 67
3.13 Individual pixel spectra. 68
3.14 Line fits on pixel spectra in semi-log-y space. 69
3.15 Slicing a 3D power spectrum. 70
3.16 t-slices of the image stack. 72
3.17 x-slices of the image stack. 73

3.18 y-slices of the image stack. 74
3.19 t-slices of the Gaussian-windowed image stack spectrum. 76
3.20 Individual flow fields shown by the McGM. 79
3.21 Mean flow fields shown by the McGM and Sun's method. 80
3.22 Circular histograms of the directions of motion estimated by the McGM and Sun's method. 81
3.23 A flow field produced by the sMcGM from a dynamic flame image stack. 83
3.24 Mean flow fields from the sMcGM; normal and edge-filtered data. . . . 85
3.25 The results of naive PCA on a dataset of monochrome flame images. . 88
3.26 Dynamic texture synthesis of flame. 90

4.1 Experiment 4.1: results. 99
4.2 Experiment 4.2: results. 101
4.3 Experiment 4.3: results. 106
4.4 Experiment 4.4: results. 110
4.5 Experiment 4.5: results 114
4.6 Experiment 4.5: individual results 115
4.7 Learning effects in Chapter 4. 119

5.1 Visual search on a continuous search space. 124
5.2 Experiment 5.1: results. 132
5.3 Experiment 5.1: accuracy depends on test/sample ratio rather than test length. 133
5.4 Experiment 5.2: results. 137
5.5 Experiment 5.3: results. 141
5.6 Experiment 5.3: measures of observer bias. 142
5.7 Experiment 5.4: results. 145
5.8 Experiment 5.5: results. 148
5.9 Experiment 5.5: signal detection measures. 149
5.10 Learning effects in Chapter 5. 153

6.1 Face matching pilot study: results. 162
6.2 Experiment 6.1: results. 164
6.3 Experiment 6.2: results. 171

6.4 Experiment 4.4: the effect of stimulus angle. 172
6.5 Learning effects in Chapter 6. 175
6.6 Intertrial dependence: Chapter 4 180
6.7 Intertrial dependence: Chapter 5 181
6.8 Intertrial dependence: Chapter 6 182

List of Tables

4.1 Experiment 4.1: results. 98
4.2 Experiment 4.2: results. 102
4.3 Experiment 4.3: results. 107
4.4 Experiment 4.5: results. 115
4.5 Learning slopes in Chapter 4. 118

5.1 Experiment 5.1: clip lengths. 134
5.2 Experiment 5.2: sample lengths and accuracies. 138
5.3 Experiment 5.5: Mean accuracies (%) by ISI and sample length. 150
5.4 Learning slopes in Chapter 5. 152

6.1 Face matching : sample lengths. 161
6.2 Experiment 6.2: correlations across trials. 167
6.3 Learning slopes from Chapter 6. 174

Chapter 1

Literature review

The human visual system is capable of effectively perceiving and comparing an enormous variety of stimuli. From coloured squares to landscape panoramas, from static form to dynamically moving faces, we are able to represent the natural world as patterns of neural activity and use these encodings to perform useful tasks.

The study of object recognition, a central visual skill, began with behavioural experiments on static images. As a result, most early theories of object recognition did not consider time. In recent years, research on dynamic faces has spurred efforts to modify or replace these models in order to explain how we encode and search for changing stimuli.

We begin this review with a survey of static theories of object recognition, looking in depth at the cases of face recognition and natural scene perception. We then introduce the dynamic case, examining work on dynamic faces and biological motion. We conclude by highlighting the problems that dynamic natural scenes pose for current theories of object recognition.

1.1 Theories of static object recognition

Recognising objects in the environment is vital for survival; a large amount of visual cortex is involved in this task[1]. Object recognition poses several specific challenges for the visual system.

The segmentation problem[2]. Cluttered natural scenes contain a profusion of objects and background distractors, and attended objects must be separated from the

scene.

The invariance problem. Objects should be recognised effectively under a wide variety of angles, retinal positions, distances, lighting directions, ambient light spectra, and structural changes. No two views of an object ever produce the same retinal image; even when matching two objects whose pixels are identical on a computer screen, the brain will receive slightly different signals. If our task is classification rather than identification, more variance is tolerated.

The binding problem. Von der Malsburg[3] pointed out that low-level feature detectors for entities such as corners and edges need to be bound to a particular high-level object, lest mis-binding generate hallucinations or illusory percepts[4]. He suggested that neural synchrony supports binding, although other mechanisms, such as sequential attention, have been proposed[2]. The binding problem is similar to the segmentation problem, but requires that features are bound to the correct object, not just separated from the background.

Despite these challenges, the visual system can detect a target in a cluttered visual scene in about 150 milliseconds[5]. This places a limit on the amount of computation which can take place - in particular, on the number of synapses which can fire in sequence before activation must pass to motor cortex so that the observer can respond. For this reason, most models of object recognition are either single-step or feedforward, without recurrent connections.

The simplest way to recognise an object is to store a holistic representation which is similar to its retinal image: a template. Templates are holistic, meaning that they represent an entire object and cannot be decomposed into smaller parts. Theories involving them are common[6, 7, 8, 9] and provide a convincing explanation for detection, but do not sufficiently address the invariance problem. An object which has changed slightly in shape, angle, or lighting condition will no longer match its template, and will not be recognised. Computers use template models to store images: a digital image is a holistic list of pixels with no innate substructure. Each pixel is represented on an equal basis, and external algorithms (such as edge detection or clustering) are needed to impose structure. This is why digital image comparison is so trivial (requiring simple pixel comparison) and cripplingly vulnerable to the invariance problem: slight translation of an object between two compared images is enough to

defeat matching by low-level, local mechanisms like pixelwise comparison.

It is important to note that template matching models do not suggest that the image is processed holistically from the start: retinal representations, optic nerve representations, and V1 representations are highly local. A chain of progressively less local representations exists as we move through the visual processing stream. Template matching models do not deny the existence of these representations, but say that they are unstructured, inaccessible, and not useful for object recognition. The great vulnerability of template matching models to invariance motivates theories that have more than one processing stage.

Most models of object recognition are hierarchical: their components are arranged in a tree-like as opposed to linear way. One of the first models of object recognition, Selfridge's Pandemonium[10], shown in Fig. 1.1, posited that each object is recognised by a tree of "demons." Each one looks out for specific features and alerts higher demons when it spots them. The many demons on the bottom layer of the tree look out for simple features in the world; those in the intermediate layers look out for conjunctions of features spotted in the bottom layer. The top-layer demons signal the presence of complete objects.

This basic layout, a tree of feature detectors, becoming less numerous and more high-level as we proceed up the feature hierarchy, is common to many models of object recognition. Pandemonium was proposed in 1956. Riesenhuber and Poggio's HMAX[12], published in 1999, is a fully-implemented computational model working along very similar lines. Its hierarchical detectors are wired together using two operations: linear weighted summing, which allows up-layer cells to respond to conjunctions or disjunctions of down-layer features; and the nonlinear MAX operation, by which a cell responds to the most active of its afferents.

HMAX deals with the problem of multiple views by having the top layer combine the output of cells in the highest subordinate layer that respond to a single view of an object (view-tuned units). View invariance is thus put off until the final processing stage. This is not the case for all models: Biederman's geons theory, which we examine shortly, proposes that 3D sub-shapes are extracted locally and independently in a view-invariant fashion.

The idea of a hierarchical object detector has had many implementations. One

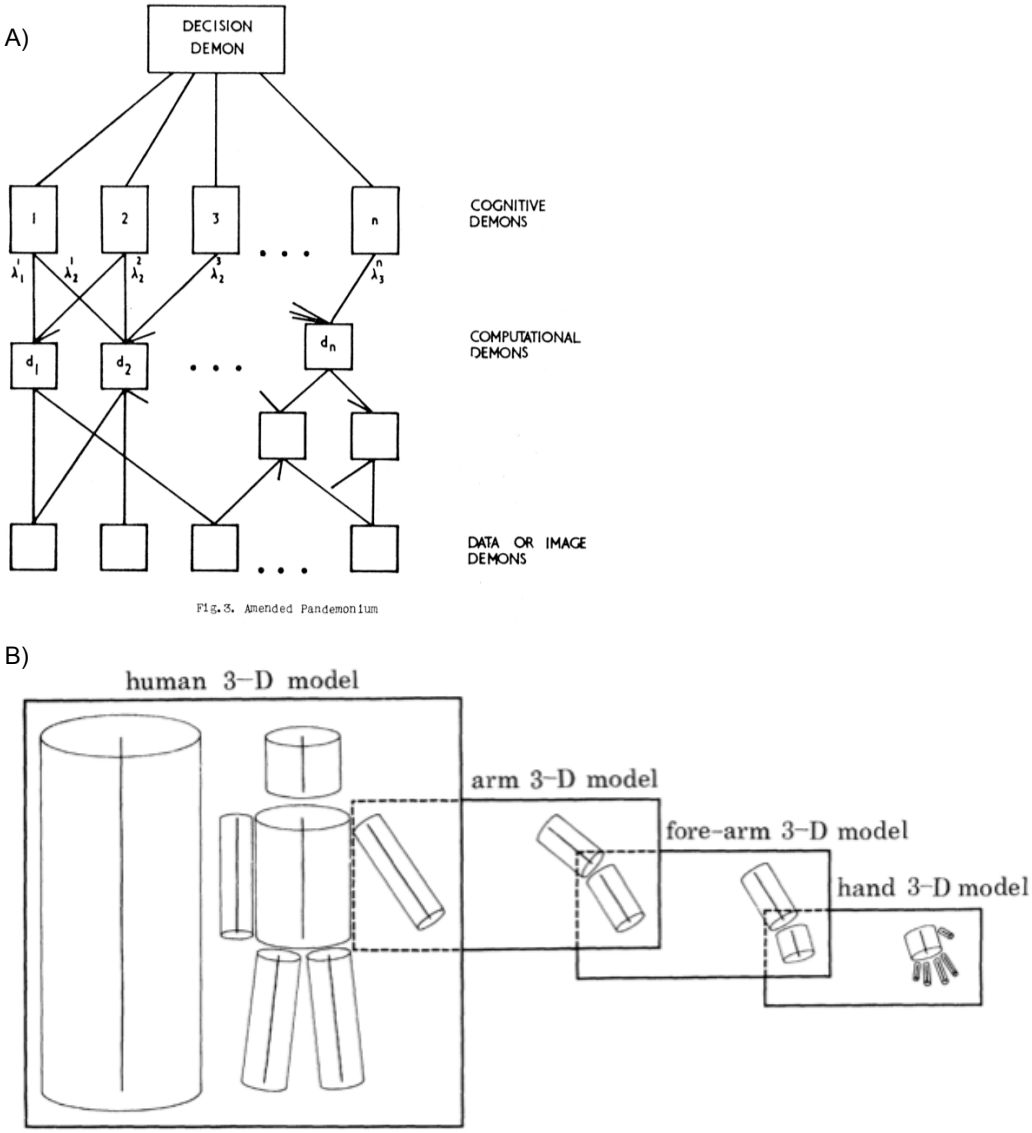


Figure 1.1: Hierarchical models of object recognition. A) Selfridge's Pandemonium[10]. B) Marr and Nishihara's structural description model[11].

of the first was Fukushima's neocognitron[13], composed of stacked layers of cells and initially aimed at recognising handwritten letters. The lowest layer corresponds directly to the input pattern; successive layers contain reduced numbers of cells, the final layer consisting of just ten and thus indicating the network's classification of an input numeral. The neocognitron is invariant to stimulus position and size because of its so-called C-cells, which respond to a group of lower-level cells detecting variously transformed versions of a feature. It can learn either in an unsupervised way (in which case it self-organises to discriminate patterns based on similarity) or an unsupervised way (in which case it attaches the correct labels to classified patterns).

Marr and Nishihara[11] suggested a hierarchical representation, using mammalian body parts as an example (See Fig. 1.1). They proposed that view invariance was delivered by a representation describing the relative location of object sub-parts described as generalised cones. Its hierarchical nature would allow objects to be compared on several different scales, allowing similar objects to be grouped together (stability) but very slightly different objects to be differentiated (sensitivity). For example, this model allows a creature's number of legs to be counted, classing it as a biped or quadruped, without requiring the precise leg morphology.

The design of artificial object recognition systems has always been motivated by neurophysiological evidence. The difference between simple and complex cells, as elicited by Hubel and Wiesel[14], is often summed up as a slight increase in the receptive field sizes of complex cells, along with the addition of position invariance. Many models of object recognition have borrowed this nomenclature, calling plain feature detectors S-cells and position-invariant feature detectors C-cells. A second detail copied from neuroanatomy is the organisation of cortex into columns and hypercolumns[15]. The neocognitron borrowed the term "hypercolumn," applying it to a set of cells that recognise different features from a similar location in the input pattern.

One way to classify models of object recognition is by the types of invariance they possess. The neocognitron, dealing with 2D input, can only be invariant to size and position; for successful detection, stimuli could not be rotated. It is of value for the human brain to be invariant to these transformations, as well as to changes in lighting, occlusion and surface texture. It can also benefit from invariance to changes in shape: it is still possible to recognise an object if it is slightly deformed, and many object labels

(such as "horse") apply to an enormous set of object instances of slightly different shapes.

The problem of invariance to rotation motivated Biederman's recognition-by-components (RBC) model[16], a structural description model descended from Marr and Nishihara's. This system imposes a 3D description on the world from an early stage, segmenting an object into characteristic sub-shapes called geons. Next, a description of which geons are present and how they fit together is produced; this is invariant to 3D rotation. RBC differs from Marr and Nishihara's model in that its library of shapes is more restricted: geons are described precisely in terms of lower-level features, and they are a subset of generalised cones[17]. The key difference between this model and HMAX is the order in which invariances are built up. HMAX builds local descriptions in 2D and then combines them into view-invariant descriptions: its view-invariance comes last. The geon model builds local descriptions which are already invariant to rotation, then combines them into full object descriptions.

The Marr and Nishihara model, as well as RBC, predicted that structural description would enable full viewpoint invariance, allowing objects to be recognised with equal speed whatever their viewpoint. Shepherd and Metzler[18], as well as Jolicoeur[19] proved otherwise, demonstrating recognition delays for unfamiliar views. Tarr and Pinker explored this effect psychophysically, attempting to differentiate between models which store multiple views and models which store a single view and then transform it, with mixed success[20]. They thus proposed that shared mechanisms cooperate in the object recognition system. Further evidence from fMRI was found by Gauthier *et al*[21], who showed dorsal (where) pathway activation during mental rotation paths and ventral (what) pathway activation during object recognition tasks.

There is mathematical support for leaving view invariance until the last moment: the structure of a 3D object can be approximated by a small number of 2D projections, as long as they are orthographic[22]. There is also psychophysical support from the repeated finding that objects are matched less well when viewed under a different orientation than that presented during learning, as tested by Bülthoff *et al*[22]. Specifically, they found support for the idea that novel views are interpolated from existing views, finding more accurate recognition for novel views between learned views (as opposed to outside learned views).

Poggio and Bülthoff [23] provided supporting evidence in the form of a computational model using radial basis functions to match novel views to learned views without mental rotation. Another approach, suggested by Ullman[9] is that of storing a model of each recognisable object and deforming it to match a newly viewed object. A viewed object V is matched with a stored object M and a transformation which brings M into correspondence with V. This process is necessarily preceded by selection of an object from the visual scene and its segmentation to produce an outline. A key feature of this recognition scheme is that it is performed in space: rather than relying on image statistics or computed features, the components of the model are actually shapes. This illustrates another judgement we may make about models: what should their primitives look like? What is the nature and complexity of their internal representations?

How do we evaluate the usefulness and power of competing models? Marr and Nishihara suggest some solutions[11] - their "criteria for judging the effectiveness of a shape representation." By a representation they mean a model of shape recognition: "a representation for shape is a formal scheme [an algorithm] for describing shape or some aspects of shape, together with rules [an algorithm] that specify how the scheme is applied to any particular shape."

Marr's representation is not simply a static description (a piece of data) but rather a set of rules (a model) for creating a representation. One could argue that a model of shape representation is not a model of object recognition, because a model of recognition outputs a classification (a linguistic tag). However, it is generally accepted that models of object recognition need not include the full machinery necessary to generate an object's name or send the motor commands necessary to write it: they must simply describe the building of a representation which is more easily classifiable than the retinal image.

Marr's criteria are:

- **Accessibility:** can the desired description be computed from an image, and can it be done reasonably inexpensively?
- **Scope:** to what class of shapes is the representation applicable?
- **Uniquenesss:** do shapes have canonical descriptions in the representation?
- **Stability:** a representation remains the same when its retinal image changes.
- **Sensitivity:** a representation changes when its retinal image changes.

Marr points to a tradeoff between stability and sensitivity: in order to obtain more of one, we must have less of the other. He gives the example of representing animal bodies using skeletons made from identically-sized sticks. If one long stick is used, we can represent only the animal's overall size and orientation. Using a larger number of shorter sticks brings more detail, allowing us to discriminate different species. Using even shorter sticks allows us to discriminate individuals of the same species.

Additional evidence from neural recordings has emerged since Marr's day. Gross and Bender[24] were the first to locate neurons (in the macaque) which showed selectivity to shape, size and colour. IT neurons were more sensitive to complex properties, while V1/V2 neurons selected for simpler properties. This supported the idea that complex features are gradually built up from more local descriptions. Rust and DiCarlo[25] deal with selectivity (equivalent to Marr's sensitivity) and tolerance (equivalent to Marr's stability). They set out to test the idea that selectivity and tolerance both increase as signals propagate through the ventral visual stream (from retina to LGN, through V1, V2 and V4, to IT). Macaques performed an object-detection task and recordings were made from 140 neurons in IT and 140 neurons in V4. A support vector machine (SVM) classifier was trained to detect objects from the population recordings. For stimuli in which the same object was presented at different sizes and in different positions, classifier performance on V4 dropped but performance on IT did not change; the IT population thus showed more stability. Selectivity was first measured by comparing classifier accuracy on each population; it was unchanged. However, IT neurons' responses were more linearly separable than those of V4 neurons. The populations were equally sensitive, but the IT population required an easier computation to decode the result.

Overall, evidence from single-neuron recordings does not offer much help in discriminating between models of object recognition. For example, mapping the combinations of features to which neurons in IT are sensitive[26] reassures us that these features are represented in the brain at the single-neuron level, but the same features are decodable from V1 at the population level. Since high-level models of object recognition do not specify whether their machinery operates at the single-neuron or the population level, this insight does not have much discriminatory power. Some observations, however, can be useful. The finding that view-dependent cells respond faster than view-invariant

cells[27] supports models in which views are represented first, followed by objects. Logothetis *et al*[28] trained monkeys to recognise shapes (natural and artificial, including monkey faces) from different angles; the monkeys eventually achieved view invariance. Considering the population of IT neurons, many (10%) were view-selective, but few (1%) were view-invariant. Booth and Rolls[29] found the same pattern in monkeys which had learned and manipulated real objects instead of images, helping to rule out overtraining and maintain ecological validity. Together, this evidence convincingly supports view-tuned recognition.

Single-unit recordings can provide only so much information. Real neurons have thousands of afferents, meaning that they can decipher population codes. Kobatake *et al*[30] recorded a total of 131 neurons from two monkeys, finding that training increased population selectivity. Tanaka *et al*[31] investigated further, smoothly altering the properties of stimuli shown to anaesthetised monkeys to find geometrical shapes which activate particular neurons as strongly as possible. These inferotemporal neurons were most strongly activated by quite complex shapes. When combined into a population code, they have even more expressive power. Variations were smaller along perpendicular electrode tracks, suggesting the existence of feature-selective minicolumns. Neurons and columns are anatomical analogies for mid-level detectors in general hierarchical recognition models such as HMAX. This view is supported by optical recordings of macaque cortex performed by Tsunoda *et al*[32], which are consistent with the representation of objects by combinations of feature columns.

After conducting an extensive study of visual agnosics, Farah[33] proposed that the visual system possesses two recognition systems: part-based and holistic. She exploits numerous patient dissociations to find support for mid-level object representations, which she calls "psychologically real parts" to differentiate them from parts which are physically real but not usually represented by the brain[34]. These dissociations show convincingly that we have not one universal object recognition system but many, specialised at the very least for faces and words. She posits a part-based recogniser (used mainly for words and not at all for faces) and a holistic recogniser (used mainly for faces and not at all for words). These recognisers are both used for general object recognition (neither faces nor words).

Evidence from the time course of perceptual processing can rule out particular

model classes. The fact that humans and macaques can deliver object classifications within 300 milliseconds[5] limits the number of synapses over which a signal from the retina may pass before reaching motor cortex. This has been interpreted to mean that object recognition is a largely feedforward process[35] because there is not time for signals to bounce back and forth between low-level areas and high-level areas. This rules out models where many alternatives are tried in succession - but not models where many alternatives are tried in parallel. However, most models of object recognition are so high-level that they do not specify exactly how their rules may be implemented on the neural level.

This difference in levels of description illustrates another dimension along which models can be judged: their level of abstraction. The original Pandemonium, for example, is a very high-level model consisting of verbal descriptions of shouting demons which represent the magnitudes of a template match signal. HMAX is more complex, possessing an algorithm which is precisely described in an actual Matlab implementation as well as in a high-level description. This multilevel description is what makes it useful: models with only low-level descriptions (such as convolutional neural networks, the state of the art in computational object recognition[36], and other neural networks like deep Boltzmann machines[37]) can perform recognition, but they do not help us understand how it is done. Recognition is trained into the network rather than designed in. Once we have trained a convolutional network, it can perform the task, but it has no high-level description or algorithm. It is described on Marr's physical level, but not on the algorithmic level. In order to find out more about how the network operates, we need to study it exactly as we do the brain.

Marr described exactly this difference in levels of description with his well-known levels of analysis: the computational level (what the system does), the algorithmic level (how the system does what it does) and the implementational/physical level (how the system is physically realised). There are many such levels: we can describe brain function in terms of the movements, of atoms, molecules, neurons and spikes, neurons and spike rates, cortical columns, or large brain areas. We can describe the function of a computer in terms of individual electrons, transistors, gates, chips, and functions. In both cases, there is a hierarchy of levels of representation.

This hierarchy need not be composed of nameable parts[38]. In faces, we have a

rich vocabulary of features such as eyes, mouth and nose. Face recognition systems, however, do not start by localising these features; they use internal features without names, such as Gabor patches[39] or jets[40]. A hierarchy of features does not imply a corresponding hierarchy of words.

So far, the models we have examined only deal with recognition: the situation in which an object appears in the visual field and is recognised according to an existing definition. In reality, object categories are plastic: they are built up gradually during development[41] or familiarisation[42]. Wallis[43] proposed a simple self-organising network model of object recognition which learns its representations from data. Taking 256×256 images as input, it performs edge detection and then feeds into two sheets of neurons. These layers are trained sequentially using Hebbian learning ("cells that fire together wire together"[44]) with lateral inhibition. This model is very simple, admitting that it does not accommodate position or view invariance (or even 3D objects) and it also ignores important results about early vision, replacing simple and complex cells by a trivial Laplacian-of-Gaussian edge detector.

Most models of object perception are general: they apply to the recognition of all objects. However there is one specific object class that has received much attention in the literature: faces.

1.1.1 Face perception

Humans and other primates use a complex arrangement of facial muscles to send signals to their conspecifics. Because of their skill in face identification and evaluation, there is a large amount of experimental and theoretical work on the recognition of faces compared to other objects.

The capabilities of human face recognition highlight those of object recognition in general. Faces can be easily recognised even in images of low spatial resolution: humans can identify blurred versions of 7×10 pixel faces at 50% accuracy[45]. Indeed, the spatial frequency band which carries the most information about faces is low, about ten cycles per face[46].

Face perception is very vulnerable to high-level after-effects, such as adaptation to age or gender. These are robust to low-level changes such as size[47], retinal position, and rotation[48], showing that they correspond to the computation of high-

level invariants. Dimensions along which adaptation is strong are often considered to form a high-level "face space"[49].

A surge in face perception research was initially motivated by the observation that eyewitness reports of face identity were very fallible[50]. A series of influential models were created to explain the burgeoning experimental results. Bruce[51] interprets them according to Marr's three levels[52]. She points out that many purely computational models are not hierarchical; they operate in a single step, transforming input data (the analogue of the retinal image) to a high-level description (the analogue of the higher representations in a hierarchical object recognition model). Conversely, she points out that once we reach the algorithmic level, we find hierarchical theories such as that due to Baron[53].

Extensive experimental results on face perception allow us to better understand the nature of the feature hierarchies featured in models of object recognition. In early models such as Pandemonium, next-level-up demons only responded to the presence or absence of conjunctions of features. Human face processing is configural: it responds not only to the presence of certain features, but to their configuration. The top half of a face is often recognisable on its own; when combined with the bottom half of another face, accuracy drops as the foil half is unavoidably included in the configural perception process[54]. This type of processing means that in any face recognition model consisting of hierarchically arranged units, higher-level units must have access to detailed information from lower-level detectors, not just a simple "I have detected this feature!" signal.

The well-known model of face recognition by principal component analysis (PCA) is "flat" in that it processes an entire image in one step. Turk and Pentland's initial application of PCA to face recognition, for example, moves directly from an image to a high-level principal component space[55]. Face images (around 100,000 numbers) are projected into a high-dimensional space (as low as 12 numbers) according to a matrix found by exploiting correlations in a large face dataset. Support vector machines (SVMs), another method obtaining among the highest computational face recognition performance[56] are also non-hierarchical. Deep convolutional neural networks, which have been successfully applied to face recognition[57], are hierarchical in that they possess stacked layers of filters. The features within these layers are not defined,

however; they are left to self-organise during training, and so a trained deep network must be studied to discover the features it detects and passes on to higher layers. Moreno *et al*[58] compared a local descriptor based on the scale-invariant feature transform (SIFT) and a hierarchical object recognition algorithm (Poggio's HMAX) on a face detection and localisation task using the Caltech faces database. HMAX was found to out-perform SIFT, suggesting that faces are better suited to hierarchical description.

Despite a scarcity of hierarchical models specific to faces, we can use evidence from neural recordings to evaluate how well general object recognition models apply to faces. In the macaque, view-tuned neurons seem to precede view-invariant ones[59]. As with general object recognition, view-invariance seems to appear last. What of selectivity; the property of responding to only one object? In the macaque, temporal lobe neurons rarely respond to individual faces[60], indicating that face identity is coded by population in this region. Recent fMRI work[61] has revealed "face patches" in the macaque with different functional roles: there appear to be two patches of view-specific neurons, one of neurons invariant to reflection, and one patch of fully view-invariant neurons. This pattern supports the sequential build up of invariance found in models of recognition, as well as the precise ordering.

There is wide agreement that faces are processed differently than other stimuli, but much debate about the question of holistic face processing: whether faces are represented as sets of individual features, or as global wholes. Piepers[62] points out a lack of agreement about what these terms actually mean. The concept of holistic perception can be traced back to that of the gestalt, a unified whole with properties that are greater than the sum of its parts[63]. Gestalts possess emergent features which cannot be computed from the parts alone [64]; the classic example is a rectangle, which is made up of four lines, but whose area we cannot calculate unless we see the lines as a rectangle.

Detectors for eyes, noses and other facial parts are presumed to output some information about these parts: size, colour, and most importantly location. Feature location gives us access to first-order configural properties (where the parts are), thence to second-order configural properties (where the parts are in this face compared to in other faces). Sensitivity to configural properties seems to start at birth[65]. Other

configural models, often from the computer vision literature[66], look at the arrangement of key points or fiducial points rather than features, since it is difficult to define the centre point of a nose or eyebrow.

Some models equate holistic and configural processing[67] whereas some differentiate between them[68]. Configural processing requires detection of anatomical features or keypoints so that their configuration may be computed. Holistic processing does not require anatomical feature detection; it simply suggests that the entire face is processed at once, as with template-matching models of object recognition.

Jiang *et al*[69] approached the problem of face detection with a hierarchical, feature-based model which pools over simple, complex and then view-tuned units in the same way as HMAX. Surprisingly, it was found to demonstrate configural effects, despite the absence of any configural information or global templates. It demonstrates the inversion effect and is sensitive to morphing in a similar way to human observers. By showing that a large amount of behavioural data can be explained by a non-configural, feature-based model, Jiang's account undermines experimental support for configural processing.

A major focus of face perception research is the creation of measures of holistic processing. The earliest was the disproportionate inversion effect: in faces, inversion impairs matching accuracy more than in other objects[70]. This effect, however, could simply indicate training for faces, and can be reduced by training[71]. Another measure is the composite task effect, in which the bottom half of one face is shown aligned with the top half of another face[54]. Observers are asked to identify one half. When the other half is misaligned, their accuracy increases. The interpretation is that alignment triggers holistic face processing, and the resulting whole-face code is contaminated by the foil half of the image. The part-whole task is similar: observers are asked to judge the identity of a particular feature inside a face, and increased accuracy under inversion or face scrambling is taken as evidence of holistic processing.

The term "configural coding" creates some problems. As shown by Jiang, a configural code cannot be differentiated from a feature-based code by behaviour alone; we must use imaging to examine the representations themselves. As pointed out by Barenholtz and Tarr[72], configural behaviour does not require low-level representations which are themselves configural. Globally, configural behaviour can be delivered

by local non-configural features which are not simply pooled over but arranged in a certain pattern. Much work simply analyses the nature of these features, neglecting the fact that they are part of a neural hierarchy which can carry information about their arrangement in other ways than pooling.

Together, these definitions suggest a simple computational interpretation: we can build a configural code from several independent local feature codes by some significant computation. We can also produce a non-configural code from several independent local feature codes by an operation such as binding. A bag of features, for example, combines local features in a non-configural way: it does not keep track of their precise location. Are configural representations built by trivially combining smaller configural representations, or by combining smaller non-configural representations in a complex way? Neurophysiological research, such as Rust & DiCarlo's measurement of the transformations occurring between V1 and IT, is beginning to characterise these computations.

Neurophysiology

Face recognition tasks increase BOLD response in many areas of cortex[73]. Haxby *et al* proposed an influential distributed model of face processing[74] according to which separate cortical areas process invariant and changeable information. In this model, early facial features are constructed by the inferior occipital gyrus. Observing these, the superior temporal sulcus (STS) constructs changeable features (such as expression change or eye movement) while the lateral fusiform gyrus (otherwise known as the fusiform face area or FFA) constructs invariant features. This arrangement is a hierarchical model of face recognition: early features as processed in the occipital gyrus correspond to low-level demons in the Pandemonium model, while identity or expression percepts correspond to higher-level features. Unlike most object recognition models, the hierarchy flows upwards in two separate directions (high-level expression features and high-level identity features) above the initial level.

1.1.2 The problems posed by natural scenes

Natural scene perception is a more challenging process than object recognition. The standard task, which DiCarlo calls "core object recognition," involves basic-level clas-

sification of a single already-segmented object. Natural scenes, however, contain many different objects of varying salience which are not segmented from the background. We can therefore examine a natural scene with various degrees of intensity, ranging from a quick glance in which we do not identify the individual objects (gist extraction) to a full sequential examination of each object. When searching for a target, the search space is continuous, as shown in Fig. 1.2. We must segment the scene ourselves instead of relying on the provided segmentations of a discrete search space. Segmentation is not always as simple as the separation of nonoverlapping convex shapes; see Fig. 1.3.

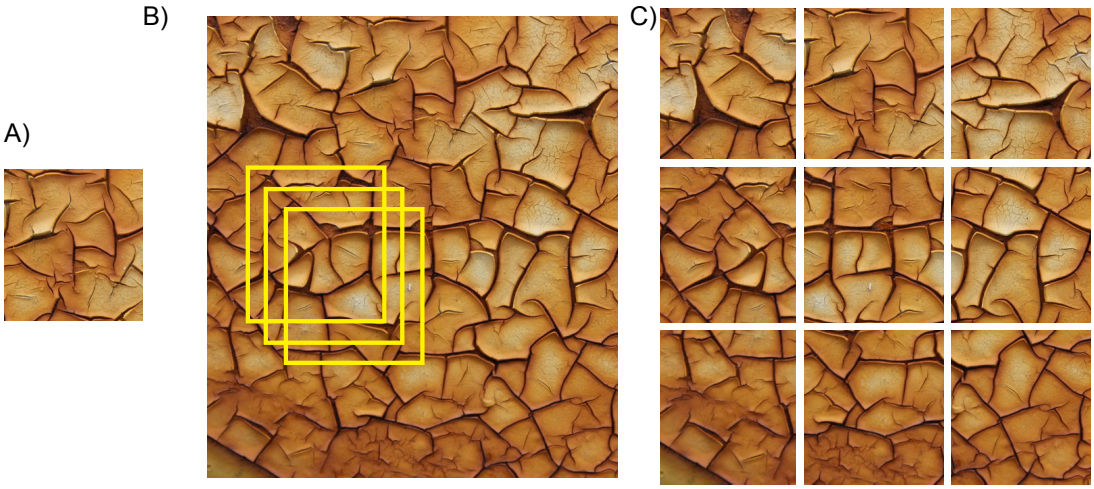


Figure 1.2: The challenges of visual search in a continuous search space. A) shows the target, which is very difficult to find in B), a continuous search space, because of the large number of overlapping potential matches (yellow rectangles). In C), the search space has been split into nine segments, making matching much easier (the target is at the centre of the top row). This illustrates the need to correctly abstract the search space, imposing the right high-level groupings, in order to match the target.

Asking observers to categorise natural scenes (for example into grassland, mountain or desert) is an example of gist extraction. Scene categorisation initially appeared to be resistant to disruption by dual-task interference[75] or inattentional blindness[76]. However, Cohen *et al*[77] recently found that, with sufficiently hard auxiliary tasks, both these types of interference reduce natural scene categorisation performance.

Perception can operate on more complex levels than simple classification. When one is physically present in a scene, or studying an image of a scene in great detail, stable object representations must be formed. It is here that theories of object recognition[12, 10, 13, 9] fall short: they assume that a single object has already been detected, localised and segmented, and that all that remains is to identify it.

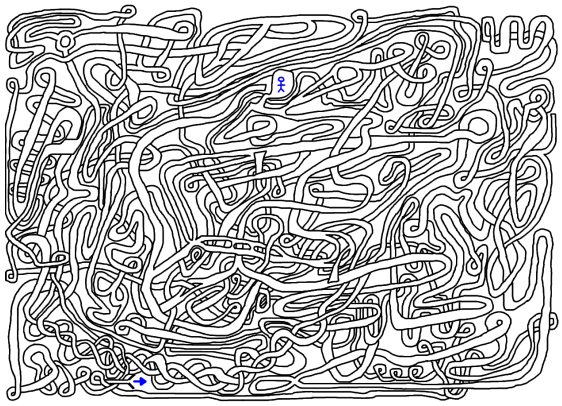


Figure 1.3: An illustration of the implausibility of automatic total segmentation: there may be multiple shapes here, but any pair cannot be preattentively labelled as "same object" or "different object" without a time-consuming attentive process of examination.

Another deficiency of current models relates to the dynamic nature of natural scenes. We often aim to detect not just the presence of an object, but a particular pattern of movement. Specific body and hand movements carry useful information[78], as do particular face movements[79]. An "object," for example a hand, is a spatially distinct part of a scene. If the scene is static, there is no temporal change; if the scene is dynamic, recognition models integrate all views of the object together, binding them under one view-invariant representation. Current models, however, do not explain how we pick out particular temporal slices of a moving object (such as a distinctive hand-wave) and represent them as events. We treat this problem in more depth in the next section.

There is symmetry here: an object is a spatially distinct part of a scene, and an event is a temporally distinct part of a sequence, usually associated with an object. An object's boundaries are measured by asking an observer to outline it[80]; an event's endpoints are measured by asking an observer to press a button when it starts and ends[81]. In neither case is there a definite solution: observers will disagree about the outlines of an object[82] and the endpoints of an event[83]. However, in both cases, observers can indicate a point which is definitely inside the object and an instant which is certainly part of the event. An object or event can thus be seen as a distribution, over space or time respectively, of the probability of being part of the object or event.

Core object recognition requires invariance to position, size, rotation, lighting, configuration, and within-class variation. In addition, dynamic recognition requires

invariance to overall pattern speed and differences in relative pattern speed (a gesture which is fast at the beginning and slow at the end should be matchable with the original). Recognition of objects in cluttered natural scenes requires invariance to the background and the number and type of other objects present in the scene.

1.1.3 The gist

The gist of a scene is a compact representation of its essential properties. Definitions of the gist are often motivated by the examples of flicking through television channels or observing rapidly presented image sequences. Oliva defines it as the representation which allows "the phenomenal experience of understanding everything at once, regardless of the visual complexity of the scene"[84]. In reality, not every detail is stored, but the observer has an impression of understanding of and familiarity with the scene. Gist is often contrasted with "clutter," the irrelevant details which can be ignored without impeding the task.

The gist can contain diverse features: low-level descriptions of colour and texture, spatial layout of the scene, scene class (such as "city" or "mountain"), salient objects, and semantic knowledge of acts and affordances relating to the scene[84]. Potter and Levy[85] were the first to demonstrate that scene-class cueing improved detection in a rapid sequential visual presentation (RSPV) task. They showed that presenting subsequent images interferes with the consolidation of the gist in memory[86].

The gist underlines another deficiency in current object recognition theories: the brain may represent a scene in multiple ways at the same time, as when a gist is available as well as a more detailed image in visual short-term memory. Existing theories of object recognition focus on feedforward recognition, not the comparison of a target object in short-term memory to a candidate object being directly perceived. Visual search tasks often involve evaluating a series of pictures for a match with a known gist. How is information routed and compared between heterogeneous representations?

To summarise, existing object recognition theories do not deal well with the problems of natural scene perception. Scene complexity means that objects must be segmented from the background and from other objects; this cannot always done as a preattentive step by an imagined segmentation module, as shown in Fig. 1.3. It is not clear how the gist and the spatial envelope could integrate with hierarchical feed-

forward recognition models. Finally, static theories do not extend well to the time domain. We now discuss the role of time and dynamics in object recognition, using the examples of motion, dynamic face perception and biological motion.

1.2 Recognising dynamic stimuli

So far, we have considered perception from the static perspective only. We now introduce time. The world is rich in dynamic, varying stimuli which the visual system is able to process and encode effectively. Most object recognition theories, however, do not treat dynamic stimuli at all; they explain only how to represent static images. How is a 3-second smile expressed in face space? We can think of it as a trajectory, but this leaves us without a dedicated high-level representation for the dynamic expression. How can HMAX represent a changing stimulus when it contains no provision for encoding multiple time points?

There is much debate about the nature of time[87, 88, 89] and its neural encoding[90, 91, 92]. It is important to distinguish between the concepts of duration and event time: duration is the amount of time which passes while an event takes place, and event time is the time at which an instantaneous event takes place. The event time of a non-instantaneous event must refer to an instantaneous point during that event, such as its beginning or middle. When looking at dynamic visual search, we are concerned mainly with attaching an event time to an encoded spatiotemporal feature. Models of duration are still worth consideration, since the event time of a spatiotemporal feature is computable as the duration between the beginning of a clip and the beginning of the spatiotemporal feature.

One central question is whether time is encoded as a quantity (a scalar value, such as firing rate or a simple population code) or whether it is not directly represented on the neural level at all, but is accessible as a high-level representation at the cognitive level. Karmarkar and Buonomano show in a neural network model that time can be encoded "in the absence of clocks"[93]. In other words, cortical networks may be able to tell time without encoding it directly on the low level. They behave as though they do, but this behaviour can be correlated only with change in their high-dimensional state, not with a single local element (or a group of synchronised local elements) which

serve as clocks.

1.2.1 Motion

The human visual system is highly adapted to compute and process motion, both on short and long timescales. Braddick[94] suggested a two-process theory of apparent motion. In this account, one process combines information over short spatial and temporal scales and cannot operate dichoptically, whereas the other is capable of detecting motion over larger spatial scales and without smooth displacement (as in the phi phenomenon).

Cavanagh *et al* support the idea of two separate systems[95]. A low-level system appears to specialise in detecting motion over short time intervals[96]. Stimuli such as illusory rotating discs[97], or drift-balanced video clips[98] do not show any motion in terms of their spacetime Fourier components; they defeat spatiotemporal energy models. Such drift-balanced stimuli were therefore supposed to be unable to activate this system, suggesting that the brain uses other methods to construct long-range motion percepts. However, it has been suggested that this mathematical difference is not enough evidence to posit separate motion systems[99]. The visual system is not a mathematically perfect system, and the blurring effects of spatial and temporal filters could give a drift-balanced stimulus the gradients necessary to activate a unified motion system. The multichannel gradient model (McGM)[99] is one such model. Benton *et al* [100] showed that spatiotemporal gradient histograms can contain enough information to recover the direction of drift-balanced texture-defined motion.

Lu and Sperling[101] used pedestalled displays (in which a moving sine wave is superimposed on a stationary sine wave of the same frequency) to muster evidence for three separate motion processing systems: a first-order system computing motion energy, a second-order system which can perceive the motion of texture-contrast modulations, and a third-order system which perceives and tracks features. The first two systems are said to be exclusively monocular, while feature tracking is suggested to operate interocularly. Feature tracking also requires higher stimulus contrast and has a lower maximally sensitive temporal frequency (3 Hz as opposed to 12 Hz for the faster systems). However, these conclusions are controversial; much evidence exists that simpler systems can handle both motion energy and texture-contrast modulations.

Complex motion percepts can mix opposing low-level and high-level motion signals. Theta motion, for example, involves a moving area of the image which is itself constituted of pixels which are moving in the opposite direction[102]. Fruit flies can perceive the correct direction of both second-order motion[103] and theta motion[104], showing that basic versions of these computations are tractable even by brains much smaller than ours.

Intensity changes are central to motion perception; in reverse phi motion, contrast inversion of one of a pair of alternating images (which would usually induce phi motion due to displacement of the objects they show) reverses the perceived direction of motion. When motion is tested psychophysically, stimuli are usually carefully constructed to yield a particular type of motion. When viewing complex natural scenes, however, the same processes that generate illusions like reverse phi may use natural, fast variations in contrast to create complex percepts. Unless we have access to motion ground truth, we cannot strictly call these "illusions."

Motion percepts can be low- or high-level, reflecting variation in the amount of information integration. Johnston[105] proposes terms for four classes of motion percept. Local motion describes a dense field of velocities at each point in an image. Object motion refers to the tagging of an object with a higher-level motion property such as "fast" or "moving upwards" which could correspond to a large number of motion fields. Object-based motion refers to change in the parameters of an object, such as the opening of an eye or the intensity of a facial expression. Finally, gestures are patterns of object-based motion expressed as an ordered sequence. These levels of motion recognition recall the ubiquitous hierarchies of object recognition models.

As soon as we impose a hierarchy, we create a segmentation problem as well. Object-based motion relies on an object representation, whose parameters may vary; gestures must be temporally segmented and picked out as part of a sequence. This is of especial importance in natural scene perception, where objects are segmented neither spatially (from the background) or temporally (by being presented along with cues to their beginning and end, such as the start and finish of a trial). Dynamic natural scenes are full of motion, and there is considerable work on the advantage that motion confers on recognition. We now survey results on dynamic face perception, dynamic object recognition and biological motion.

1.2.2 Dynamic face perception

Much early experimental work on face perception involved only static images. Real faces, however, are always moving, both rigidly (by rotation and translation) and internally (through the action of the facial muscles). A video of a moving face contains much more digital information than one of a static face. How does the brain use this additional information?

Bassili[106] provided initial evidence that the motion signal alone is useful. He attached white dots to a face, recording the motion of local skin patches, then presented observers with the dot motion alone. They were able to classify the six fundamental expressions[107] better than chance. His further work[79] attempted to measure the information carried by the upper and lower halves of the face in each expression.

One 1997 study[108] indicated a slight recognition advantage when a face was learned from dynamic video and tested against static images. Similarly, Knight and Johnston[109] found that motion conferred an advantage when recognising celebrities' faces, but not when they were inverted. This was confirmed by Lander *et al*, who found that motion conferred a recognition benefit when images were degraded[110, 111, 112]. The role of motion information thus appears to depend both on familiarity and image quality. Subsequent work, however[113, 114] found no such advantage for newly learned faces.

Motion may aid face perception by allowing the brain to construct a better 3D model of a face: the representation enhancement hypothesis[115]. Conversely, motion details may provide extra information which is secondary to static recognition: the supplemental information hypothesis[112]. Haxby's distributed face recognition model supports the view that motion information can be processed separately, as does a clinically observed double dissociation[116]. Statistically, separating the two classes of information seems to make sense: principal component analysis of moving faces shows that identity and expression can be mostly coded by separate axes in PCA space[117, 118].

O'Toole *et al* extended Haxby's model and linked it with the dual streams model[115]. They propose that static information is processed along the ventral stream according to Haxby's model, while motion information is extracted in a general low-level way by area MT (along the dorsal stream) and then interpreted in a higher-level,

face-relevant way by the STS. This "supplemental motion backup system" has been supported by double dissociations in prosopagnosics[119, 120, 121] as well as the observation that natural speaker mannerisms make it easier to match video clips of a speaking person to the corresponding audio[122].

Compared to static faces, moving faces possess an additional time dimension. Features which change in time can be hierarchically organised in time just like parts and subparts of an object are organised in space. For example, a smile is made up of a mouth-opening movement and a mouth-closing movement; a wave comprises a sequence of hand movements in opposite directions; and a conversation can be broken into individual sentences, individual words, and individual phonemes. We use this framework when splitting a video recording into separate portions (segmentation). Hill *et al* attempted segmentation using chin position extrema as section points[123]; this is similar to using extrema of curvature to segment an object in space[11].

Do we have access to high-level representations of dynamic expressions? Curio *et al* approached this question by testing for high-level adaptation to dynamic expressions generated from 3D face scans[105, 124]. In each trial, observers were shown a dynamic adaptor and then asked to classify a dynamic expression (disgusted or happy). The adaptors, which were anti-expressions (anti-disgusted or anti-happy) were found to shift judgement towards the corresponding positive expression. In an attempt to rule out low-level adaptation, the authors temporally reversed some of the adaptors, reversing their low-level motion fields. This had no effect on classification, suggesting high-level adaptation of representations describing facial features. The authors interpret this as evidence for high-level dynamic expression spaces.

As soon as we gain the ability to segment a low-level stream of features, such as a video of facial movement, we impose a higher-level description. This gives us the ability to say that two different instances of a high-level feature, such as two slightly different smiles, are the same on the high level (they are both a smile). How do we describe the low-level ways in which a higher-level feature can vary? We use various types of invariances. One is temporal constancy: the ability to say that two expressions are speeded-up or slowed-down versions of one another[125]. Another is intensity invariance: the ability to say that an intense smile and a weak smile are the same kind of expression. Finally, we have temporal invariance: the ability to detect an

expression at different points in time.

Temporal invariance is the equivalent in time of position invariance, a key concept in object recognition[126, 127]. Temporal constancy (or speed invariance) is, similarly, the equivalent in time of size invariance. Little attention has been given to dynamic temporal invariance in object recognition. This issue will be addressed in more detail later.

1.2.3 Biological motion

The ability to recognise other animals has great survival value, but often has to make do with very poor information: predators may be occluded, prey may be camouflaged, or conspecifics may be distant. The visual system has learnt to exploit sparse information to recognise biological motion, as shown by Johansson's point-light walkers[128], which enable recognition of walking human figures from a small number of dots at key anatomical points.

If dots are placed at stable anatomical points (usually the limb joints), a large amount of information can be expressed by a simple model. Troje[129] used a linear discriminant analysis model to capture 98% of point-light walker variance using only four components; an axis in this 4D space which corresponded to gender was found. If dots were moved around on the body between frames, human performance was still above average[130], indicating that human biomotion perception involves more than just local motion.

This was confirmed by Neri *et al*[131], who presented stimuli in which dots were shown not continually but for two frames at a time. This dynamic temporal undersampling ensured that observers did not have access to a continual global motion signal, forcing them to integrate information across space and time. They were still able to effectively perform both walker detection and walking direction judgement, despite the presence of hundreds or thousands of dots (this study used QUEST adaptive thresholding to vary the noise level). Observers' sensitivity also increased with longer walker presentation. According to probability summation theory[132, 133], the authors interpret this as evidence that biological motion is processed by a network of different mechanisms with varying efficiency.

Is biological motion internally represented as motion or form? Casile and Giese[134]

produced a stimulus intended to contain only motion information. Statistical analysis of point-light walkers showed common patterns of opponent motion; these were exploited to produce walkers in which pairs of dots showed the correct opponent motion, but were not arranged according to the form of real limbs. In naive observers, these stimuli still produced an impression of a walking human body, suggesting that motion information alone is sufficient for this percept.

Lange and Lappe challenged this theory using a model in which static, stick-figure templates alone permitted recognition[135]. They used three types of point-light walkers: Johansson's original walker[128], a version whose dots change position between frames[130] and the walker due to Casile and Giese in which all dots move randomly except those located near the hands and feet, which have a defined vertical motion component[134]. The model was then made to perform direction discrimination and direction coherence (between the top and bottom of the walker) tasks.

The model works by searching in parallel across all stored templates and attempting to match each dot to a limb. Showing the same pattern of results as human psychophysical observers conducting the same tasks, it suggests that static form information can be effectively used to process biological motion.

1.2.4 Dynamic object recognition

Faces and bodies are not the only moving objects we are able to recognise. Animals, plants and artificial objects are often able to deform or articulate, producing specific patterns of shape change. We are able to imagine and recognise the distinctive patterns of motion shown by moving water, clouds, and flames. Even if an object is not able to change its shape, an observer can alter its retinal image by moving herself or the object.

Recognition works on sequences of images produced as an object moves in relation to the retina. One observation is key: the order of these images in time can help link the various view-specific representations of an object. The brain does not learn view-independent representations from a series of randomly ordered views, but exploits their order in time. Bülthoff *et al* proposed a computational model of how this might be done[136]. Features are detected in an initial view of an object; as it rotates, they are tracked, and when a sufficient number of features have disappeared (80% in this

study), a new set of features is established and tracked again. The first frame, and the frames in which features are re-detected, are called keyframes and correspond to stored views. The process of tracking features between a keyframe and a real view corresponds to view interpolation.

There is psychophysical support for models of this type: Stone[137] showed that 3D shapes, learned while rotating clockwise, are harder to recognise when shown rotating anticlockwise. Blanz and Vetter found that if faces are learned while they are simultaneously rotating and morphing between two identities, the two identities are more highly confused[138].

Computational models which take the time domain into account exist. Rolls' VisNet is a hierarchical neural network object recognition model whose input is a spatiotemporal sequence, not a series of static images[139]. The model stores a temporal trace of recent input, which allows it to perform Hebbian learning[140]. This process hints at how translation invariance might be performed: since objects often translate across the retina, representations of an object's retinal images will be temporally close and will be learned. Hawkins' Hierarchical Temporal Memory (HTM) takes a similar approach, using a stack of nodes which perform learning and inference[141].

1.2.5 Time and neural binding

Binding is a property of neural systems which was proposed by von der Malsburg[4] as a solution to the "binding problem": the fact that object part representations must be connected to object wholes, otherwise features may be confused between objects and unreal feature arrangements may be hallucinated. The predominant theory is that binding is implemented by temporal synchrony: high-level holistic object neurons and low-level local feature neurons oscillate together, behaving as a unified neural ensemble. The existence of this ensemble signifies the perception of an object.

Many theories of binding only consider the case of static object perception. However, it is well known that neural ensembles are easily entrained by dynamic or rhythmic stimuli, as when watching movies[142] or faces which change in identity, generating steady-state visual evoked potentials (SSVEPs)[143].

If temporal synchrony is to explain binding, it must thus also deal with changing objects. This adds another layer of complexity to the problem, because high-level

object files must be bound with spatiotemporal features which no longer exist on the retina or in early visual areas because the stimulus has changed. Perceiving dynamic stimuli correctly requires representation of features which are no longer present.

1.2.6 Extending object recognition theories to the temporal domain

Most object recognition theories deal with "core object recognition:" the classification of uncluttered static images of a single object. Numerous attempts to extend these theories to dynamic stimuli have been made.

It has been suggested that object appearance learning takes place by associating a succession of views[144]. This was tested by Wallis and Bülthoff[138], who asked observers to learn sequences of heads which rotated in 3D and morphed in identity at the same time. They then performed a delayed match-to-sample task, judging whether two static faces were the same. Performance was lower for identity pairs which had been morphed together, showing that temporal association caused confusion. The authors' interpretation is that temporal association links identity representations together. Two control experiments convincingly eliminate the effects of morphing alone (by presenting morphs at the same time) and viewing the morphs in the wrong order.

It is therefore clear that the temporal domain has a strong effect on object recognition. How can static recognition theories be extended to the temporal domain?

One strategy is to use a local feature model to track identifiable image patches through time. Bülthoff *et al*[136] tested a classifier in which small features are isolated in the first frame of a rotating face sequence, then tracked as the face rotates. Features drop out or disappear due to occlusion or shape change; when enough features are lost, a new "keyframe" is defined and a new set of features are detected. This computational model provides evidence that local spatiotemporal features can represent a moving object.

In general, how do we extend a static model to the temporal domain? The simplest way of encoding a dynamic stimulus is to ignore its dynamism and take a static sample: a snapshot. This, of course, does not relate to a strict instant in time, since even at the retinal level neurons perform temporal integration. Even taking a single snapshot

can aid recognition of a dynamic sequence, since the visual system is often able to construct correlations between a single frame and the rest of the sequence.

A more powerful approach is to use the model to sample temporally from the input sequence, producing separate representations for each time point, as does the Bülthoff model. We then face the problem of how to represent the time at which each snapshot was taken. It may be enough to store snapshots without this information at all, producing an unordered set of representations.

The next level of sophistication is to store multiple snapshots and also to represent their order. If this is done by direct representation, as opposed to an emergent process as described in[93], then there are several ways order may be stored. Firstly, the order of snapshots may be stored, but not the delays between them. Secondly, the order may be stored, along with the spacing between the snapshots. Thirdly, we may store the order along with the offset of each snapshot from a single point in time, usually the beginning of a stimulus. Wallis and Bülthoff's rotation study provides psychophysical evidence that order is taken into account when learning face representations, but does not address the question of inter-snapshot delays.

Consider motion direction maps, which exist in macaque visual cortex[145]. Neural activation in a motion map is a representation of a changing retinal image, not a static sample. High-level motion representations provide another example; a common facial expression such as a smile is unlikely to be represented by a series of static snapshots, which would lead to much redundancy of duplicated information. Such central motion patterns are better represented by object-based motion (change in the high-level parameters of an object description).

It is unlikely, then, that all dynamic stimuli are represented by sequences of static samples. Some models perform temporal integration, which means that they compute new representations which are more than bags, sets or lists of static representations.

An example from face perception helps to illustrate the point. Consider a 5-second dynamic facial expression. Equipped with a PCA model which describes static face images low-dimensionally (along the lines of [117, 146], we have two ways to encode our dynamic stimulus. Firstly, we may represent each movie frame by its coordinates in low-dimensional PCA space ("face space"). This gives us independent representations for each frame. Secondly, we may arrange the low-dimensional coordinates into a

vector which describes a dynamic expression. We may then use other 5-second dynamic expressions, described in the same way, to build an expression space by applying PCA again. This gives us a space in which each point corresponds to a dynamic expression; here, temporal samples (frames) are not represented independently.

Some of these encodings of time are independent of the neural substrate's position on the emergence-vs-direct-clocking continuum. However time may be encoded, and whether it is stored locally and explicitly or not, we can still differentiate between an ordering and a spaced ordering (which records order and inter-snapshot duration). The final case, however, specifies the encoding: theorising that snapshot times are stored as offsets from the beginning of the stimulus specifies that they must be scalar quantities.

These issues show that the encoding and processing of event times, durations, and dynamic stimuli are not very well understood. In the rest of this thesis, we use psychophysics to investigate the visual system's ability to encode and search for video clips of dynamic natural scenes.

1.3 Our experimental approach

We have seen that existing theories of dynamic natural scene perception pose significant challenges for models of object recognition. In order to test the human visual system's ability to encode, match and search for complex patterns of natural motion, we performed psychophysical experiments using two natural stimuli: dynamic flame and dynamic faces. Our flame stimuli were recorded from a hearth fire, whereas face stimuli were captured from human subjects.

Our experiments are the first to measure the visual system's ability to encode and match sequences of dynamic flame; previous work on dynamic natural scenes has focussed on the extraction of descriptions and affordances from video clips, not their encoding and matching.

Much attention has been paid to natural stimuli, including sunsets[147] and water[148]. However, most of this work has focussed on judging material properties[149], such as the viscosity of a liquid[150] or the glossiness of a surface[151]. Such a property judgement is a mapping from an enormous space of low-level visual

percepts into a small space encoding the property in question, which may be one-dimensional (in the case of temperature) or multidimensional (as in the case of surface texture).

On the other hand, limited attention has been paid to our ability to encode the dynamic form of rapidly-changing natural stimuli: to remember a particular exemplar, to match it to other exemplars and to determine whether it forms a temporal part of a longer stimulus. This is a much more taxing task, since it requires a low-level pixel stream to be matched to another low-level pixel stream. A useful description of the first stimulus, not just its position in temperature space or viscosity space, must be encoded, maintained in memory, and matched with the description of the second stimulus.

We chose to investigate dynamic flame, the pattern of light produced by burning gas. This stimulus is appropriate for several reasons. It is highly dynamic, posing an encoding challenge to the brain's motion system. It contains both sharp edges and areas of continuous, smooth texture. It is associated with a strong high-level percept of upwards motion, and it has an upright orientation, which may have allowed the development of specialised encodings, as for dynamic faces. It is a stimulus to which the human visual system has been exposed for a very long time: human control of wildfires dates from as long ago as 1.8 million years, with frequent and certain use in cooking and agriculture from 400,000 years ago[152]. Its importance may also have led to specialised encodings.

Our experiments aimed to answer several questions:

- How well can observers perform temporal visual search on dynamic natural scenes (specifically, patterns of moving flame)?
- What invariances do observers possess? Which low-level features help observers perform matching (sensitivity) or do not impair matching when they are disrupted (invariance)?
- How do observers encode and match dynamic sequences of varying lengths? What are the effects of varying sample length and target length? Do observers show target position invariance or search space size invariance?
- Is the adult human visual system specialised for the representation of dynamic flame?

- How does dynamic flame matching ability compare to dynamic face recognition ability?
- What directions of motion do observers perceive in dynamic flame, and how spatiotemporally local are these motion percepts?

We now discuss (Chapter 2) our stimulus acquisition methods and experimental setup, before conducting an analysis of the image statistics of dynamic flame (Chapter 3), then moving to a psychophysical evaluation of observers' ability to match and encode this complex example of dynamic form (Chapters 4, 5 and 6). We return in Chapter 7 to a discussion of our findings and their implications.

Summary

- The human object recognition system is often described using hierarchical models.
- Well-established models of object recognition concentrate on still, pre-segmented images free from background or distractors.
- Humans are very good at face perception, but models of this process are less hierarchical.
- There is still controversy concerning whether biological motion models are based on motion features or static templates.
- There is a paucity of models which treat the encoding of dynamic stimuli.
- Models of static object recognition do not sufficiently explain our ability to recognise moving objects in cluttered scenes.

Chapter 2

General methods

As seen in the previous chapter, a wide range of object recognition models do not satisfactorily explain our ability to recognise, match and search within complex dynamic natural stimuli. Our experiments on dynamic form used two datasets of recorded video (facial expressions and hearth fire) and a novel visual search paradigm: delayed match-to-sample on clips of different lengths. This chapter details stimulus capture, experimental setup, and the general paradigm. The following chapters describe each experiment in more detail.

2.1 Recording and processing of stimuli

2.1.1 Fire dataset

A continuous 45-minute recording was acquired from a hearth fire using a Sony HXR-NX5E digital camcorder recording at 50 Hz with a shutter speed of 1/150 (the shutter was open for 6.67 ms). The scene was lit by a mixture of natural and artificial light and no CCD gain was applied. Video was saved directly to the compressed AVCHD format at an initial resolution of 1024 $\times$ 768. Before presentation, stimuli were cropped to 564 $\times$ 641 pixels, removing the background and most of the fireplace. Individual frames were decompressed and saved as bitmaps (see Fig. 2.1). Our experiments used either a 1,000-frame (20-second) or 10,000-frame (200-second) subset of this corpus; these datasets were short enough that there was little variation in the background, preventing matching on easily-percieved static features such as displaced logs.

Figure 2.1: Fire dataset.

(a) HD image as recorded

(b) Cropped image as displayed to observers (564 $\times$ 641)

2.1.2 Face dataset

The facial movement dataset was recorded from four subjects using the same Sony HXR-NX5E digital camcorder recording at 50 Hz. We asked subjects to read out identical passages of text in order to obtain synchronised recordings of facial motion. Firstly, subjects were asked to speak freely. Secondly, subjects were asked to speak along with recordings played through their headphones. We used recordings of well-known nursery rhymes. From each of the four subjects, we recorded one free-speaking clip and three nursery rhyme clips.

Clip lengths (slightly different across subjects due to approximate trimming) were:

Rhyme	**Mean length**
Free-speaking	2k frames (40 seconds)
Hot Cross Buns	1k frames (20 seconds)
The Grand Old Duke Of York	1k frames (20 seconds)
Twinkle Twinkle Little Star	1k frames (20 seconds)

Examples are shown in Fig. 2.2.

2.2 Choice of subjects

We recruited observers 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.

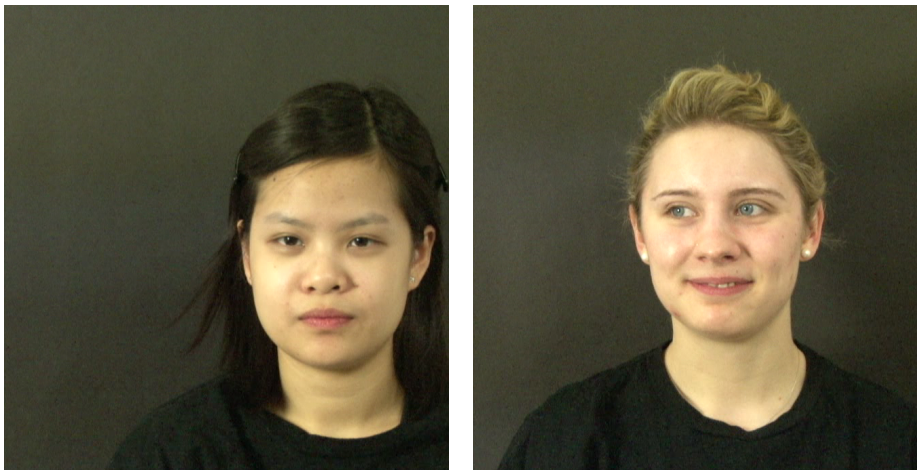
During each experiment, subjects were screened for accuracy during a training phase, which was designed to be an easy task (average accuracy on piloting was found to be 80% or higher). We followed the policy of rejecting subjects if they could not perform better then 75% correct during the training phase. However, this rarely occurred; only two subjects were rejected during the entire course of experiments.

2.3 Experimental set-up

Experiments were programmed in Matlab using Psychtoolbox[153]. Video was displayed by loading bitmaps into video memory and programmatically displaying each

Figure 2.2: Face dataset. Cropped images as displayed to observers (655 x 686)

(a) Subject 1 (b) Subject 2



(c) Subject 3 (d) Subject 4

frame to the screen using Psychtoolbox's texture management functions. This allowed precise control of frame rate.

Stimuli were displayed with a frame rate of 50 Hz on a Mitsubishi DiamondPlus 230 SB CRT monitor with a refresh rate of 85 Hz and a resolution of 1280 $\times$ 1024. The active video area subtended a visual angle of 14 degrees; subjects used a chin-rest at a distance of 57 cm from the screen. They were asked to keep their heads upright and still and their heads vertical. Subjects were not requested to fixate any specific point on the stimuli and therefore could scan the video if they wished and time allowed. The experiments took place in a darkened room.

We dealt with the mismatch between monitor refresh rate and stimulus frame rate by scheduling screen redraws at the appropriate time using Psychtoolbox's **Flip()** function. As opposed to requesting a screen redraw at the next available opportunity,

this strategy ensured that new frames appeared as close as possible to their ideal arrival time, and that if a screen redraw was missed the rest of the stimulus presentation was not delayed.

We logged the number of frames whose presentation missed the required deadline; this happened on under 0.1% of presented frames, which is acceptable performance with current video hardware.

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

2.4 The task: delayed match-to-sample

We used a delayed match-to-sample task, presenting a sample clip first and a test clip second. In 2AFC tasks, we presented two candidate test clips. Where tests were close in length to samples, observers were performing a matching task; where tests were much longer than samples, they performed a visual search task.

Visual search is the process of finding a sample stimulus (or target) in a test stimulus (search space or group of distractors). When targets and search spaces are dynamic, the search space can vary in duration (ratio of test length to sample length) as well as size (ratio of test size to sample size). We term "temporal visual search" a search task in which the target is contained in a sequence of a longer duration than the target clip, but both videos are the same size.

One form of temporal search is rapid sequential visual presentation (RSVP). Here, however, the search space is not continuous, the target item is static rather then time varying and stimuli are separated by gaps. RSVP is a series of comparisons with separate objects rather than a search operation, in a continuous space, for a dynamic sequence.

A huge body of literature describes the behaviour of visual search under changes in search space size[154]. How search for a dynamic event performs under changes in search space duration, however, is relatively uninvestigated.

2.4.1 Trial structure

An example Yes/No trial in a visual temporal search experiment is shown in Fig. 2.3. The sample clip is shown first, followed by an ISI and then a longer test clip. The observer hits the up arrow if they think the target is present, the down arrow if absent.

The sample clip was either contained in the test clip (as a pixel-perfect copy), or was chosen randomly from some other location in the corpus. This was implemented by first randomly picking the test, then randomly picking the sample within it (for true trials) or picking the sample from the whole database (for foil trials). Random numbers were sampled from a uniform distribution. We also used 2AFC trials in some experiments (Fig. 2.3). The observer hits the left arrow if she thinks the sample was present in the first test clip, and the right arrow if present in the second test clip. In this case we pick two test clips, then pick a sample randomly from one of them. The most important independent variables are sample length, test length, test/sample ratio, and inter-stimulus interval (ISI).

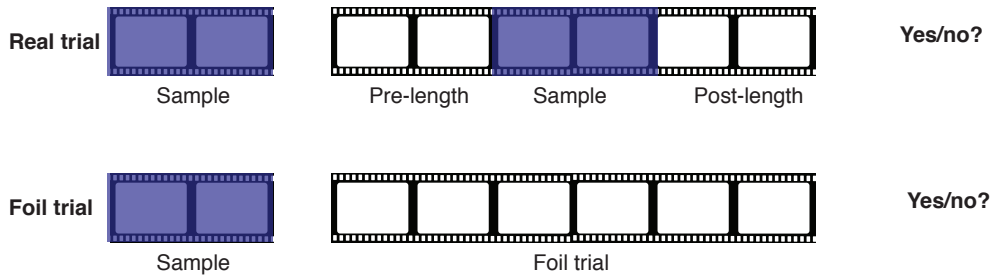
For some experiments we also produce a modified trial by manipulating or transforming sample or test clips. Example transformations are inversion, filtering and colour manipulation.

2.4.2 Instructions

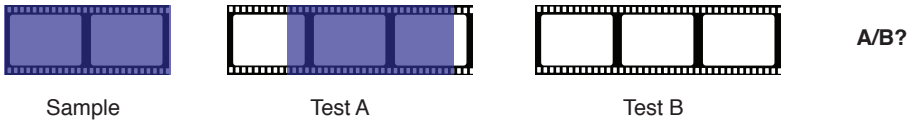
Subjects occasionally found this trial structure difficult to grasp initially, so it was important to ensure a consistent and effective set of instructions. We used the following verbal instructions, along with the appropriate diagram from Fig. 2.3.

> "In each trial, you will see a short clip and then a longer clip. The clips will be the same physical size on the screen. We ask you to decide whether the first clip was present in the second clip, or not. If the first clip is present in the second clip, it could be anywhere: from the beginning to the end. In other words, the first clip might be a randomly picked section of the second clip - or it could be a completely unrelated clip."

Figure 2.3: Trial structure in temporal search



(a) Yes/No: a sample is followed by a longer test, which may contain the sample (blue). Real trials are those in which the tests contain the sample; in foil trials, the test is from a different part of the dataset than the sample. Observers respond Yes or No, allowing the application of signal detection theory.



(b) 2AFC: a sample is followed by two tests, one of which contains the sample (blue). Observers respond A or B, meaning that naive signal detection theory cannot be applied.

2.4.3 Iconic memory

We noticed during piloting that percepts and memories of the first and last frames of each clip are enhanced. As we wished to study the encoding and perception of dynamic form rather than iconic memory for snapshot images, we ensured that first and final frames never co-occurred across sample and test. This would have provided an easy route to matching using iconic memory. Instead, we ensured that samples were always picked to begin or end at least one frame away from the endpoints of the test clip.

2.4.4 Training

We usually used some training trials (around 20, see individual experiments) before the main body of the experiment. The example condition we used varied across experiments, but it was always an easy version of the main task. During training, subjects were shown the mean accuracy of their last ten trials. We required their accuracy to be above 75% correct in order to continue to the main experiment. This

ensured a) that subjects could perform the task and b) that they were paying attention. We did not give feedback after each trial or block during the main experiment.

2.4.5 Block structure

Blocks were timed to be 10-15 minutes long. Subjects were asked to have a 2-minute break between blocks, and offered refreshments. There are several aspects of block structure which are worth noting. Firstly, a factor (for example test length) held steady within a block allows the observer's visual system and task-set to acclimatise to that condition. Secondly, we did not give feedback during the main experiment in order to reduce observers' opportunities to test and rely on high-level strategies.

Summary

- We used a dynamic flame dataset, which poses a natural and complex dynamic form encoding problem to the visual system.
- A delayed-match-to-sample paradigm was used: a sample was encoded and compared against one (Yes/No) or two (2AFC) tests.
- To prevent iconic memory providing an easy matching cue, we ensured that the first frame of the sample was never the same as the first frame of the test. The same was true for the last frame of the sample and test.
- When the test/sample ratio was only slightly higher than 1, we have a matching task. Higher test/sample ratios correspond to a temporal search task.

Chapter 3

Image domain analysis

Dynamic flame is a familiar part of the natural world, but its image statistics have not been studied in detail. The computer vision literature mentions algorithms which detect fire[155, 156, 157, 158, 159, 160, 161], but these all focus on detecting flame in video feeds (often from security cameras), not examining its statistics or representing it in a biologically plausible way. In this chapter, we use a variety of methods to investigate dynamic flame in the image domain.

We can characterise the challenge which video stimuli pose to the brain by analysing the raw pixels that make them up. Even without a human observer, image-based techniques can reveal identifying statistics which hint at the strategies the visual system may use to recognise and classify images. This chapter reports our analyses of dynamic flame in the image domain using averaging, Fourier analysis and motion analysis.

How do we study natural scene videos in the image domain? When we have access to a set of images from the same category, we can simply align and average them; this idea goes back to Galton in the 1870s[162]. Provided that each image possesses features which can be aligned with the others, this technique can show whether the class possesses a common global structure. Torralba *et al* [163] used this method to produce average images for various types of natural scene.

We can also apply mathematical transformations to individual images. The Fourier transform is particularly useful. The mapping of a 2D image into the frequency domain shows us the distribution of spatial frequencies present in that image (its power spectrum). A series of studies[164, 165, 166] have observed that the power spectrum of natural images is often of the form $1/f^a$ with $a \approx 2$. An orientation bias is also

often present, with vertical and horizontal orientations occurring more frequently than oblique ones[167]. What kinds of power spectra are present in images of dynamic flame? We address this question using the 2D and 3D Fourier transform.

Most of the literature on natural scenes describes only static images, which are much easier to acquire, store and work with. With video sequences, such as our dynamic flame dataset, we have access to a fundamental percept: motion. We report conflicting estimates, returned by two modern algorithms, of the low-level motion present in dynamic flame.

Finally, we discuss efforts to model dynamic flame using three widely-applied techniques: principal component analysis (PCA), PCA with a shape-detecting morph model, and dynamic texture modelling.

Throughout this chapter, we analyse a 5000-frame (20-second) flame dataset which was also used in our psychophysical experiments; see Chapter 2 for details of capture and preprocessing.

3.1 Image statistics

Figure 3.1 shows four randomly chosen images from the 1000-frame dataset. No amplification or stabilisation was used. Frame rate is 50Hz, corresponding to one frame every 0.02 seconds. We chose a high shutter speed (1/150 s), meaning that each frame integrates information from a period of the stimulus lasting 6.67 ms.

We note immediately that the flames shown in each frame have a well-defined shape. Their edges are distinct and fairly sharp. Occasionally, flames lack well-defined edges, fading smoothly into the background.

There is hardly any variation in the appearance of the logs; this 1000-frame dataset was chosen to keep the log position constant so that observers could not use it as a cue in matching tasks. We note (see the first and third frame) the occasional bright, upwards-moving spark.

Figure 3.2 shows four sequential frames. We can see that there is high correlation between successive frames; the two separate flame peaks in the first frame are shown merging into a single peak in the last frame. The stimulus has been sampled frequently enough to capture the similarities between successive frames. Can we still detect

Figure 3.1: Four randomly chosen images from our dynamic flame dataset. Frame rate is 50 Hz.

similarities between frames which are further apart?

Figure 3.2: Four sequential images from our dynamic flame dataset.

3.1.1 Similarity measures

To investigate, we measured three metrics of image similarity: absolute pixel difference, Euclidean distance, and the structural similarity (SSIM) index. Pixel difference was calculated as the absolute value of the sum of differences between images I_1 and I_2 over the n pixels:

$$D = \sum_{i=0}^{n} I_1(i) - I_2(i). \tag{3.1}$$

Euclidean distance (the square root of the sum of the squared differences) was

calculated between the points in image space corresponding to each frame. We note that, in the absolute value of the sum of differences, positive and negative differences will cancel, whereas in the Euclidean distance they will add as squares are taken:

$$D = \sqrt{\sum_{i=0}^{n} (I_1(i) - I_2(i))^2}. \tag{3.2}$$

The SSIM index is a simple but perceptually-motivated measure of image similarity[168] with less dependence on noise. It uses separate contrast and luminance comparisons to generate a similarity measure, aiming to separate structural information from illumination (algorithm flow chart shown in Fig. 3.3). Although its perceptual validity has been debated[169], it is useful as a comparison metric. The SSIM between two windows x and y is

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3.3}$$

where $\mu_x$is the average of x, μ_y the average of y, σ_x^2 the variance of x, σ_y^2 the variance of y, σ_{xy} the covariance of x and y, $c_1 = (k_1 L)^2, c_2 = (k_2 L)^2$ two variables to stabilize the division with weak denominator, L the dynamic range of the pixel-values (typically $2^{\text{bits-per-pixel}} - 1$), and finally $k_1 = 0.01$ and $k_2 = 0.03$.

Figure 3.3 shows how these similarity measures decrease as we compare images which are increasingly separated in time. Across all metrics, as separation increases, similarity decreases very quickly, reaching a minimum at a separation of 10 frames (0.2 s). Similarity then remains constant at this minimum. The similarities detectable by these indices, then, are very local in time. We see no peak in similarity after a particular time; these indices do not show any periodicity in the stimulus. These functions do not fit well with a straight line on either log-log or semi-log plots.

The structure of flame, then, is transient and rapidly changing. What structure do we see if we increase our temporal integration period by averaging over time?

3.1.2 Average images

Ever since Francis Galton's work with the average faces of criminals and law-abiders in the 1890s[170], producing the average image of a dataset has been used to show

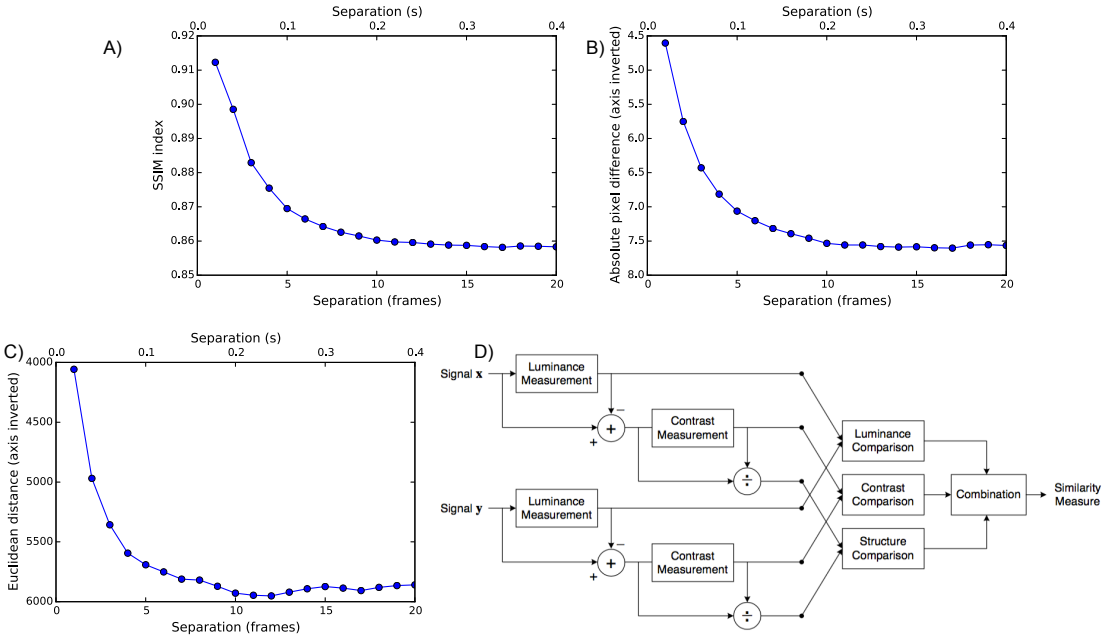


Figure 3.3: A,B,C): Image similarity indices decrease with frame separation in time. Each of these curves is the mean similarity-by-separation function of 625 different frames. A similarity curve was produced for each of the 625 frames; the mean of these curves is shown. Axes are inverted for the absolute pixel difference and Euclidean distance, making values nearer the top of the Y axis reflect similarity instead of difference/distance. All three indices reach a minimum after approximately 10 frames (0.2 seconds), showing that image similarities are very local in time. D) Flow chart showing the operation of the SSIM algorithm.

global structure. Does fire have a global structure? To investigate, we produced the mean image of the first 100, 1000 and 5000 frames of our flame dataset.

The results are shown in Fig. 3.4. An average of the first 100 frames shows clearly that the stimulus has two main components: a static background made up of logs, and a semitransparent dynamic flame component. The background never changes significantly, as we chose a dataset with no variation in log position. Observers must therefore use information from moving flames in order to match images.

An average of the first 1000 frames shows the same overall structure as the first 100, but is more smooth. It reveals two "flame columns" side-by-side. However, this is a very different average structure from that revealed by Galton's faces. The human face has a resting shape which is stretched and warped into an individual expression by muscle movements; this resting shape is revealed in an average image. Each frame of the flame stimulus, however, is made up of shapes which have little relation to those present in distant frames. They are not warped or deformed versions of the average.

A) First frame

B) First 10 frames

C) First 100 frames

D) First 1000 frames

E) Random 5 frames

F) Random 10 frames

G) Random 50 frames

Figure 3.4: Mean images, showing stimulus structure across time. A) The first frame. B) Mean of the first 10 frames. C) Mean of the first 100 frames. D) Mean of the first 1000 frames. E) Mean of 5 random frames. F) Mean of 10 random frames; there is less coherence than in B. G) Mean of 50 random frames.

3.1.3 Variance

The mean images suggest that frames in the dataset tend not to change around the border, but are highly variable in the centre. We can confirm this by looking at the variance of each pixel across the 1000-frame dataset, as shown in Fig. 3.5. Most of the image has very low variance; the areas corresponding to the two flames have higher variance. We note two areas of very high variance: the small flame on the left and the horizontal area under the front log.

Analysis of individual images, image similarity and variance shows that flame stimuli are locally self-similar in time and show a common structure when averaged, but that the shape of each individual frame is not a deformation of the average shape. The decrease in image similarity with frame separation suggests that flame does not have a periodic, repeating structure. To look at this issue further, we use Fourier analysis. Our 3-dimensional dataset can be investigated with three types of Fourier transform: on the 1D average brightness signal of the whole dataset (or of an individual pixel), on individual 2D images and on the entire 3D image stack.

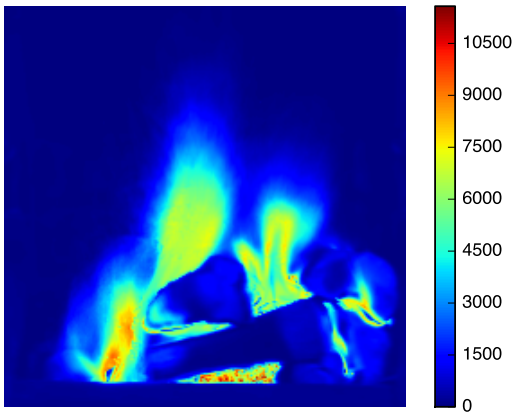


Figure 3.5: Variance image of the central area of the dataset. Each pixel shows the variance of the corresponding pixel across the 1000-frame dataset. We note two areas of very high variance: the small flame on the left and the horizontal area under the front log.

3.2 Fourier transforms

3.2.1 1D Fourier transforms

Objects are often blurred (as with a TV screen on which one is not focussing correctly) or seen in the periphery (as with a traffic light seen from the corner of one's eye). In these cases, the visual system can downsample a complex percept to a one-dimensional brightness signal. We can treat our fire dataset as a one-dimensional signal $x(t)$ by calculating $x(t)$ as the mean luminance of the frame $f(t)$.

This signal is shown in Fig. 3.6. It exhibits variation on both small and large temporal scales.

Performing a fast Fourier transform (FFT) on this signal gives the power spectrum shown in Fig. 3.6. The dataset's sampling frequency is 50Hz; we show the spectrum up to the Nyquist limit of 25Hz, above which aliasing may occur[171]. We do not show the linear plot, since the large DC component at 0 Hz swamps the rest of the spectrum. On a semi-log-Y plot, the power spectrum appears approximately straight, except for the values very close to the DC component. On a log-log plot, the power spectrum fits well to a straight line up until about 4 Hz, when it drops below the line.

We note a sharp peak in the spectrum at approximately 16 Hz. Due to its sharpness and artificiality, this appears to be a compression artefact due to the AVCHD codec used to encode the stimulus, and is not intrinsic to the stimulus. 16 Hz is likely to be the frequency at which the codec describes an entirely new frame as opposed to encoding the difference from the previous frame (a keyframe)[172].

We used base 10 logarithms throughout. A straight line in semi-log-Y space with equation $\log_{10}(y) = mx + c$ corresponds to an exponential curve in linear space with equation $y = 10^{mx+c}$ (see Fig. 3.7). Thus, the power spectrum is well approximated by an exponential curve of the form

$$\text{power} = 10^{m*\text{frequency}+c} \tag{3.4}$$

For the 1D spectrum fit, the parameter values are

$$\text{power} = 10^{-0.106*\text{frequency}+4.823} \tag{3.5}$$

A wide variety of phenomena in neuroscience and the natural world have power spectra that follow a $1/f$ curve matching the spectrum of pink noise[173, 174, 175]. This spectrum is thought to indicate the presence of complexity and interactions on multiple spatial and temporal scales[176, 175]. The spectrum of the 1D average luminance signal from dynamic flame approximates a $1/f$ curve (which plots as a straight line in log-log space) from 0 to 4 Hz, then drops below it. Overall flame luminance therefore shows less power in the high frequencies (4-25 Hz) than pink noise.

The presence of an exponential spectrum suggests that dynamic flame is produced by a system which is not completely correlation-free (as is white noise) but does not exhibit as much multi-scale complexity as a system exhibiting the classic $1/f$ power spectrum.

We lose a great deal of information by creating a time series from the mean luminance of each frame. To look more closely at the spatial structure present in each frame, we produced 2D Fourier transforms from individual frames.

3.2.2 2D Fourier transforms

We begin by taking the individual 2D FFTs of a set of representative images. Three individual spatial spectra are shown in Fig. 3.8; the mean of 5000 spectra is shown in Fig. 3.9. The vertical line present in each spectrum is due to edge effects; we shortly describe a re-analysis which uses a Gaussian window to eliminate these. In some of the individual frames, we note an asymmetry between power in the vertical directions and the horizontal directions. The mean spectrum displays an "X" pattern, with the legs of the X tilted closer to the vertical than the horizontal. This indicates that dynamic flame contains spatial frequencies oriented at an angle close to the vertical: the video contains periodic patterns closely aligned with the vertical.

3.2.3 2D Fourier transforms with Gaussian window

Our previous Fourier analyses used unaltered images, meaning that edge effects may appear on the power spectra. In order to differentiate between edge effects and features of the stimulus, we repeated the 3D FFT after applying a circular Gaussian window to each frame.

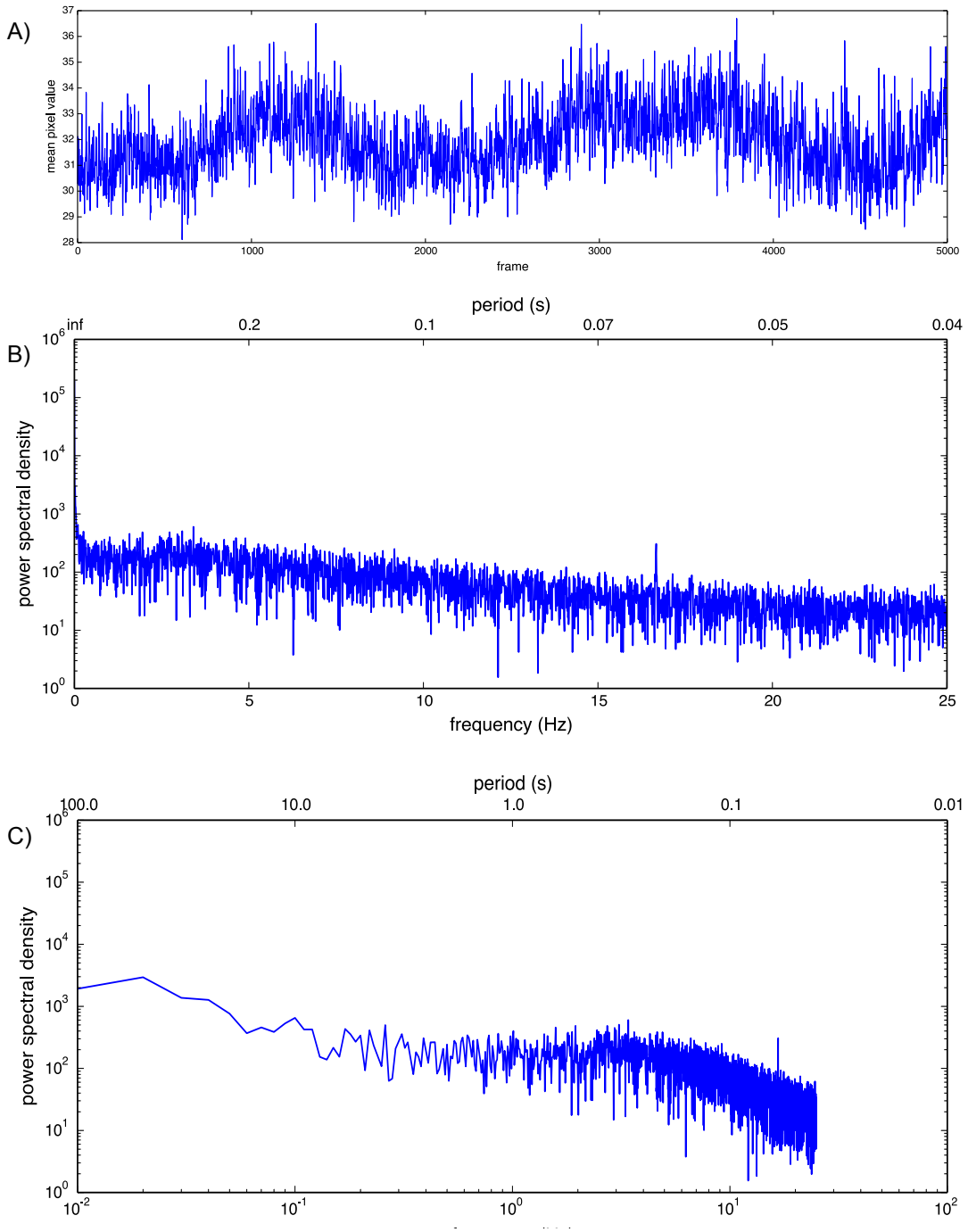


Figure 3.6: A) The mean luminance signal from the first 5000 frames (100 seconds). B) This signal's power spectrum in semi-log-Y space approximates a straight line, indicating an exponential power distribution. C) This signal's power spectrum in log-log space. In both cases we note a peak at circa 16 Hz which appears to be due to video compression.

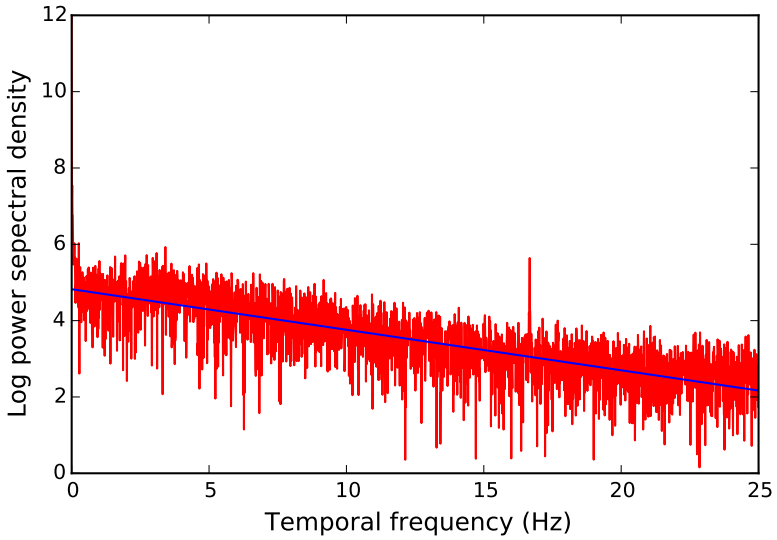


Figure 3.7: Power spectrum of the 1D brightness signal, plotted in semi-log-y space with a line fit. The line's equation is $\log 10(p) = -0.106 * f + 4.823$ or $p = 10^{-0.106*f+4.823}$.

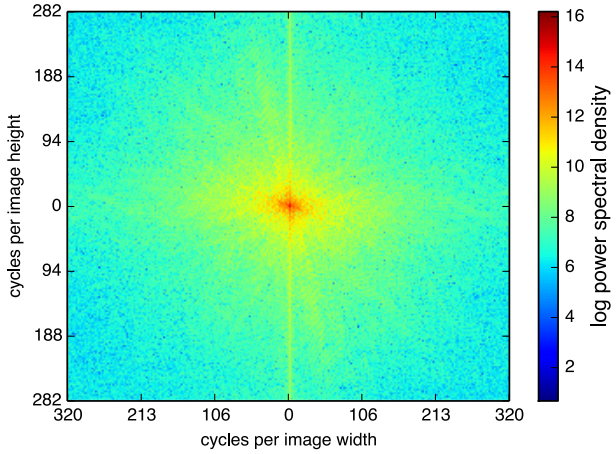
We analysed a 450×450 pixel region centred on the middle of the video. The standard deviation (100) was chosen so as to zero the values at the edges of the image, eliminating edge effects. In this dataset, cropping and windowing also has the effect of removing most of the background from the image, allowing the analysis to focus on the perceptually important elements: the flames at the centre.

Fig. 3.10 shows three individual-frame power spectra and the mean of 5000, calculated from the same dataset as the previous analysis. The vertical lines are absent, confirming that they are due to edge effects. The 2D mean power spectrum has a very different shape, with less power near the edges and a more coherent pattern near the DC component. There are three prominent streaks emerging from the DC component, oriented off the vertical; this indicates that there are three directions in which spatial oscillations are more likely to fall.

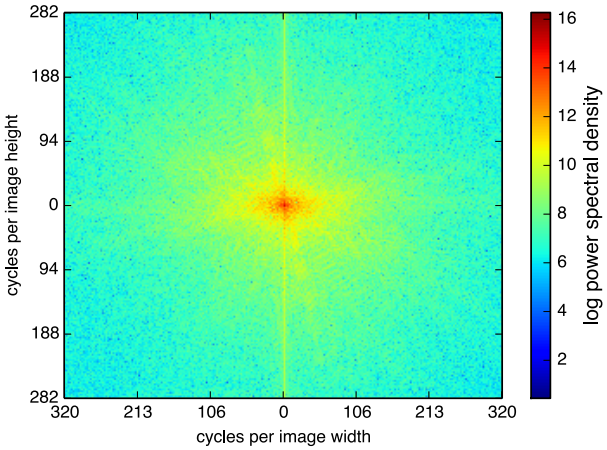
3.2.4 Individual pixel Fourier transforms

The FFT also allows us to analyse the global structure of the flame dataset. By treating each pixel as a 1D signal, we can perform an FFT on the time-course of each pixel and investigate the frequencies present in different parts of the frame. Fig. 3.12 shows the spectrum of an individual pixel near the centre of the frame. This particular

A)



B)



C)

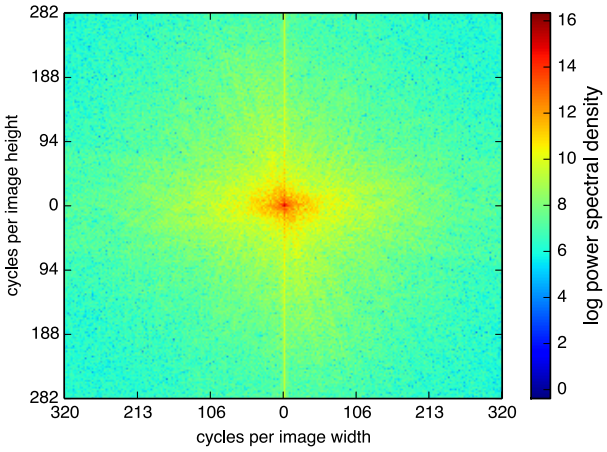


Figure 3.8: Power spectra from three individual frames, produced by 2D FFTs. The space axes are linear and the power is colour-coded in log space (otherwise, only the high DC peak is visible). There is a strong vertical line due to edge effects. The first two frames show power peaks angled slightly off the vertical.

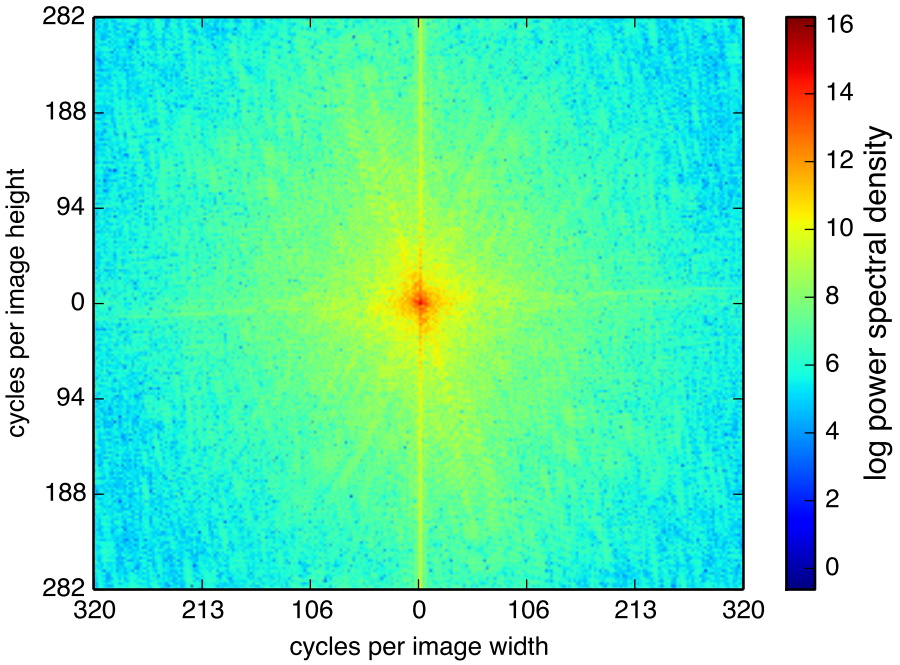


Figure 3.9: The mean of 5000 individual-frame power spectra. The space axes are linear and the power is colour-coded in log space (otherwise, only the high DC peak is visible). There are no obvious peaks in any particular direction. There is a strong vertical line due to edge effects.

pixel's spectrum is well fit by a straight line in semi-log-y space.

How do frequency domain characteristics vary as we look at different parts of the stimulus? To investigate, we fit a line to each spectrum in log-log space and recorded its slope and intercept. Since we have one line fit per pixel, we can plot the slope and intercept as an image, as shown in Fig. 3.13. Lower slope means relatively more low frequencies than high frequencies, which we see in the static parts of the image. The slope increases as we move from the logs to the top of the image, indicating relatively more high frequencies. Higher intercept means more power in the lowest frequencies, which we see again in the static parts of the image.

We also note a heavy gridding effect; this is due to the AVCHD compression codec, which splits the image into small parts for improved coding. Compression effects are more pronounced in the high frequencies, where miscoding is not as perceptible to the human eye.

These slope and intercept maps show the relative frequencies present in different parts of the image, showing a clear pattern of low frequencies just above the logs and

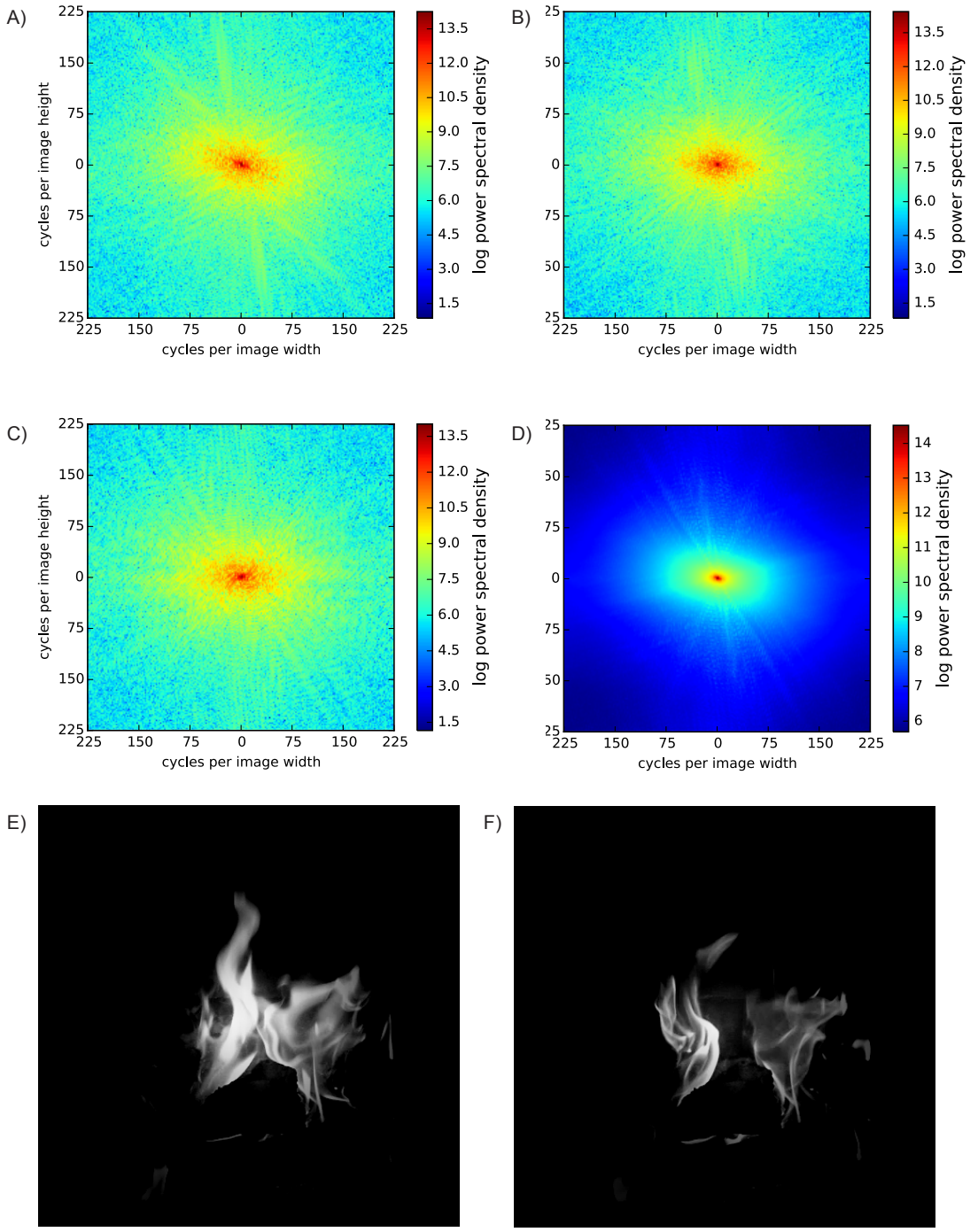


Figure 3.10: A,B,C): Power spectra of three individual frames, produced by 2D FFTs. These frames have been cropped to 450 x 450 px, removing the static background, and a Gaussian window applied (standard deviation 100 px). The mean spectrum is much cleaner than that of the uncropped, unwindowed data, showing three definite streaks and an off-vertical envelope. E,F): Example monochrome images with a Gaussian window applied in order to eliminate edge and background effects.

higher frequencies moving upwards. The encoding challenge for the visual system is clear: a complex mix of static form (the logs in the centre of the frame) and both slow and fast temporal oscillations.

We repeated the line fits on the same data in semi-log-y space, since we observed better fits than in log-log space. We also restricted the fit to the spectrum below 15 Hz, to avoid any compression effects above the artificial peak at 16 Hz. The results, which are very similar and show fewer compression artefacts, are shown in Fig. 3.14.

Having access to the power spectra for each pixel, we generated a series of images showing the power of each pixel at a particular frequency, from the DC component to the Nyquist frequency of 25 Hz. Each of these images shows the power of each pixel signal at the key frequency. The entire video is available on the accompanying CD (**PixelFFTFrequencyVideo.avi**). A selection of still images are shown in Fig. 3.11, as is the colour legend corresponding to the video. We can see that at low frequencies, there are peaks in power near the logs and in the gaps between them (the source of combustion). At higher frequencies, most power is present in the hot gas rising from the logs. This pattern shows that different frequencies dominate in each area, a pattern which the visual system could exploit when scanning a stimulus for encoding. At each frequency, we note a smooth variation of power in the flame area.

3.2.5 3D Fourier transforms

Reducing the dataset to a 1D signal allows us to characterise its temporal spectrum as exponential; expressing individual frames in the Fourier domain shows their spatial structure. In order to look at both spatial and temporal oscillations, we perform a 3D Fourier transform of the 5000-frame dataset. Images were first converted to greyscale using MATLAB's **rgb2gray()** function. The Fourier transform produces a 3D volume which we can render to 2D in various ways: t-slices, x-slices and y-slices, as shown in Fig. 3.15. A t-slice at a particular time point shows us the spatial frequencies present at that time; a slice at a particular x-coordinate shows us the frequencies present in the temporal domain and the vertical domain at that x-coordinate, and slicing at a particular y-coordinate shows us the temporal and horizontal frequencies at that y-coordinate. We note that pixels near the centre of the x- and y- slices do not correspond to information near the centre of the original images, but to information

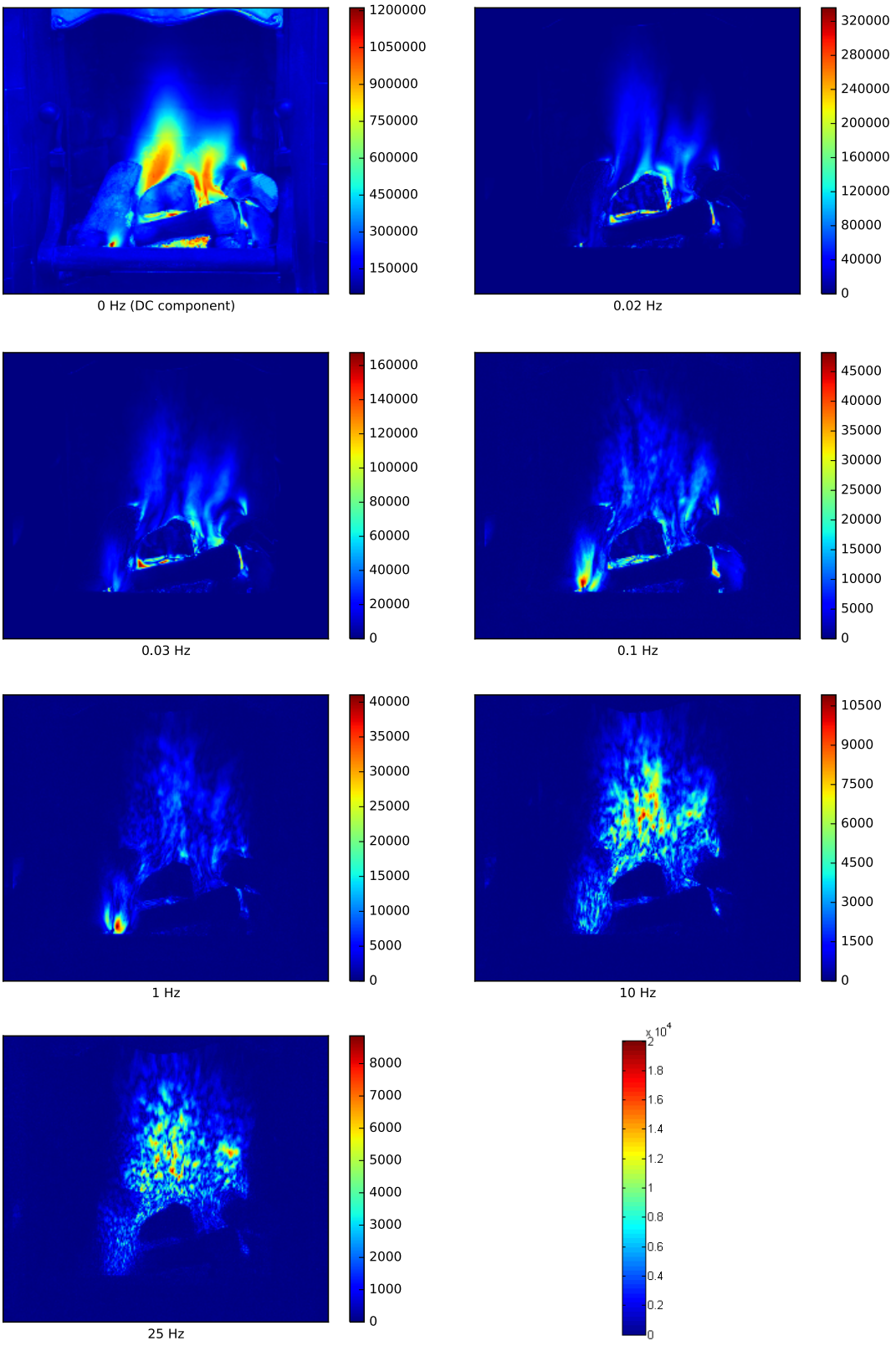


Figure 3.11: A selection of still images showing the power of each pixel at specific frequencies. These were generated from the individual-pixel FFTs. The DC component corresponds to the average image. At low frequencies, power is concentrated around the edges of the logs and in the gaps between them. As we move up to higher frequencies, power is located more towards the hot gas rising from the logs. Bottom right: colour legend for the video **PixelFFTFrequencyVideo.avi**.

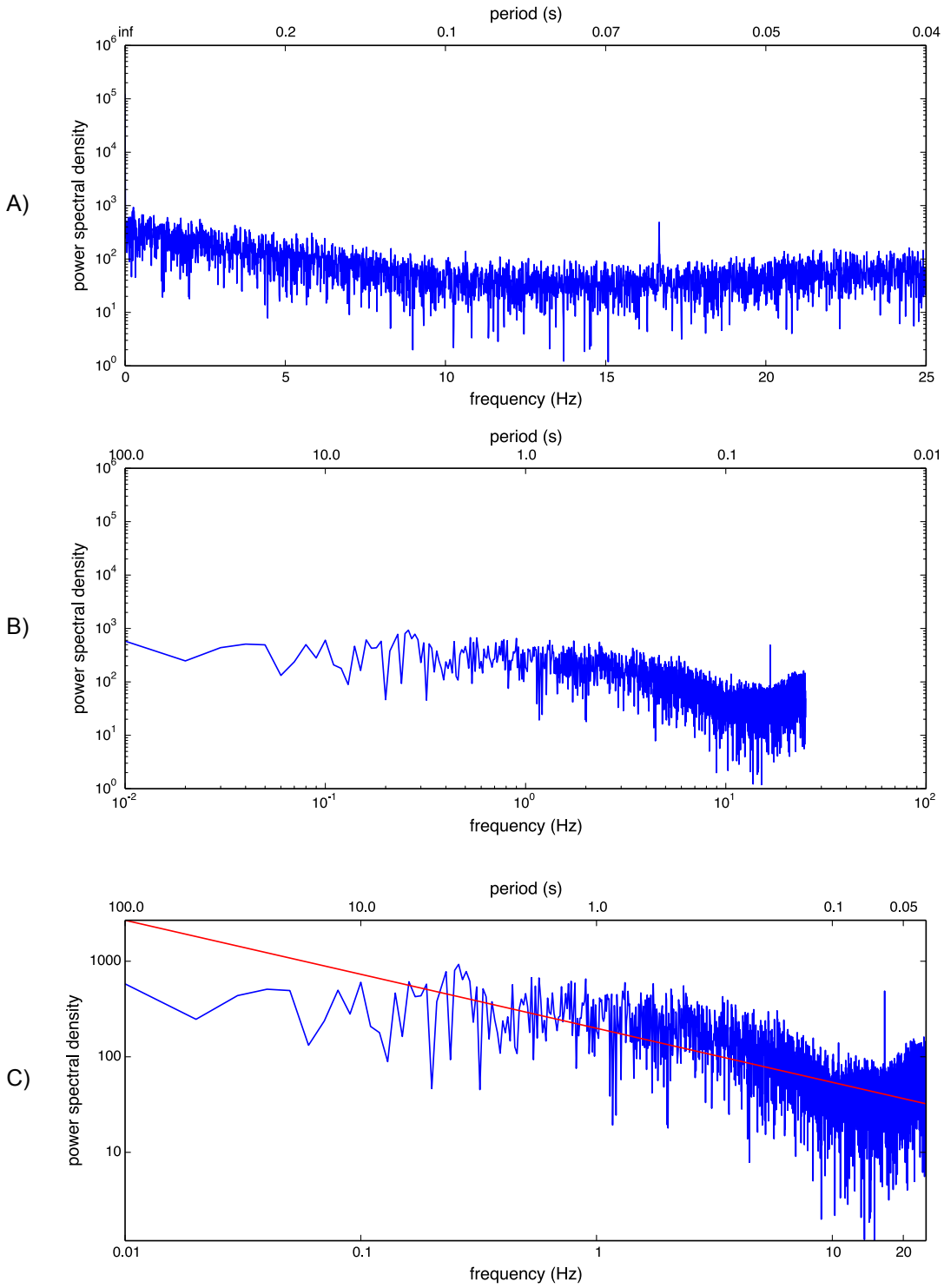


Figure 3.12: A) Power spectrum of the FFT of an individual pixel near the centre of the frame, in lin-log space. B) The same spectrum in log-log space. The spectrum is approximately linear up to about 12 Hz, fitting with $1/f$ noise. Between 12 and 25 Hz we see more power than we would expect for $1/f$ noise. There is a peak at about 17 Hz which is likely a characteristic of AVCHD compression. C) A line fit to the spectrum in log-log space: the fit is not as good as for the lin-log plot between 0 and 10 Hz.

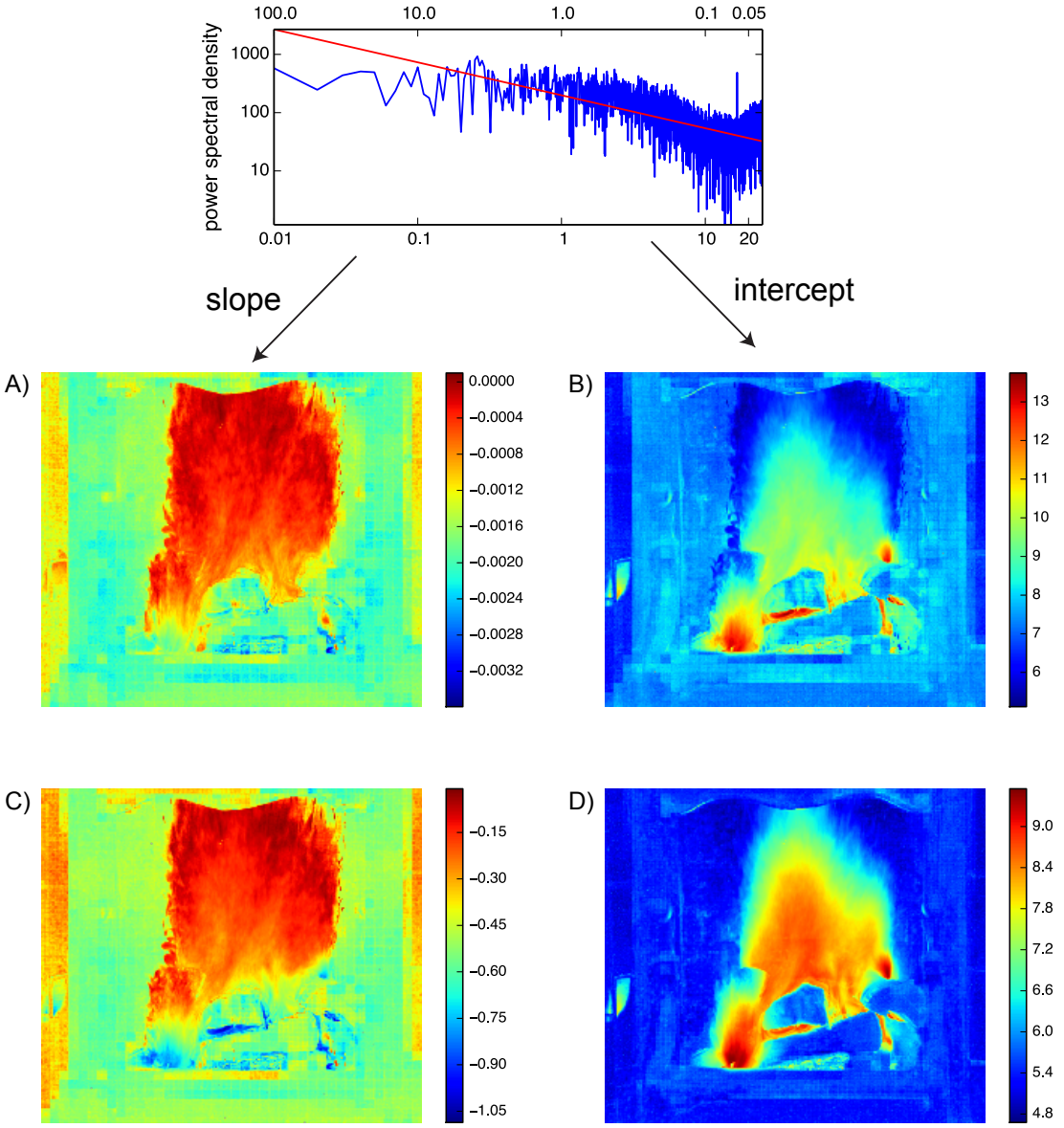


Figure 3.13: A 1D FFT was performed independently on each pixel of the 5000-frame (100-second) dataset. A line was fit to the power spectrum (0 - 25 Hz) in log-log space and its slope and intercept were recorded. A) Slope: lower slope means relatively more low frequencies than high frequencies, which we see in the static parts of the image. B) Intercept: higher intercept means more power in the lowest frequencies, which we see at the base of the flames. There is gridding due to AVCHD compression, which has more effect in the high frequencies. Thus, a line was fit to the first half of the power spectrum (0 - 12.5 Hz) in log-log space. C) The slope of this line shows a similar pattern. D) The intercept of this line shows less effect of compression. The base of the flame shows more power in the lowest frequencies.

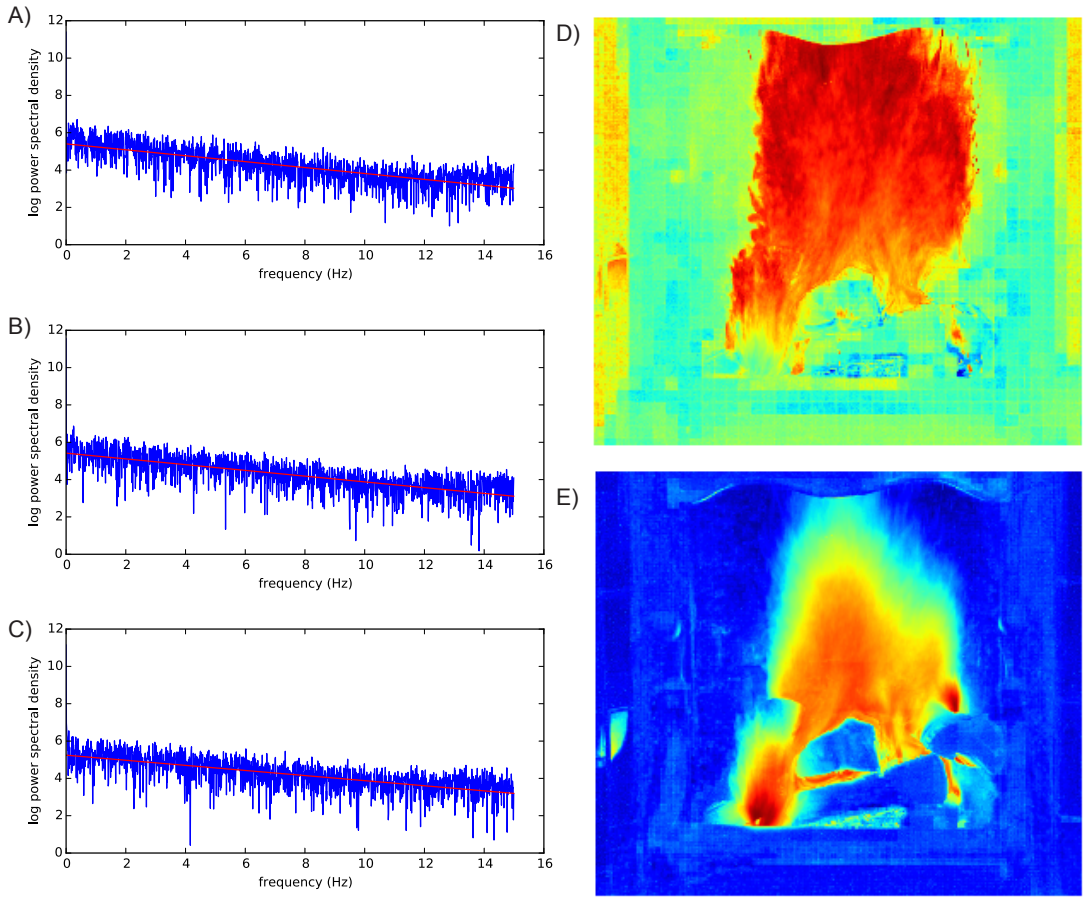


Figure 3.14: Line fits on the individual pixel spectra were repeated in semi-log-y space rather than log-log space. A,B,C): Three example pixel spectra are shown. D) Slope image. E) Intercept image. The same patterns are apparent.

about low vertical and horizontal frequencies respectively.

The 3D FFT is most useful because it allows us to see how spatial frequencies vary across different temporal frequencies, rather than looking at individual time points and averaging them up. Fig 3.16 shows a selection of characteristic spatial spectra (t-slices). The DC component shows us the temporally static spatial frequencies: we note a strong vertical line due to edge effects. This is only present in the DC component, however: the two spectra on either side of the DC component (showing the most slowly oscillating frequencies) do not show this line. They do however show the asymmetric X pattern previously noted, showing more power in directions slightly offset from the vertical and horizontal. As we move away from the DC component, looking at spatial frequencies which oscillate faster temporally (slices 3750 and 1250), we note a degradation of this pattern; these spectra do not show the X, and have more power in the horizontal directions. The same pattern is found at the highest

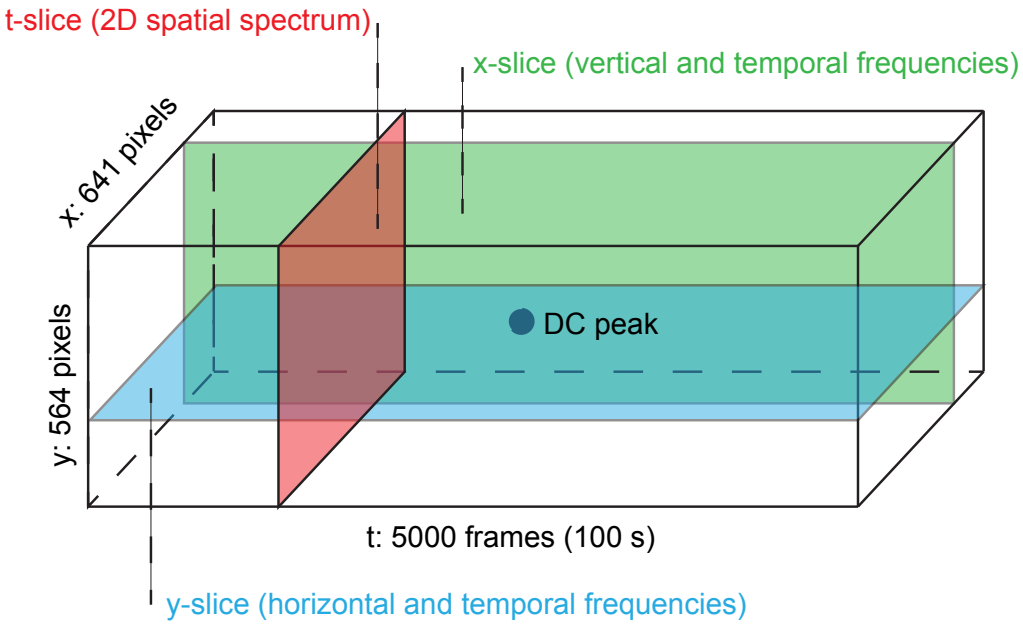


Figure 3.15: Result of a 3D Fourier transform of the image stack. The DC peak is in the centre of the volume; we can produce either y-slices (which show the temporal and horizontal frequencies at a particular vertical frequency), x-slices (which show the temporal and vertical frequencies at a particular horizontal frequency) or t-slices (which show the horizontal and vertical power at a particular temporal frequency).

frequencies (slices 1 and 5000, at the Nyquist frequency of 25 Hz).

We produced three videos (see CD in Appendix A) which run through the t-slices, x-slices and y-slices:

FFT_tSlices.avi The x axis shows horizontal power, the y axis vertical power. We can see that power is mainly horizontal at high temporal frequencies, but shows a characteristic x-shape at low temporal frequencies.

FFT_ySlices.avi The x axis shows horizontal power, the y axis temporal power. There is a horizontal line at the temporal DC component, which is an artefact of the lack of temporal windowing. We can see that at low vertical frequencies, power is concentrated near low horizontal frequencies.

FFT_xslices.avi The x axis shows vertical power, the y axis temporal power. There is a horizontal line at the temporal DC component, which is an artefact of the lack of temporal windowing. We can see that at low horizontal frequencies, power is concentrated near low vertical frequencies.

This analysis characterises dynamic flames as a spatially and temporally complex stimulus, with no particular frequencies carrying most of the power. To remove spatial

edge effects, we now repeat this analysis with a circular Gaussian window applied to each frame.

3.2.6 3D FFT with Gaussian window

Since the Fourier transform treats input images as if they are tiled on an infinite plane, unaltered images can cause edge effects: frequencies appear in the spectrum which correspond to the "false edges" produced by placing two copies of an image next to each other. To eliminate these effects, we faded out the edges of the images. We applied a circular Gaussian window to each image in the 3D image stack and performed a second 3D Fourier transform. Each frame was windowed individually and there was no windowing in the time dimension. Fig. 3.19 shows the windowed stimuli and results of this analysis. We note an absence of vertical lines in the spectrum, confirming edge effects as their source. We make the same observations as with the non-windowed stimuli: at high frequencies, more horizontal power is present, but at low frequencies power is spread in two near-vertical directions (as shown by an X shape in the spectrum).

As with the previous 3D FFT, three video clips showing the $t$slices, x-slices and y-slices are included on the accompanying CD.

The videos confirm the observations made with the unwindowed 3D FFT, confirming that none of them are due to edge effects.

FFT_Gaussian_tSlices.avi The x axis shows horizontal power, the y axis vertical power. We can see that power is mainly horizontal at high temporal frequencies, but shows a characteristic X shape at low temporal frequencies.

FFT_Gaussian_ySlices.avi The x axis shows horizontal power, the y axis temporal power. There is a horizontal line at the temporal DC component, which is an artefact of the lack of temporal windowing.

FFT_Gaussian_xslices.avi The x axis shows vertical power, the y axis temporal power. There is a horizontal line at the temporal DC component, which is an artefact of the lack of temporal windowing.

Fourier transforms give us information about the power present in different areas of the spatiotemporal frequency domain; we analysed the a large 5000-frame dataset. Human observers, on the other hand, perceive restricted areas of the stimulus (up

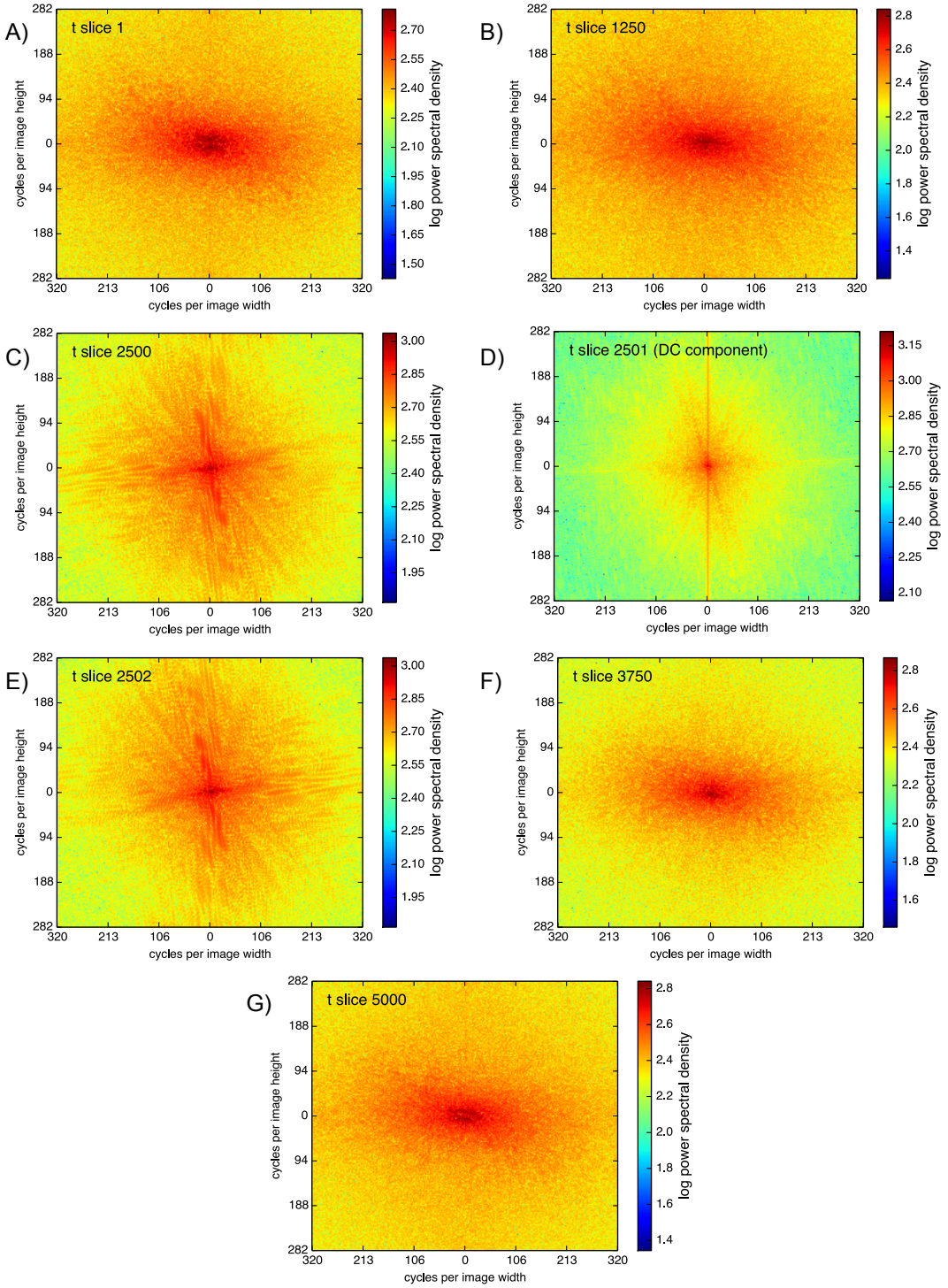


Figure 3.16: t-slices of the image stack FFT, showing the spatial frequencies at a particular temporal frequency. A) At the highest temporal frequency, horizontal power predominates. B) Halfway between the highest frequency and the DC component. C) Close to the DC component, at very low temporal frequencies, we rediscover the tilted cross pattern, showing more horizontal and vertical frequencies. D) The DC component shows the constant spatial frequencies. The X pattern is maintained, along with a vertical line which is due to edge effects. E,F,G): the same observations are repeated.

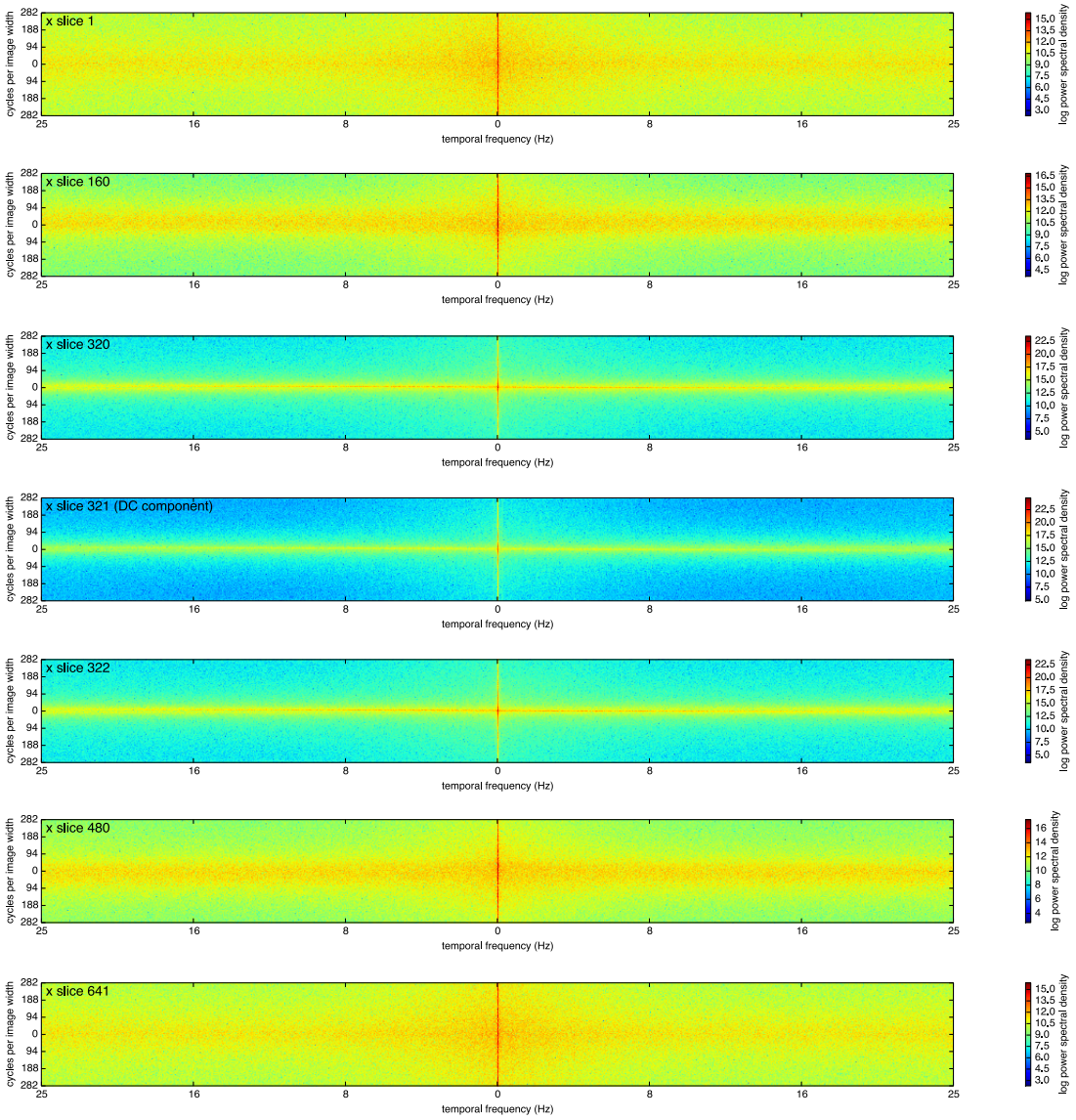


Figure 3.17: x-slices of the image stack FFT, showing the spatial and vertical frequencies at a particular horizontal frequency. At low h-frequencies, near the DC component, power is more tightly concentrated around low v-frequencies. At high h-frequencies, power is more spread out across higher v-frequencies and there is a higher temporally static component (red vertical line).

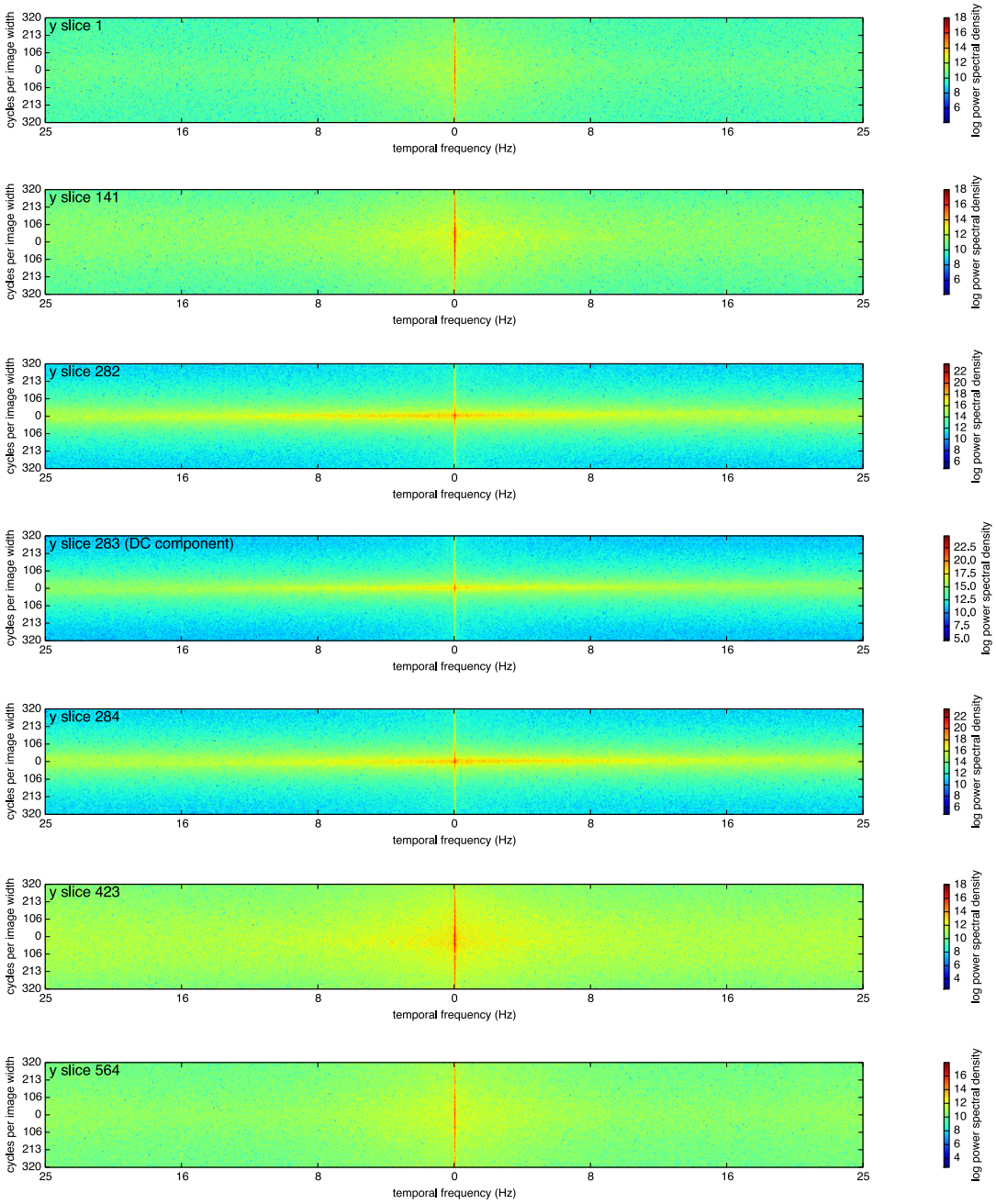


Figure 3.18: y-slices of the image stack FFT, showing the spatial and horizontal frequencies at a particular vertical frequency. We find a similar pattern to the x-slices (for low v-frequencies, there is more power in low h-frequencies). Near the DC v-frequency, there seems to be more power in the low t-frequencies than we find in the low t-frequencies near the DC h-frequency (see x-slice figure, where the horizontal marks extend further outwards). This indicates that long vertical oscillations move faster in time than long horizontal oscillations, which fits with the vertical displacement of gas.

to several seconds) and extract a limited amount of information which is useful for matching. In dynamic flame, motion percepts are particularly strong. In the next section, we look at the results of applying motion algorithms to dynamic flame.

3.3 Motion

Motion is one of the core percepts of visual experience. Moving shapes carry very useful information, whether they are facial features[106] or body parts[128]. There are several brain areas strongly implicated in motion processing[177, 178]. Information about motion appears to be a key part of the representation of visual scenes.

Natural scenes, which are likely to contain animals and non-rigid plants, are full of motion. Interestingly, not much attention has been paid to the motion statistics or percepts of natural scenes[179]. This is partially because dynamic natural scene stimuli are difficult to acquire; most studies are performed using still images.

The same is true for theories of object recognition; most models deal with the classification of still images and do not deal with moving features or changing objects. The representation of moving objects poses a challenge for current theories of recognition.

Flame is a stimulus typical of natural scenes: it is chemical in origin and complex in form. In this section we investigate the motion properties present in dynamic flame, a complex, fast-moving phenomenon made up of rapidly shifting shapes.

What do we perceive when we mentally picture a flame? Observers often report upward motion, and this is corroborated by features whose motion is easy to measure: sparks, which are trivial to track as they move upwards. These details are relatively rare, however, and a substantial motion percept is also available from the luminance gradients in the flames and the moving edges formed by their outlines.

3.3.1 Optical flow

When watching a video of flame, observers predominantly report upwards motion. Does this match with the motion fields produced by optical flow techniques? We investigated using two models of motion perception: the multichannel gradient model (McGM) and a method due to Sun, Roth and Black[180].

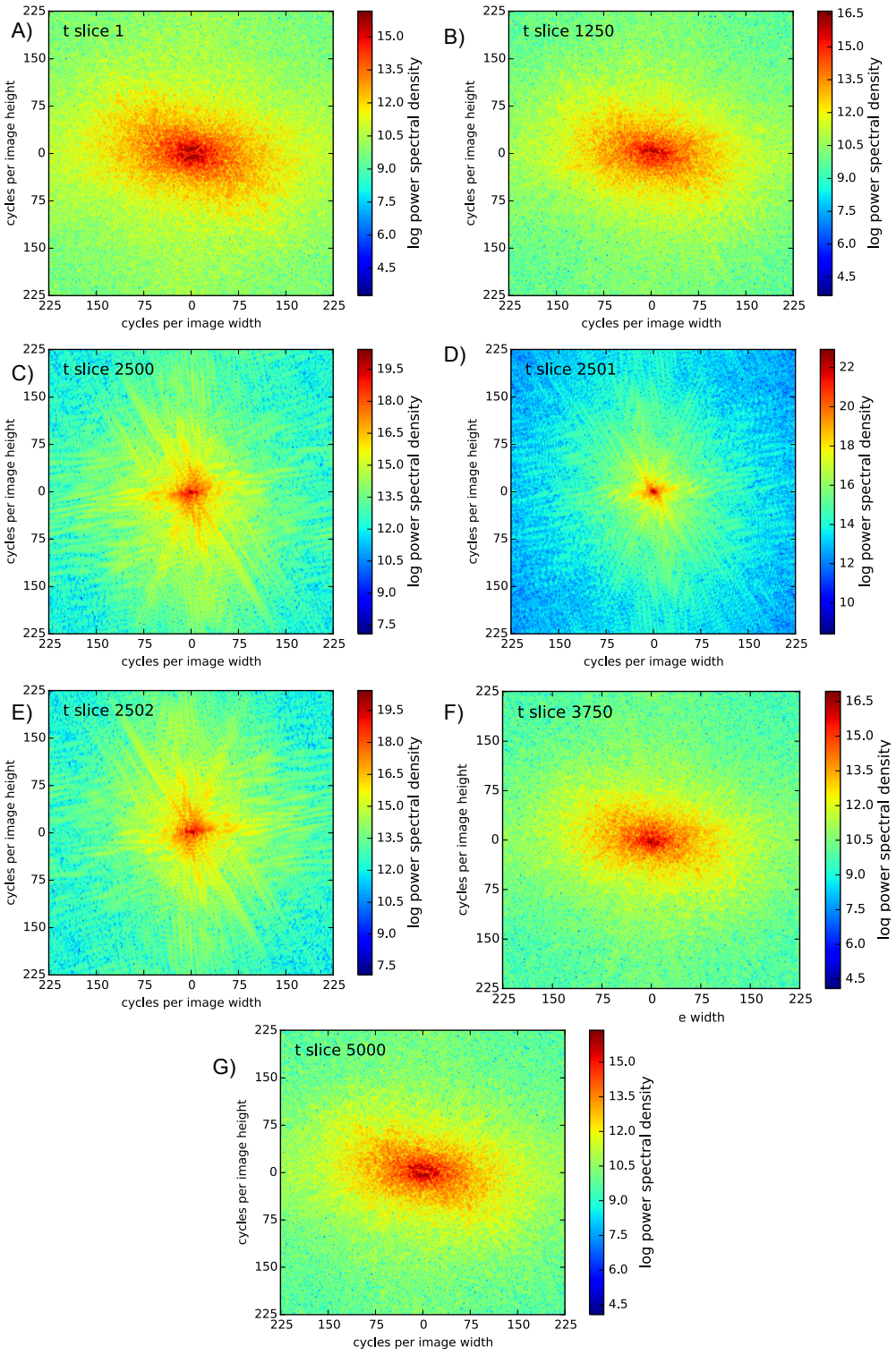


Figure 3.19: t-slices of the Gaussian-windowed image stack FFT, showing the spatial frequencies at a particular temporal frequency. A) At the highest temporal frequency. B) Halfway between the highest frequency and the DC component. C) Close to the DC component. D) The DC component shows the constant spatial frequencies. E,F,G): the same observations are repeated, since FFTs on real numbers are symmetrical. We note the same patterns as for the non-windowed images, but without vertical lines, which confirms edge effects as their cause.

3.3.2 Multichannel gradient model

The McGM is a bio-inspired spatiotemporal gradient model[181, 182] which has performed well under changes in illumination and interference from static patterns[183]. Here we employed a reduced form of the standard McGM which employs a log Gaussian temporal filter and its first and second derivatives. We used a dataset which contained 1000 consecutive flame images 564 pixels high and 641 pixels wide. The McGM was applied separately to each pair of sequential images, creating 999 full optical flow fields of the same size as the images.

3.3.3 Sun et al model

This model, described fully in[180], uses the basic optimisation methods of Horn and Schunk[184] and Black and Anandan[185]. Given two images, the components u and v of the optical flow field are chosen to minimise an objective function. Sun *et al* use several additional techniques including a coarse-to-fine estimation pyramid, interpolation and median filtering. This model was at the top of the Middlebury optical flow technique rankings[186] in 2010.

This model assumes (after Horn and Schunck) brightness constancy and spatial smoothness. It does not assume oriented smoothness, rigidity constraints or image segmentation[180]. Regularisation is employed (adjusted for good results on the Middlebury evaluation) and median filtering is used in post-processing.

3.3.4 Results and model comparison

As with the McGM, we applied the Sun *et al* model to each consecutive pair of frames of a 1000-frame dataset, producing 999 optical flow fields of the same size as the images. Some example fields are shown in Fig. 3.20. We use the same direction-magnitude colour coding as the Middlebury motion evaluation[187].

The McGM is characterised by a detailed map of small areas moving in different directions, with nearly as much activity around the edges as in the centre. Sun's method, on the other hand, detects most motion in the middle, with hardly any around the edges. It tends to report a larger blob of consistent motion.

The mean flow fields produced by the two methods are shown in Fig. 3.21. The

two methods agree about the main direction of motion, both reporting predominantly upwards motion.

To investigate further, we binned the motion directions into 40 equally sized directional bins around the unit circle. We then produced two histograms: one showing the number of pixels whose motion direction belonged to that bin, and one showing the total magnitude of all the motion vectors belonging to that bin. Each histogram was computed across all 999 optical flow fields, once for the McGM and once for Sun's method.

The results are shown in Fig. 3.22. The McGM detects motion in all directions, with a bias towards the four cardinal points and a very slight bias towards the horizontal directions over the vertical. This pattern is borne out when the count histogram is scaled by the magnitude of each vector.

In terms of the number of vectors associated with each direction, Sun's model detects more motion in the downwards direction. However, when we scale by the magnitude of each vector, there is an overwhelming bias towards upwards motion and rightwards-upwards motion. The difference between the count histogram and the magnitude-scaled histogram shows that Sun's method detects many downwards-pointing motion vectors with very low magnitude.

Why are these two models inconsistent? Sun's method works by iteratively solving equations which assume brightness constancy and spatial smoothness, which are not valid assumptions for flame. The McGM, on the other hand, works in a more biologically inspired manner: spatial filters produce a Taylor expansion of the local derivatives of image points, and a number of direction-sensitive detectors operate separately. The McGM is not set up as an optimisation problem: it is purely feedforward, and uses little post-processing apart from thresholding. Despite these differences, they produce similar results on simple visual stimuli such as gratings or stereo pairs.

The inconsistency between two highly-performing methods, the McGM and Sun's method, show that optical flow methods are not well-suited to describing the complex motion percepts obtained from fire. With artificial optical flow stimuli, we can obtain the motion ground-truth by recording object displacement or imaging hidden fluorescent textures[187]. Since fire is not a rigid moving surface, however, there is no ground-truth available for its motion.

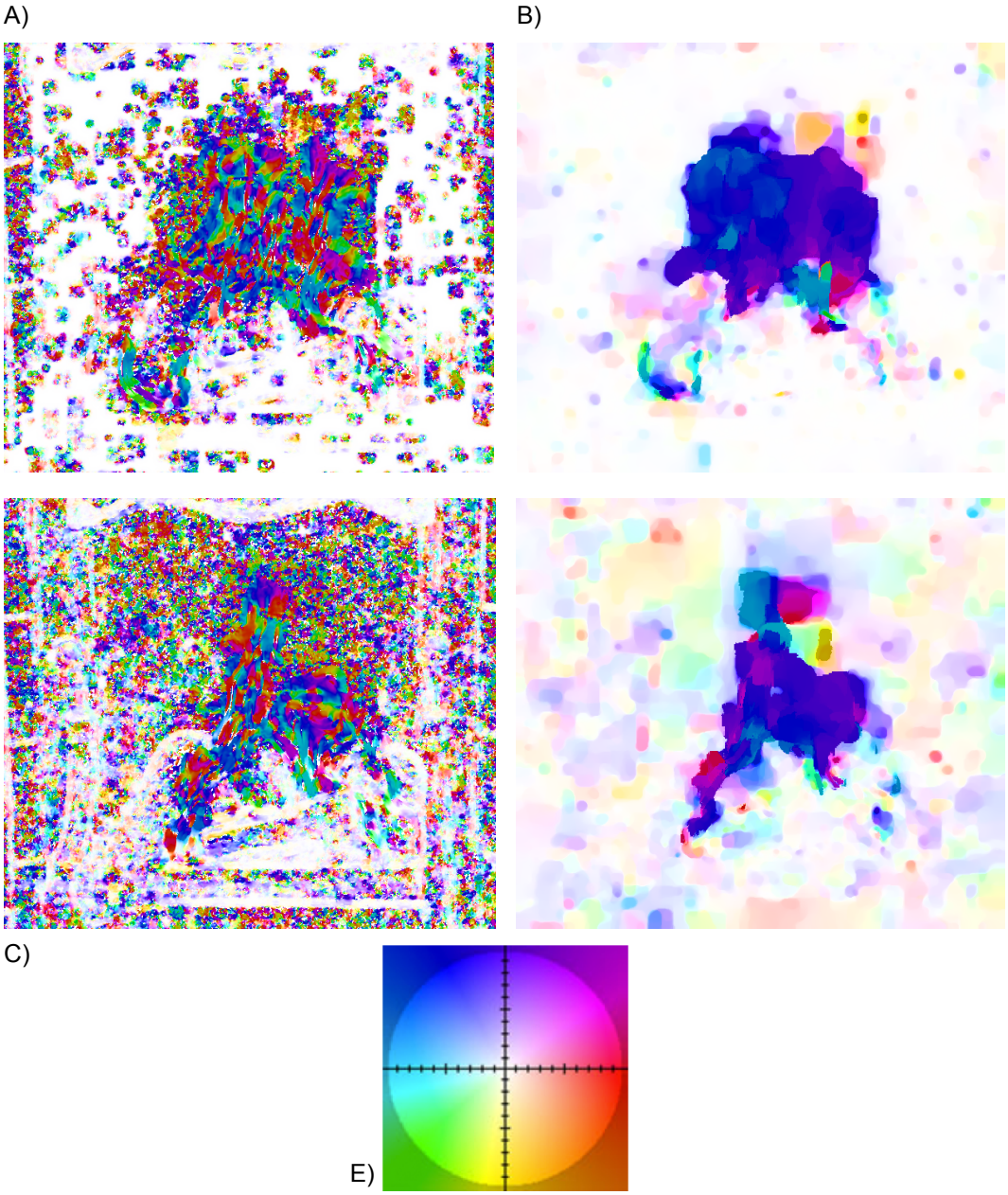


Figure 3.20: A,C) Flow fields computed between two pairs of consecutive images by the McGM. This method is sensitive to noise (near the edges of the frame) but detects separate motion directions for the flames near the centre of the image. B,D): Flow fields for the same pairs of images, computed by Sun's method. There is less sensitivity to noise in the edges of the frame, but this method tends to show a common direction of motion for the flame patterns near the centre. E) Colour wheel showing direction as hue and speed as intensity (after the Middlebury motion evaluation project).

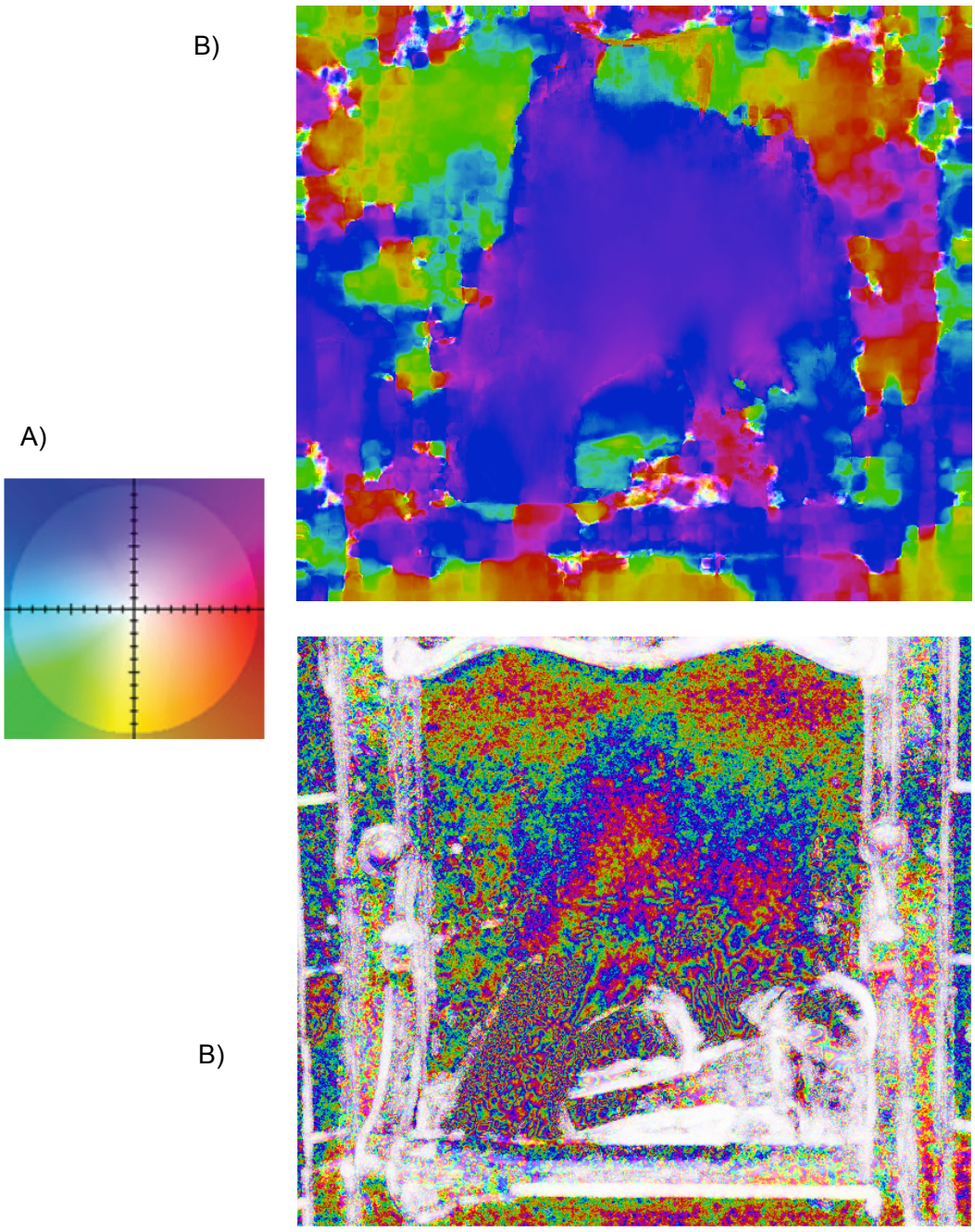
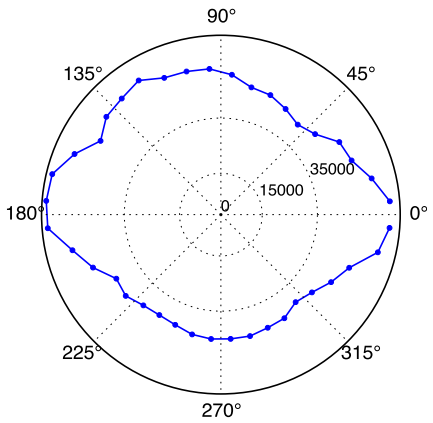
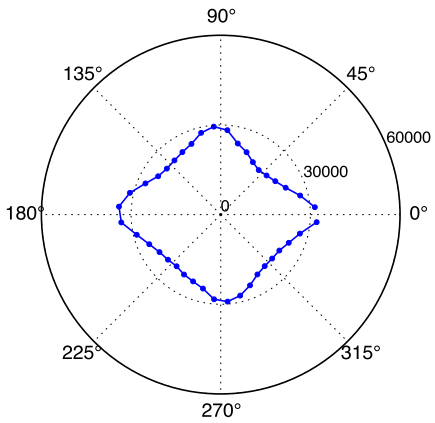


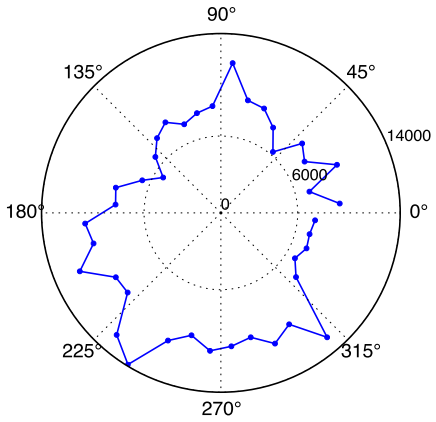
Figure 3.21: Mean flow fields produced by two motion algorithms from 1000 consecutive frames (999 motion fields). A) Colour wheel showing direction as hue and speed as intensity (after the Middlebury motion evaluation project). B) Mean flow field produced by Sun's method. There are plenty of areas which agree in their motion estimation, indicating that this result may depend on the static structure of the image. B) Mean flow field produced by the McGM. Local areas of motion in different directions have cancelled out, giving an overall slight downwards motion. We note some areas of motion downwards and to the left (green) around the borders of the logs. The McGM is also influenced by the background: the logs and fireplace are perceptible as static (white) areas.



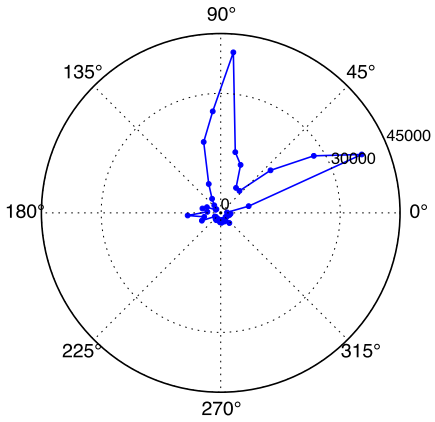
A. Histogram of motion directions (counts only) computed by the McGM



B. Histogram of motion directions scaled by magnitude, as computed by the McGM



C. Histogram of motion directions (counts only) computed by Sun's method



D. Histogram of motion directions scaled by magnitude, computed by Sun's method

Figure 3.22: Circular histograms of the directions of motion estimated by the McGM and Sun's method over a 1000-frame dataset. Pixels were sorted into 40 bins by direction of motion. A,C) The graph shows bin counts only, showing the distribution of directions. B,D) The graph shows the total magnitude of all vectors in each bin, showing the quantity of overall motion in each direction. Overall, the McGM finds slightly more horizontal motion than in other directions, while Sun's method finds much more motion upwards and to the right. The two algorithms disagree.

3.3.5 Four-dimensional McGM model

Both the McGM and Sun's method take only two images as input; they are thus unable to make use of temporally non-local information, as can the human visual system. To overcome this problem, we applied a version of the McGM which operates on an image stack: the sMcGM. This operates in a similar manner to the two-frame McGM, with more than two images available as input to the temporal filter operations used to calculate the velocity of each pixel.

We applied the sMcGM to a stack of 500 images, using a temporal filter size of 23 frames (contrast to two frames previously). An example flow field is shown in Fig. 3.23; a video is present on the accompanying CD (**sMcGMDynamicFire.m4v**). The top left panel shows blurred input images; the top right shows direction view (speed is ignored); the bottom left shows a speed view (direction is ignored); the bottom right shows a combined view with hue indicating direction and saturation indicating speed.

Fig. 3.24 shows the mean flow field found by the sMcGM. It is fairly coherent in terms of direction, consisting mainly of a smooth flow field pointing upwards.

3.3.6 Applying the four-dimensional McGM to edge-filtered images

Edge filtering is a key task of the early visual system. We show experimentally in Chapter 4 that humans are capable of encoding and matching an edge-filtered version of the flame database; see Fig. 4.3 for example images. This indicates that the visual system is encoding dynamic form as well as local, low-level motion signals. We wondered how well the sMcGM, whose temporal integration window makes it the most capable model we studied, responds to edge-filtered images.

We repeated the previous analysis, running the sMcGM on a series of 500 edge-filtered images. The average flow field is shown in Fig. 3.24; black areas represent pixels where NaN was returned, while white areas represent pixels with very small motion vectors.

The edge-filtered data contain enough motion signals to give a central zone with mainly upwards motion. The edges of this area, however, show motion towards the right (red band). We did not observe this in any other motion analyses; it does not

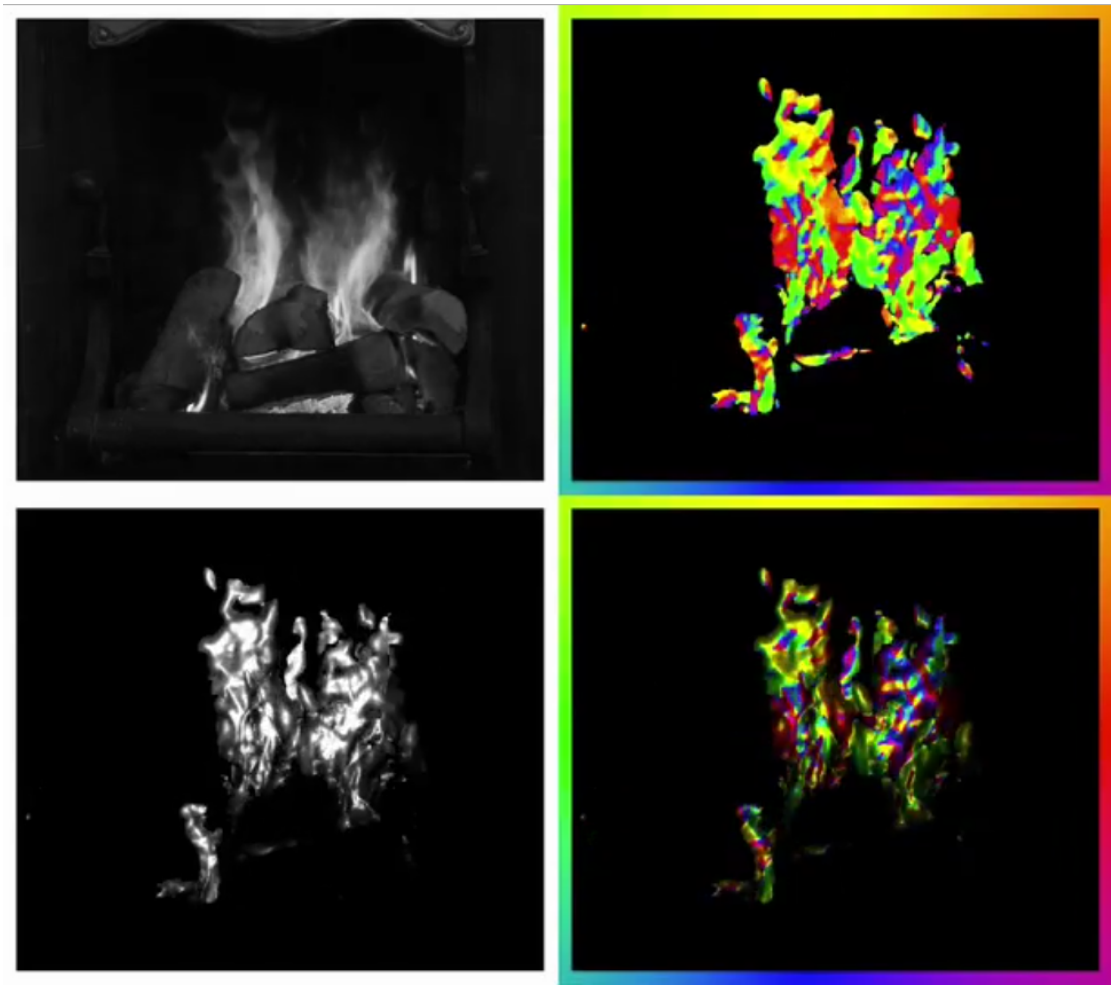


Figure 3.23: An example flow field produced by applying the sMcGM to a stack of standard flame images. Top left: temporally filtered version of the stimulus. Top right: a single flow field (from one frame to the next, but taking into account temporally local frames); direction is shown. Bottom left: speed of the same field. Bottom right: direction scaled by speed, for the same field.

appear to be an artefact of the sMcGM, since it was absent from the previous analysis, and it cannot be introduced by the edge filter since we convolved the original images with a horizontally symmetric filter (see Chapter 4).

This analysis shows that a low-level motion evaluator is able to recover the correct overall motion direction from edge-filtered dynamic flame. Edge information alone can induce a reasonable motion percept, and since our two motion algorithms do not explicitly treat displaced form, this result does not suggest that the brain must rely on computing displaced form to obtain a directional motion percept from flame.

3.3.7 Is dynamic flame drift-balanced?

Drift-balanced stimuli are usually created to defeat the correct perception of motion by motion energy models. Precisely, a drift-balanced stimulus is one whose power in the frequency domain is symmetric with respect to temporal frequency: if every spatially oriented oscillation is equal in power to its oppositely-oriented counterpart. A microbalanced stimulus is one in which every spatiotemporal region, considered separately, is drift-balanced[98]. The human ability to perceive motion in non-drift-balanced stimuli has been proposed as evidence for multiple motion processing systems[95], although some authors disagree[99]. Observers certainly perceive motion in dynamic flame; can we address the question of whether it is drift-balanced?

The 3D power spectrum produced by a full-stack Fourier transform shows no obvious asymmetry to the eye, but it is irrelevant as it is produced from a 5000-frame sequence which observers never see in full. Given the large amount of oriented gradients present in dynamic flame, it is difficult to imagine that it is drift-balanced. Such stimuli require care to construct and it is extremely unlikely they will come about by chance in the natural world.

In Chapter 4 we create edge-filtered versions of flame stimuli which contain no luminance gradients. Although not explicitly drift-balanced, these videos are much closer to displaced form stimuli and would not be expected to activate traditional motion energy models very strongly. They nevertheless elicit motion percepts in both human observers (who report a sense of upwards motion as in unfiltered images) and a gradient-based motion model, the sMcGM.

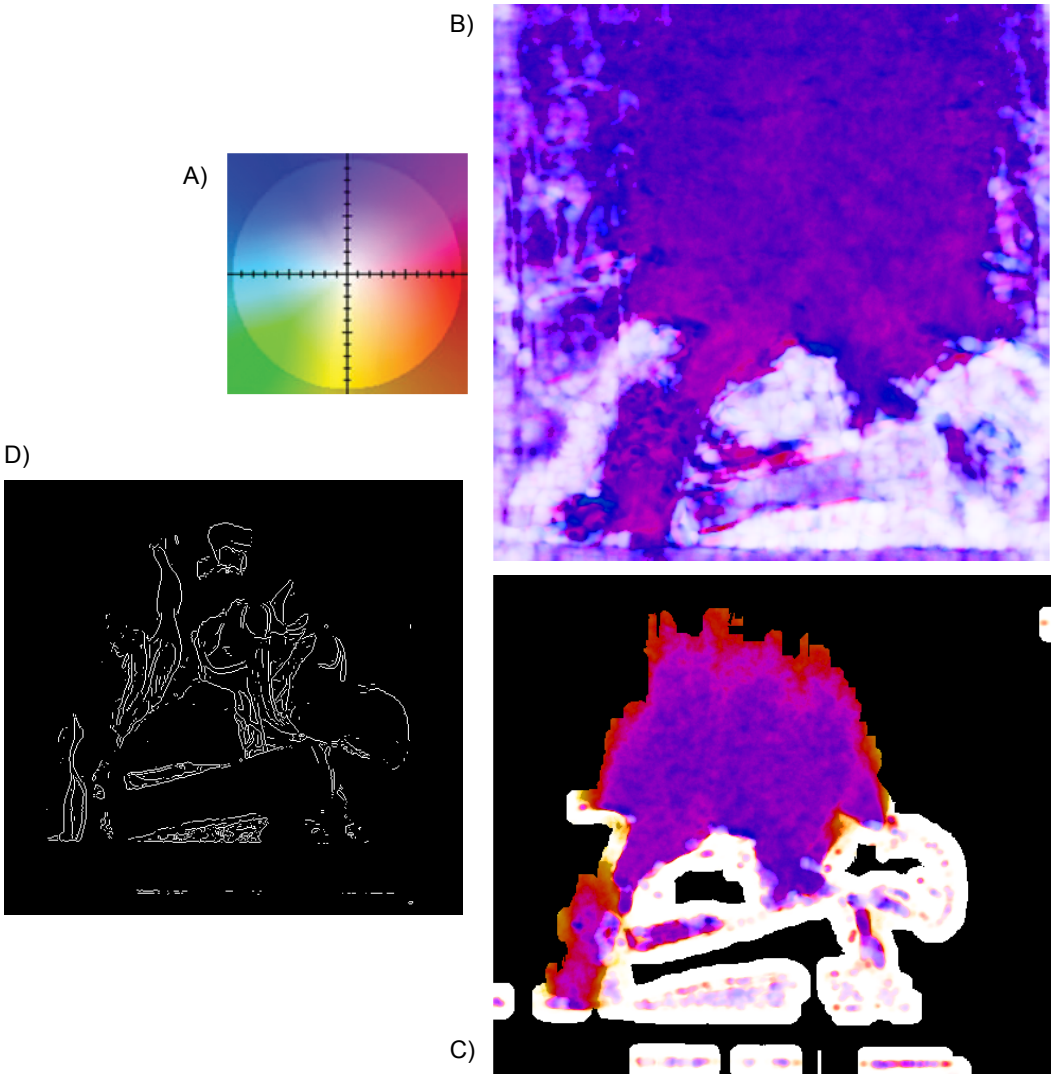


Figure 3.24: Mean flow fields produced by the sMcGM, which operates on a stack of images. A) Colour wheel showing direction as hue and speed as intensity (after the Middlebury motion evaluation project). B) Mean flow field produced by the sMcGM from a standard dynamic flame video. There is very little motion around the logs, the mean flow field is smooth, and the main direction of motion is up. C) Mean flow field produced by the sMcGM from edge-filtered images. Because most of an edge-filtered image consists of black pixels, the algorithm has not returned any motion estimations for the edges of the image. Around the logs, the mean flow field consists of very small vectors (white areas). The algorithm has determined a sensible (upwards) direction of motion from the dynamic flame area (purple); there is a red crest, which may indicate rightwards motion of the flames' top edges. D) An example edge-filtered image (see Chapter 4 for filter details).

3.4 How can we encode fire?

We can learn about a new stimulus by trying to build a computational model which effectively encodes it. Here we review our efforts to effectively represent clips of dynamic flame in high-level spaces using PCA, a morph model, and a dynamic texture synthesis algorithm due to Doretto[188]. Finding an effective way to reduce the dimensionality of a stimulus supports the idea that the brain may do this as well, as is the case with the popular PCA models of face perception (face spaces). PCA, although a linear technique, can construct a low-dimensional face space allowing near-photorealistic face generation. When computational efforts fail, they can inform us about the challenges posed by the stimulus.

Any video clip can be effectively encoded as a plain list of pixels. The kinds of encodings we are really interested in are those which allow us to compress a stimulus, to reduce its dimensionality; since this is also the job of the visual system, these encodings suggest ways in which a stimulus might be represented neurally. The successful application of PCA to co-registered face images, for example[55], suggested that face perception may employ a mean-centred, low-dimensional face space.

3.4.1 Principal component analysis

We began by applying plain PCA to a 1000-frame dynamic flame database. Results were not encouraging; reconstructed images were blurred and did not look much like real flame; see examples in Fig. 3.25. We observed similar results even when reconstructing using a large number of principal components (fifty). This highlights an important difference between fire and faces: faces have a general structure which is deformed by the facial musculature to produce an individual expression, whereas the structure of flame is much less constrained.

When PCA is applied to faces, the first few components often match well with high-level characteristics of the image: identity[189], gender[190] or age[191]. It is not surprising that this is not the case for flame, since its variation is mainly within-exemplar as opposed to within-category. However, the observation that the first few flame PCs differ very little is key: the algorithm is not able to effectively extract useful high-level descriptions at all.

PCA also does not support the automatic generation of video sequences in a reasoned way; this must be done manually by moving a point through PC space and generating a series of images from the trajectory.

3.4.2 Morph space PCA

Morph space PCA is a motion-based morph model which decomposes an image into a texture and a warp field before applying PCA[192, 193, 194]. We applied this morph model to our hearth fire dataset. This method gave blurred and unrealistic results, since the morph component was unable to find correspondences between pairs of images, which display very different structure as they contain multiple changing flames. Reconstructed images were similar to those produced by naive PCA. This confirms that PCA could detect no general structure even with the ability to warp and deform textures.

3.4.3 Dynamic texture synthesis

A more complex encoding is found in Doretto's dynamic textures algorithm[188]. This system begins with the observation that a dynamic texture has high-level percepts which are stable over time. Moving clouds or flowing water have rapidly changing low-level features, but their high-level interpretation (clouds or water) does not change; the nature of the texture is stable, even though its pixels are changing. Doretto models this by making a texture a stationary stochastic process on the high level. The low level (synthesized images) are generated from this high-level representation.

Since this model does not depend on storing a common shape and deforming it, and uses a time series rather than generating a sequence of independent images, we hypothesised that it might encode and reconstruct flame more effectively. We used the dynamic texture synthesis model (with default parameters) to produce a series of flame images; results are shown in Fig. 3.26. The images are more realistic, especially when viewed as a video clip; the structures they contain deform smoothly over time. The output still does not look much like real flame, however; it lacks the sharp edges and well-defined shapes of the real flame dataset.

The failure of these three techniques to produce realistic images suggest that more powerful techniques are required to construct a good high-level representation

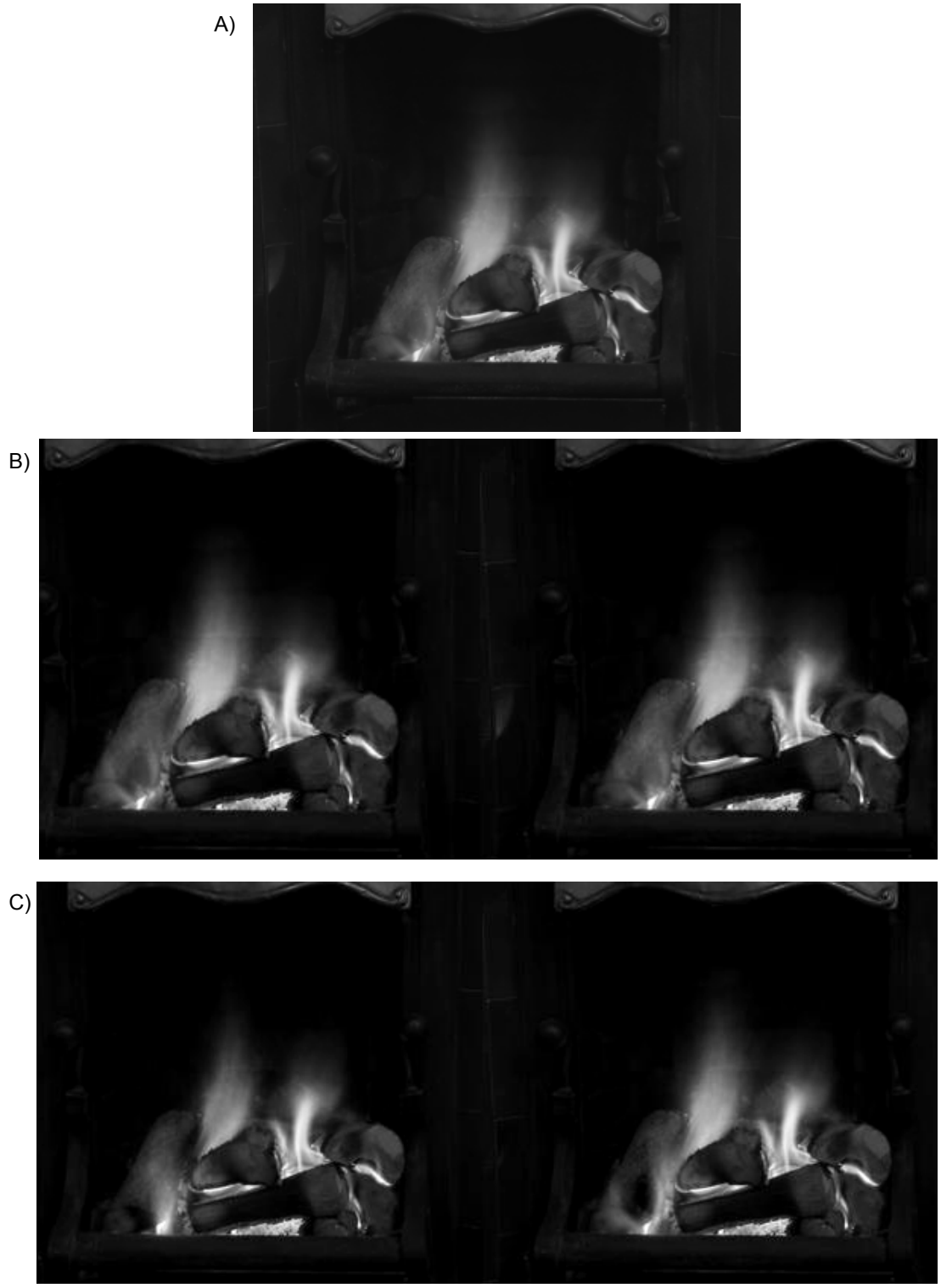


Figure 3.25: The results of naive PCA on a dataset of monochrome flame images. A) The mean image, showing average form. B) Images produced by setting the first two component loadings to -1 standard deviation, with other loadings at zero. C) The corresponding image at $+1$ standard deviation. The principal components are extremely similar and not very informative.

of dynamic flame. There are several low-level constraints such a system would have to satisfy: flames have sharp edges, and successive images tend to be very similar. Linear techniques such as PCA, and fairly unconstrained generative techniques such as dynamic texture synthesis, cannot fulfil these requirements.

Low-quality reconstruction does not prove conclusively that fire cannot be encoded simply, but they provide convincing evidence that finding a high-level description of authentic low-level flame sequences is a difficult problem. In later chapters, we use psychophysical matching experiments to ask whether the brain has access to such a representation.

3.5 General discussion

We have used motion and Fourier analysis to characterise dynamic flame as an unpredictable, rapidly-moving stimulus with little global form and few long-range correlations. Physically, a fire can be decomposed into a static element (logs) and a dynamic element (flames). The static element is of little interest; although logs do move, they do so very little, and we chose the dataset used in our psychophysical experiments to be free of such movement.

Producing the mean of a set of flame images gives us an "average flame" which becomes smoother and less sharp as we increase the number of images. Dynamic flame thus has an average structure; but it is not a base structure in the sense that it can be deformed to accurately reconstruct a single flame. Classes of stimuli such as faces often form an equivalence class; each one can be transformed to accurately represent the mean, and the mean can be transformed to reconstruct each instance. This is not the case for fire; frames which are distant in time are so different that encoding them as a transform from a mean does not save much information over encoding them as single images.

What kinds of visual features can we construct from fire? We used Fourier analysis to express the 1D global brightness signal, individual pixel signals, individual 2D frames, and the entire 5000-frame 3D image stack in the frequency domain. Natural scenes often show a $1/f$ power spectrum, which plots as a straight line in log-log space; the 1D brightness signal appears to be exponential, which shows less power in the high

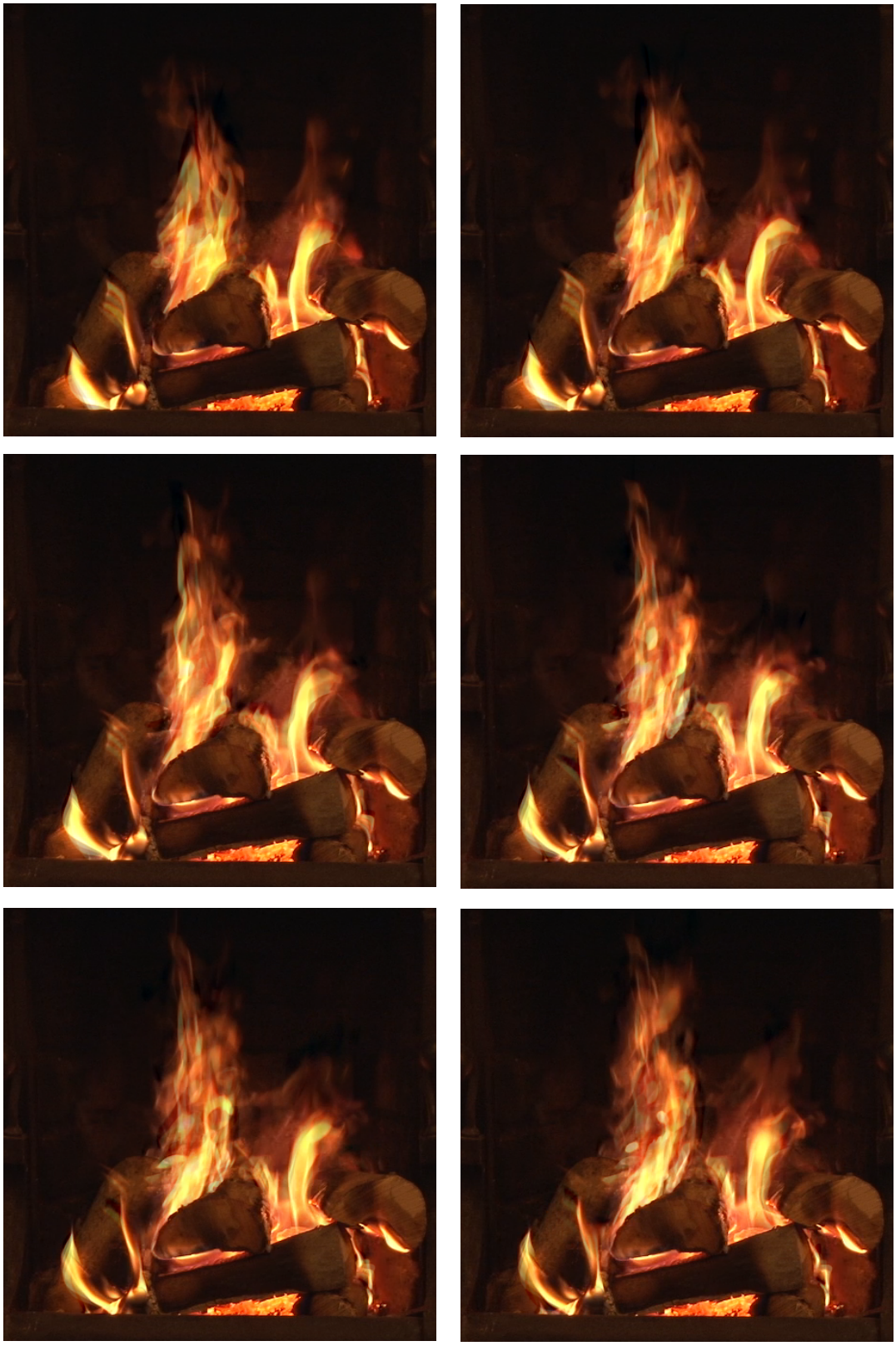


Figure 3.26: Six consecutive frames produced by Doretto's dynamic texture synthesis algorithm. Reconstruction is low quality; flames seem to be overlapping transparencies rather than shapes with definite edges, they have a very similar overall shape (which is maintained even across nonconsecutive images) and an sense of upwards motion is not present.

frequencies (4 to 25 Hz) than a $1/f$ spectrum.

We usually find $1/f$ spectra in complex systems with structure at multiple scales. Exponential spectra have not been reported in natural scenes or any stimuli used in psychophysical experiments, although they have been described in high-energy plasma[195, 196]. The high energies present in flame and plasma may explain why their spectra are more similar to each other than to those produced from lower-energy complex systems which usually produce $1/f$ spectra.

An exponential spectrum displays a smooth variation in power across the frequency range. There is thus no particular spatial or temporal frequency band which contains more oscillations than other areas (apart from a peak at 17 Hz due to AVCHD video compression). This poses a challenge for the visual system, which cannot concentrate on a particular frequency band, as it can for spatial frequency in face perception[46, 197].

What motion features do we find in dynamic flame? We applied two motion evaluation algorithms, the multichannel gradient model (McGM) and Sun's method, a model descended from that of Horn and Schunk. Like the Fourier transform, these methods measure dynamic changes in the image stack, but they are much more local (considering only a pair of frames) and much more perceptually inspired. The two methods disagree, with the McGM constructing small patches of global motion and Sun's method finding larger patches of motion, with little change around the edges of the image, although they find the same overall motion direction. Circular histograms of the directions of motion associated with each pixel also disagree, with Sun's method returning two main directions of motion.

These methods are optimised for slightly different tasks (see [181, 182] for the McGM and [180] for Sun's model), so it is not surprising to find slightly different results. The fact that the results are so different underlines the fact that there is no motion ground truth present in fire. There are no rigid moving surfaces or deforming objects, only a 3D gas cloud which is projected, via the recording equipment, on to the observer's retina.

We applied three modelling techniques which have shown excellent results at reconstructing faces (PCA and morph space PCA) and dynamic textures (DTS). PCA produces reconstructions which do not look much better than the dataset's average

image. DTS captures flame textures well, but fails to properly reconstruct the form of the flame's edges or its dynamics. The failure of these techniques shows that flame is a complex stimulus which is difficult to reduce in dimensionality, and that it has no common structure. This suggests that the brain cannot represent dynamic flame as a deviation from a prototype.

Analysis in the image domain allows us to build up a picture of dynamic flame as a fast-moving, complex stimulus containing fragmented areas of motion in different directions, a layer of static form, smooth texture variations, and shapes with defined edges. Such a stimulus poses a difficult encoding challenge for the visual system. In subsequent chapters, we test observers' ability to match and evaluate clips of dynamic fire against long-standing category representations and short-term memories of similar clips.

Chapter summary

- Flame images have no general structure which can be smoothly deformed to produce individual frames.
- Frames close in time are very self-similar, but become very different after about 0.2 s according to low-level metrics (Euclidean distance, absolute pixel difference and SSIM).
- Dynamic flame sequences show an exponential frequency spectrum.
- Individual flame images, and 2D spectra at low temporal frequencies, show a characteristic spatial spectrum with most power close to the horizontal and vertical.
- Two modern motion algorithms, applied to dynamic flame, disagree in the sizes of patches of coherent motion which they report. They agree, however, on the overall direction of motion.
- PCA, morph space PCA and dynamic texture synthesis are not able to produce effective high-level encodings of dynamic flame.
- Overall, image-based analysis characterises dynamic flame as free of long-range spatial or temporal correlations.

Chapter 4

The features of dynamic flame

When we look at a natural scene, we are able to extract diverse visual features: the type of landscape[198], the objects present[199] and the gist[200]. Some, like the gist, are holistic: they apply to the whole image, and cannot be mapped to a particular position. Other features, like colour, luminance and edge orientation, are more local, and can be evaluated for a small, specific area of an image (as small as a pixel for colour, or a 3×3 pixel patch for edge orientation). Dynamic features, such as motion percepts, are nonlocal in time. In this chapter, we describe an experimental investigation into the features of dynamic flame, aiming to measure their relative importance for matching. We examine colour, orientation, and temporal order, before looking at the importance of motion in observers' category representation of dynamic flame.

What is a "feature?" There are two common usages in the literature. The first, which we term a *local feature*, refers to a spatially restricted (local) part of an image ("the red square is a distinctive feature", and "facial features" such as the eyes and nose), as used in feature integration theory[201]. Pixels are local features, although they are not always individually perceptible. The second, which we term a *nonlocal feature*, refers to general property which can be computed from an image and is not restricted to a local part (such as colour or texture). Colour, luminance and motion are examples; they may be attached to a local feature, such as a red square, but also refer more generally to aspects of the stimulus as a whole, or specific percepts which may be computed from local image patches. In this chapter, we are interested in the importance of colour and motion (nonlocal features) to dynamic flame matching. We are also interested in whether the matching process is invariant to the spatial

arrangement of any local features which may be computed, and to their temporal arrangement.

In this chapter, we are interested in matching, not search; this means that in the delayed-match-to-sample tasks we set observers, the first clip (the sample) is close in length to the second clip (the test). The test is always slightly longer so that first and last frames do not co-occur, allowing an easy route to matching through iconic memory. In the next chapter, we use much longer test clips, allowing us to characterise the search process.

How do we rank nonlocal features in order of importance for matching? When an observer perceives a still image, we call this importance judgement salience[202]. Regions of distinctive colour, for example, are highly salient in natural scenes[203]. Bottom-up salience is computed without reference to task goals or a search target[204]. When an observer performs a matching or search task, however, top-down salience is informed by the representation of the target.

One way to measure the salience of a feature is to examine the effect of distractors. Initial work by Neisser[205] and later by Treisman[201] established a two-stage theory of visual processing: the generation in parallel of a set of basic features, followed by a higher-level serial process. When the target is surrounded by similar objects, some features (such as unique colour or orientation) cause it to pop out independently of the number of distractors[206]. Higher-level features, such as conjunctions of two simpler features, often require a serial search strategy.

How similar must a stimulus be to the target in order to act as a distractor? This question was investigated by Julesz, who conjectured[207] that identical second-order statistics were sufficient to render two textures indistinguishable without close examination. Although eventually proved false [208], this conjecture motivated Julesz' division of visual processing into an automatic, preattentive system and a computationally intensive, spatially local process, focal attention[209].

Natural scenes pose a special challenge here. Since they are not easily decomposable into conjunctions of simple features, like an artificial display of coloured dots, it is harder to characterise the features involved. Complex scenes are also harder to synthesise convincingly. Accordingly, most experiments on natural scenes use recorded images rather than synthesising their own stimuli.

Here we take a hybrid approach: we record authentic video of a natural scene, then manipulate it to alter or completely remove certain nonlocal features. For example, we can remove colour information from a clip by rendering it in monochrome. We then ask an observer to match an original, unaltered clip with an altered clip. We reason that if the observer can accurately decide whether the two clips are the same, the missing feature does not provide evidence which helps with the decision. If, however, the altered clip cannot be accurately compared with the original clip, the missing feature must carry important information which is useful for matching. In this way, we can use a matching task, along with a feature manipulation, to evaluate the importance of that feature for matching.

Which features should we examine? Much of the low-level visual search literature concentrates on colour and shape. Since our scenes are not synthesised, we cannot directly alter the shapes they contain. We can, however, alter the colour: after expressing clips in hue-saturation-value (HSV) space as opposed to the native red-green-blue (RGB) space[210], we can rotate the hue value by 180 degrees, creating a clip whose luminance pattern is perceptually similar to that of the original but whose colour has been radically altered. This representation also allows us to easily alter luminance, another basic nonlocal feature. By inverting the value (V) channel, light areas of an image can be rendered dark (and vice versa) while the hue percept of each pixel is unchanged.

In Chapter 3 we examined, computationally, the motion signals present in dynamic flame. Motion is a broad concept applicable to different levels of processing, as discussed in Chapter 1. Low-level motion can be computed from a very small retinal area by simple computations (Reichardt detectors) or more complex neural processes (cells acting as spatiotemporal filters). Instantly displaced form can constitute motion, as in the phi and beta illusions. Finally, smooth change in the parameters of a morphable object (such as the intensity of a smile or the colour of a feature) can give rise to high-level object-based motion. Which of these kinds of motion form a useful part of observers' representations of dynamic flame?

Another alteration which can tell us much about object perception is inversion: playing a clip upside down. This can be done either by a 180 degree rotation or by a reflection about the horizontal axis. Here we use a 180 degree rotation, for consistency

with the 90 and 270 degree rotations we also apply to stimuli. The study of inversion effects has a long history in face perception research. First investigated by Yin[70], decreased face recognition performance under inversion has been interpreted as a sign of specialisation (whether innate or acquired) to upright faces.

There are really two separate issues at work here. We see the first in the problem of viewing an inverted face, remembering it, and matching it to an inverted face. A performance drop here indicates a lower-quality representation of inverted faces. The second issue is at work in the problem of viewing an upright face, remembering it, and matching it to an inverted face. In this case, accurate comparison requires the observer to be able to compare an upright face to an inverted face: they must have access to a representation which is stable across inversion, or use a process of transformation (mental rotation).

A study of the effect of inversion is also applicable to natural scenes. The first case (both sample and test inverted) is discussed in Chapter 6. In this present chapter we examine the second case: the sample is inverted and the test is upright. Observers' performance here can inform us about the object recognition strategies being used. If we find a small or nonexistent accuracy drop when the sample is inverted, observers are using a representation which is stable across inversion. At one extreme, this could mean that the representation is intrinsically insensitive to inversion (the 1-dimensional mean luminance signal of a video clip, for example, has this property). At the other extreme, it could mean that the observer can perform a computation which effectively compares an inverted representation with an upright one. Alternatively, performance might be compromised by such a process of mental rotation, which would lead to an accuracy drop.

The three manipulations just described (colour alteration, luminance inversion and spatial inversion) can be applied equally to static images or dynamic video clips. We apply these manipulations to clips by independently altering each frame and reassembling the resulting frames into a clip. As such, these manipulations cannot tell us anything about dynamic object perception: how we represent and compare moving objects. As we have seen in Chapter 1, theories of object perception principally consider static images, and dynamic stimuli pose problems for these theories. How well do these theories explain the perception of dynamic flame?

To address this question, we perform a fourth manipulation: temporal inversion, or backwards playback. By asking observers to compare an original clip with its reversed equivalent, we examine whether their representations are sensitive to the temporal order in which frames are displayed. A small or null drop in accuracy here would indicate the involvement of a representation that is not sensitive to temporal order. As with spatial inversion, this could either be due to the representation being unaffected by ordering (a static frame produced by averaging over the frames in a clip has this property, as it is the same whatever the order of the frames) or to the observer's ability to effectively compare the representations of a forwards clip and a backwards clip. On the other hand, a large accuracy drop would indicate either that the representations are intrinsically sensitive to playback direction, or that the observer cannot perform a computation that compares the two representations.

To investigate the effect of the four manipulations described above, we performed a 2AFC delayed match-to-sample experiment. In each trial, the sample clip (presented first) was altered, while the two test clips (presented after the sample) were untransformed. We used 1 second samples (50 frames) and 1.2 second tests (60 frames).

4.1 Experiment 4.1: Feature manipulation on long clips

4.1.1 Methods

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

Materials We presented stimuli on a CRT monitor in a darkened room as described in Chapter 2 (General methods). Observers used a chin-rest. Stimulus clips were taken from a continuous 1000-frame corpus of flame video.

Design We used a 2AFC delayed match-to-sample task. We manipulated the sample characteristics (no manipulation, colour change, luminance change, inversion, or reversal) within subjects.

Procedure 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 contained the sample using the keyboard. The sample length was 50 frames (1 second). The test length was 60 frames (1.2 seconds). Prior to the experimental trials, 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 24 training trials with dynamic, unaltered samples and tests of identical length. After training, we presented 5 blocks (one corresponding to each manipulation) in random order. We used 80 trials per block, making for a total of 400 trials.

4.1.2 Results

That data from this experiment are shown in Fig. 4.1. In the natural condition (untransformed sample), observers' accuracy was 60%. Random responding would give an expected score of 50%.

Accuracy was significantly above chance under each condition (one-sample t-tests, see Table 4.1 for p-values and accuracies) showing that observers could still perform the matching task under each of the manipulations. However, accuracy was not significantly different across each of the manipulations (repeated measures ANOVA, $F(4,36)=0.44$, $p=0.78$).

Manipulation	Accuracy (%)	p
None	60	0.01
Negative	59	0.01
Chromatic	61	0.01
Reversed	58	0.04
Inverted	59	0.01

Table 4.1: Results from Experiment 4.1, showing manipulation, matching accuracy and p-value of single-sample t-tests comparing to the chance level (50%).

4.1.3 Discussion

These results show that observers are capable of matching an altered sample to the standard test. Overall accuracy, however, was quite low, around 60%. It appears that observers' ceiling level in this task was quite low, leaving little room for changes in manipulation to influence accuracy.

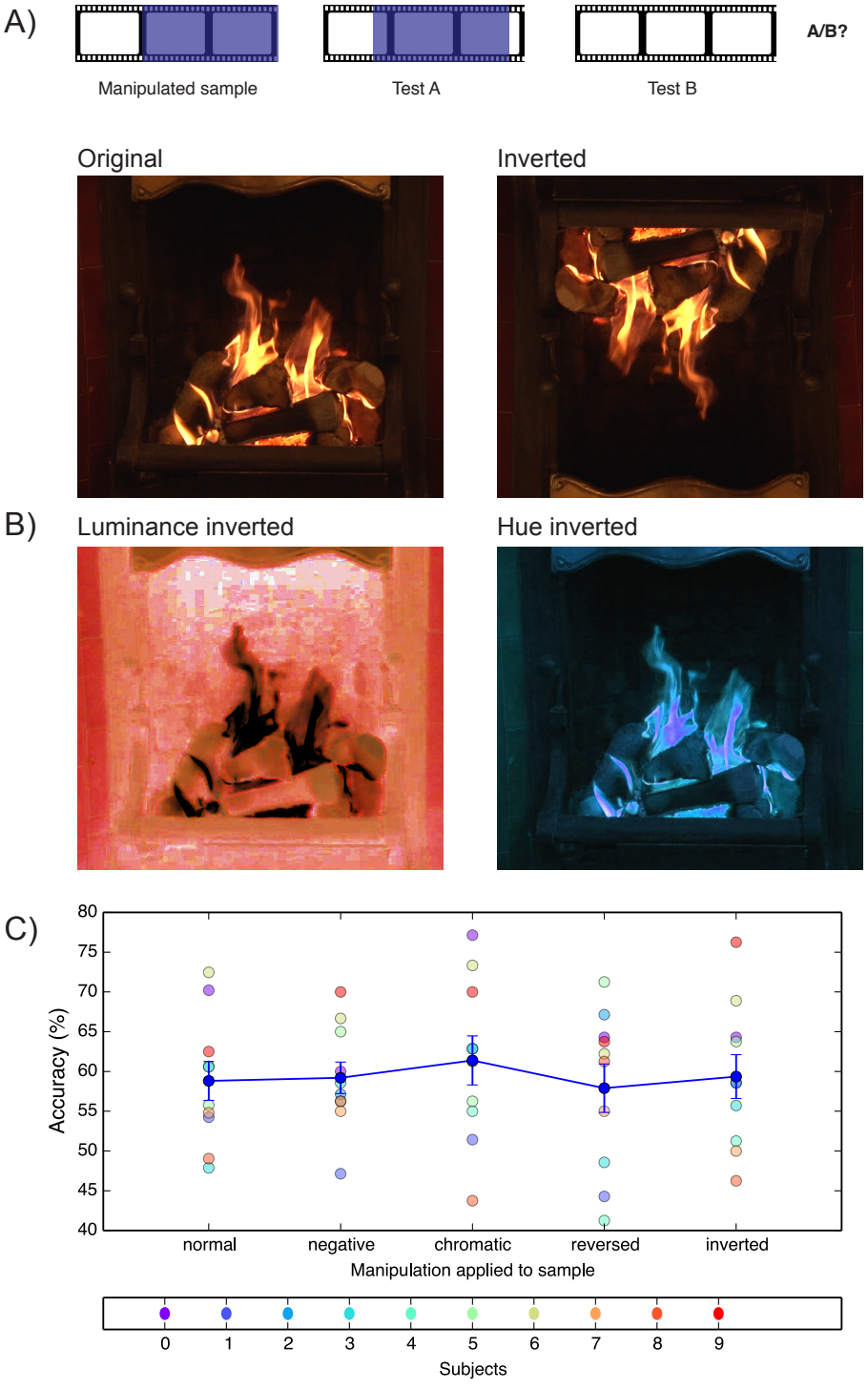


Figure 4.1: Experiment 4.1. A) Trial structure: a 1-second (50-frame) altered (for example, inverted) sample was followed by two untouched 1.2-second (60-frame) tests, one of which contained the sample. B) An original frame and three altered versions. C) Matching accuracy for each of the alterations. Detection was above chance under all manipulations, but was too low to discern a contrast between the effects.

In order to attempt to increase the overall recognition rate we repeated Experiment 4.1 with shorter clips (samples of 0.2 seconds and tests of 0.3 seconds), which appeared in pilot studies to improve performance.

We also noted easily visible artefacts in the luminance-inverted clips (see Figure 4.1). These artefacts are caused by video compression. Compression artefacts are also physically present in the standard clips, but are not perceptible due to decreased sensitivity to luminance change at high luminance (following Weber's law). Since these artefacts could provide easy matching cues, we removed the luminance-inversion condition from the next experiment.

4.2 Experiment 4.2: Feature manipulation on short clips

Our second experiment investigated observers' ability to match shorter clips under the same manipulations as Experiment 4.1, except for luminance inversion, which was removed due to the presence of artefacts. In this experiment, samples were 10 frames (0.2 seconds) and tests were 15 frames (0.3 seconds) long.

4.2.1 Methods

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

Materials We presented stimuli on a CRT monitor in a darkened room as described in Chapter 2 (General methods). Observers used a chin-rest. Stimulus clips were taken from a continuous 1000-frame corpus of flame video.

Design We used a 2AFC delayed match-to-sample task. We varied the manipulation applied to the sample clips (none, colour-inverted, backwards, or spatially inverted) within subjects.

Procedure 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 contained the sample using the left arrow (first sample) and right arrow (second sample) keys. The sample length was 10 frames (0.2 seconds).

The test length was 15 frames (0.3 seconds). Prior to the experimental trials, 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. In the body of the experiment, we used 4 block types (corresponding to each manipulation) with 4 repetitions of each block (16 blocks total) in random order.

4.2.2 Results

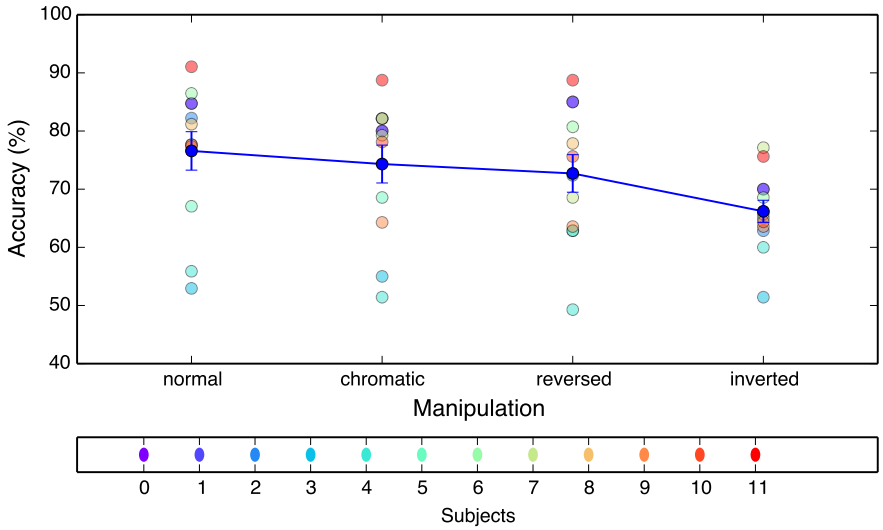


Figure 4.2: Experiment 4.2 used the same procedure as Experiment 4.1, but with shorter samples. A 0.2-second (10-frame) altered (for example, inverted) sample was followed by two untouched 0.3-second (15-frame) tests, one of which contained the sample. Chromatic alteration did not produce a significant drop in accuracy. Inversion produced the greatest accuracy drop, followed by reversion and colour shifting.

The data from this experiment are shown in Fig. 4.2. There was a significant effect due to choice of manipulation (repeated-measures ANOVA, $F(3,33)=9.4$, $p<0.001$). With shorter clips, observers are no longer at ceiling and we find a measurable effect of type of manipulation.

In all of the conditions accuracy was greater than chance ($p < 0.001$, single-sample t-tests), showing that observers can still perform the matching task under each manipulation.

Overall accuracy in the standard condition was 77%. Other accuracies, and drops

in accuracy from the standard condition, are shown in Table 4.2. The p-values of paired-sample t-tests between each manipulation and the standard condition are also shown.

The largest drop was associated with the inversion condition, in which accuracy dropped to 66%. We also note a significant difference between reversal and inversion (p=0.03, paired-samples t-test).

Manipulation	Accuracy (%)	Difference (pp)	p
None	77	0	-
Chromatic	74	2.2	0.11
Reversed	73	3.9	0.05
Inverted	66	10	0.01

Table 4.2: Results from Experiment 4.2, showing manipulation, matching accuracy, accuracy drop in percentage points, and p-value of single-sample t-tests comparing to the unmanipulated condition.

4.2.3 Discussion

In each condition, the observer must attempt to reconcile information in the altered sample with information in the unaltered test. Accuracy in the unaltered condition serves as a benchmark. If an alteration causes a large drop, the observer finds the comparison problematic; a small or null drop indicates the comparison is still possible.

Colour While hue reversal drops the mean accuracy by 2 percentage points, a paired-samples t-test shows a low probability that the data are due to difference from the mean (p=0.11). Observers do not require the correct colour in order to match fire samples. There is therefore no evidence that colour forms a key part of the representation of dynamic flame used in this task.

Reversal Backwards playback drops the mean accuracy by 3.9 percentage points, and is associated with a significant drop in performance compared to unchanged clips (paired-samples t-test, p=0.05). Reversing a video clip alters many of its simple motion properties (such as direction of motion). It does not, however, 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 reversal is associated with a 3.9 percentage point drop in accuracy which is marginally significant (p=0.05). This shows that observers' matching processes are not invariant to playback direction. If observers are sampling discrete spatiotemporal events, their order is encoded, but may not be required; if observers are constructing a gist-style summary, this result suggests that motion forms a part of this representation.

Inversion alters the local processing of motion features found in fire clips. We used a 180 degree rotation, which transforms upwards motion into downwards motion and leftwards into rightwards. It does not, however, alter any of the temporal properties of the clip; features which are not mapped by location (such as the global mean brightness signal) are not altered.

Inverting a video clip alters the spatial location of all the features contained therein. The signature of a fire clip may not consist just of a set of unlocalised features; each feature may linked to its location in space ("a flare in the upper left of the frame"). This information is disrupted by inversion. Here inversion is associated with a 10 percentage point drop in accuracy, much larger than the drop caused by reversal. This drop is highly significant (paired-samples t-test, p <0.001). For the clip size and duration used here, then, the spatial location of the represented details is more important than their temporal location.

The accuracy drop under inversion (10 p.p.) is much larger than that under reversal (3.9 p.p). There is a significant difference between these two conditions (paired-samples t-test, $p < 0.05$). For clips of length 0.2 s, then, spatial arrangement of features is much more informative than temporal arrangement of features.

If observers are encoding spatiotemporally localized features, as opposed to creating gist-like compressed representations of larger portions of the clip, what do these results indicate? As inversion impairs matching more than reversal does, subjects are helped much more by knowing where features are in the image than when they occur in the sequence. This suggests that spatial location plays a more important part in the stimulus representation than temporal location.

Colour alteration, because it can be applied independently to each pixel in an image, can be characterised as a local manipulation. Changing the colour of a clip alters its representation in early visual structures (the retina, the optic nerve, and

V1). However, the small drop in accuracy induced by changing colour shows that this alteration does not significantly affect the test clip's representation at a high level - its correspondence with the sample clip. We can conclude that performance in the task is effectively invariant to colour.

The manipulations we have used so far are simple to compute and do not depend on object, texture or edge detection. This means that they cannot influence mid-level features, such as textures or edges. If we want to remove from an image all its convex edges, or all occurrences of the letter "H," we will not be able to do so with a manipulation which we apply independently to each pixel, or a spatial or temporal inversion. The visual system constructs mid-level representations by combining information from groups of pixels. The next experiment examines the effect of a higher-level transformation: highlighting the edges in the scene.

Edges are a key component of complex visual scenes. We find edge-responsive cells early on in the visual processing stream[14], and in the case of natural scenes, independent component analysis suggests that edge filters represent a scene efficiently[211]. Edges are also key representations for most theories of object recognition, which often employ an initial layer of oriented edge-detection filters[12], and many object recognition experiments employ black-and-white line drawing images, which consist mostly of edges.

An edge representation does not capture all the information present in an image, however; edges cannot convey textures or luminance gradients. In Experiment 4.3 we asked observers to match an edge-filtered sample with an untransformed test. The sample, being filtered to show edges only, contained no prior information about texture, colour or gradient.

4.3 Experiment 4.3: Edge filtering

Luminance gradients are a central part of many natural images. They are key for extracting age and gender properties from faces[212] as well as perceiving 3D shape[213]. Here we investigated their importance for dynamic flame matching by removing all luminance gradient information from the test using an edge filter, creating an altered clip built from binary images in which a white pixel represents an edge.

In half the trials, the sample was manipulated using a Sobel edge filter[214]. The test clips were left unchanged. Each frame was convolved with the 3×3 filter

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

to give an estimate of the image derivative. All pixels above a certain threshold absolute value in the derivative image were returned as edges. The implementation used was MATLAB's **edge()** function. See Figure 4.3 for an example.

4.3.1 Methods

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

Materials We presented stimuli on a CRT monitor in a darkened room as described in Chapter 2 (General methods). Observers used a chin-rest. Stimulus clips were taken from a continuous 1000-frame corpus of flame video.

Design We used a 2AFC delayed match-to-sample paradigm with altered samples. In half the trials, the sample was manipulated using a Sobel edge filter. The test clips were left unchanged.

Procedure In each trial, a sample was presented first, followed by two tests. In half the trials, a manipulation was applied to the sample; the tests were always unchanged. Subjects indicated which test they thought corresponded to the sample using the left arrow (first sample) and right arrow (second sample) keys. The sample length was 10 frames (0.2 seconds). The test length was 15 frames (0.3 seconds). To begin the experiment, 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. In the main part of the experiment, there were 2 block types (normal sample and edge-filtered sample). We presented each block 7 times (in random order), giving a total of 14 blocks. We presented 40 trials per block, making for a total of 560 trials.

4.3.2 Results

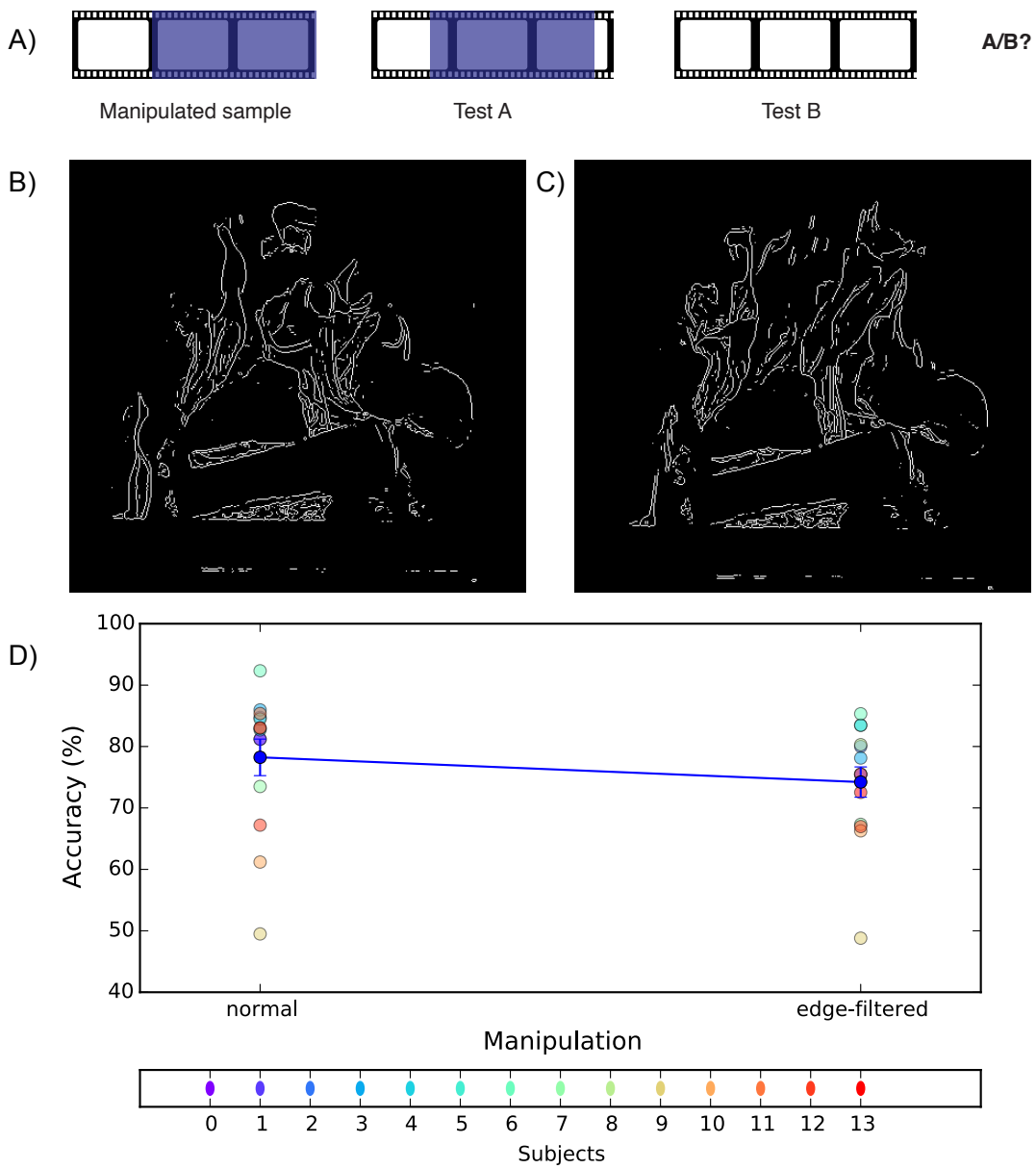


Figure 4.3: Experiment 4.3. A) Trial structure: a 1-second (50-frame) edge-filtered sample was followed by two untouched 1.2-second (60-frame) tests, one of which contained the sample. B,C) Two edge-filtered images. D) Edge filtering induced a slight drop in performance which was highly significant.

Results are shown in Fig. 4.3. Edge-filtering the sample induced a 4 percentage point drop in accuracy compared to the standard condition. Mean accuracies are shown in Table 4.3. Under both conditions, matching was significantly better than chance (one-sample t-tests, p <0.001 in both cases).

This difference was significant (paired-samples t-test, p <0.005) indicating that

there was some loss of performance when the visual stimulus was limited to edges. However, performance with on samples containing only dynamic edges induced a drop of only 4 percentage points and was still significantly better than chance, showing that observers can still perform the task effectively.

Sample	Accuracy (%)	p
Normal	78.2	0.001
Edge-filtered	74.2	0.001

Table 4.3: Results from Experiment 4.3, showing sample, matching accuracy, and p-value of single-sample t-tests comparing to the chance level (50%).

4.3.3 Discussion

We made a significant change to the sample clips, replacing most of their structure with black pixels and leaving only edge information. We expected this alteration to impair matching but were surprised when accuracy only dropped by 4 percentage points: a large change in the image induced only a small change in matching performance.

This result limits the importance of two signals. First of all, the global 1D luminance signal (the mean luminance of each frame) hardly varies at all in edge filtered clips, so cannot be of great importance. Secondly, edge-filtering completely removes gradient information (shading, colour, and texture). The small drop shows that matching can be done without this information, and thus indicates that our representation of dynamic flame is mostly built from dynamic form information.

Edge filtering preserves the dynamic form of the main segmentable objects in the scene: flames. Light against a black background, these are easy to segment; they have a clear outline, and although they move quickly, our use of a high shutter speed (1/150 s) ensured that edges were captured clearly. To be precise, what is preserved after edge-filtering is the shape of the flames and the motion of their edges.

Edge-filtered images and their originals induce different motion percepts. The original, unaltered stimulus contains a closely-packed field of pixel-level information, allowing computation of dense motion vectors. We generated some of these algorithmically in Chapter 3. The edge-filtered stimulus, however, is much sparser: dense motion fields cannot be generated, since most of the pixels are black. By applying

spatial and temporal filters, the visual system may still be able to generate low-level motion percepts from the edge-filtered stimuli (as can the sMcGM motion algorithm), but it must rely mainly on perceiving the displacement of form. This is a higher-level process as the visual system must match two instances of a particular form across time in order to perceive form displacement.

The surprisingly small drop in accuracy induced by edge-filtering shows that observers can still perform matching effectively using only edge information. Most pixel information allowing gradient-based motion perception has been removed. This suggests that dynamic form plays an important role in matching individual exemplars of dynamic flame.

Does dynamic form also play a part in observers' category representations of flame? We address this question by asking them to detect reversed playback rather than testing their invariance to it. One of the properties that sets the motion of flame apart is that there is no ground truth. Videos of natural scenes are not generated by translating objects in a controlled way; as a result, we cannot say that any patch of flame has a particular canonical motion direction. Are observers still able to percieve a particular motion direction and use it to cue their classification of a long flame video as normal or reversed?

4.4 Experiment 4.4: Can observers detect backwards playback?

When matching a sample and test clip, backwards playback of the sample has been shown to reduce matching accuracy on short clips (Experiment 4.2). The visual system is thus not invariant to playback direction; it must induce a change in the representation of the stimulus. Are observers able to access this information by explicitly detecting playback direction?

When performing a matching task, an observer must encode the sample and compare it against the test. Internal models and specialised representations, like the ones we use to encode faces[215] can help with this task, but the key challenge is comparing two presentations. When we ask an observer to indicate whether a single clip's playback is forwards or backwards, however, we only present one stimulus per trial.

The observer's challenge is to encode this clip and match it against information that they already know, comparing it against their internal model of dynamic fire.

To investigate how well untrained observers can perform this task, we displayed single clips at various frame rates, manipulated playback direction, and asked for a forwards/backwards judgement. We also manipulated the angle at which clips were displayed, using four different angles to test for orientation dependence.

4.4.1 Methods

Observers 13 observers were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Materials We used a 1000-frame corpus of consecutive fire images, displayed using the equipment described in Chapter 2 (General methods).

Design We employed the method of binary choice.

Procedure In each trial, a 2-second clip was played. The observer then indicated whether they thought the clip was being played forwards or backwards, using the keyboard. We manipulated the angle at which the clip was played (0, 90, 180 or 270 degrees) and the frame rate (50, 25,16.7, 12.5, 10, 8.3, 7.1, 6.3 ,5.6 or 5 Hz). These frame rates correspond to interframe durations which are multiples of 0.02 seconds, the standard interframe duration. We did not vary clip speed or length, only frame rate. We varied frame rate within blocks and orientation across blocks, giving 4 block types. There were two repetitions of each block type, giving 8 blocks in total.

4.4.2 Results

Figure 4.4 shows the data from this experiment. Accuracy drops very rapidly as frame rate decreases, reaching chance by 10 Hz (when each frame is displayed for 0.1 s). This effect is confirmed by a two-way ANOVA over orientation and frame rate, which shows a highly significant effect of frame rate (F(5,16) = 2.2, p = 0.016). There appears to be no effect of orientation (F(11,16) = 8.13, p =0.146); the accuracy curves depending on orientation are shown in Fig. 6.4.

Observers are much more likely to correctly judge the direction of a forwards clip than a reversed clip (paired-samples t-test, p=0.001). However, the slower the frame rate, the more likely observers were to indicate forwards playback. This is confirmed by

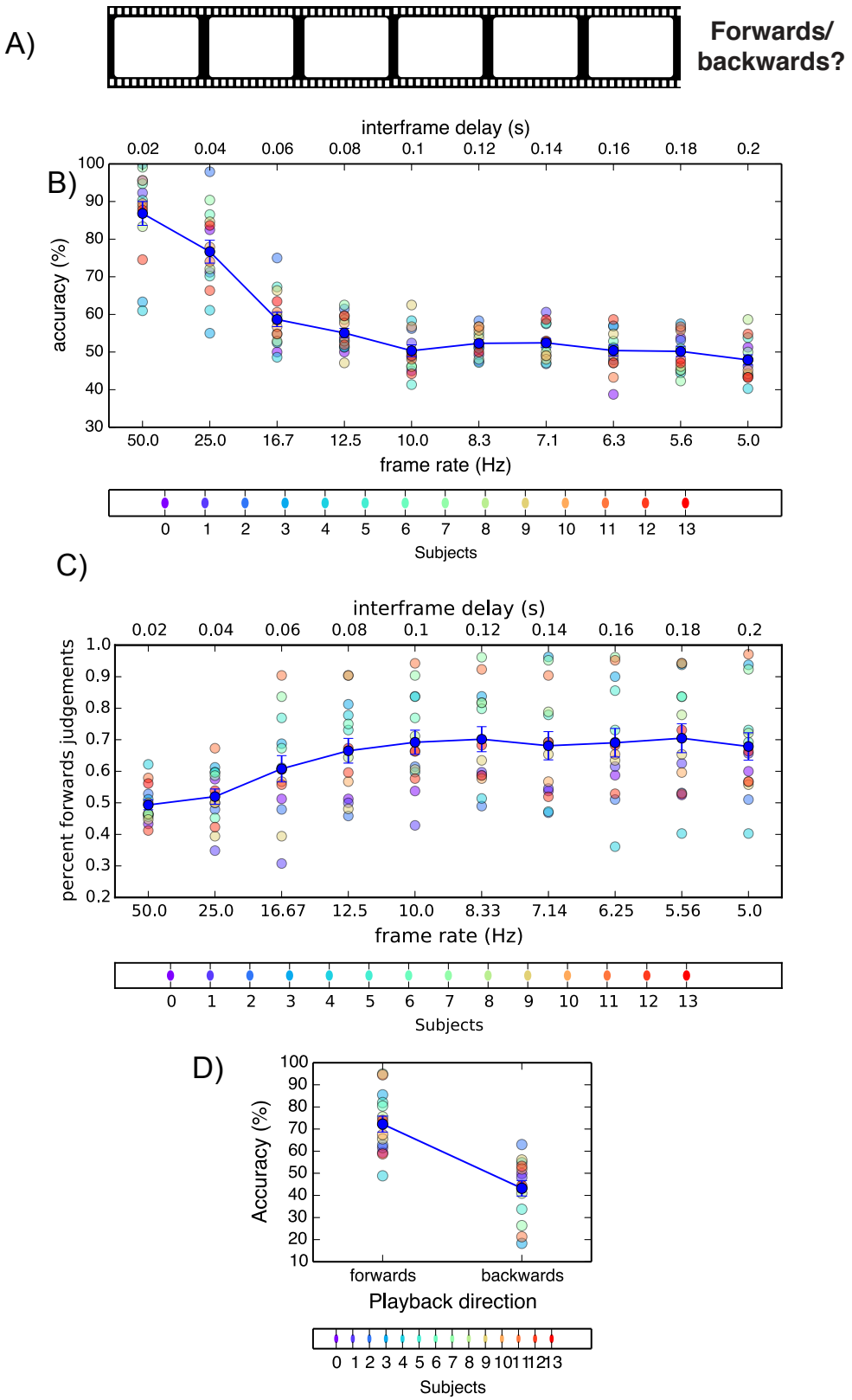


Figure 4.4: Experiment 4.4: backwards detection. A) On each trial a 2-second clip was played; observers then indicated whether they thought playback was forwards or backwards. B) Accuracy dropped quickly down to chance at a frame rate of 10 Hz. C) As frame rate decreased, the amount of "forwards" judgements increased. D) Observers were much more accurate at judging the clips which were played forwards.

a one-way ANOVA on observers' responses, with frame rate as a factor ($F(1,9)$=11.91, $p < 0.001$).

4.4.3 Discussion

It is clear that low frame rates greatly impair observers' ability to compare the stimulus against their existing category representation of forwards-moving dynamic fire. At frame rates below 10 Hz (interframe delay 0.1 seconds), observers' judgements are at chance. This means that, to detect forwards or backwards playback, observers are dependent on temporally local patterns and correlations. They are not able to detect or exploit long-range patterns in the stimulus.

This fits well with the results of our correlation analysis in the image domain. As shown in Fig. 3.3, three measures of image similarity reach a minimum after a duration of 0.2 seconds. This shows that these measures cannot detect any similarity between frames more than 0.2 seconds apart. This experiment shows that the visual system cannot extract any information which is useful for forwards/backwards judgement from frames more than 0.1 seconds apart.

This task relates more to the category representation of fire rather than the encodings of individual frames. The task asks "are you observing forwards fire?" which is very similar to the question "are you observing fire?" It asks the observer to compare their internal representation of what moving fire should look like with the presented stimulus. The slower the frame rate, the more likely observers are to make a "forwards" judgement. This suggests that a high fidelity stimulus is required for a judgement of backwards-moving fire and any degradation leads to a "default" forwards judgement. This strongly suggests that the category representation of dynamic flame contains motion representations.

The key information for this task is local: it is integrated within a 0.1 second window. Are observers computing low-level motion or high-level displaced form? It is certainly possible to obtain low-level motion stimuli algorithmically from edge-filtered flame, as we showed in Chapter 2, although motion fields are sparse due to the predominance of black pixels. After an 0.1 second period, metrics of image similarity (Euclidean distance and SSIM) drop nearly to the minimum value that they reach after 0.2 seconds. It is likely, however, that the human visual system is better at

matching displaced form than the SSIM, so this drop in similarity metrics does not allow us to argue for an absence of form information.

Low-level motion detection does not require expertise, since it is implemented by early neural computations which have not been shown to adapt to particular stimulus classes. Displaced form, however, is a higher-level motion percept, and can be aided by expertise: matching a flame to a slightly distorted flame occurring 0.2 seconds later could benefit from a knowledge of the ways in which flames are likely to deform. Since fire is highly asymmetric, there is much opportunity for direction-specific learning and we would expect displaced form detection to be more effective on upright flame. Fig. 6.4 shows that this is not the case; upright playback confers no advantage on backwards detection, and rotation through 90 or 270 degrees does not impair the task. This suggests that backwards detection is done by an orientation-invariant process, favouring low-level motion over dynamic form.

What kinds of motion percepts do observers construct from dynamic flame stimuli? In the next section, we investigate by showing observers small areas of dynamic flame and asking them to evaluate the main motion direction they perceive.

4.5 Experiment 4.5: Motion direction percepts in dynamic flame

4.5.1 Methods

Observers We recruited 4 subjects, one of whom was the author. All reported normal or corrected-to-normal vision.

Materials We presented stimuli on a CRT monitor in a darkened room as described in Chapter 2 (General methods). Observers used a chin-rest. Stimulus clips were taken from a continuous 1000-frame corpus of flame video.

Design We used the method of adjustment, asking observers to rotate a moving grating to match the direction of motion perceived in a small sample of dynamic flame.

Procedure In each trial, a disc-shaped patch of dynamic flame was sampled from the 1000-frame dataset, rotated by a random amount, and displayed to the observer on the left side of the screen. On the right side, we simultaneously displayed a moving

grating whose direction of motion could be adjusted using the arrow keys. Observers were asked to rotate the grating so that its direction of motion matched that which they perceived in the flame sample.

A circular Gaussian window was applied to the flame sample to prevent the occurrence of illusory contours at the circular border. We manipulated the diameter of the flame disc, which was either 10, 20, 30, 40, 50, 60 or 70 pixels. Flame stimuli were not scaled. We sampled from a location in the frame which gave no non-motion clues to the direction of motion, either by the appearance of static features (such as logs) or asymmetry in the average luminance (sampling from an area too high in the flame would have resulted in more black space and less flame near the top of the sample).

On each trial, a clip 4 seconds in length was looped; observers were given as much time as necessary to make their decision, and instructed to be as quick and accurate as possible.

4.5.2 Results

The results of this experiment are shown in Fig. 4.5. We calculated the absolute error e in degrees as

$$e = [(a - b + 180) \mod 360] - 180 \tag{4.1}$$

where a is the angle to which the vertical has been rotated and b is the observer's response. This calculation ensures that e always reflects the absolute value of the smaller angle between the rotation angle and the motion direction judgement (not the larger angle). Since the direction of motion is usually reported as up, a is an estimate of the "true" direction of motion for this flame disc. When a =0, the clip was untouched (upwards motion); when a =180, the clip was inverted (downwards motion). Because e may range from 0 to 180, a mean value of 90 would be expected if an observer was performing at chance. Mean absolute errors are shown in Table 4.4.

Absolute error decreased strongly as flame size increased. To investigate individual observers' response profiles, we show their responses on a scatter plot (Fig. 4.5).

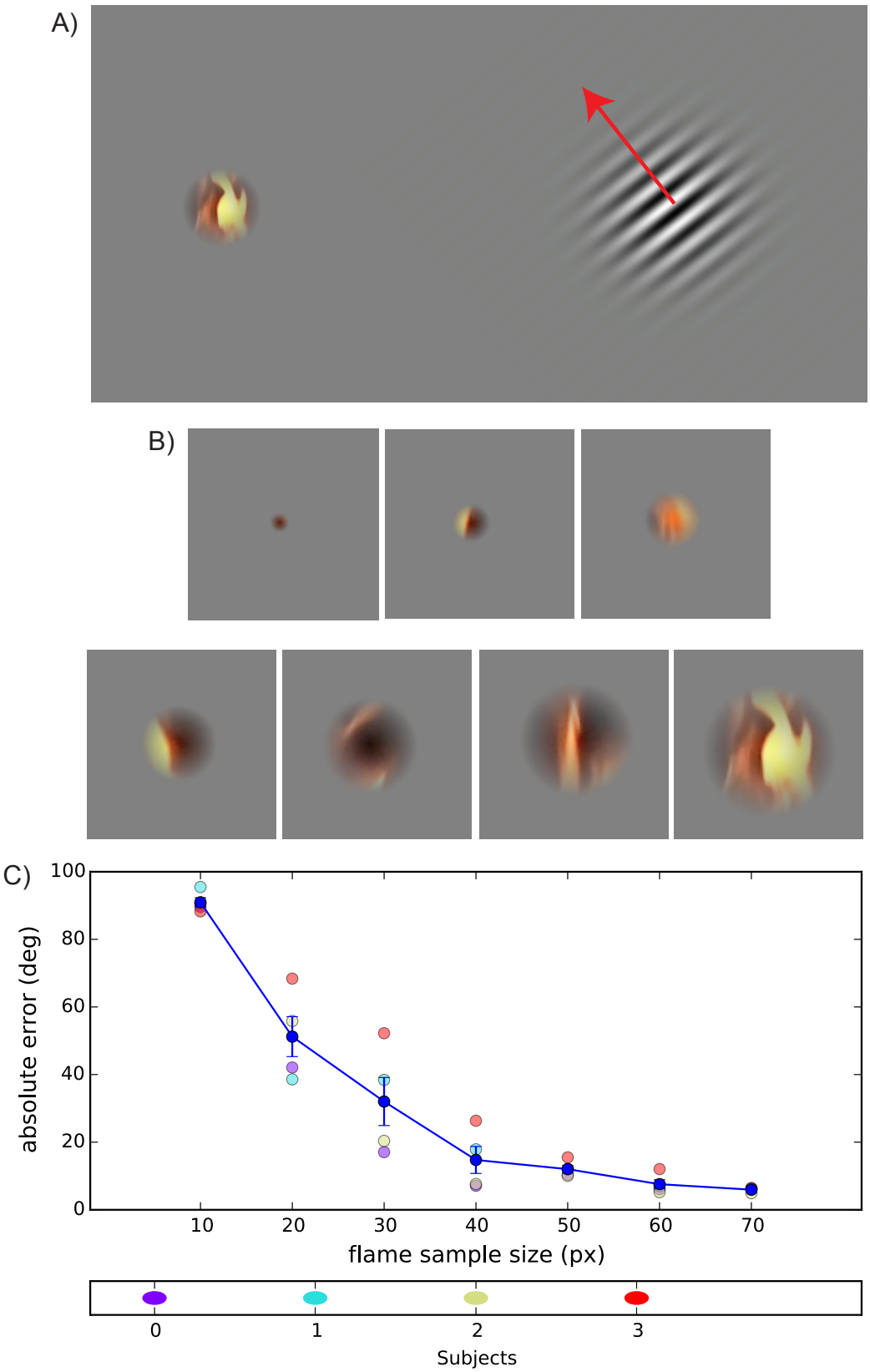


Figure 4.5: A) We presented a circular disc sampled from an area near the centre of our dynamic flame dataset. During each trial, this clip (duration 4 seconds) was looped, while observers adjusted the motion direction of a moving grating to match the direction of motion they perceived in the flame disc. Motion direction is marked here by a red arrow. All flames shown here are non-rotated. B) The seven flame disc sizes, from 10 to 70 pixels in width. C) Error decreased slowly, from the chance level (90°) to a minimum of 5.9°.

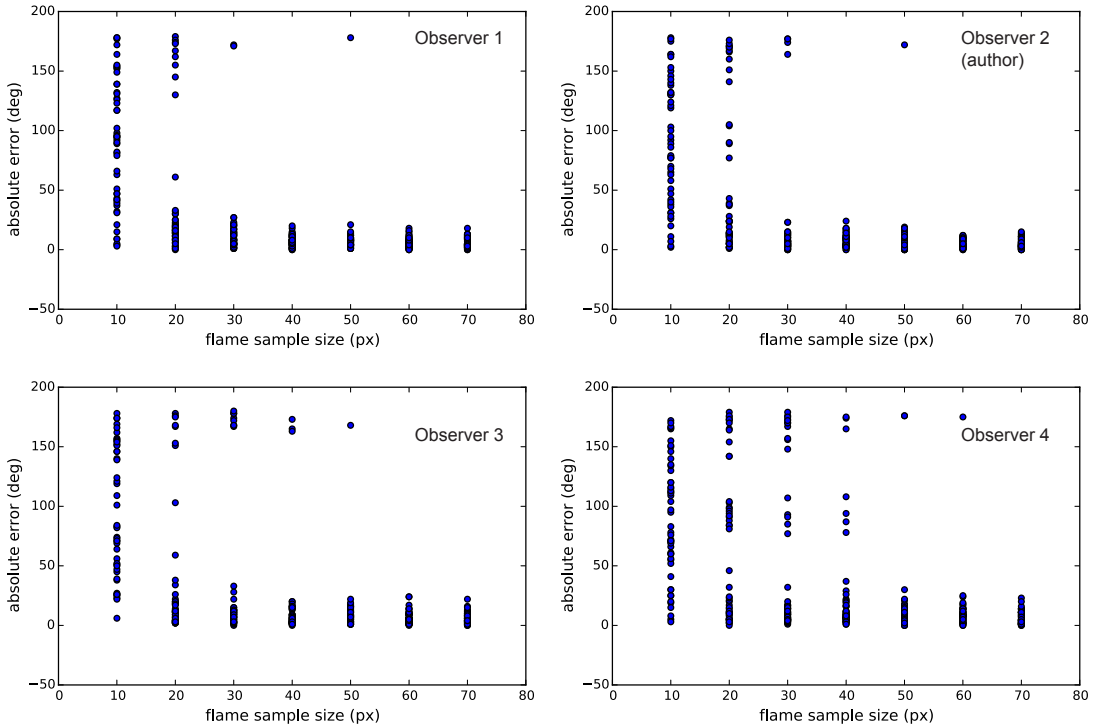


Figure 4.6: Experiment 4.5: individual results. Each observer's responses are shown as points on a scatter plot depicting absolute error e against flame sample size. A consistent pattern, from high error to very low error, is shown. Medium clips are often erroneously labelled with the opposite motion direction (180° error); observer 4 also makes a number of 90° errors.

Flame size (px)	e
10	91.4
20	54.3
30	37.0
40	17.3
50	12.6
60	8.0
70	5.9

Table 4.4: Experiment 4.5: mean absolute error e by flame size.

4.5.3 Discussion

All observers show a smooth progression in mean error from 91° to 5.9° with remarkably low variance. This suggests that motion mechanisms which vary little between observers are being employed. On patches 10 pixels wide, performance is at chance. Looking at the scatter plots, we find a bimodal error distribution for most observers: errors cluster around 0° and 180°. A small patch of dynamic flame is a visual metamer for a patch of flame moving in the opposite direction. For the largest patch size, all errors are close to 0°.

On large clips, error is very small: 5.9 degrees on average. Because all clips were rotated from the vertical, this result indicates that an upwards motion percept is indeed generated by normal flame clips, and that observers agree on this. It is notable that observer 2, the author, did not show any accuracy gains over any of the other participants- who had never seen these flame stimuli before.

How do these results correspond with our computational motion analyses of dynamic flame (Chapter 3)? Analysis of individual frames using the McGM revealed small patches of upwards motion interleaved with small patches of downwards motion. Sun's method, which included a regularisation stage, did not show these small patches, but larger areas of consistent motion in one direction. Experiment 4.5 shows that small patches of dynamic flame also elicit motion percepts in multiple directions. Observers never reported downwards motion from full flame clips, however, or from the largest patches; how do we explain this contrast? When combined together into a full flame clip, the overall percept is one of upwards motion; but this could be due to top-down inhibition of percepts which are not in accordance with the high-level knowledge that flame is supposed to move upwards. High-level knowledge can affect low-level motion perception, as in the spinning dancer illusion[216]. It is either the case that downwards motion percepts are not generated from complete flame clips at all, or that they are generated by low-level motion detectors and later filtered out, swamped by upwards motion signals, or integrated into a dynamic texture.

Judging the motion direction of an isolated local patch poses different challenges than judging its direction when it is integrated into a larger scene. In particular, small patches suffer from the aperture problem[217]: it is mathematically impossible to judge the movement direction of a sine grating displacing behind a circular aperture,

since the angle of the grating cannot be dissociated from its direction of motion. It is possible that this experiment simply did not provide observers with enough information to make a motion judgement. In this case, they may be performing close to optimality, which explains the low inter-observer variance in absolute error.

For isolated local patches, however, these results suggest that the motion percepts present in fire are constructed by low-level mechanisms which are stable across individuals and do not show much inter-observer variation.

Learning effects can indicate whether a decision process uses high-level or low-level percepts. In the next section, we evaluate subject performance over time in the previous four experiments.

4.6 Learning

Representing and matching dynamic fire is a novel task for our observers. It is natural to ask whether they learn; in other words, whether their representations improve as they progress through an experiment.

The experiments presented in this chapter lasted between 1.5 and 2 hours. Over the course of several hundred trials, observers were exposed to a large quantity of dynamic flame clips. During half of these clips (the samples) they needed to encode and during the other half (the tests) they needed to match. Because clips were randomly sampled from large datasets (either 1,000 frames or 10,000 frames), there was little opportunity to learn the specific characteristics of individual clips or frames. There was, however, scope for observers to learn general mid-level spatiotemporal features commonly occurring in fire clips.

As a proxy for learning, we used the change in accuracy as subjects progressed through an experiment. For each experiment, we arranged the trials in the order in which they were presented, blocked them into groups of 20 using a sliding window, and calculated the average accuracy for each block.

To check for an improvement in mean accuracy, we fitted a line to the sequentially arranged data. Calculated slope values are shown in Table 4.5.

The results of this sliding-window approach are shown in Figure 4.7. Note that the curves are smooth due to sliding-window averaging; this does not indicate a smooth

Experiment	Slope (percentage points per trial)
4.1	-1.20×10^{-3}
4.2	2.02×10^{-4}
4.3	1.15×10^{-3}
4.4	-3.40×10^{-4}
4.5	-1.8×10^{-2} (e/trial)

Table 4.5: Learning slopes in Chapter 4.

trend in observers single-trial accuracy, which is binary. None of the experiments show a consistent improvement, except for Experiment 4.3 (in which samples were edge-filtered). Observers appear to be learning to match the novel edge-filtered stimulus with normal video, but do not show any evidence of learning in any of the other experiments. On the motion direction matching experiment (4.5), a clear decrease in error is notable in the first 50 trials; there appears to be no trend during the rest of the experiment. Please note that this graph shows absolute error, not accuracy, so good performance corresponds with a lower y-coordinate.

Observers did not show evidence of continual learning in our matching experiments. In the backwards detection and motion direction evaluation experiments, there is only evidence of learning in the first 50 trials. This lack of improvement suggests that observers are using low-level mechanisms which are not trainable.

4.7 General discussion

Experiments on 0.2 second clips show that observers are tolerant of variations of colour between the sample and the test, but less so to variations of temporal arrangement and even less so to variations of spatial arrangement. High performance on edge-filtered samples provides convincing evidence that dynamic flames are represented mainly as moving form.

How do these results sit with established theories of object perception? We recall the problems dynamic natural scenes pose for object recognition. Natural scenes are dynamic; they are composed of many parts; and they are not naturally segmented.

Most models of object recognition only consider static images. Applying traditional object coding theories to dynamic video, then, means thinking about movies as a set of static snapshots. Each frame could be represented either as a hierarchy of compo-

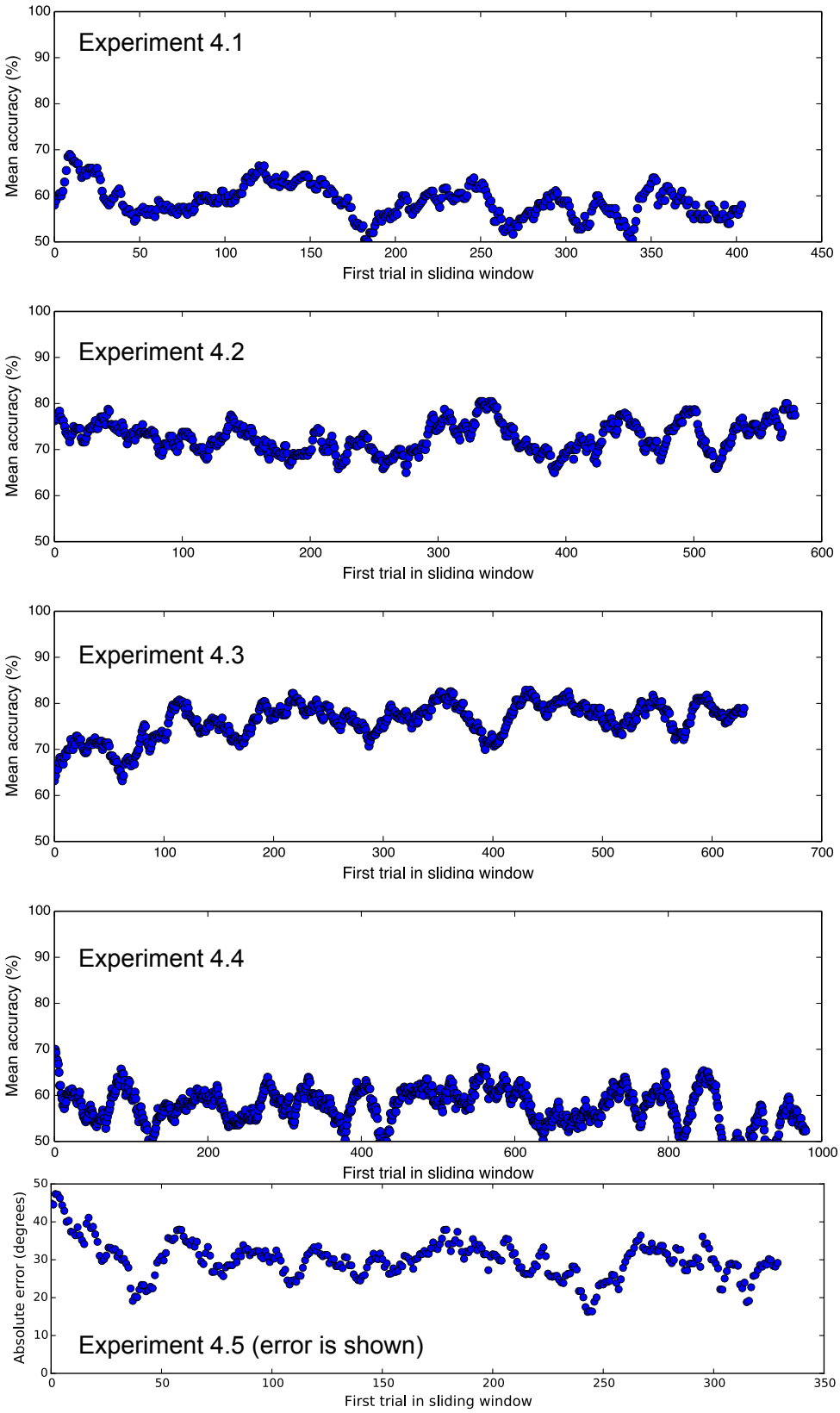


Figure 4.7: For each experiment, trials were aligned in order of presentation and a sliding average applied to show how accuracy changed during the experiment. There was no overall trend of increasing accuracy except in Experiment 3 (edge filtering). This indicates that observers may be learning the unfamiliar edge representation.

nents (as with Biederman's geons or Selfridge's demons) or an atomic representation (a point in a feature space, such as a gist). Could these frame representations be stored sequentially in order, or as an unordered bag of snapshots? Our results suggest that backwards playback impairs matching performance. This suggests that, if frame snapshots are used, they are ordered in time. We also find accuracy impairment due to spatial inversion; this suggests that, if individual features are represented on a sub-frame level, they are either individually orientation-sensitive or orientation-invariant but represented at a specific location in the stimulus.

Perceptual processes can be characterised by the invariances they possess. Flame matching is differentially invariant to reversal and inversion: we note respective accuracy drops of 3.9 and 10 percentage points, and a paired-samples t-test shows a significant difference between these means. For 0.2 second samples, then, spatial configuration is more important than temporal configuration.

What kind of motion percepts are generated by fire? Texture patches which change smoothly in intensity are common, meaning that first-order motion must play a key part. Drift-balanced stimuli, usually designed to evoke second-order motion and no first-order motion, only occur accidentally. On frames which are close together, flame contours may remain sufficiently self-similar to induce phi motion. It is also possible that the combination of a wide flame displacement (changing an area from high to low contrast) and a change in local form may create the correct combination of displaced form and contrast inversion necessary to evoke reverse phi motion. One kind of motion percept which we can confidently rule out is long-range displaced form; our image-based analysis suggests that form is not preserved for more than a few frames, and our backwards detection experiment shows that no motion direction cues are available over long temporal intervals.

Observers have a category representation of dynamic flame, which they are able to use to detect backwards playback without any prior training; indeed, they do not learn to increase the accuracy of this judgement even after 2 hours' practice. Although they often report a reliance on upwards motion, they do not differ in accuracy on rotated stimuli, suggesting that the mechanisms they use to detect this oriented motion are orientation-invariant. When they do make correct judgements, a frame rate of at least 10 Hz is required, showing the temporal locality required to judge playback direction.

From flame patches 70 pixels in width, observers report a motion direction closely matching the patch's degree of rotation from the vertical. This shows that, from non-rotated patches, observers would report an upwards motion direction. For patches 10 pixels in width, observers are clearly at chance. Patches between these sizes generate either a directional percept close to the original vertical, or a 180° error. This shows that medium-sized flame patches are visual metamers for patches depicting gas moving in the opposite direction.

There are two clear conclusions from this series of experiments. Firstly, observers are severely impaired when trying to match inverted samples of dynamic fire to their upright counterparts, which shows that space figures importantly in their representations. If spatiotemporal features are being sampled, their location is recorded and used for matching. If low-level motion fields are calculated, they may be represented in a map which is used for matching. If a gist is computed and used for matching, it integrates spatial information as opposed to throwing it away. Secondly, dynamic edges (containing no luminance gradient, colour gradient or texture information) allow effective matching at nearly full accuracy. Since this manipulation alters local motion signals, it suggests that observers encode dynamic form.

Summary

- Observers can still match tests to samples which are hue-inverted, reversed or inverted.
- Under both reversal and inversion, performance is significantly impaired, showing that observers are sensitive to these transformations.
- Inversion impairs performance more than reversal, showing that (for 0.2 second samples and 0.3 second tests) the spatial arrangement of features is more important than their temporal arrangement.
- Observers do not show an increase in accuracy during this experiment, suggesting that they are not improving their representations.
- Edge filtering of the sample, which removes most of the information and all the texture and gradient information, induces a 4 percentage point drop in performance. This surprisingly small impairment shows that gradients, and motion features derived from smooth luminance variations, are not key for representations

of dynamic flame. Observers also progressively improve in accuracy throughout the experiment; they appear to be learning edge-based representations.

- At frame rates below 10 Hz, judgements of whether a flame clip is being played forwards or backwards are at chance. Observers' category representations of forwards-moving flame are thus very local in time; observers are not able to exploit long-range correlations.
- Observers are able to report the correct motion direction from large flame clips (70 pixels in width) but not small flame clips (10 pixels in width). The error distribution is bimodal for patches between these sizes; errors close to 0° or 180° are common. Small flame patches are visual metamers for patches moving in the opposite direction.

Chapter 5

Visual search for dynamic flames

Most theories of object recognition focus on the task of "core object recognition" [35], the classification or identification of a static image containing a single object against a blank background. In reality, however, natural scenes contain multiple objects and complex backgrounds[179], meaning that spatial segmentation is not always straightforward. Natural scenes are also dynamic; they change constantly in time, and can be segmented temporally into events. In this chapter, we characterise the visual search process on dynamic flame stimuli. We use the same delayed match-to-sample protocol as in the previous chapter, but pose a search challenge to observers by increasing the test/sample ratio.

The well-established field of visual search[154] deals with the location of targets in cluttered static images. Typically, a visual search task has two parts: the presentation of a small target, followed by the presentation of a larger search space including target and distractors[218]. In most search stimuli, the target and distractors are already segmented from the background. When distractors are sufficiently similar to the target, the observer cannot use the bottom-up effect of pop-out to locate the target without conscious effort[219]; however, even in this case, the targets are spatially discrete, with obvious borders that make them distinct from each other and from the background. Treisman[220] pointed out the dependence of search duration on search space size for conjunctions of features on discrete targets.

What would visual search feel like if the targets were not discrete? An example is shown in Fig. 5.1, where we ask you to search for a small texture patch in a much larger texture patch. Here the search space is continuous rather than discrete, and

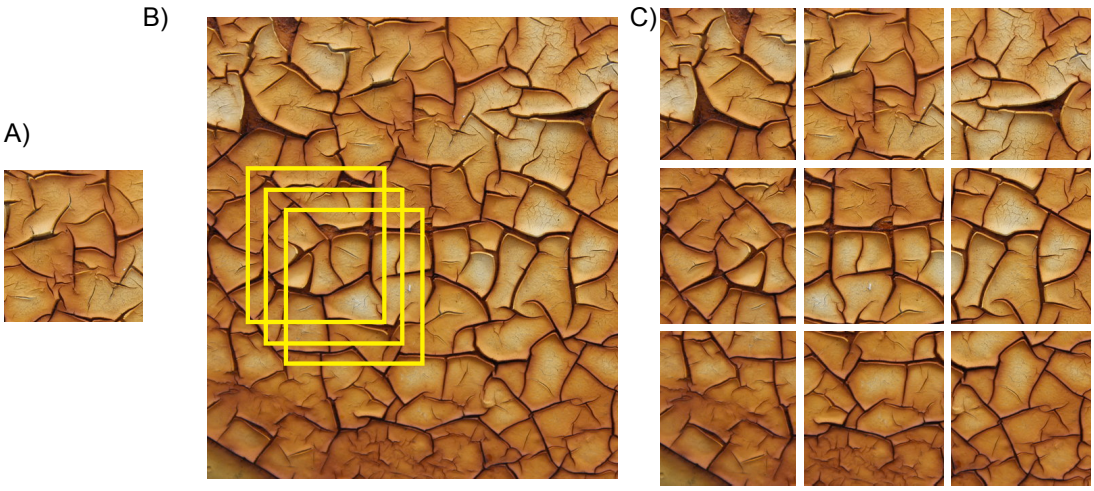


Figure 5.1: Visual search space is easier when the search space is discrete. A) Target texture. B) In this search space, the visual system must generate its own segmentation by dynamically attending to groups of features. A large number of overlapping candidate matches (yellow boxes) must be evaluated. C) In this pre-segmented search space, there are only nine candidate matches, and search is much easier - provided that the target is one of the blocks into which the search space has been segmented.

the task appears harder than in the discrete case (shown alongside). The visual search literature has neglected continuous search, focussing on tasks in which the search space consists of a set of individual objects[221].

The same divide between discrete and continuous occurs in the temporal domain. When a percept is changing rather than static, we may abstract it into a sequence of events (consider a slowly-played sequence of musical notes). On the other hand, we may see it as a constant, unbroken stream of information (consider white noise, or the pattering of rain). The divide is not sharp but gradual. Just as discrete pixels become a texture when we zoom out far enough, a sequence of individual events can be perceived as a continuous stream when we speed it up or add more clutter.

Visual search can also take place in the temporal domain. By "temporal visual search," we mean a search task where the target and the test clip are both dynamic video clips of the same size, the duration of the test clip is longer, and the observer's task is to judge whether the target is present in the test clip. Because the stimulus has an additional temporal dimension, the observer only has one chance to attend to the correct area (either in space or in time) and to perform a useful segmentation. The term "temporal visual search" contains a tacit assumption that the search space also has a spatial dimension; if it did not, our stimuli would have to be points or areas

of constant luminance and colour. We also note that the term "visual search" does not imply that the subject always has to report the space/time location of the target, merely detect its presence.

Temporal visual search on natural scenes, then, presents several difficulties: the stimulus has temporal extent as well as spatial extent, and the scene has not been pre-segmented in either dimension. This makes the task much more challenging than traditional visual search on simple stimuli, and creates problems for existing theories of visual search, which do not need to deal with segmentation. How well can the visual system perform this difficult task? In this chapter we evaluate the visual system's ability to perform temporal search on dynamic clips of flame.

Invariances As we have seen in Chapter 1, theories of object recognition can be characterised by the invariances they possess. Rotation invariance deals with changes in an object's orientation relative to the eye. Position invariance deals with changes in the object's position on the retina, and is one of the most extensively modelled problems in recognition. Training a neural network to recognise an object in a single location is easy; building a position-invariant recogniser is harder[222], but possible[223]. Is there a temporal equivalent to position invariance?

When we say that an ability (such as object recognition or face detection) is position-invariant, we refer to retinal position: the location of the retinal image relative to the fovea. In tasks with fixation, it refers to the ability of the visual system to detect an object regardless of its position on the retina. In tasks where fixation is not required, however, position invariance is not so well-defined, because the fovea can be pointed towards any observed object. Nevertheless, a certain amount of position invariance is still required, because two fixations will never bring an object into exactly the same position on the retina - and it is not possible to bring two slightly different objects into retinal alignment in any case. We can also measure position invariance in some context: relative to another object, such as the screen on which objects are shown.

Temporal invariance can also be defined in relation to a temporal context, a high-level event made up of several low-level events. Within the context of a ten-second video clip, temporal invariance refers to the ability to detect an event whether it occurs at the beginning, middle or end of the clip. This is similar to the ability to detect an object anywhere on the screen. There is, however, no equivalent to requiring fixation

in the temporal domain. There is no anatomical "centre" of the current perception of time. The spatial "centre" of visual perception is the fovea; the temporal "centre" of perception is the instant of "now," which is constantly moving through time. It thus does not make sense to discuss temporal invariance in relation to the instant of "now," and we restrict ourselves to its meaning in relation to a temporal context: a longer temporal event, such as a clip, which contains the feature in question.

It is in this context that the well-established effects of primacy and recency in short-term memory research are described. Events closer to the beginning or end of a sequence are recalled better[224]; here our temporal context is the sequence, not the subjective "now," which does not have a defined position except for the observer experiencing the experiment. What we mean here by "temporal invariance" is similar to a lack of primacy and recency effects in a memory task. When remembering a sequence of objects, our accuracy is higher for objects near the beginning or end of the sequence[225]. Our reference point here is not the subjective "now" but the sequence itself.

"Temporal invariance," then, is the ability to recognise an event whether it takes place at the beginning, middle or end of a longer event in which it is embedded. In our experiments, observers watch a short video clip of dynamic flame (the sample) and then look for it in a longer flame clip (the test). The spatial size of the clips is the same.

Another type of invariance is the focus of much of visual search literature: the number of distractors, or the search space size. This type of invariance is demonstrated when a red bar is easily recognised among a field of blue bars. In this case, we call the effect pop-out or the result of high bottom-up salience. In other cases, such as the search for an X among crosses, invariance is weak or absent and we must work harder to find the target. In some cases, there is no pop-out at all and we must conduct a sequential, conscious scan of our candidate objects in order to locate the target. We can discern which case we are presented with by measuring observers' invariance to the amount of distractors: their search space size invariance.

This chapter describes experiments investigating position and search space size invariance on video of dynamic flames. By studying search space size invariance, we are asking whether observers can detect sequences of dynamic flame in longer

sequences. By studying temporal invariance, we are asking whether the timing of the target event within a longer clip affects observers' ability to find it.

Mechanisms of object recognition in visual search What mechanisms could the visual system use to perform temporal search? We reviewed various models of object recognition in Chapter 1, considering how they might be extended to the temporal domain. In natural scenes, object recognition is very similar to search; it requires ignoring the background and separating the attended object from its distractors. There are a number of potential strategies the visual system may use in detecting a temporal target contained in a temporal sequence, which are considered below.

When presented with a dynamic sample, the simplest strategy is not to encode the temporal component at all. Observers may represent a single, static snapshot sampled from the target, then search for this in later stimuli. In this case, we would not expect better performance for longer targets.

The visual system may instead use multiple static snapshots. In this case, performance may increase with target length, but there may also be limits on the capacity to store snapshots and compare each frame against multiple targets. This strategy may also be more vulnerable to false matches at the frame by frame level, a problem which could be mitigated by matching against a temporal sequence.

Template matching has been frequently proposed as an object recognition mechanism[7, 226, 227]. In this process, a template is "moved" over a viewed object (the template is compared to visual input at a variety of positions in the visual field). The template position which causes a peak in similarity is taken as evidence of the presence of the target at some position in the scene. This method has been used extensively in the computer vision literature[228, 229, 230] and bears much resemblance to the operation of cross-correlation on functions or 2D images.

How could template matching apply to temporal search? A dynamic template generated from the sample (a representation with a temporal component) could be scanned along the extent of the test clip. Precisely, this means that the entire dynamic template of length a is progressively matched with a succession of test clip chunks of length a. These chunks are taken incrementally from the test using a sliding window. Conceptually, the process is similar to finding the cross-correlation of two functions, or convolving a filter with a larger image. In this case, as long as the template were

accurate enough, it would be equally well matched with the target regardless of its position in the test. We would not expect matching accuracy to depend on the length of the test.

Template matching may not be serial: a template may be matched in parallel with different parts of the test stimulus. Reaction time data show this to occur for classic parallel search for a single feature. In temporal search, templates could be matched in a temporally parallel way: the visual system could compare a dynamic template to the beginning and end of a test clip at the same time, but this would require the whole stimulus sequence to be encoded and accessed at the same time, which is unlikely in a biological system.

The previous two models assume that the observer has a good representation of either a static snapshot or a dynamic template. A "snapshot" connotes an accurate low-level representation rather than a heavily processed and downsampled gist. Often, however, information appears at too fast a rate for accurate representation. Due to the attentional bottleneck[231], and the capacity of visual working memory[232], there are limits on the amount of information which can be perceived and then represented.

One strategy the visual system could adopt is to represent a small number of spatial features accurately. According to this strategy, each feature is accurately represented, but most of the information in the stimulus is thrown away. A feature in this context is a set of attributes (such as colour, a shape descriptor and location relative to the stimulus frame) which is bound together as an object representation. This is a very similar entity to the object file proposed by Kahneman[233, 234].

We call this the set-of-features (SoF) strategy. It is characterised by the tendency to encode small spatiotemporal patches with high fidelity, rather than processing information from the entire stimulus and compressing/downsampling it to fit into visual working memory. It takes its name from the similar bag-of-features strategy in computer vision[235, 236], which involves keeping orderless collections of local image descriptors[237]. The approaches are similar in nature but not the same: we mean to convey the accurate representation of a small number of sparsely picked local stimulus patches, rather than the holistic processing of an entire image. We do not apply this constraint here: our usage of SoF specifies the accurate encoding (with or without location) of a small set of local features and the discarding of information from most

of the stimulus.

The set-of-features model is applicable to static images, and can deal in various different ways with dynamic stimuli. The first approach is not to encode time at all: a set of features contains no information about the time of each feature. Alternatively, the time at which each feature occurs could be coded relative to the start and finish of the sample clip. An example is "this distinctive yellow flash occurred 1 second from the start of the clip." This is the temporal analogue of coding spatial location relative to the stimulus frame.

Finally, the times at which each feature in the set occur could be coded relative to each other, but not relative to the stimulus frame. An example is "this yellow shape in the top right occurred before this red shape in the bottom left." This is the temporal equivalent of coding the relative locations of a set of objects in the visual field, but not their absolute locations.

The SoF model can thus be divided into three sub-models: orderless SoF, absolute SoF and relative SoF. These refer to three different ways of imposing structure on separate spatiotemporal features; in other words, binding them into a holistic representation. Orderless SoF does not do this at all, which places it at the local end of the temporal continuum.

Another way to deal with the bottlenecks of representation and storage is to attend to as much stimulus information as possible and then downsample it to construct a gist. In this case, rather than accurately representing small areas of the scene, an observer uses the whole scene to construct a holistic but highly compressed representation. Fast natural scene perception appears to rely on the construction of gists[167, 238].

What is the difference between constructing a gist and constructing a template? They are both holistic; they use information from the entire image, as opposed to accurately sampling a small part. Template models in object recognition usually connote a fairly accurate representation of the stimulus[135, 7, 226], whereas the gist is a more lightweight, compact descriptor[200, 239] and can contain useful high-level information such as its spatial envelope (volume, perspective, openness and level of clutter)[238].

The snapshot models differ from the feature models in the degree of holistic encoding. Snapshot models encode a representation of the entire stimulus (such as the

global shape of the flame outline). Feature models encode only a local patch, but with greater accuracy.

We can draw an important distinction between object recognition models: whether their representations are atomic or not. An atomic code is one which cannot be deconstructed and which does not allow the visual system to access or match its components: it can only be compared or introspected as a whole. The gist is an atomic representation, as is the point in face space which represents a face (it can be deconstructed into loadings on the axes of the face space, but these do not correspond to physical parts of the face: there is no isomorphism between an axis in face space and the retinal image). Snapshot models, both static and dynamic, are also atomic.

Non-atomic representations are those which can be effectively split into pieces which correspond to spatiotemporal areas of the stimulus. The set-of-features models fall into this category: whether densely or sparsely sampled, each feature corresponds to a part of the stimulus.

The distinction between atomic and non-atomic representations is central: an atomic representation can allow introspection of its features ("I recognise that distinctive flare in the bottom left of the screen") and allows matching based on individual features. It also implies less computation, requiring no transform into a high-level space, only the sampling of low-level features.

Finally, we note the multidimensional stimulus space models. Computational techniques such as PCA and ICA often involve the dimensionality reduction of stimuli and their expression as points in a low-dimensional space, for example a face space[240] or an emotion space[241]. When matching dynamic flame, a video clip could be embedded either as a sequence of points in a static stimulus space (each point corresponding to a sample of the entire stimulus) or as a single point in a dynamic stimulus space.

We summarise the models as follows:

- **Static snapshot.** The visual system encodes a representation of a single frame in a holistic manner.
- **Orderless snapshots.** Several holistic snapshots are encoded, but their temporal order is not stored.
- **Ordered snapshots.** Several holistic snapshots are encoded, in order.
- **Set of timeless features.** Several spatiotemporally local features are encoded,

but the visual system is unable to access information about their relative order.

- **Set of relatively ordered features.** Several local features are encoded, along with their relative order (either a simple ordering, or an ordering as well as the delays between feature appearance).
- **Set of absolutely ordered features.** Several local features are encoded, along with their temporal offset from the beginning of the sample clip.
- **Static stimulus space.** Each clip is represented as a series of samples in stimulus space, each one corresponding to an instant of the stimulus.
- **Dynamic stimulus space.** Observers have access to a dynamic stimulus space, each point of which represents a complete clip.

Our experimental approach These approaches represent points in a space of models rather than mutually exclusive descriptions of the temporal search process. The visual system may operate in a regime between models; it may also implement different models in different functional areas. We used visual search experiments to profile the capabilities of the visual search process on dynamic flames and to investigate the underlying mechanisms. Firstly, are humans capable of encoding and matching the complex dynamic forms of fire at all? How much information (how long a sample) do we need to form a matchable representation, and in how big a search space (how long a test clip) can the sample be found?

While the internal mechanisms that implement object recognition are difficult to characterise directly, we can use behavioural measures to profile these mechanisms and examine their capabilities. We began by setting observers a visual search task designed to measure search space invariance, the ability to detect a target in search spaces of varying size. This experiment is an analogue of Treisman's classic visual search tasks, but with several important differences: the stimuli are natural rather than artificial, they are dynamic, and we adjust the duration of the test clip rather than its size.

This experiment used the apparatus and stimuli described in Chapter 2. Each trial involved the presentation of a sample followed by two tests; the observer was asked to indicate which test contained the sample. We manipulated the duration of the sample and test clips and measured observers' matching accuracy.

5.1 Experiment 5.1: Matching dynamic flame samples

5.1.1 Methods

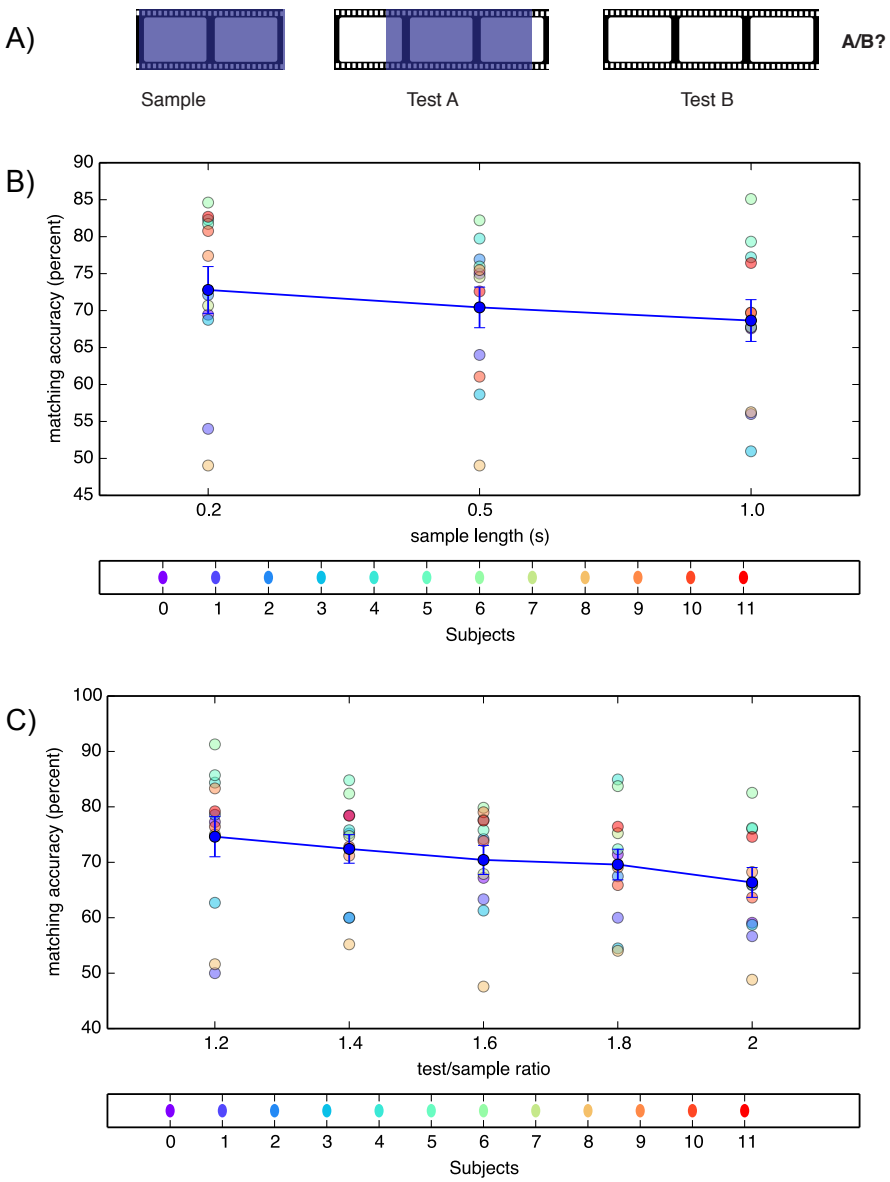


Figure 5.2: Experiment 5.1: visual search in dynamic flame clips. A) Trial structure: a sample is followed by two tests, one of which contains the sample. Observers indicated which test they thought matched the sample. B) Accuracy against sample length; no significant effect was found. C) Accuracy against test/sample ratio: a significant effect was found.

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

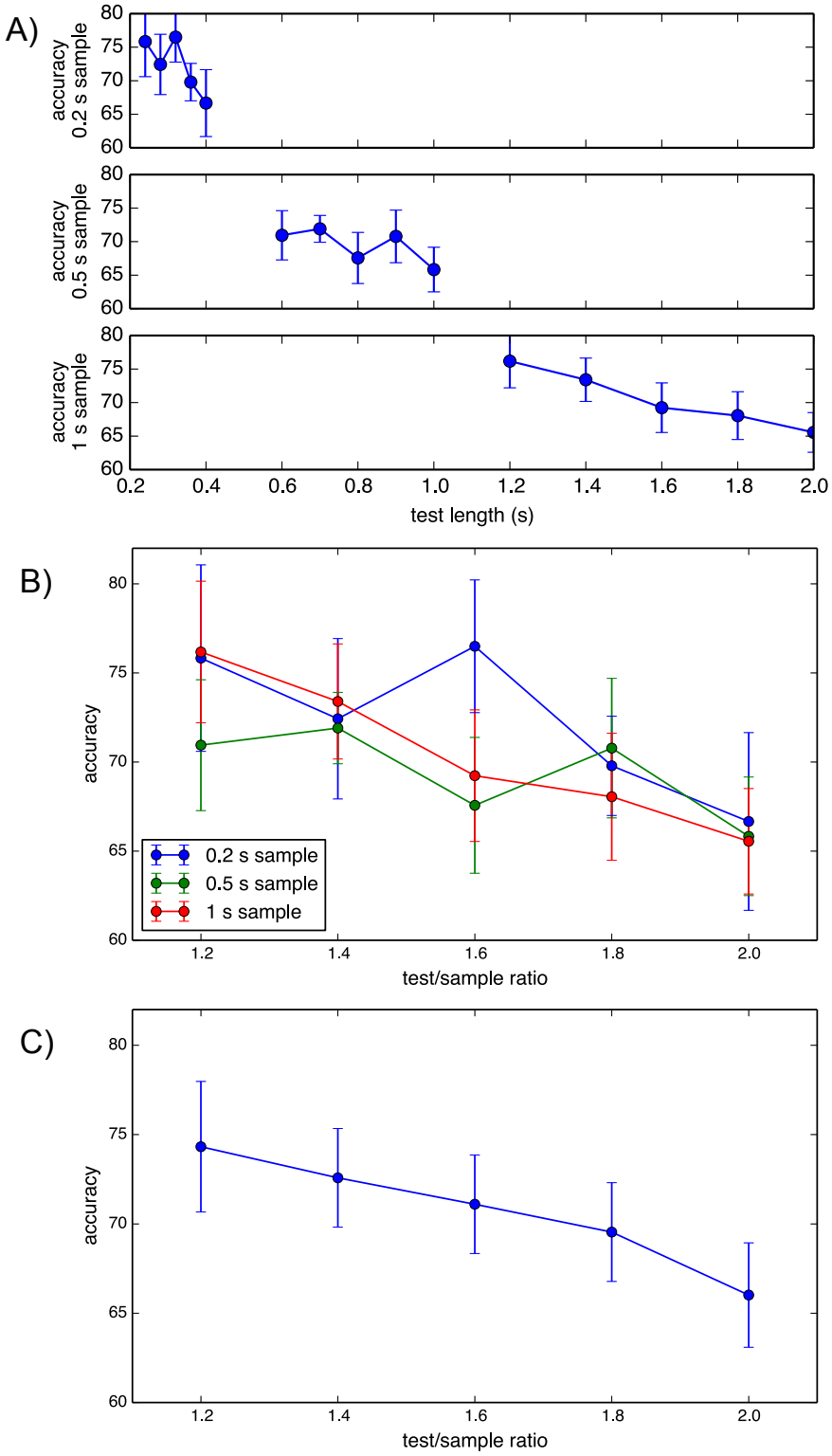


Figure 5.3: Experiment 5.1: accuracy depends mainly on test/sample ratio. Error bars are 1 SEM. The three test/sample ratios correspond to a wide range of different test lengths. B) Accuracy depends on test/sample ratio rather than test length. C) Accuracy against test/sample ratio, averaged over the three sample length conditions.

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

Table 5.1: Experiment 5.1: clip lengths.

Materials We presented stimuli on a CRT monitor in a darkened room as described in Chapter 2 (General methods). Observers used a chin-rest. Stimulus clips were taken from a continuous 1000-frame corpus of flame video.

Design We used a 2AFC delayed match-to-sample paradigm.

Procedure In each trial, a sample was presented first, followed by two longer tests. Using the keyboard, subjects indicated which test they thought contained the sample. Sample length was 10, 25 or 50 frames, equivalently 0.2, 0.5 or 1 second. We varied the ratio of test duration/sample duration (comparable to search space size), setting it to 1.2, 1.4, 1.6, 1.8 or 2. This corresponded to a different set of test clip lengths for each sample length, according to Table 5.1. Each sample lasted 20 ms.

This design provided $3 \times 5 = 15$ conditions. We presented 3 blocks, one corresponding to each target length. We showed 25 training trials first. In this experiment, we varied test length within blocks and sample length across blocks. This meant that on each trial, subjects were aware of the current sample length, but did not know the current test length.

5.1.2 Results

Figure 5.2 shows the data from Experiment 5.1. Chance in this experiment is 50%. We can see from the means that accuracy decreases as the ratio between the test and the sample increases (as the sample lengthens relative to the test). This effect is confirmed by a 2-way repeated measures ANOVA, which 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).

Interestingly, for samples between 0.2 and 1 seconds in length, accuracy does not appear to depend on sample length. This implies that sample encoding occurs in the same way for a range of sample lengths.

Figure 5.3 explores the effect of test/sample ratio in more detail. We see that the three levels of sample length, together with the five levels of test/sample ratio,

generate a wide range of test lengths, which do not overlap between sample lengths. Despite this wide range of test lengths, accuracy ranges between similar extremes and follows the same general trend. Accuracy depends heavily on test/sample ratio.

5.1.3 Discussion

This experiment tested observers' ability to decide which of a pair of test clips contained (temporally) the sample clip. Observers can perform this task effectively, showing that the visual system is capable of representing dynamic flame well enough to perform matching and visual search.

Interestingly, accuracy did not depend on sample length. This means that limits on search performance were not due to observers' not being able to extract enough information from the sample. The effect of test/sample ratio, however, was highly significant (p <0.0001). This means that accuracy depends on the relationship between the amount of information extracted from the sample, and that extracted from the test. In this task, test/sample ratio is a proxy for the amount of distractor video present in the test clip, compared to the length of the sample. On flame clips, temporal visual search is not invariant to search space size: it is highly sensitive.

This is also a difficult task: even at maximum accuracy (with a test/sample ratio of only 1.2) observers only perform at 74%. Observers show high variance in accuracy; as we can see from the individual subject results in Fig. 5.2, some perform consistently well and some consistently badly. We did not reject any subjects *post hoc*; we followed the procedure described in Chapter 2 (General methods), only rejecting subjects if they failed a pre-screen matching test.

5.1.4 Evaluation of models

In this experiment, we find no significant effect of sample length. For 0.2 to 1 second clips, we therefore have no evidence that observers are able to better represent longer templates. This is consistent with two theories: either observers are representing information from a single temporal slice of the sample, or they have reached maximum memory capacity with clips of this length. We investigate this in the next experiment, which uses shorter clips.

The strong dependence of accuracy on search space size suggests that observers are not using a process of dynamic template matching. If an accurate template produced from the sample was simply being scanned along the length of the test, we would not expect such low accuracy with a test/sample ratio of 2.

Our results could be consistent with the matching of a low-quality template, which could cause false positives to occur. In this case, the longer the test, the more potential false positives; this could reduce accuracy. The template, however, is sufficiently good to yield 77% accuracy under the (ratio=1.2, sample length= 0.2 s) condition.

In this experiment, we used samples between 10 and 50 frames (0.2 and 1 second). Are subjects still capable of matching when faced with shorter clips? To investigate, we conducted a similar experiment using samples between 1 and 12 frames (0.02 and 0.24 seconds) in length.

5.2 Experiment 5.2: Matching shorter flame samples

Experiment 5.1 showed us that the visual system can effectively discriminate the complex patterns of motion and form found in fire. To estimate recognition accuracy in shorter clips, we performed a very similar 2AFC delayed match-to-sample experiment, but with shorter clips lasting between 0.02 and 0.24 seconds (1 to 12 frames).

5.2.1 Methods

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

Materials We used a 1000-frame corpus of consecutive fire images, displayed using the equipment described in Chapter 2 (General methods).

Design We employed a 2AFC delayed match-to-sample paradigm.

Procedure In each trial, a sample was presented first, followed by two longer tests. Using the keyboard, subjects indicated which test they thought contained the sample. Sample length was 1,3,6 or 12 frames, equivalently 0.02, 0.06, 0.12 or 0.24 seconds. Test length was one of 15, 20 or 40 frames, equivalently 0.3,0.4 or 0.8 seconds. We

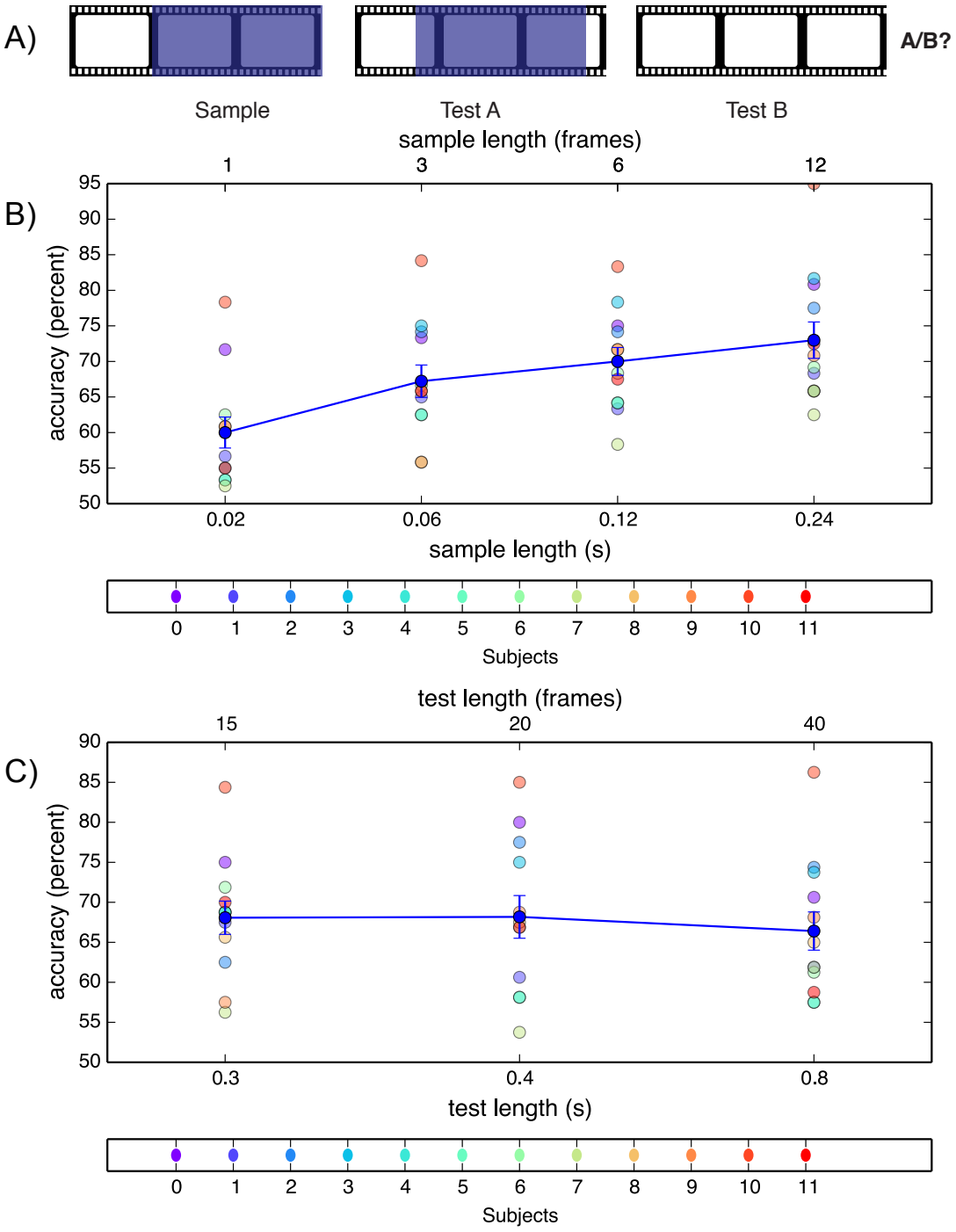


Figure 5.4: Experiment 5.2: when subjects perform a matching task with very short samples (1 to 12 frames), the effect of sample length predominates. A) A short sample was followed by two longer tests, one of which contained the sample. B) Mean accuracy against sample length (averaged over levels of test length), shown in frames and seconds. Subjects match longer samples with greater accuracy. C) Mean accuracy against test length (averaged over levels of sample length), shown in frames and seconds. There is little effect of test length. Error bars are 1 SEM.

Sample length (s)	Mean accuracy (%)
0.02	60
0.06	67
0.12	70
0.24	73

Table 5.2: Experiment 5.2: sample lengths and accuracies.

varied test length across blocks, and sample length within blocks. This meant that subjects were aware of the next trial's test length, but not its sample length. We presented 24 training trials at the beginning of the experiment. There were 3 $\times$ 4 $=$ 12 conditions. We presented 3 blocks, one corresponding to each test length, in random order. 480 trials (40 per condition) were presented in total.

5.2.2 Results

Figure 5.4 shows the data from Experiment 5.2. Chance in this experiment is 50%. We can see that mean accuracy increases as the sample lengthens. This effect is confirmed by a two-way repeated measures ANOVA, which reveals a significant effect of sample length (p <0.0001) but not of test length (p =0.652), with no interaction between the two (p <0.395).

Table 5.2 shows mean accuracy by sample length. Even when presented with a sample consisting of a single frame (0.02 s), accuracy was greater than chance (mean accuracy 60%).

5.2.3 Discussion

This experiment covers a shorter range of sample lengths than Experiment 5.1: 0.02 to 0.24 s, as opposed to 0.2 to 1 seconds. Here, we are investigating observers' ability to extract useful information from very short dynamic samples.

Even when presented with a sample consisting of a single frame (0.02 s), accuracy was greater than chance (mean accuracy 60%). Observers can thus extract useful information from a single frame. This shows that they do not require a long temporal integration period to build useful representations, which supports bag-of-features theories rather than dynamic snapshot theories. The more frames available in the sample, the more useful information observers can extract and the higher the accuracy. This

suggests that observers are not using a static snapshot model, as in this case, we would not expect an effect of sample length.

The 0.24 s sample, 0.8 s test condition in this experiment is very close to the 0.25 s sample, 0.8 s test condition in Experiment 5.1. Respective mean accuracies across subjects are 71% and 69%. The small differences in accuracy across these conditions, provides a useful confirmation of our data. In terms of protocol, in Experiment 5.1 we vary sample length across blocks, whereas in Experiment 5.2 we vary it within blocks. Observers, noticing this constancy, had prior information as to the length of the next test (except at the beginning of each block). We varied sample length within blocks, so the length of the next sample was not knows. Not having this information does not appear to limit observers' ability to perform the task.

In Experiment 5.1 we demonstrated that for 0.2 to 1 second samples, accuracy depends on test/sample ratio and not on sample length. Here, for shorter clips of 0.02 to 0.24 s, we find the opposite pattern: accuracy depends on sample length and not test length. This is consistent with a matching strategy which samples continuously from the sample clip as opposed to encoding a static snapshot, but reaches saturation for the longer set of clips. In Experiment 5.1 the longer clips appear to have filled observers' memory capacity: observers gain no benefit from longer clips. In Experiment 5.2 the clips are shorter, meaning that varying their size allows more information to be encoded. Here, there is no effect of test length, which indicates that the smaller search space does not provide a sufficient challenge to drop performance.

This experiment demonstrated observers' ability to recognise very short flame clips, down to a single frame, presented within longer clips. Because the sample clip was randomly placed within the test clips, this experiment could not examine the second type of invariance we highlighted: invariance to target position. In order to investigate this, we performed another matching experiment in which we explicitly manipulated the duration of distractors played before and after the target.

5.3 Experiment 5.3: Visual search in more detail

Here we used a similar matching task to that of Experiment 5.1. Sample length was fixed at one second (50 frames), requiring the visual system to encode a significant

amount of information. We did not directly control test length, but rather manipulated the length of the clip played before the target (the pre-length) and that of the clip played after the target (the post-length); see Fig. 5.5. Test clips were always continuous, since on each trial the test clip was picked first and the sample clip was selected from within from the test clip. To allow more trials and in order to use signal detection theory, we used a yes/no instead of a 2AFC design.

5.3.1 Methods

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

Materials We used a larger 10,000-frame corpus of consecutive fire images, displayed using the equipment described in Chapter 2 (General methods).

Design We used a Yes/No delayed match-to-sample paradigm.

Procedure In each trial, a sample was presented first, followed by a longer test. Using the keyboard, subjects indicated whether they thought the test contained the sample, or did not contain the sample. Samples were all one second (50 frames). True test clips consisted of the sample, temporally surrounded by a pre-clip and a post-clip, which could both be of length zero. Foil clips consisted of a randomly-chosen clip equal in length to (1 + prelength + postlength) seconds. The minimum-length test was therefore also one second in length. The lengths of the pre-clip and the post-clip (which we term prelength and postlength) were either 0, 25, 50, or 100 frames, equivalently 0, 0.5, 1, and 2 seconds. Each factor thus had four levels, giving us 16 conditions. We presented 30 training trials whose samples and tests were all 1 second in length. We varied both pre-length and post-length within blocks, and the experiment was divided into 10 blocks. There were 400 trials in total (25 trials per condition).

5.3.2 Results

Figure 5.5 shows the data from Experiment 5.3. We can see that accuracy drops as total test duration increases; in other words, as the potential for the presence of a distractor increases. This effect is confirmed by a one-way repeated-measures ANOVA, which reveals a highly significant effect of total time (p <0.001).

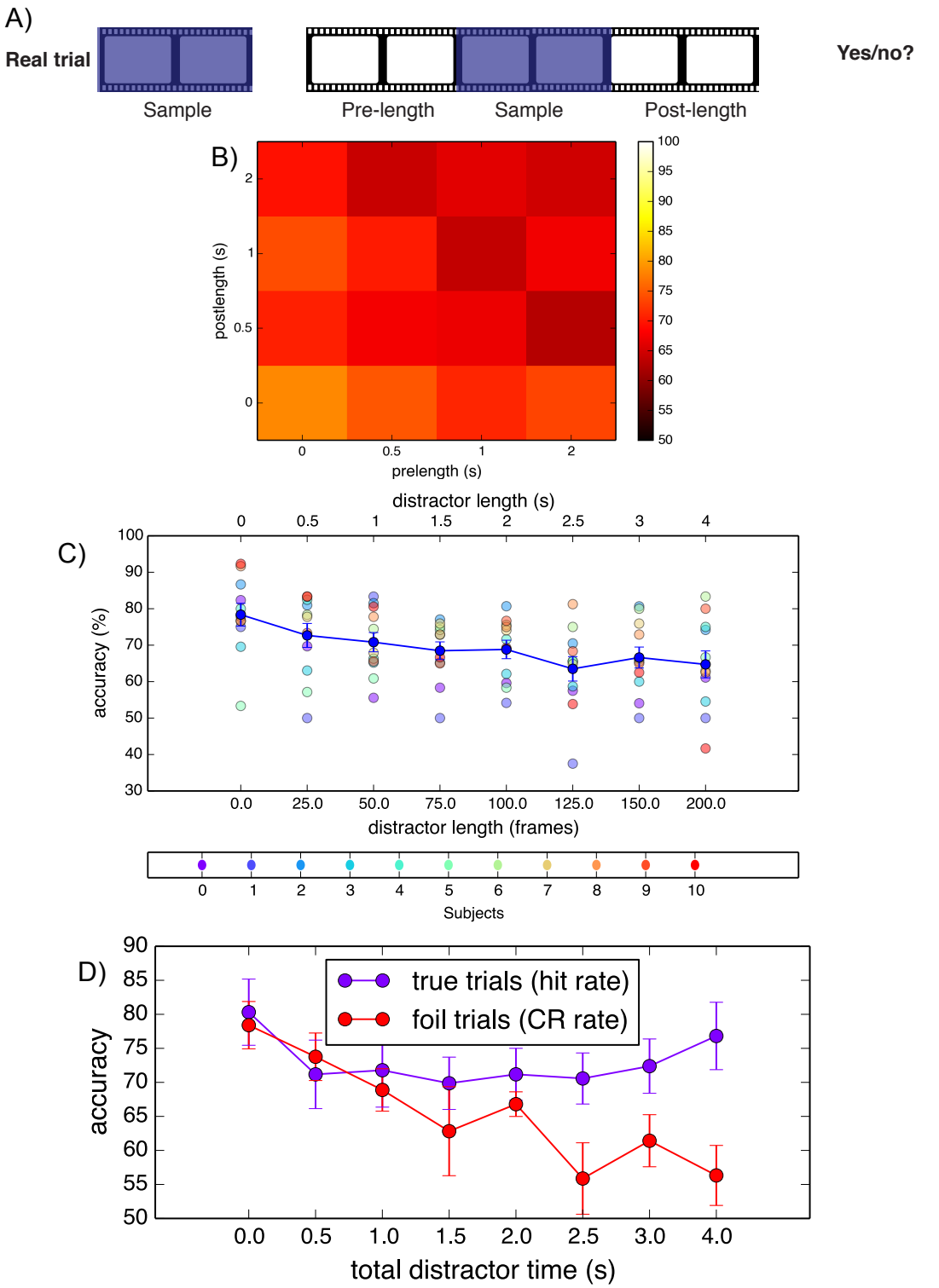


Figure 5.5: Experiment 5.3: we manipulated the lengths of the clips played before and after the target, the pre-length and the post-length. A: trial layout. B: heat map showing accuracy in function of pre-length and post-length. C) Accuracy drops as total distractor time rises; this effect is highly significant. D) Hit rate and correct rejection rate against total distractor time. The hit rate remains constant, whereas the correct rejection rate drops, showing that observers are better at noticing the sample than rejecting a foil test. Error bars are 1 SEM.

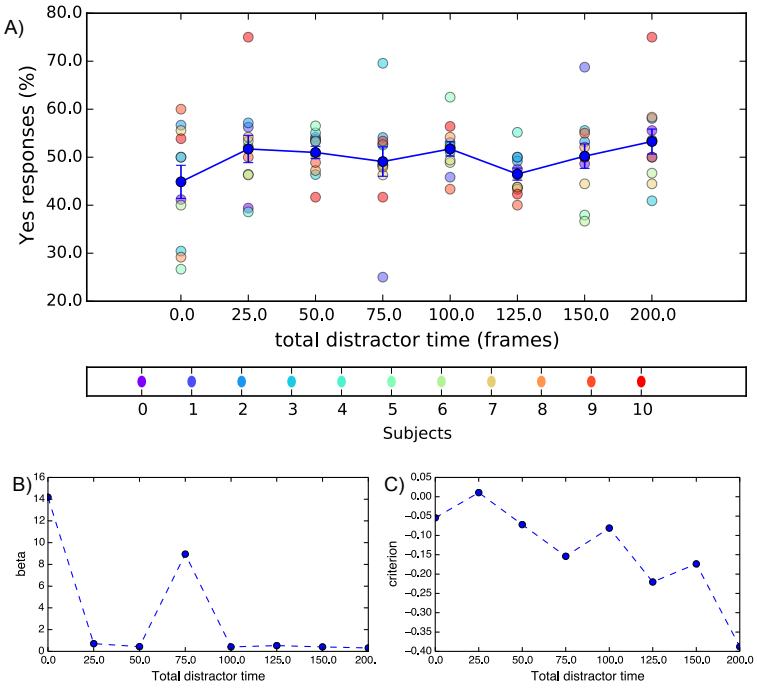


Figure 5.6: Experiment 5.3: measures of observer bias. A) Proportion of Yes responses against total distractor time. B) Bias (β) in function of total distractor time. C) Criterion distance c in function of total distractor time. Both measures indicate that observers are mostly biased towards responding Yes.

The effects of pre-length and post-length on accuracy are shown in Fig. 5.5B. Only the true trials are included in this calculation, since foil trials have no target and thus pre-length and post-length are not defined. With pre-length and post-length as factors, a two-way repeated-measures ANOVA on the true trials shows no significant effect of prelength (p =0.414) or postlength (p =0.373).

This experiment used a Yes/No paradigm, allowing us to classify each trial as a hit, miss, correct rejection or false positive. Figure 5.5D shows this data. We see a strong difference between the hit rate (accuracy on true trials) and the correct rejection rate (accuracy on foil trials). As total distractor time increases, the hit rate stays constant, while the correct rejection rate drops. This means that decreasing accuracy is explained mostly by observers' mistaking foil tests for true tests (false alarms), rather than missing true tests. Observers are much better at detecting a real target than rejecting a false target: they tend to confuse distractors for the target.

Criterion distance(c) and bias (β) measures were calculated for each observer. Their mean values in function of total distractor time are shown in Fig. 5.6. Values of β are below 1 except for two points, showing a bias towards responding Yes. The

c measure estimates the distance between the criterion and the neutral point (the point at which Yes and No responses are equally likely). All values of c except for one are negative, confirming that observers are biased towards responding Yes. This bias appears to increase as distractor time rises. Also shown is the proportion of Yes responses in function of total distractor time, confirming that observers are more likely to respond Yes as the distractor load increases.

5.3.3 Discussion

In the 1-second test condition, observers' matching accuracy is 80%, showing that they can match a clip with its counterpart with high precision when no distractors are present. Accuracy drops to 65% for 5-second samples with 4 seconds of distractor video. This figure is much lower but still above chance (one-sample t-test, $p <$0.0001). This experiment uses a much larger test/sample ratio than the previous two, showing that observers are still able to perform the task when they must deal with four times as much distractor video as sample video.

The difference between the hit rate and the correct rejection rate shows that observers are very good at detecting present targets; they show a hit rate of nearly 80% even with four seconds of distractor video. The correct rejection rate, on the other hand, drops quickly to 55% under the same conditions: observers cannot accurately reject foil clips, and mistake them for clips containing the target.

5.3.4 Evaluation of models

There was no significant effect of pre-length, which gives us no evidence that targets at the beginning of the sample clip were better recognised. This gives us no support for models whose temporal coding is start-relative, since such models predict easier matching and higher accuracy for targets at the beginning of the test.

The asymmetry between hit rate and correct rejection rate is interesting, but has limited power to discriminate between models. It could be due to a high-level cognitive strategy, such as "respond Yes if you see the target, and if you do not see the target, respond Yes anyway." We can see from Fig. 5.6 that observers have an increasing bias towards Yes responses as the distractor load increases. If we assume that observers have access to a level of confidence in their decision, and that their confidence level

decreases as distractors increase, then this pattern is also consistent with the strategy "respond Yes if you are not sure."

With no distractors, the hit rate and correct rejection rate are the same; we only see a difference with increasing distractor length. This indicates that the asymmetry has something to do with the matching process rather than the representation process.

This experiment manipulated prelength and postlength separately, showing that the recognition of dynamic flame is highly sensitive to search space size, but not to how much of the clip was played before and after the target. As in previous experiments, however, we manipulated the amount of distractor video. In the next experiment, in order to examine sensitivity to target position and not search space size, we hold the test length constant and vary the target's position.

5.4 Experiment 5.4: Position dependence

5.4.1 Methods

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

Materials We used a 10,000-frame corpus of consecutive fire images, displayed using the equipment described in Chapter 2 (General methods).

Design We employed a Yes/No delayed match-to-sample paradigm.

Procedure In each trial, a 1-second sample was presented first, followed by a 3-second test. For true trials, we manipulated the position of the target in the test. For foil trials, we displayed a test clip which did not contain the sample. Sample offset (the delay between the start of the test and the start of the sample) was either 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 or 100 frames, equivalently 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8 or 2.0 seconds. Observers indicated whether they thought the sample was present in the test using the keyboard. We ran the experiment in short blocks, varying the offset within blocks.

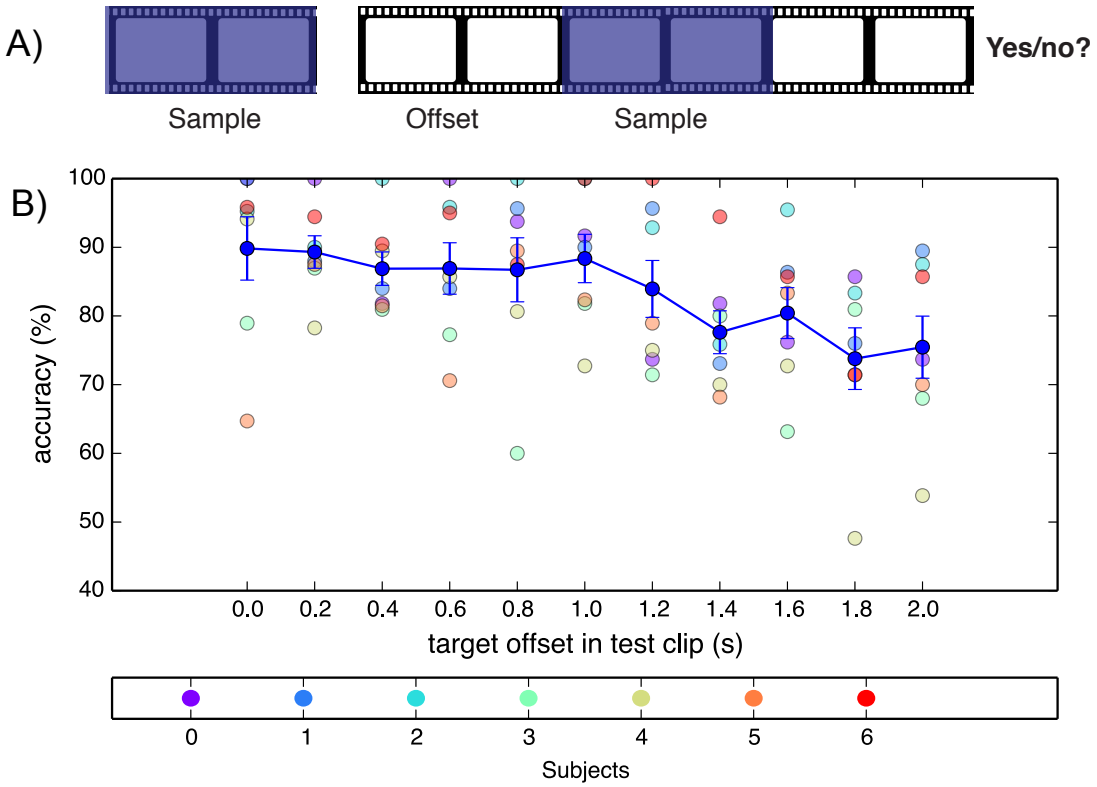


Figure 5.7: Experiment 5.4. with sample and test length held constant, the target's position in the sample (offset) was varied. We used a 1-second sample and a 3-second test, allowing offsets of up to 2 seconds. These data are taken from the true trials (those which contain the target) as offset is not defined for foil trials. There is highly significant effect overall, but accuracy does not drop until an offset of 1 second, indicating that detection is position invariant for small offsets from the beginning.

5.4.2 Results

Figure 5.7 shows the data from Experiment 5.4. As with the previous experiment, the offset is only known for the true trials, which make up half of the experiment. Looking at the true trials, we find a smooth decrease in accuracy as the target approaches the end of the test. This effect is confirmed by a one-way repeated-measures ANOVA on the true trials (p <0.0001). Within-subjects contrasts indicate that the trend is linear (p <0.005).

Offset is only known for the true trials. When analysing these data for effects due to offset, we cannot count correct rejections or false positives- only hits and misses. This means that signal detection theory is not applicable here.

5.4.3 Comparison to Experiment 5.3

Experiment 5.3 manipulated the total distractor time, with the target played at various positions during the test. We observed a decline in the correct rejection rate and a steady hit rate.

Experiment 5.4 kept the total distractor time constant, manipulating the position of the target in the test. We noted a decrease in the hit rate and a steady correct rejection rate.

In Experiment 5.4, the (constant) length of the test was always known in advance. Observers therefore had the opportunity to realise they were nearing the end of the test and alter their response bias.

5.4.4 Discussion

There is a strong effect (p <0.0001 by one-way repeated-measures ANOVA) of target position on accuracy. This confirms that dynamic flame matching is position-sensitive: when the target occurs later on in the test, it is more difficult to detect.

Why is a later target more difficult to detect? One reason could be that the sample's representation decays over time, independently of any new visual information. In this case, we would expect to find a drop in accuracy if a long pause was introduced between sample presentation and test presentation. Another possibility is that viewing the test interferes with the representation of the sample. In this case, we would expect

no drop in accuracy when introducing a pause (during which the observer is not viewing any stimuli) between sample and test. We test this case in the next experiment.

5.4.5 Fit to models

This experiment showed a strong effect of position dependence: detection of dynamic flame sequences is not position-invariant. If we assume that the sample does not decay during the task, this supports models in which time is encoded relative to the beginning of the clip. Temporal information coded in this way would not need to be transformed when matching targets at zero offset, but would need to be transformed when matching targets later in the clip.

In Fig. 5.7, accuracy appears to remain high until offset reaches 1 second, then drops. This could be consistent either with a temporal coding scheme which is invariant until it has to transform representations of time by a certain amount, or with a representation which does not begin to decay until 1 second into the test presentation. It is difficult to approach this question psychophysically, since we do not have access to representational quality data on a trial-by-trial basis.

Lowered performance when the target is later in the clip could be due to a decaying representation of the sample. To test this possibility, we performed a further matching experiment in which we varied the length of the interstimulus interval (ISI) between the sample and the test.

5.5 Experiment 5.5: Memory

5.5.1 Methods

Observers 7 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Materials We used a 1000-frame corpus of consecutive fire images, displayed using the equipment described in Chapter 2 (General methods).

Design We employed a yes/no delayed match-to-sample paradigm.

Procedure In each trial, a sample clip was presented first, followed by a variable interstimulus interval (ISI) and finally a test clip. Observers indicated whether they

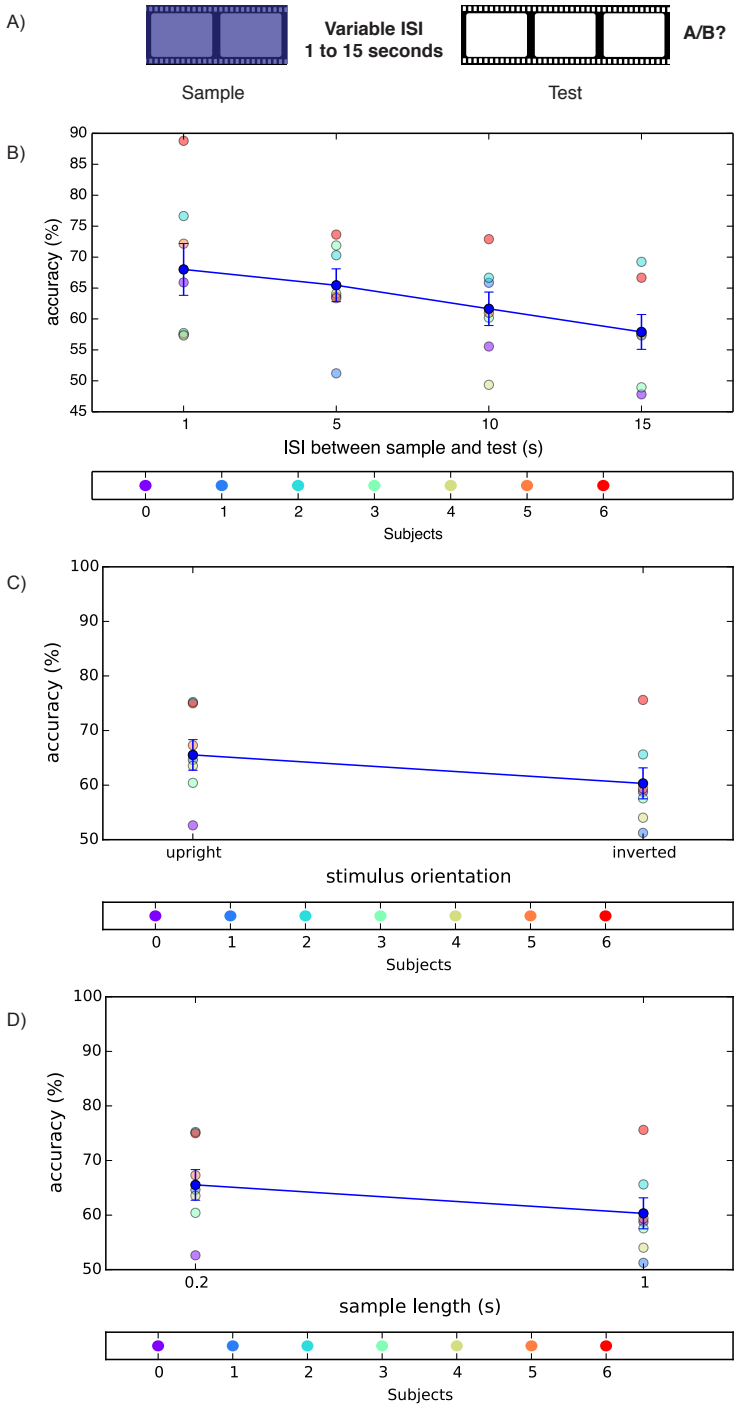


Figure 5.8: Experiment 5.5: the effects of a variable-length blank-screen ISI on matching accuracy. Error bars are one SEM. A) A sample (0.2 or 1 second) was followed by a blank-screen ISI (1 to 15 seconds) and then a test clip (0.24 or 1.2 seconds). B) Mean accuracy decreases rapidly, approaching chance for the 15-second ISI. C) Mean observer accuracy plotted against stimulus orientation; there is no significant drop due to inversion. D) Mean observer accuracy plotted against sample length, which caused no significant effect.

thought the sample was present in the test using the keyboard. In half the trials, both clips were inverted; in the other half, they were both upright. We used samples of either 10 frames (0.2 s) or 50 frames (1 s). Test length was 1.2 times the sample, corresponding to 12 frames (0.24 s) or 60 frames (1.2 s). On true trials, the sample was randomly placed within the test; on foil trials, the sample did not contain the test. The ISI was either 1, 5, 10 or 15 seconds. Sample length was varied across blocks, while the ISI was varied within blocks. We used three repetitions of each type of block (6 blocks total). There were 36 trials per condition (a total of 288 trials).

5.5.2 Results

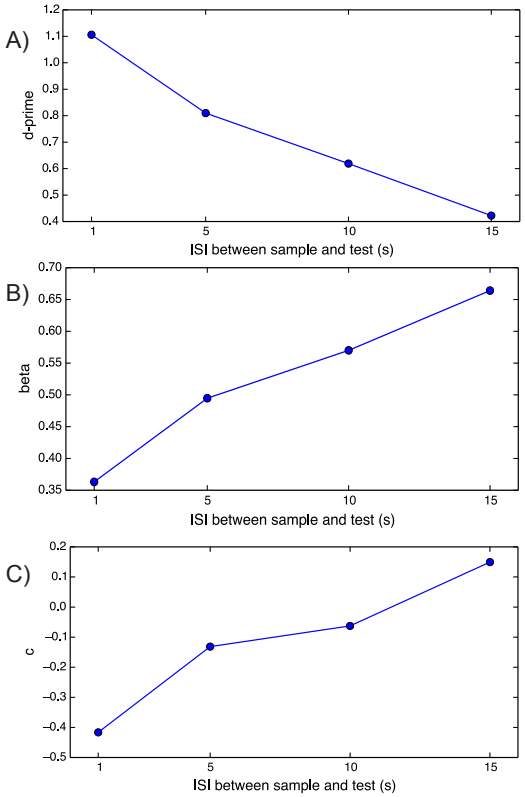


Figure 5.9: Experiment 5.5: the effects of a variable-length blank-screen ISI on observer sensitivity and bias. Measures from signal detection theory were calculated separately for each observer; their mean values are shown here. A) d-prime decreases rapidly as the ISI increases. B) β, a measure of bias, increases. Here the mean β across observers is shown. C) c, an estimator of observers' decision criteria. Both bias measures indicate that a "no" response is more likely as the delay increases.

Figure 5.8 shows the data from Experiment 5.5. Accuracy drops severely as the ISI lengthens; this effect is confirmed by a one-way repeated-measures ANOVA (F(1,10)

Sample length	ISI = 1 s	ISI = 5 s	ISI = 10 s	ISI = 15 s
0.2 s	0.72	0.63	0.60	0.66
1 s	0.62	0.67	0.60	0.52

Table 5.3: Experiment 5.5: Mean accuracies (%) by ISI and sample length.

= 3.43, p=0.001). Trend analysis using within-subjects contrasts suggests that the trend is linear ($p = 0.009$). Mean accuracies by ISI and sample length are shown in Table 5.3.

There is little effect of sample length: a two-way repeated-measures ANOVA shows no effect of sample length ($F(1,6) = 0.2$, $p = 0.67$) but a significant effect of ISI ($F(3,18) = 4.496$, $p = 0.016$) and no interaction between the two factors ($F(3,18) = 1.32$, $p = 0.299$). Trend analysis using within-subjects contrasts suggests that the trend is linear ($F(1,6) = 11.58$, $p = 0.014$).

A two-way repeated-measures ANOVA on inversion and ISI revealed no effect of inversion, ($F(1,6)$=2.171, p=0.19), but confirmed an effect of ISI, ($F(3,18)$=3.575, p=0.35).

Fig. 5.9 shows measures of sensitivity and bias from signal detection theory. We calculated an individual d-prime value for each observer. We can see that observers' mean d-prime drops as the ISI increases, showing that their sensitivity depends on the ISI. This trend is supported by a one-way repeated-measures ANOVA performed on observers' individual d-prime values. This test reveals a significant effect of ISI ($F(3,18) = 4.66$, $p = 0.014$) which, according to trend analysis with within-subjects contrasts, appears to be linear ($F(1,6) = 11.54$, $p = 0.015$).

Under the condition with the longest ISI (15 seconds), observers' mean accuracy is 57% and their mean d-prime is 0.42. A one-sample, one-tailed t-test shows that mean accuracy is still above chance in this condition ($n = 7$, $p = 0.04$).

We calculated criterion distance (c) and bias (β) measures for each subject; their mean values as a function of ISI are also shown in Fig. 5.9. In Experiment 5.3, these measures indicated that observers were more likely to respond Yes with increasing distractor load. Here we find the opposite pattern: the longer the ISI, the more likely observers are to respond No.

A decayed representation biases observers towards a No response; increasing distractors bias observers towards a Yes response. This indicates that separate mecha-

nisms may be at work. In both cases, observers should have less confidence in their decision; the cognitive strategy "respond Yes if your confidence is low" does not explain the pattern found in this experiment.

5.5.3 Discussion

Observers are capable of encoding both 0.2 s and 1 s samples and effectively matching them with tests of the same length, over delays of between 1 and 15 seconds. With a 15-second ISI, accuracy drops to 57%, but is still significantly above chance. This low accuracy and high p-value ($p = 0.04$) indicates that 15 seconds is close to the maximum ISI with which observers perform above chance. The observation that the accuracy curve drops linearly supports this hypothesis; its gradient is not likely to lessen, which would delay the curve's drop to the 50% level.

This rapid drop in accuracy shows that representations of dynamic fire decay very quickly, even in the absence of interference from new dynamic stimuli (observers were presented with a grey screen during the ISI). The information retained by observers' visual systems, then, is not stable enough to survive for much longer than 15 seconds.

5.5.4 Comparison to Experiment 5.4

In Experiment 5.4, we found that targets positioned later in the test clip were detected less effectively. What do the results of Experiment 5.5 mean in this context?

In Experiment 5.5, accuracy is at 89% when the target is presented at the beginning of the test, but drops to 76% when observers have seen 2 seconds of test before the 1-second target is displayed. In Experiment 5.5, there are two conditions: one using a 1-second target and the other using a 0.2 second target. For the 1-second target, accuracy is at 62% with a 1-second ISI, 67% for a 2-second ISI and 52% for a 15-second ISI.

These conditions are not directly comparable, as observers are performing a slightly different task: search in a test 3 times longer than the test (Experiment 5.4) as opposed to search in a test 1.2 times longer (Exp. 5.5). In Experiment 5.4, sample and test lengths are constant; observers have more opportunity to learn that particular configuration.

However, Experiment 5.5 provides convincing evidence that sample representations take much longer than 2 seconds to decay substantially. In Experiment 5.4 a 2-second offset induces a 13 pp accuracy drop compared to an 0-second offset. We only find a 10-pp drop in Experiment 5.5, even with a 15-second ISI. This indicates that the accuracy drop in Experiment 5.4 is due mostly to the effect of distractors, not simply the delay between seeing the sample and seeing the test. There appears to be an interference effect, not simply a decay effect.

5.5.5 Learning

In the same way as in Chapter 4, we examined whether observers were learning new representations by asking whether their matching accuracy increased as they progressed through the experiment. For each experiment, we arranged the trials in the order in which they were presented, blocked them into sequential groups of 20, and calculated the average accuracy for each block.

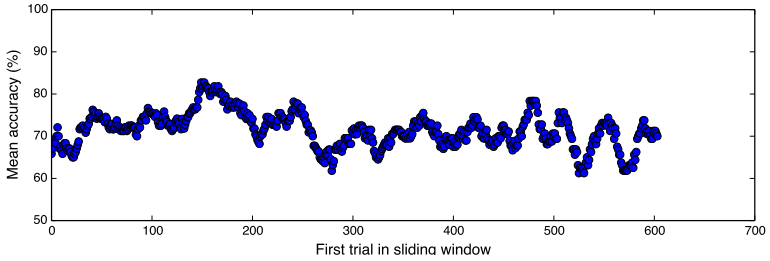
The results of this sliding-window approach are shown in Figure 5.10. To check for an improvement in mean accuracy, we fitted a line to the sequentially arranged data. Calculated slope values are shown in Table 5.4.

Experiment	Slope (percentage points per trial)
5.1	-8.20×10^{-4}
5.2	-6.09×10^{-4}
5.3	1.63×10^{-4}
5.4	-4.04×10^{-3}
5.5	-6.14×10^{-4}

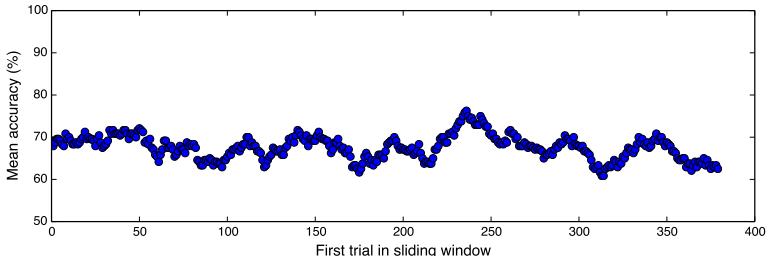
Table 5.4: Learning slopes in Chapter 5.

These slope values are all too small to suggest a consistent increase in accuracy across the entire experiment. Most of the experiments show a gradual and consistent increase in accuracy over approximately the first 30 trials. Observers are not capable of performing the task at full accuracy immediately.

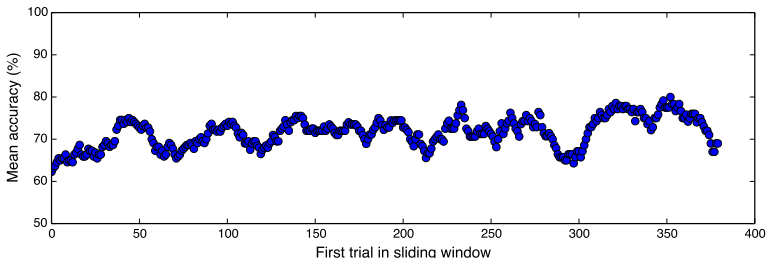
None of the experiments, however, show a consistent increase in accuracy over the entire sequence of trials. Observers do not appear to be learning the task after the first 30 trials.



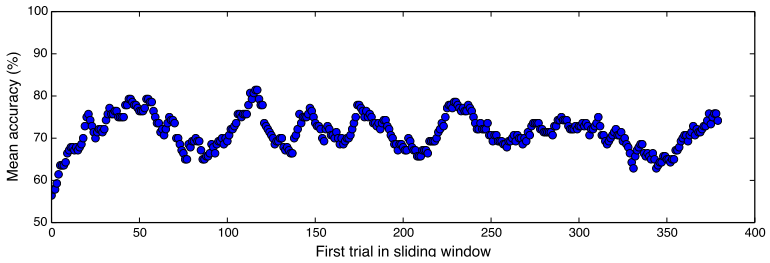
(a) Experiment 5.1



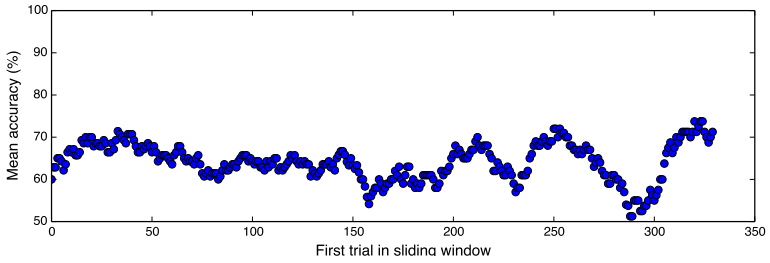
(b) Experiment 5.2



(c) Experiment 5.3



(d) Experiment 5.4



(e) Experiment 5.5

Figure 5.10: For each experiment, trials were aligned in order of presentation and a sliding average applied to show how accuracy changed during the experiment. There was no overall trend of increasing accuracy, suggesting that observers did not learn or improve any useful representations during the course of the experiments. Experiments 6,7 and 8 showed gradual improvement over the first 50 trials, but not after this.

5.6 General discussion

We explored five aspects of the flame matching process: its dependence on test/sample length ratio, sample length, search space size, target position (amount of distractors seen before the target), and time elapsed between test and target.

Experiment 5.1 asked subjects to match samples between 0.2 and 1 seconds with longer tests. Accuracy was found to depend mostly on the test/sample ratio, not on sample length or test length. Under these circumstances, the limiting factor is the relationship between the amount of information in the sample and the amount of information in the test. These results suggest that these clips saturate observers' encoding capacity, and show that their detection is not invariant to test length.

Experiment 5.2 repeated a near-identical procedure but used shorter samples (0.02 to 0.24 seconds). In this case, accuracy depended more on sample length than on test length. Observers are still capable of matching 1-frame samples with above-chance accuracy. However, with very short samples, the limiting factor appears to be the amount of information which can be extracted from the sample. These results suggest that clips of this length do not saturate observers' working memory and that matching is invariant to search space size for clips this short.

Experiment 5.3 used 1-second samples and explicitly manipulated the length of the distractor video played before the target and after the target. These factors had no significant effect individually, but increasing total test time lowered accuracy significantly. Here a Yes/No paradigm revealed a striking asymmetry between hits and correct rejections: for longer tests, there were many more hits than correct rejections. Observers have a tendency to mistake a foil for the target, but not vice versa.

Experiment 5.4 measured the (temporal) position invariance of dynamic flame matching by varying the position of a 1-second target in a 3-second test. Although there was no significant effect on accuracy, investigation of the hit rate and correct rejection rate revealed position-dependent changes. As target offset increased, observers made fewer hits and more correct rejections.

Experiment 5.5 measured the stability of sample representations over time. Observers' accuracy descended close to chance after 15 seconds, showing that representations of flame decay quickly.

Sequential trial analysis revealed that, during each of these experiments, observers'

accuracy does not increase during the course of our experiments; they do not appear to learn more useful representations as the task progresses, except for a short acclimatisation period in the first 50 trials.

Together, these results allow us to build up a picture of the temporal visual search process for dynamic flame. It is a difficult task, with high variance in accuracy between observers. Search is highly vulnerable to distractors, with samples between 0.2 and 1 seconds being found with 66% accuracy when the test is only twice as long as the sample. Search is also highly position-dependent, with targets which are placed earlier on in the test being found with higher accuracy. According to the decay rate of memory representations, which we measured by proxy in Experiment 5.5, this is likely to be due more to the effect of distractors than loss of representational quality.

What do these data enable us to say about the models of dynamic visual search described at the beginning of this chapter?

Static snapshot If the visual system were encoding a single snapshot from each sample, then scanning the input stream for each snapshot, we would not expect accuracy to depend on sample length. According to Experiment 5.2, 0.24 second samples are matched at much higher accuracy than 0.02 second samples, which allows us to reject the static snapshot model.

Dynamic snapshot This model involves encoding the entire sample as a dynamic snapshot, then scanning this over the test as it is presented. In this case, we would not expect accuracy to depend on test length; according to Experiment 5.1, test/sample ratio has a key effect on accuracy.

Set of features According to this model, a small number of spatiotemporal features are encoded with high precision; information from most of the sample is discarded. We will consider this model alongside the gist model, which posits that observers compute an atomic representation of a stimulus; the gist is a rapidly constructed high-level description of a scene.

In Chapter 4 we showed that much of the information which is useful for matching is contained in dynamic edges. This indicates that local, high-frequency information is being used; but this could just consist of the first stage of processing, with the gist later forming a compressed, further processed representation of the sample.

One key difference between the set-of-features model and the gist model is that the

SoF model relies on accurately encoding small spatiotemporal regions of the sample. We know from the analysis of Chapter 3 that there are few long-range correlations in dynamic flame, so these features are likely to be local, both in space and time. This means that, when viewing the test, it is crucial to attend to the right spatial regions at the right time. Otherwise, the features encoded while viewing the sample will be missed. The SoF model provides a natural explanation for errors due to misses, but not errors due to false positives.

Now consider the gist model. Here, the visual system observes the entire stimulus (a large amount of local information) and generates a relatively low-dimensional summary. The gist is generated quickly and only contains enough information to make coarse, high-level judgements about a scene, such as whether it shows a city or a countryside landscape[239]. The gist is not well-suited to making fine, local discriminations between similar images.

The gist model offers a natural explanation for the observation that errors are mostly due to false positives. On a "no" trial, when faced with a sample clip and a test clip which are subtly different, the visual system has to compare two very similar gists. They will never be as different as, say, the gist of a mountain scene and that of a desert scene. It is therefore likely that observers cannot distinguish between the true target and a similar sequence, a tendency which would result in an overall bias towards false positives. This is exactly what we observed in Experiment 5.3.

A gist is an atomic representation; its load on memory is not directly influenced by the length of the video it is created from. It is therefore tempting to say that if observers represented flame clips by gists, this could not explain the dependence on sample length noted in Experiment 5.2. The set-of-features model samples local spatiotemporal features until its memory capacity is exhausted, allowing it to easily explain this dependence: longer samples allow more information to be stored, at least until maximum capacity is reached, as appears to the case in Experiment 5.1. However, even though the gist may not contain any more information, it may contain more accurate information. The gist of a longer clip could be a more precise representation without causing any additional memory load, in the same way that face space codes can differ in accuracy even though they consist of the same number of coordinates.

The gist is usually reported to contain useful natural scene features like environment

(mountain or city) or the affordances[84]. It is therefore more useful for differentiating between very different natural scenes rather than very similar exemplars of the same scene, as with dynamic fire. The gist is therefore unlikely to carry the precise, detailed information necessary to match flame clips.

Space codes possess the same problem: they are special-purpose, with axes related to specific stimulus attributes. Face spaces use configural measures (such as eye separation) or attributes such as skin colour or gender[240]. Personality factor spaces use traits such as emotional intelligence[242]. If dynamic flame were represented by a space code, this code would have to be either general-purpose or specialised for flame. It is difficult to imagine a space code with axes suitable for representing arbitrary stimuli. Observers' lack of learning effects, combined with their ability to perform flame matching with very small amounts of training, suggest that a special-purpose code is not being acquired during our experiments. It is also implausible that a special-purpose code already existed, given that flame matching is not a common task.

Our results therefore favour the set-of-features model. The sequential encoding of local spatiotemporal features fits well with our sample length effects, as well as requiring less computation. It encodes low-level visual information directly, as opposed to transforming it into a compact scene representation; this fits well with our characterisation of flame as a locally correlated stimulus with few long-range spatial or temporal correlations which gist construction could exploit.

Summary

- We used delayed-match-to-sample tasks to examine the effect of varying search space size on search performance for clips of dynamic flame.
- Experiment 5.1: Search accuracy depends mainly on the ratio between test length and sample length.
- Experiment 5.2: Longer samples were matched more effectively than shorter samples, showing that the visual system is not simply matching a static snapshot.
- Experiment 5.3: Matching was not significantly sensitive to the length of distractor clip played before the target; this suggests that representations of features' temporal location are not coded as an offset from the beginning of the stimulus.
- Experiment 5.4 replicated this lack of dependence on the amount of distractor

video played before the target. Accuracy declined for targets later in the clip, however.

- Experiment 5.5 measured the effect of lengthening the ISI on accuracy, in order to enquire whether the previous result was due to distractors or decay of the sample representation. Results suggest that interference by distractors is responsible for the dependence on search space size shown in Experiments 5.3 and 5.1.

Chapter 6

The decision process in face matching and flame matching

In Chapter 4, we looked at the importance of colour and edges to the dynamic flame matching process. In Chapter 5, we looked at the influence of search space size and target position in flame matching. In this chapter, we compare observers' ability to match dynamic flame stimuli with their ability to match dynamic face stimuli. We also pool data from our previous experiments to address two general questions: whether observers show an inversion effect for dynamic flame, and whether they treat each trial as a separate decision problem (as opposed to being influenced by stimuli or responses from previous trials).

The specificity of visual working memory is a long-standing research question. When studying perception, we can point out visual areas which are relatively unspecialised (such as the retina or V1) or those which are tuned to particular stimuli (such as V5/MT). The same is true when studying memory: we may isolate general, nonspecific memory abilities which do not rely on specialised processing mechanisms (such as the retinal afterimage or iconic memory) or those which rely on the ability to perceive a particular stimulus and represent it with a specialised high-level code (such as short-term memory for faces, which shows an inversion deficit). In previous chapters, we have characterised observers' ability to match and remember dynamic flame stimuli. Are they using a specialised memory store? In this chapter, we compare observers' ability to encode and recall dynamic flame with their ability to match a frequently-studied moving stimulus: dynamic faces.

Neurophysiological studies of monkeys and fMRI results in humans both suggest that both prefrontal cortex and inferotemporal cortex are involved with working memory maintenance[243]. Visual and auditory working memory are functionally separate, and an area of human superior frontal sulcus appears to be specialised for spatial working memory[244]. Do we find specialisation within vision as well? Is visual working memory subdivided into stimulus-dependent components rather than one general store? This problem can be approached by comparing working memory capacity across stimulus classes, but we lack a general capacity measure.

Luck and Vogel [232] found that unifying feature conjunctions into objects increases working memory capacity beyond what is measured for individual, unbound features. This suggests that high-level representations are able to unify low-level encodings and maintain their activity.

Alvarez and Cavanagh[245] used a change detection task to measure memory capacity for line drawings, shaded cubes, random polygons, Chinese characters, letters, and coloured squares. They also used a visual search task to estimate the amount of visual information present in each stimulus class, using search rate as a proxy for information. This demonstrates that memory capacity varies across classes. This could be due, however, to complexity variations between classes, rather than the recruiting of different working memory systems.

Moving faces and dynamic flames are very different stimuli. Faces have been extensively characterised as highly evolved communication channels which use sets of muscles to slowly alter their form[246]. As our image domain analysis shows, dynamic fire is a rapidly-changing stimulus with no common global form. It is not a complex, evolved system but an uncontrolled chemical reaction. How, then, does the human visual system differ in its ability to represent and match stimuli of these two classes?

Consider an experiment in which participants encode and match two stimuli on each trial: a face clip and a flame clip. If faces and flames were encoded using the same memory resource, we would expect two effects: a high across-subject correlation between face matching accuracy and flame matching accuracy, and an effect of capacity limitation. Firstly, if using the same processing resource, observers who are skilled at matching faces would also be expected to be skilled at matching flame. We would therefore expect observers who are good on average at matching faces to also be good

at matching fire. On trials where face matching was accurate, observers must have encoded a high-fidelity representation of the test face; this would leave less capacity in the limited resource for a useful flame representation. Within trials, we would therefore expect face performance and flame performance to be negatively correlated (trials with a correct face judgement would be less likely to show a correct flame judgement).

Before comparing observers' matching performance for faces and fire in a dual-task experiment, we asked them to perform temporal search on faces alone in order to estimate appropriate stimulus durations.

6.1 Pilot study: Delayed match-to-sample on faces

This experiment aimed to replicate Experiment 4, using face stimuli instead of flame stimuli. We presented a short dynamic expression clip, then asked observers to match it to a longer clip. We used larger maximum test/sample ratios than Experiment 4, hypothesising that face matching accuracy would exceed flame matching accuracy.

6.1.1 Methods

Observers 2 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Materials We used the face dataset described in Chapter 2 (General methods).

Design We used a 2AFC delayed match-to-sample paradigm.

Procedure In each trial, a sample was presented first, followed by two longer tests. Using the keyboard, subjects indicated which test they thought contained the sample. Sample length was 10, 25 or 50 frames (0.2,0.5 or 1 second). We varied the ratio of sample to test (an indicator of search space size), setting it to 1.2, 1.6, 2, 2.4, or 2.8. This corresponded to a different set of test clip lengths for each sample length, according to Table 6.1.

Sample length	**Test lengths (s)**
0.2	0.24,0.32,0.4,0.48,0.56
0.5	0.6,0.8,1,1.2,1.4
1	1.2,1.6,2,2.4,2.8

Table 6.1: Face matching: sample lengths.

This design contained 3*5 = 15 conditions.

6.1.2 Results

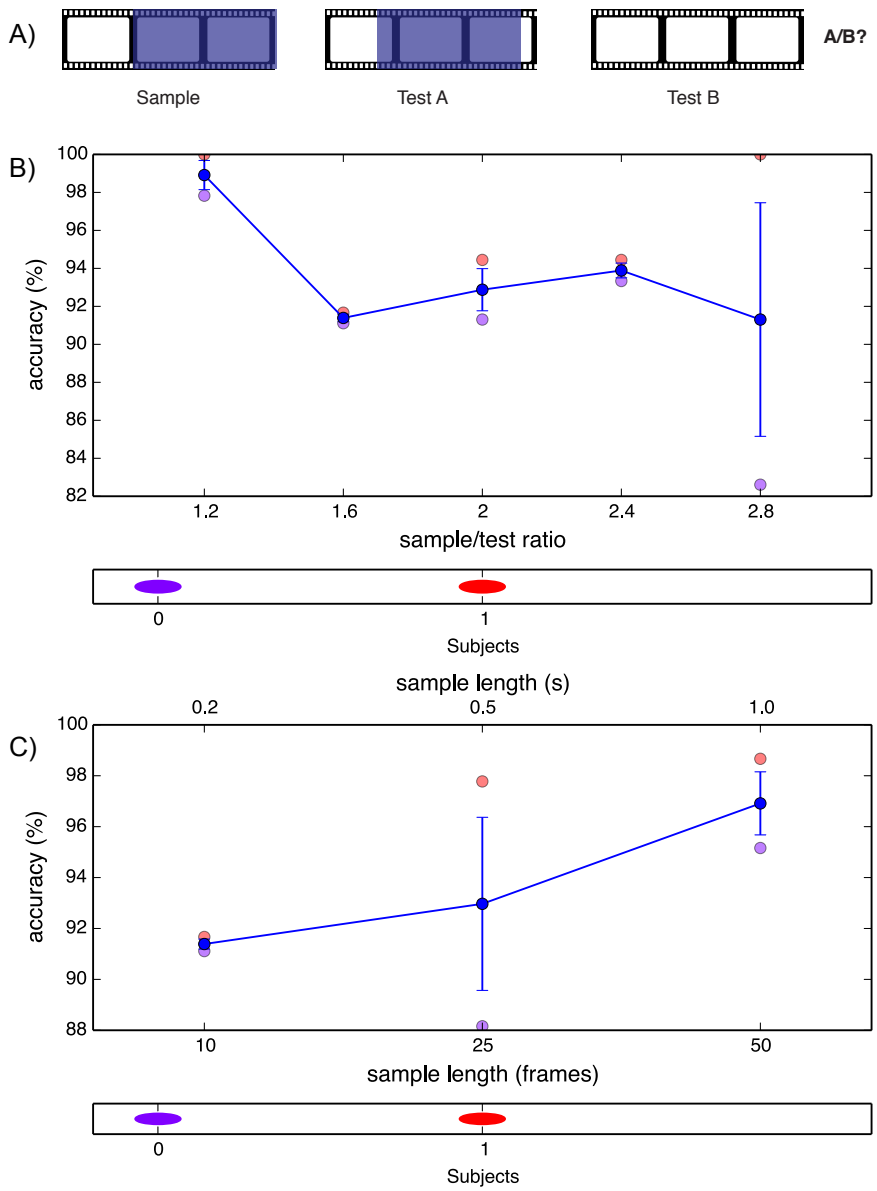


Figure 6.1: Face matching pilot study. A) Observers were shown a sample clip of a moving face, then two test clips. We asked them to indicate which test corresponded to the sample. B) Accuracy against test/sample ratio: both subjects' accuracy was consistently high. C) Accuracy against sample length: observers seemed to match longer samples with higher accuracy.

The data from this experiment are shown in Fig. 6.1. Both observers' mean accuracies were consistently high, approaching 100% for a test/sample ratio of 1.2.

6.1.3 Discussion

Although only two participants were tested, this experiment suggests strongly that observers are capable of matching dynamic face clips with much more accuracy than for dynamic fire clips. Given the small n, this experiment is of most utility as a pilot study to show that dynamic face matching is possible with our stimuli.

In order to address the question of specificity, we used a parallel loading task. In the next experiment, observers were asked to encode both a dynamic face clip and a dynamic flame clip before matching each one to a test clip of the same category.

6.2 Experiment 6.1: Parallel loading with faces and fire

In this experiment, observers were asked to encode a face and a flame clip in parallel, then match them. In order to correctly match both clips, they needed to perform a dual encoding task and make two sequential matching judgements. We used a dual encoding task to test for interference between two stimulus classes. Similar dual encoding tasks have been used to demonstrate the independence of sensory working memory systems, as by Hitch and Baddeley[247], as well as to investigate working memory maintenance using EEG measures[248].

We manipulated the order of presentation, keeping the order of matching judgements the same. If both stimulus classes were using the same memory resource, we would expect a recency effect due to interference between the two stimuli. Because measured face matching accuracy in the previous pilot study was higher than measured flame matching accuracy in our dynamic flame matching experiments, we always presented the flame test first and the face test second. This was done to minimise interference at test time, as we are more interested in interference at encoding time.

6.2.1 Methods

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

Materials We used the faces dataset and the 1000-frame dynamic flame dataset

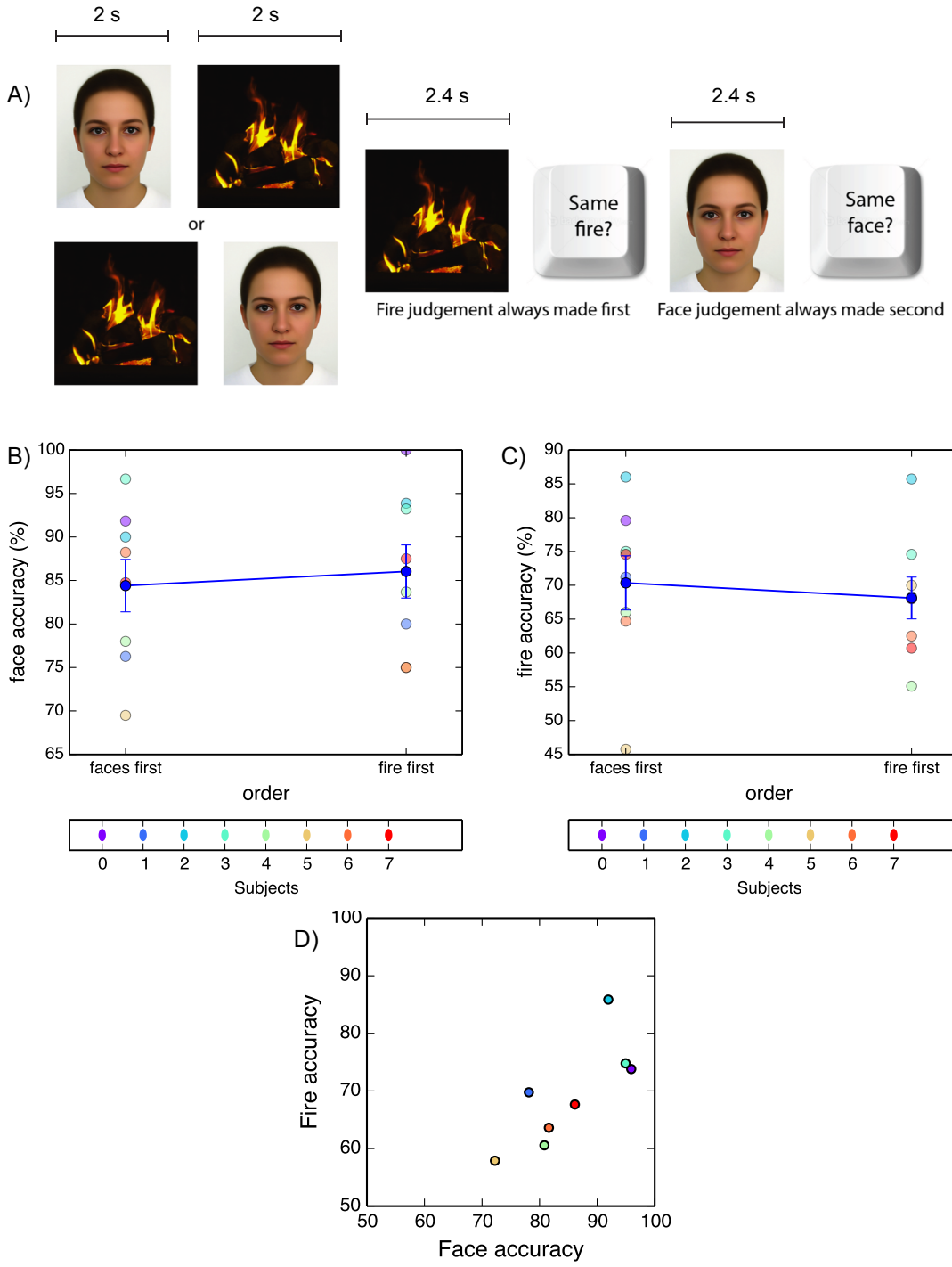


Figure 6.2: Experiment 6.1: comparing performance on fire and face matching. A) Observers were shown a fire and a face (we manipulated the order) and then tested on a fire, then a face (always in that order). B) Face matching accuracy, shown by order. Accuracy is higher when faces followed fire, but the difference is not significant. C) Fire matching accuracy, shown by order. Accuracy is higher when fire followed faces, but the difference is not significant. D) Across subjects, there is high correlation between face matching accuracy and fire matching accuracy; Pearson's $r = 0.78$.

described in Chapter 2 (General methods).

Design We used a yes-no dual encoding delayed match-to-sample task, asking observers to encode and match a test fire clip and a test face clip with a sample fire clip and a sample face clip.

Procedure In each trial, we presented two samples (one fire clip and one face clip), followed by two tests (one fire clip and one face clip). We manipulated the order in which the samples were displayed, but the tests were always displayed in the same order: fire first, then face. 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).

6.2.2 Results

The data from this experiment are shown in Fig. 6.2.

A pair of one-way ANOVAs showed no significant effect of order on either face accuracy, $F(1,7) = 0.448$, $p = 0.525$; or fire accuracy, $F(1,7) = 0.278$, $p = 0.614$.

Fire matching performance and face matching performance were highly correlated, with a Pearson's r of 0.78.

6.2.3 Discussion

Mean accuracies show the trend we would expect if a shared working memory resource was being used to encode both fire and faces. However, there is no significant difference in the means, so we cannot say a recency effect is present. For clips of this length and size, observers do not appear to be performing near their working memory limits. Face and fire representations are either dealt with by separate systems, or by one system which is not operating near its capacity limit in this experiment.

Matching performance correlation How do we explain the high correlation between face matching performance and fire matching performance? There are several trivial explanations: this pattern could have been due to differences in general visual skill, or observer motivation, or observer effort. It could also have been due to a latent

variable which varied across observers, such as motivation or fatigue. We now examine these explanations more closely.

There is a large amount of work on the relationships between general cognitive skills[249]. The general factor or g factor, attributed to Spearman[250], a latent variable related to general intellectual ability, generally accounts for 40 to 50 percent of the variance found in the results of ability tests. There is, however, little work on the correlations between psychophysical tasks in vision. Tibber *et al*[251], for example, measured observers' ability to judge size, orientation and numerosity, then analysed the correlation of these measures to a measure of general mathematical ability; they did not look at how the ability measures correlated with each other.

Consider a latent variable v which measures "general visual ability"; the equivalent of Spearman's g for visual tasks. It is tempting to say that our observed correlation is due to different levels of v across observers. However, any strategy for measuring v will depend on measuring an observer's performance on several visual tasks, then combining them (using a weighted average if v is defined using factor analysis). Explaining a correlation in accuracy across tasks, by a variable calculated by integrating accuracy across tasks, is subject to challenges of circular reasoning. Factor analysis shows us correlations between performance at different tasks, but does not give us a biological explanation of these correlations. Spearman's g and the hypothetical v explain variance, but not mechanism.

If we discount the effects of "general visual ability," and any other confounding variables, what could this correlation mean? Observers who perform well at matching fire tend to perform well at matching faces, and vice versa. This suggests that observers are using common processing systems to perform both tasks. Matching performance may depend on the ability to extract common low-level features which are used by a specialised face processing system and also to process dynamic flame. A low correlation would suggest that separate systems are being used. A negative correlation would suggest that fire processing and face processing shared a common neural substrate in a zero-sum manner (neural resources used by the face processing system were rendered unavailable to the fire processing system).

Within-subject correlations In the absence of confounding variables, across-subject correlations can tell us whether fire and face matching systems share a mutual

processing substrate. We now look at within-subject correlations, placing trials on scatterplots for each observer. For each observer, we have a time series of two binary variables: success or failure on the face and fire matching part of each trial. Since these are binary variables, we use Pearson's phi (also known as the mean square contingency coefficient), the equivalent for binary variables of Pearson's r.

High values of phi would indicate that success on individual fire trials was associated with success on individual face trials, suggesting that the level of attention paid (or effort made) during the trial was responsible for the success of both judgements. Negative values of phi would indicate that on trials where the fire judgement was successful, the face judgement was likely to fail (suggesting a common, limited attentional resource shared between the two faculties). Values of phi close to zero would indicate no connection between the two judgements.

All observers' values of phi were very close to zero, as shown in Table 6.2.

Observer	phi
1	-0.01
2	0.01
3	0.09
4	0.04
5	0.08
6	0.12
7	-0.19
8	-0.06

Table 6.2: Experiment 6.2: correlations across trials.

This again supports the idea that fire and face representations are either managed by separate systems, or one system which is not near its capacity limit.

We come to the interesting conclusion that fire and face performance are highly correlated across subjects, but not correlated at all across trials. An observer's face performance predicts their fire performance - but a trial's face success does not predict its fire success. This means that whatever explanation we posit is either a confounding variable which changes very slowly (such as level of wakefulness on that particular day) or an innate property of that observer's visual system.

Dynamic flame and dynamic faces are useful test stimuli because they are so different: observers are practiced and expert at matching faces, but naive at matching fire. Faces have a common structure and limited possible deformations; flames have no

common structure and a very large set of possible deformations. One of the strongest pieces of evidence for specialisation at face perception is the inversion effect, which consists of reduced matching accuracy on upside-down faces. We now examine data from our previous experiments, several of which contained upside-down dynamic flame stimuli, for evidence of an inversion effect.

6.3 Do observers show an inversion effect for dynamic flame?

The face inversion effect, first pointed out by Yin[70], is a strong and well-replicated effect whose meaning is frequently debated in the context of face recognition. The effect refers not simply to an impairment associated with inversion, but to a differential impairment: inverted faces are matched less well than inverted objects[252].

The face inversion effect is heavily task-dependent. As shown by Valentine[253], the effect is more often found when face stimuli are compared to faces in long-term memory than to faces in short-term memory. This constitutes the difference between a "recognition" task and a "matching" task. Although the dividing line is blurred, recognition tasks usually assume that the recognised pattern will still be in memory after the task has finished (as in the case of a famous face), whereas matching tasks assume that the observer will forget the stored pattern before the next trial or block (as in the Cambridge face test). Whether accuracy is affected when matching an upright face to an inverted face does not appear to have been investigated.

No effect of inversion was found in matching experiments performed by Bruyer[254], or Valentine[253]. Shepherd[255] observed a similar pattern in the other-race effect, finding it to be present in recognition but not in matching. This may indicate a difference in processing between long-term face memory and short-term face memory. Shepherd's theory is that local feature cues, which are resistant to the inversion effect, can be extracted from short-term memory representations; however, long-term memory representations efficiently discard low-level information, leaving only a high-level face-specific encoding which is orientation-tuned.

We note that there are two kinds of inversion effect: a *crossed* inversion effect is a matching accuracy drop when the sample is inverted and the test is upright (or

vice versa), whereas a *pure* inversion effect is an accuracy drop when both the sample and test are inverted. A crossed inversion effect shows that it is difficult to compare a representation of an inverted stimulus to that of an upright stimulus: it indicates that particular exemplar representations are orientation-dependent. A pure inversion effect, however, shows that the stimulus as a whole is more difficult to encode when it is upside down; this indicates that the mechanisms which construct these representations are, as a whole, orientation-dependent. An inverted random dot pattern, for example, is difficult to compare to an upright copy of itself (crossed inversion effect) - but two inverted random dot patterns are just as easy to compare as two upright random dot patterns (no pure inversion effect).

It does not make sense to perform a within-category recognition task with dynamic fires, since particular exemplars of this stimulus are not encoded in daily life, and would also be easily differentiable based on clues from their background and fuel shape. Do matching tasks on dynamic flame exhibit an inversion effect? Several of our experiments included conditions in which some stimuli were inverted. We now review these results together.

Experiment 4.1 measured the effect of various sample manipulations on matching to a test clip of very similar length. One of these manipulations was inversion, but there was no significant difference between their accuracies; this motivated the repetition of this experiment with shorter, more tractable clips. However, accuracy was significantly above chance in all conditions including inversion, showing that crossed inversion does not cripple observers' ability.

Experiment 4.2 repeated Experiment 1 with shorter clips. Once again, the accuracy associated with each manipulation was significantly above chance. Inversion was associated with a matching accuracy of 66%, a drop of 10% from the mean. This constitutes a crossed inversion effect, showing that mental rotation is not a trivial process and flame representations are not orientation-invariant. Accuracy was affected more than under backward playback. This result indicates that spatial feature correspondence is more important than temporal feature correspondence: observers show less invariance to inversion than to temporal order.

Experiment 4.4 tested observers' ability to decide whether a clip was being played forwards or backwards. This required the evaluation of a stimulus against their category

representation of fire rather than against a previously seen clip. As shown in Fig. 6.4, there was no effect of orientation (two-way ANOVA on orientation and frame rate, $F(11,16) = 8.13$, $p = 0.146$). This suggests that observers' category model is not tuned to upright flame either. Although this task did not include matching, this suggests the absence of a pure inversion effect.

In Experiment 5.5 we asked observers to perform a matching task and varied the ISI between sample and test. In half the trials, both clips were inverted; in the other half, they were both upright. A two-way repeated-measures ANOVA on inversion and delay revealed no effect of inversion, $F(1,6)=2.171$, $p=0.19$, but confirmed an effect of delay, $F(3,18)=3.575$, $p=0.35$. This supports the idea that the matching process is not impaired when both clips are inverted.

These three results show that matching an inverted clip with an upright clip is a challenging task, but judging the playback direction of an inverted clip is just as easy as in the case of an upright clip. What does this say about observers' flame representations? Comparing an upright clip to an inverted clip is difficult: this suggests flame representations are hard to transform in this way, which means they cannot be orientation-invariant. Judgements of luminance, colour and symmetry are invariant to rotation; flame representations are not. However, when encoding a single clip and judging its direction, orientation does not matter. This means that the representations used for this task are not sensitive to orientation. It appears that inversion only causes an issue when comparing two flame encodings, not judging the realism of a single one. This suggests that individual representations for flame are not orientation-tuned (not better at representing upright flame), simply that they are not orientation-invariant (not identical for upright and inverted flame). To test this explicitly by looking for a pure inversion effect, we set observers a matching task in which both sample and test were inverted.

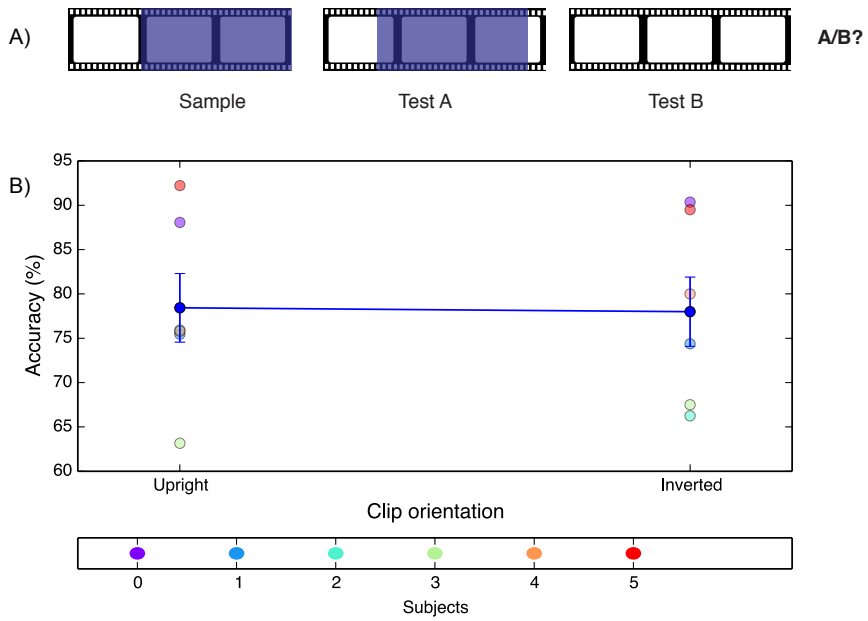


Figure 6.3: Experiment 6.2: matching inverted tests to inverted samples. A) A sample was followed by two slightly longer tests. All stimuli were inverted. B) Mean accuracies were very close and showed no significant difference, indicating that inversion does not impair matching at all when observers do not have to compare an upright test and an inverted sample.

6.4 Experiment 6.2: Looking for a pure inversion effect in dynamic flame

6.4.1 Methods

Observers 6 subjects were recruited using a mailing list operated by University College London. All reported normal or corrected-to-normal vision.

Materials We used the face dataset and the 1000-frame dynamic flame dataset described in Chapter 2 (General methods).

Design We used a 2AFC matching design.

Procedure In each trial, we presented a sample clip followed by two slightly longer tests. In half the trials, all three clips were upright; in the other half, all three clips were inverted. Observers were asked to indicate whether they thought the sample corresponded to the first or second test.

6.4.2 Results

Data from this experiment are shown in Fig. 6.3. Baseline performance was 78.4%; under inversion of all three clips, performance descended to 78.0%. A paired-samples t-test revealed no significant difference between the means (p=0.85). Observers do not show a pure inversion effect when matching dynamic flame stimuli.

6.4.3 Discussion

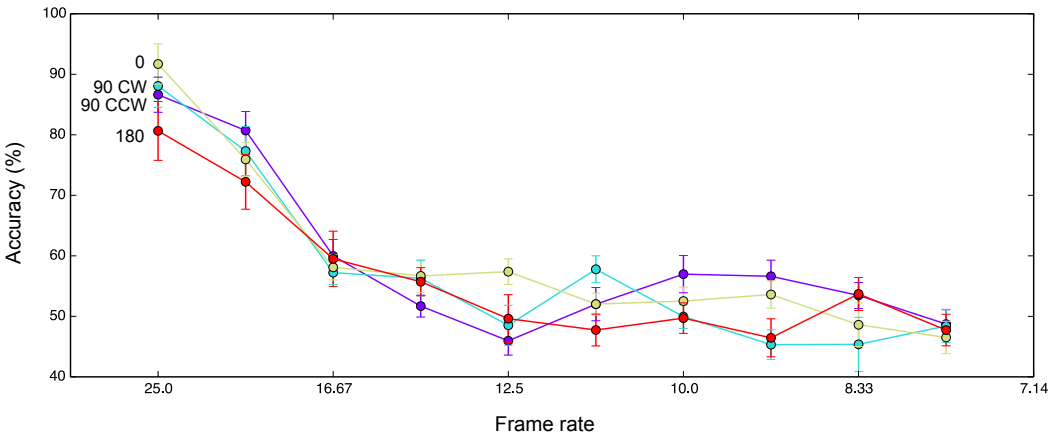


Figure 6.4: Experiment 4.4: the effect of stimulus angle on backwards detection. On each trial a 2-second clip was played; observers then indicated whether they thought playback was forwards or backwards. B) Accuracy dropped quickly down to chance at a frame rate of 10 Hz. C) As frame rate decreased, the amount of "backwards" judgements increased. D) Mean observer d-prime also dropped quickly to a minimum at 10 Hz. E) Observers were much more accurate at judging the clips which were played forwards.

There was no discernible difference between matching accuracy on inverted flame and upright flame. In this case, observers did not have to perform a conversion between an upright percept and an inverted representation, which induced a 10% drop in accuracy in Experiment 4.2.

In face perception, the pure inversion effect has often been interpreted as evidence of a high-level representation (whether specific to faces or not). A template-matching system cannot show a pure inversion impairment unless it produces templates in a different way under inversion - that is to say, if it shows orientation tuning. The lack of impairment for inverted flames is consistent with a low-level matching model whose basic feature detectors are not sensitive to inversion. It is not consistent with a high-level model which preferentially encodes upright stimuli.

This result was surprising, since a strong percept of upwards motion is dominant when viewing flame. As seen in Chapter 4, observers disagree when judging the motion direction of small patches of flame. We also know from Experiment 4.4 that frame rates of 10 Hz or above are required for backwards playback detection. Since all observers are fully aware that forwards playback corresponds to upwards motion, this means that a 10 Hz frame rate is required to detect upwards motion as well (if lower frame rates were sufficient, upwards motion could easily be used as a cue for forwards playback). This points to the involvement of highly temporally local motion detectors, agreeing with the evidence from the lack of inversion impairment found here.

In Chapter 4 (Experiment 4.2) we reported a serious accuracy impairment when observers had to match an upright dynamic flame to an inverted one: a crossed inversion effect. How does this fit with the lack of overall inversion effect found here? It is a different effect: the classic inversion effect in the face perception literature is a property of a category, not an individual trial. It refers to the overall lessening of accuracy when matching two stimuli in that category which are both inverted. Experiment 4.2 showed an individual-stimulus inversion effect, in which two stimuli are less well matched when one of them is inverted. One type of effect does not imply the other: in the pure inversion effect, no transformation is necessary since both stimuli are inverted. In the crossed inversion effect, a transformation (effectively a mental rotation) is required, which makes the task more difficult. For example, it could be found that observers have no difficult matching a pair of inverted clock faces, but perform very badly at matching an upright clock face with an inverted one. Our results, thus, do not constitute a contradiction: they indicate that the transformation required to match an upright flame to an inverted flame is challenging, but that the general representational system used to encode fire operates just as well on inverted flame as on upright flame.

Overall, our results show that inversion only impairs matching when the visual system has to compare an upright representation and an inverted representation, not two inverted representations. We have found no evidence that flame is encoded by specialist orientation-tuned representations.

Experiment	Slope (percentage points per trial)
Face pilot	2.70 e-3
6.1 (face)	1.46 e-4
6.1 (fire)	9.50 e-3
6.2	2.23 e-3

Table 6.3: Learning slopes from Chapter 6.

6.5 Learning

As in previous chapters, we looked for learning effects in the pilot study and the two experiments performed in this chapter. This was done by arranging the trials in the order in which they were presented, blocking them into groups of 20 using a sliding window, and calculating the average accuracy for each block. The results of this sliding window approach are shown in Fig. 6.5.

To check for an improvement in mean accuracy, we fitted a line to the sequentially arranged data. Calculated slope values are shown in Table 6.3.

These experiments showed slopes which were very small but positive. The largest slope was associated with the fire accuracy metric from Experiment 6.1. As with our previous experiments, a discernible improvement was shown during the first 10 trials of Experiment 6.1 and the first 50 trials of Experiment 6.2. This rapid improvement could have been due to observers' learning the apparatus or growing more comfortable with the experimental setup. None of these learning curves provide convincing evidence that observers' representations are improving in quality throughout the experiment.

6.6 Intertrial dependence

Before an observer depresses the key which records their response, they must perform a complex decision-making process. This can be interpreted in terms of deciding between two distributions (signal detection theory[256]) building evidence for two hypotheses[257] or the evaluation of posterior probability by Bayes' rule[258]. In the previous chapters, we have used the decision process as a tool to study the underlying representations of dynamic flame. We now consider it separately.

Most of the experiments we have presented set a delayed match-to-sample task: a sample clip was presented first, followed by a test clip which was longer but the

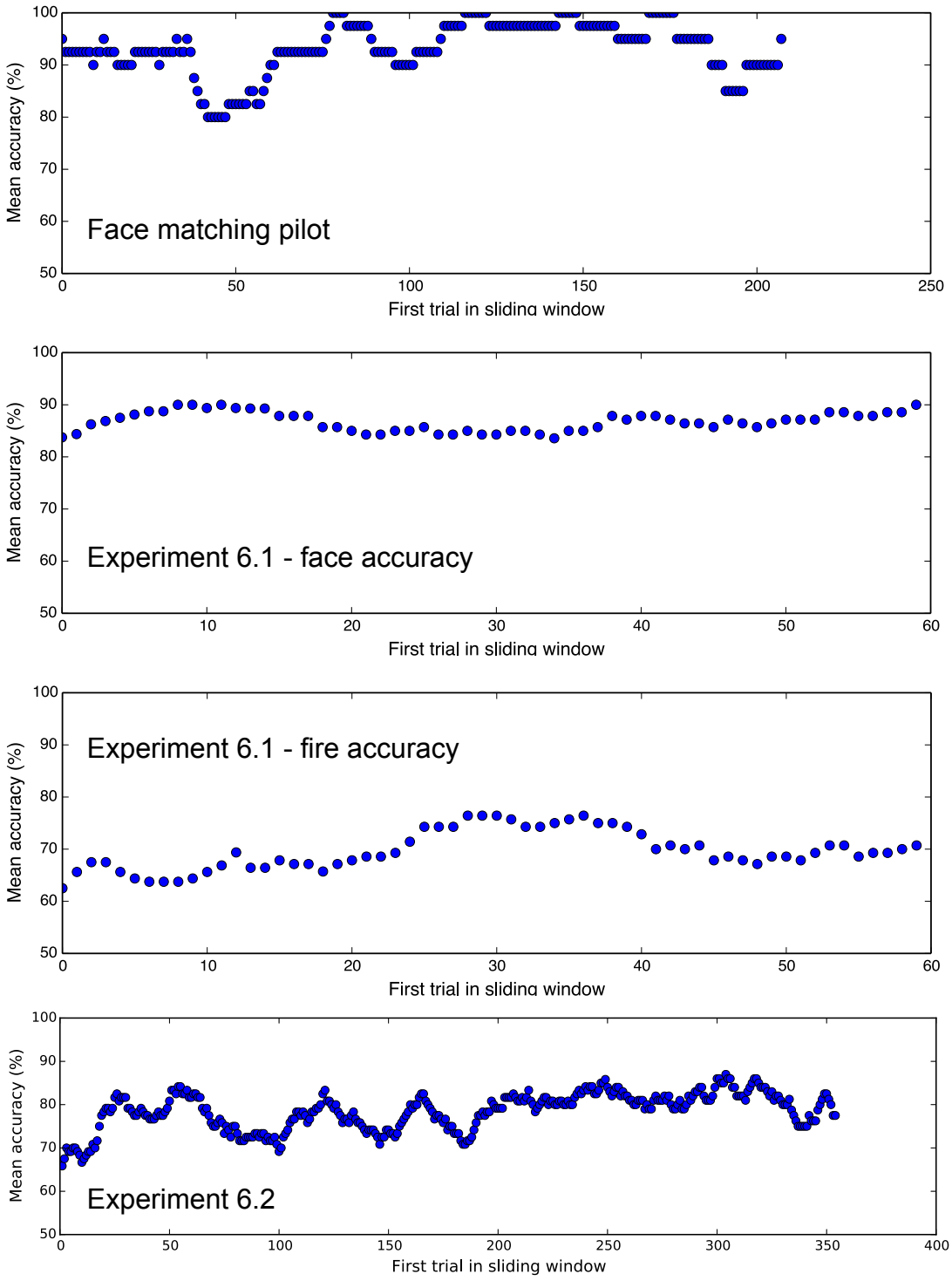


Figure 6.5: For each experiment, trials were aligned in order of presentation and a sliding average applied to show how accuracy changed during the experiment. There was no sure overall trend of increasing accuracy. A slight increase in fire matching accuracy may appear in the first 30 trials of Experiment 6.1, but it does not continue.

same spatial size. It is tempting to consider each trial as a separate unit; however, low-level mechanisms may not be affected by the conscious knowledge that one trial has finished and the next has begun. How does the history of previous decisions affect the next decision? Can the visual system make a fresh judgement on each trial, or is it clouded by previous responses?

In general psychology, the tendency to rely on the first piece of evidence considered is known as the anchoring effect[259]. This effect is mainly considered in cognitive and social psychology, although it has been discussed in the context of psychophysics[260]. It is usually taken to refer to dependence on the initial piece of evidence - not the first response, or the previous response.

Another effect which creates intertrial dependence is adaptation[261]. For example, an observer can be adapted to classify androgynous faces as female after exposure to a series of male faces. To properly describe an adaptation effect, one must say either that it biases perception towards a certain kind of stimulus(making that percept more likely) or away from a certain percept (making it less likely). We thus require a stimulus description space (usually a continuum or low-dimensional space) in which to describe the effect; for example, we can adapt in a lightness continuum or in a multidimensional face space.

Adaptation describes percepts, not responses; observers are not adapted towards certain motor responses or certain judgements. We aim here to investigate the effect of the previous judgement on the current judgement; are observers more likely to report a match if they did so on the previous trial? This effect has been called intertrial dependence[262] and has recently been treated by Frund *et al*, who modelled it statistically[263] using a 14-parameter model influenced by the observer's entire response history.

6.6.1 Method

We take a simpler approach here, looking at the probability of responding Yes or No given that the previous response was Yes or No. Let R_n be an observer's response on trial n. Our delayed-match-to-sample experiments required either a Yes response or a No response ($R_n = Y$ or $R_n = N$). Let T_n be the truth for trial n: whether there was actually a match ($T_n = Y$) or a non-match ($T_n = N$). Since true trials were selected

randomly, then over all n, $P(T_n = Y) = P(T_n = N) = 0.5$. From here onwards, we use $P(T_n) = x$ to refer to the probability, across all trials, that the response is x.

Observers have an internal bias; signal detection theory calculates the bias for an experiment as a whole, not each individual trial. We can measure this "static bias" by looking at $P(R = S)$ for the entire experiment. Observers may also have intertrial bias: the tendency to respond either the same way as in the last trial, or the opposite way. If there is no intertrial bias, we would expect

$$P(T_n = Y \mid T_{n-1} = Y) = P(T_n = Y \mid T_{n-1} = N) = P(T_n = Y) \tag{6.1}$$

and

$$P(T_n = N \mid T_{n-1} = Y) = P(T_n = N \mid T_{n-1} = N) = P(T_n = N), \tag{6.2}$$

whereas if we find that

$$P(T_n = Y \mid T_{n-1} = Y) < P(T_n = Y \mid T_{n-1} = N), \tag{6.3}$$

this observer is less likely to respond Yes if they responded Yes on the last trial. Conversely, if we find that

$$P(T_n = Y \mid T_{n-1} = Y) > P(T_n = Y \mid T_{n-1} = N), \tag{6.4}$$

the observer is more likely to respond Yes if they responded Yes on the last trial.

We define the previous-trial dependence D_1 of an observer as

$$D_1 = P(T_n = Y \mid T_{n-1} = Y) - P(T_n = Y \mid T_{n-1} = N). \tag{6.5}$$

If D_1 is positive, a Yes response is more likely after a Yes response than after a No response (tendency to respond the same way). If D_1 is negative, a Yes response is more likely after a No response than after a Yes response (tendency to respond in a different way).

Looking further back in time, we define the kth-previous-trial dependence D_k of an observer as

$$D_k = P(T_n = Y \mid T_{n-k} = Y) - P(T_n = Y \mid T_{n-k} = N). \tag{6.6}$$

We looked at values of D_k for each of our experiments in which observers had to make a binary choice. For each value of k, we calculated D_k for each observer and, looking back in time, for values of k from 1 to 19. We did not treat Experiment 4.5, since its responses were continuous direction estimates, not category judgements.

6.6.2 Results

Results of this analysis are shown in Figs. 6.6 (Chapter 4), 6.7(Chapter 5) and 6.8 (Chapter 6). To check for effects, we do not perform t-tests, since the number of tests performed would lead to an unreasonably high chance of a Type 1 error. Because D represents a difference between probabilities, it is not clear whether the assumptions of analysis of variance are satisfied. We nevertheless note clear decreasing trends for two experiments: 4.4 (backwards playback detection) and 6.1 (dual loading with dynamic faces and dynamic flame).

There is a clear distinction here: these two experiments are the only two which involve more than a delayed-match-to-sample task (either Yes/No or 2AFC) on dynamic flame stimuli. Experiment 4.4 is a playback direction task for which a primary cue is motion direction. Experiment 6.1 requires observers to match a face clip in addition to a flame clip.

6.6.3 Discussion

A positive value of D_k means that an observer is more likely to respond the same way as on the kth previous trial. A negative value of D_k indicates a tendency to respond the opposite way as on the kth previous trial; this could be evidence of adaptation. In both the experiments in which we find a trend, D_1 is positive and D_k drops towards zero as k increases.

Why do these trends appear in Experiments 4.4 and 6.1, but not in our standard delayed match-to-sample experiments? In Exp. 4.4, observers were asked to detect backwards playback; an easy way to perform this task is by judging the overall direction of motion. Motion adaptation could easily explain a negative value of D, but not

the observed positive value, which shows that the classifiers observers are using have "inertia:" they are more likely to make their previous classification[1].

We propose the following hypothesis: dedicated, specialised perceptual systems, such as the motion processing system and the face processing system, are subject to inertia, but the mechanisms which discriminate dynamic flame are not affected.

In Experiment 6.1 we find large positive values of D_k, reaching 0.5 for fire accuracy and 0.3 for face accuracy. D_1 is positive for face accuracy, which fits with our hypothesis, since face perception is a highly specialised ability which recruits dedicated areas of cortex. However, this does not explain the very high value of D_1 for flame accuracy. Why does this pattern appear in a delayed match-to-sample task on flame only when observers are also performing a face matching task? The face decision could be entraining the fire decision. This is not reflected by the low values of Pearson's phi we found for trial correctness. We repeated Pearson's phi on Yes/No responses instead of correctness values, and found correlations very close to zero, which does not suggest that the face matching decision entrains the fire matching decision. It is possible that the face matching task induces observers to use a different decision strategy than they would employ if asked to match dynamic flame stimuli alone.

6.7 General discussion

We began this chapter by confirming that observers can perform temporal search on faces with much higher accuracy than they can on flame. Test/sample ratios were much higher than in our experiment on flame, as were accuracies.

We then asked observers to encode both stimuli in parallel, before presenting two tests. Presentation order (flame first or faces first) had no effect, showing that for 2-second tests, observers were not relying on transient early visual representations of flame which would be destroyed by attending to a face clip. There was however a high correlation (r=0.78) across subjects between face matching accuracy and flame matching accuracy. This result suggests that similar mechanisms limit the encoding and matching of both stimuli, and is inconsistent with separate specialised systems for

[1]The term "perceptual inertia" has previously been used to refer to a lack of sensitivity directly after stimulus onset, but our usage is more in keeping with the definition from physics: maintaining the same quality.

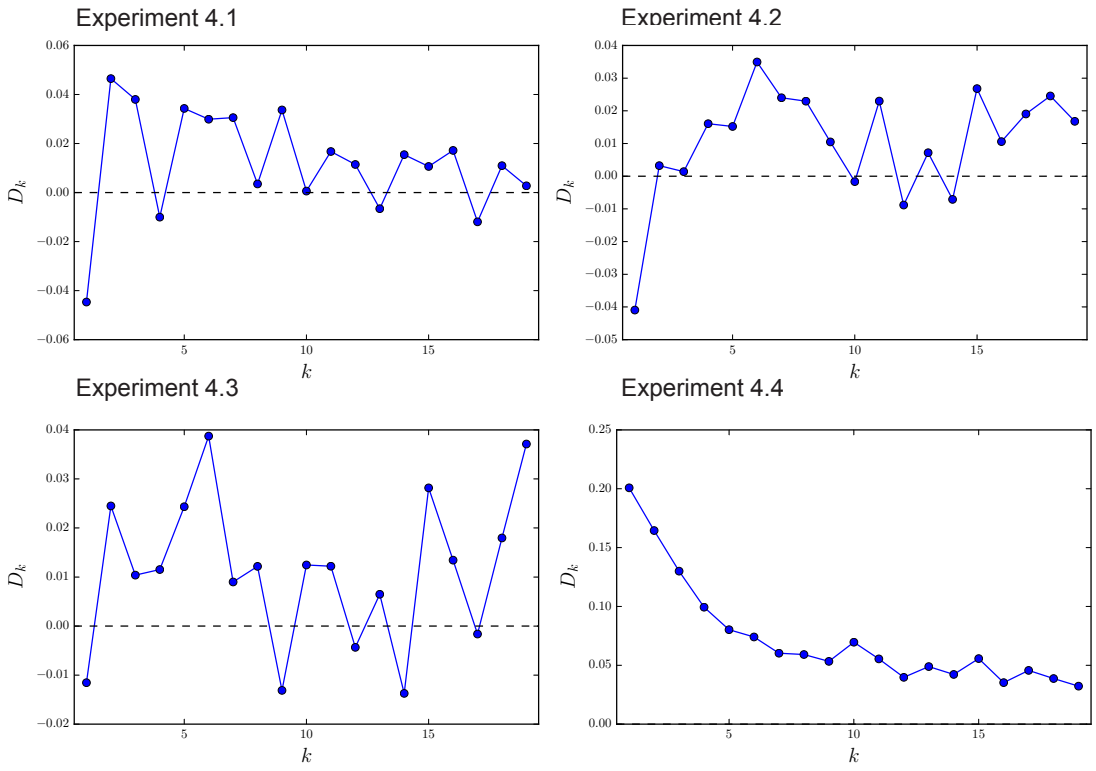


Figure 6.6: Intertrial dependence for experiments described in Chapter 4. No conclusive pattern is found except in Experiment 4.4 (backwards playback detection). Positive values of D indicate that observers are biased to respond in the same way as on previous trials; the dependence of trial n on trial $n - k$ decreases as k grows.

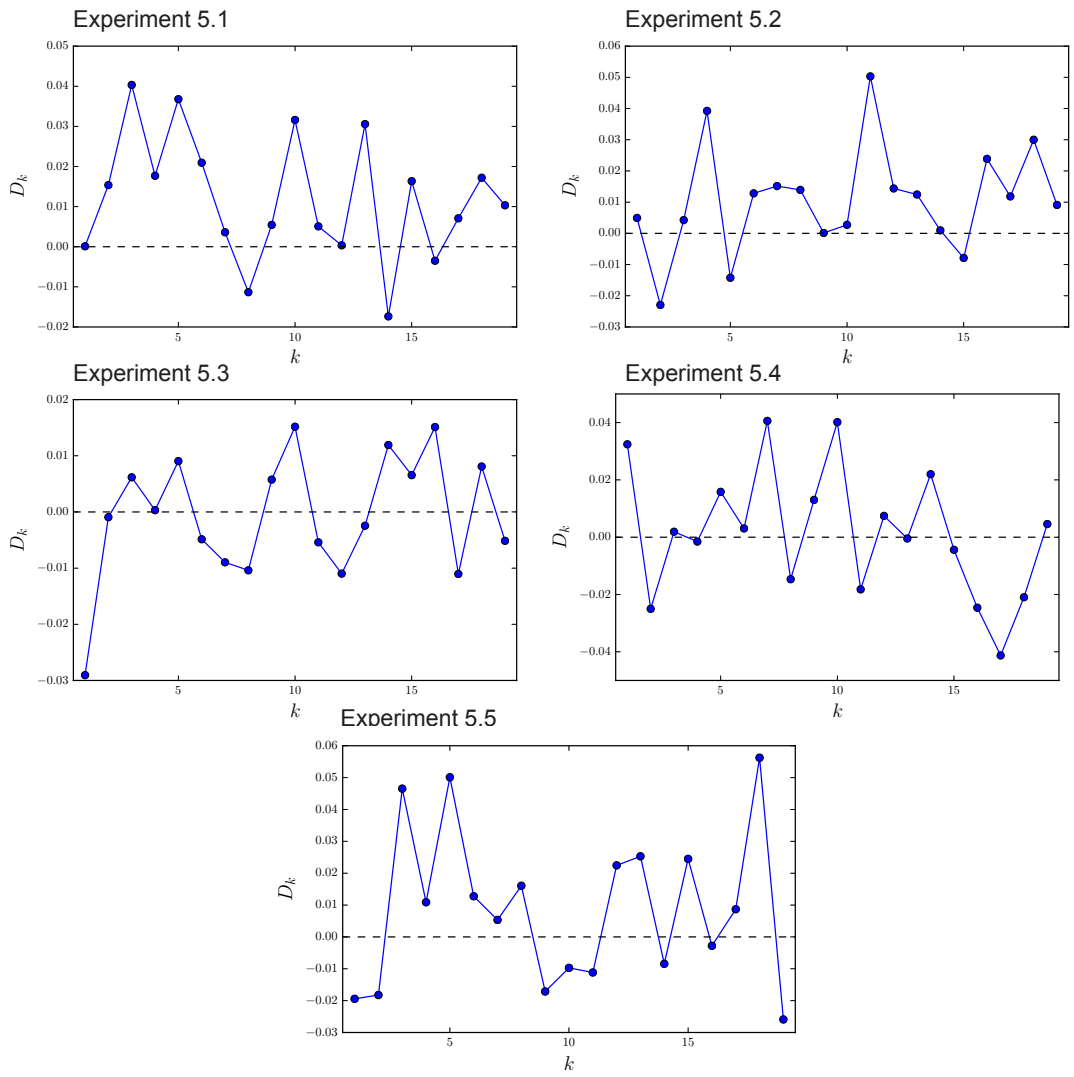


Figure 6.7: Intertrial dependence for experiments described in Chapter 5. No conclusive pattern is found for these visual search experiments.

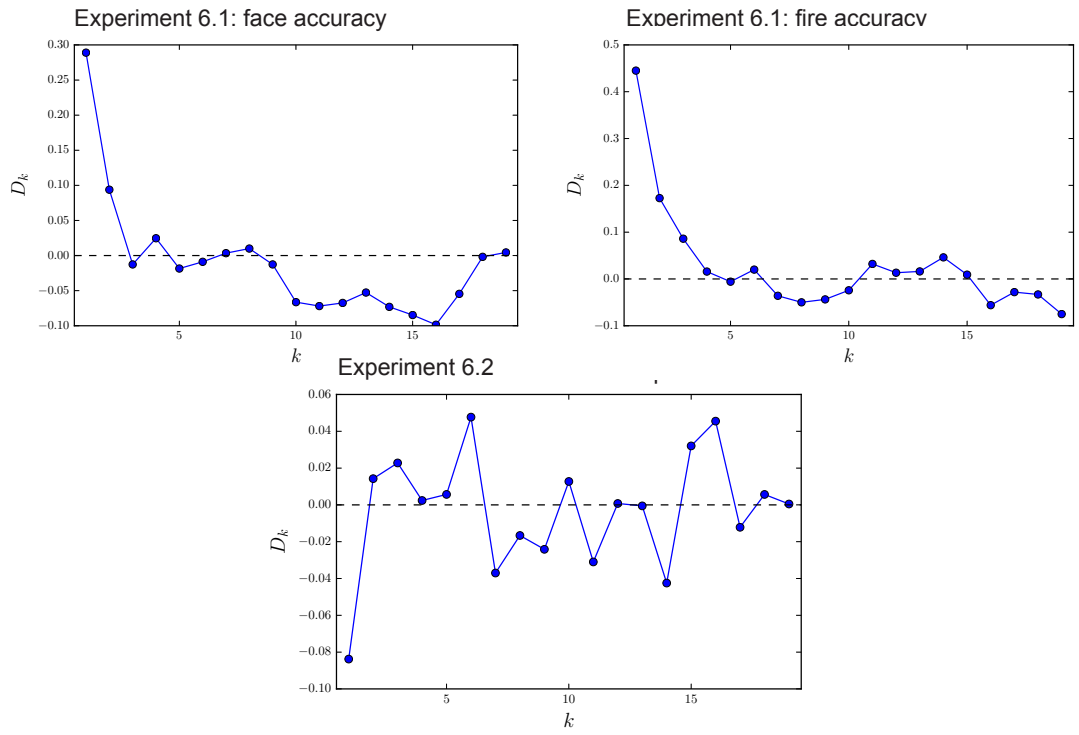


Figure 6.8: Intertrial dependence for experiments described in Chapter 6. Exp. 6.1 (dual loading with faces and fire) shows a steadily decreasing value of D for both face matching accuracy and flame matching accuracy. Exp 6.2 (testing for a pure inversion effect) shows no clear pattern.

each phenomenon. We may choose to attribute the correlation to a latent variable measuring "general visual ability," but this only characterises the correlation and does not explain it.

We then reviewed the evidence concerning an inversion effect in dynamic flame. Across a wide series of experiments, matching accuracy was only disrupted when observers had to compare an upright test to an inverted sample. Even the category representation for flame, which we tested with a backwards detection task, was not affected by inversion. This suggests that the representations we used to encode flame are not specialised or orientation-tuned.

Finally, we examined the time series of responses given by each of our experiments to check for intertrial dependence, either in the form of adaptation (a tendency to make a different response than the previous one) or perceptual inertia (a tendency to make the same response as the previous one). We found clear evidence of perceptual inertial, decreasing steadily over time, in the only two experiments which did not just set observers a flame matching task. This suggests that motion direction judgement and face perception are qualitatively different than flame matching.

Summary

- When performing temporal search on faces, observers are much more accurate and tolerant to much larger search spaces than in the case of dynamic flame.
- Observers can encode a 1-second dynamic face clip and a 1-second dynamic flame clip in parallel and match them well. Sample presentation order has no effect on accuracy, showing that clips of this length are either encoded by separate systems or by a single system operating below its capacity limit.
- Across observers, face matching performance and fire matching performance are highly correlated. This indicates either that the two stimulus types are encoded by a single resource, or that the correlation can be attributed to a latent variable signifying general visual ability.
- Across four experiments with an inversion condition, an inversion deficit was only found when matching an inverted sample to an upright test. No inversion deficit was found when matching two inverted stimuli or evaluating the playback direction of a single inverted clip. There is no evidence of category specialisation

arising from an inversion effect.

- For flame matching, observers show little or no intertrial dependence: responses on a particular trial do not depend on the responses made to the previous trials.

Chapter 7

General discussion and conclusions

How can we characterise the human visual system's representations of dynamic flame? This chapter reviews our experimental and analytical results and discusses their implications. Finally, we assess our contributions and suggest further work.

One of the most important questions we can ask about a stimulus class is whether the visual system uses a specialised high-level encoding to express it, as opposed to using lower-level, more general representations. For complex stimuli, this is the same as asking whether observers have expertise for those stimuli: expertise is characterised by specific, high-level representations which are able to deliver high accuracy.

7.1 Summary of results

In **Chapter 1** we summarised the history of object recognition models, the challenges posed to them by natural scenes, and the difficulty of accounting for the representation, remembering and matching of dynamic stimuli using static object recognition models. We pointed out two main classes of model: those which represent stimuli as points in a stimulus space, and those which sample spatiotemporal areas of the stimulus and encode them as parts of a unified representation. Models can also be differentiated by the order in which they build up their invariances (integration of entire views into a rotation-invariant model, or integration of local rotation-invariant descriptions into a whole-object model). We introduced dynamic flame as a form- and motion-rich stimulus which is representative of the encoding and matching challenges posed by dynamic natural scenes.

In **Chapter 2** we described the recording and presentation of dynamic flame and face stimuli, as well as the trial structure and the details of the tasks observers were set. We described our general delayed match-to-sample task, consisting in each trial of a sample clip followed by a longer test clip. Typically, participants were required to perform a temporal visual search, with tests longer than samples. When tests were close in length to samples, this constituted a matching task. To prevent an easy route to matching via iconic memory, we ensured that first and final frames did not co-occur between samples and tests.

In **Chapter 3** we analysed dynamic flame in the image domain. In flame clips, frames which are close in time are very similar according to image space distance and structural similarity measures. This similarity drops very quickly, reaching a minimum after 0.2 seconds. Averaging and the application of a morph model show that flame does not have a general structure which can be deformed to give individual images (a high-level prototype). We showed that flame contains a wide spread of spatial and temporal frequencies. In terms of spatial frequencies, power is concentrated near the vertical and the horizontal. In terms of temporal frequencies, power spectra of the overall clip brightness and the individual pixel brightnesses are best approximated by an exponential function, similar to the $1/f$ distribution found ubiquitously in natural images, but with less power in the high frequencies. A 3D Fourier transform found the same pattern of spatial frequencies at low temporal frequencies (near the DC peak) but a more random distribution of spatial frequencies at the high temporal frequencies.

Two different algorithms provided very different estimations of the motion present in flame images, although they agreed in overall motion direction. The McGM computed small patches of motion in varying directions, while Sun's model found larger patches of more coherent motion in two particular directions. This is probably due to the regularisation stage of the Sun model, which promotes spatial uniformity. The motion perceived in flame is not trivial to compute, and since there is no ground truth there is no correct answer to this problem.

We attempted to model fire using three algorithms which produce high-level image encodings: PCA, morph space PCA and Doretto's dynamic texture synthesis (DTS). The first two techniques produced blurred images with incorrect low-level structure, no sharp edges, and no temporal structure. DTS produced videos which changed

realistically through time and showed more convincing structure, but still lacked the definite shapes and sharp edges of real flame. This showed that sharp edges are key to realistic flame.

In **Chapter 4** we used a feature manipulation delayed match-to-sample paradigm (altering some information in the test, then asking observers to match it to an unaltered sample) to evaluate the importance of colour, spatial arrangement, temporal arrangement, and edges. Changing the test's colour did not significantly affect performance, showing no evidence that it is useful for matching. Inverting short tests impaired performance more than reversing them, showing that observers are using the spatial location of features, not just the 1D luminance signal. Edge filtering, despite removing most of the pixel-level information in the test, only induced a 4 percentage point drop in matching accuracy, showing that moving edges are indeed a key feature for flame matching.

We used a backwards playback detection task to estimate the duration of useful spatiotemporal features accessible to observers' category representation of dynamic flame. Observers could only reliably say when the clip was moving forwards if they could see frames which were less than 200 ms apart, showing that short features made up of temporally local information are required. We found no evidence that the task could be performed using less temporally local features.

We then asked observers to judge the direction of motion of small flame patches. Judgements were very accurate for patches 70 pixels in width, but at chance for patches 10 pixels in width. For patches between these sizes, 180° errors were common, showing that small flame patches with opposing motion directions are visual metamers for each other. Accuracy was very consistent across observers, suggesting that low-level motion mechanisms with little interobserver variation were being used. Since it is mathematically impossible to extract the true motion direction of gratings viewed through an aperture, there may not have been sufficient pattern available in the small patches to make a better judgement: we suggest that observers were performing close to optimality.

In **Chapter 5** we used a temporal visual search paradigm to evaluate observers' ability to store and search for representations of dynamic flame in temporal search spaces. When the search space ratio is close to 1 (sample and test are nearly the same

length), performance is quite high (75-80%). For longer tests, however, performance drops rapidly; it is highly dependent on search space size, and thus there is no temporal pop-out. Although pop-out is not a sure sign of a high-level representation, its absence shows that temporal search for dynamic flame is a challenging task. On short clips (0.02 to 0.24 seconds), observers showed an accuracy gain for longer samples, ruling out the possibility that they are using single static snapshots for matching. For longer clips (0.2 to 1 seconds), accuracy was no longer dependent on sample length, suggesting that encoding capacity had been reached. Observers do not appear able to effectively encode long periods of dynamic flame, which is not the case for dynamic faces.

We then separately manipulated the length of the distractor clip preceding the target and that following the target. Accuracy was once again heavily dependent on search space size, but not on pre-clip length or post-clip length. This allowed us to reject the hypothesis that the temporal position of features is coded as an offset from the beginning of the clip. Our results are consistent with an encoding of relative timing or simply relative order.

To eliminate the effect of search space size, we then manipulated the target position without changing the clip length at all. Later targets caused a drop in accuracy, but this drop did not begin until 0.8 seconds into the clip, which also argues against temporal offset coding. To investigate whether this drop was caused by distractors or a decaying representation, we next varied the ISI between target and test. The slow decay we found suggested that sample representations did not decay much during the test, suggesting that distractors were responsible. Such slow decay rules out encoding by iconic memory or rapidly-decaying forms of visual working memory.

We also found, across several experiments, a significant effect of search space size. This caused an accuracy drop exceeding the accuracy drop due to an ISI of the same length as the stimulus; this indicates that the decline in performance was interference-like, not decay-like.

Manipulating search space size and target position allowed us to evaluate various models of dynamic object perception. Unlike static models, dynamic recognisers must keep track of the relative temporal position of any spatiotemporal features they produce in order to search for matching configurations. Observers were not sensitive to small offsets of the target from the beginning of the test clip, suggesting that feature timings

are not coded absolutely (by offset from the beginning of the clip);

In **Chapter 6** we compared observers' performance on the delayed match-to-sample task with two stimulus classes, dynamic fire and dynamic faces. Face matching accuracy was much higher, with performance better on average and less sensitive to search space size. Across observers, we found a high correlation between fire matching accuracy and face matching accuracy (r=0.78); this suggests that observers are using the same perceptual faculties to represent fire and faces. Within observers, however, there was very low correlation across trials, giving no evidence that a limited resource is being used to encode both fire and faces. We also found no evidence of a pure inversion effect (a drop in accuracy when matching a test and sample which are both inverted, as opposed to matching an upright sample with an inverted test).

We pooled data from previous experiments to look for effects of intertrial dependence: either adaptation (the tendency to respond differently from on the last trial) or perceptual inertia (the tendency to respond the same way). We found no convincing evidence of adaptation, but two of our experiments showed inertia: Experiment 4.4, in which observers judged playback direction, and Experiment 6.1, in which observers encoded flames and faces in parallel. This result suggests a qualitative difference between the processes used in dynamic flame matching, and those used in motion direction judgement and face matching.

We now integrate these results and sum up our investigations into the dynamic flame matching process. We address two main questions. Firstly, how is dynamic flame represented? Secondly, how are pairs of representations matched?

7.2 How is dynamic flame represented?

Representation begins at the retina. Here we find the ultimate low-level code: an array of photoreceptors and basic filters[264] which code for small areas of the stimulus. Even the retina, however, does not correspond to a map of the stimulus: as saccades move the fovea around, detailed information is sampled from a small area. Most codes described in the literature assume that stimuli are static and possess no time dimension; exceptions are the specific codes proposed for biological motion and dynamic face perception.

We propose that dynamic flame is represented by a series of spatiotemporally local, low-level samples. Small parts of the stimulus, local both in time and space, are encoded in visual working memory; the visual system then attempts to detect these features in the test clip. This approach fits with our image-based analysis, which suggests that dynamic flame possess few long-range correlations; it is therefore difficult to unify disparate areas of the scene into one representation.

We found that PCA, morph space PCA and dynamic texture synthesis were unable to produce realistic high-level encodings of dynamic flame. While this does not prove such encodings are impossible, it suggests that dynamic flame is much harder to encode than dynamic faces.

Experiment 4.4 shows that observers require temporally local information in order to judge motion direction, and cannot do this task using long-range temporal information from more widely separated frames. It is difficult to imagine that long-range information can be used for matching if it is not useful for the much simpler task of motion direction judgement.

Observers' ability to remember these features is capacity-limited, as shown by the lack of dependence on sample length with long clips in Experiment 5.1, contrasted with the marked effect of sample length with shorter clips in Experiment 5.2. This can be explained if long clips saturate the memory store, but short clips do not.

The alternative to low-level sparse sampling is representation in a high-level space. This requires processing and the transformation of local detail into a compressed code, such as a face space or a gist. Our results suggest that this is unlikely; there is no common prototype in relation to which we can represent a dynamic flame scene, and the stimulus contains few long-range correlations. We find no evidence for a pure inversion effect on flame, and thus no evidence of orientation specialisation.

7.3 How is dynamic flame matched?

In computer vision, matching images is usually done by representing two stimuli in the same way and then comparing their codes. The human brain may not take the same approach; it may operate differently during sample processing than during test processing, without an explicit "comparison" phase. This idea is supported by observers'

reports that they often know whether the test clip is a match or not before it has finished; if they needed to wait until after test presentation to compute a representation, this would not be possible.

Traditional theories of object recognition assume that, when a recognition trial begins, the sample is already encoded in long-term memory. When recognising a car, for example, we already possess a category description which was built from cars witnessed during our youth (and probably modified by more recently perceived cars). Models such as Selfridge's Pandemonium and Poggio's HMAX attempt to explain the recognition of already known objects, not the matching of two new objects. The problem, however, is essentially the same: in traditional "core object recognition"[35] the sample is a category, perceived through a series of exemplars; it can then be matched to any object in that category. In a delayed match-to-sample experiment, the sample is a single video clip, but it still defines a category: the category of video clips which are similar enough to the sample to elicit a "Yes" response.

Our results suggest that dynamic flame is encoded as a series of low-level spatiotemporal features. If the test were encoded in the same way as the sample, the visual system would face an abstraction problem: how could it be sure to encode the same spatiotemporal features in the test as in the sample? If we were to sample from the bottom right of the beginning of the sample and the top left of the end of the test, the locality of flame would mean that these zones would be uncorrelated and thus unmatchable. It is easy to sample in the same spatial locations, but not the same temporal locations, since the temporal location of the sample in the test is *a priori* unknown. This fits with the results of Experiment 4.2, which showed that both spatial and temporal inversion affect matching performance. It is also supported by Experiment 5.4, in which we varied target position but not search space size: matching was much better at the beginning of the test, where the offsets (from the beginning of the sample) of any spatiotemporal features would be the same as their offsets from the beginning of the test.

How does the brain decide which areas of a complex dynamic stimulus to sample? This process is usually referred to as the computation of salience, and is bottom-up, involving early visual areas and low-level processing. Having encoded a series of features, how does the visual system ensure the same features will be encoded again?

If they are arbitrary features, it would be necessary to prime the early processing stages with details of these features, ensuring their selection. This would require extensive feedback connections. Instead, the matching process could simply select the most salient spatiotemporal features in both the sample and the test. For small search spaces, it is likely that the same features would be attended to; for larger search spaces, salient distractors could reduce the possibility of obtaining a match. This approach does not require giving early processing stages feedback about what to expect, relying on its existing learned notions of salience.

The flame search process is highly vulnerable to distractors; when tests are not much longer than samples, accuracy is high, but longer tests have a crippling effect. By characterising the slow decay of flame representations during a blank ISI, we concluded that this effect was due to interference as well as decay. Flame representations are thus subject to temporal crowding. As shown in Experiment 5.3, this was mainly due to false positives, not misses.

Finally, we note that observers show no accuracy improvement effects during flame matching, except after the first 50 trials. There is no evidence of consistent perceptual learning.

7.4 Contribution to knowledge

We conducted the first psychophysical experiments on the motion properties of dynamic flame and on observers' ability to match examples of this stimulus. We also, to our knowledge, conducted the first temporal visual search experiments on dynamic natural scenes. Research in this area has focussed on the recognition of static objects or dynamic objects which did not require segmentation from a background.

We conducted extensive image analysis of a dynamic flame dataset which was representative of the stimulus class, characterising the motion fields which can be computationally recovered from it and the spatial and temporal frequencies present. Dynamic flame possesses a temporal frequency spectrum which is exponential rather than of the form $1/f$.

We compared flame matching ability to face matching ability and found no inter-trial correlation in response accuracy, which suggests that faces and flame are matched

by different processes, not by a shared resource. In summary, we found no evidence of dedicated, trainable, or orientation-tuned processes which may be used to match flame. Matching appears to be done by existing low-level processes.

7.5 Aesthetics

Fire is often renowned for its beauty and aesthetic value. Our hypothesis that dynamic flame is represented as a set of spatiotemporally local features, and that it does not contain any long-range correlations, provides a natural explanation. When viewing flame, the visual system appears to rapidly encode short spatiotemporal features in different areas of the stimulus. These features suddenly appear, but do not last long enough to hold attention for very long. Attention thus rapidly switches from area to area. This constant stream of rapidly appearing and vanishing features holds attention on the stimulus and renders it visually interesting.

7.6 Further work

The perception of dynamic natural scenes brings to mind two main avenues of research: the nature of the representations employed to encode and compare scenes, and the challenges involved in comparing dynamic stimuli.

One of our main points is that dynamic flame is represented by sparse sampling of low-level spatiotemporal features. To further investigate this result, it would be useful to compare observers' matching behaviour to that shown by a computational classifier. A classifier capable of doing pixelwise comparison would find the job trivial; we would require a biologically plausible classifier with initial stages modelled on the visual system (as in the banks of oriented filters used by Freeman and Simoncelli[265]).

Dynamic flame has few long-range correlations, so information captured near the fovea could be of paramount importance. To test this, an eye tracker could be used during a delayed match-to-sample task to record the areas of the sample clip to which the observer is attending. The same could be done with the test clip, producing two smaller "retina's-view" clips which track the images projected around the fovea. These could then be submitted to classification, aiming to predict the Yes/No responses of

human observers.

Storing small, spare spatiotemporal features requires attending to the same areas, at the same time, in both sample and test clips. If observers are sampling from just one area of the sample, it is easy to attend to the right place in the test. If, however, a more complex sampling strategy is used, perhaps informed by bottom-up salience ("I am encoding the bright flash in the centre of the screen, followed by the distinctive flame in the top right, followed by a spark in the bottom left"), the same areas must be attended to in the test to allow good matching. This is easy if the sample and test start at the same frame; if, however, the observer does not know the location of the sample in the test, sampling the test correctly is much harder. An effective strategy would be "look for the first spatiotemporal feature I encoded, then, once it is found, shift attention to the location of the next one". Conversely, it could be that bottom-up salience is relied upon to ensure the same sampling pattern. Discussions with observers suggested that high-variance areas of the stimulus are preferentially sampled. A psychophysical experiment could differentiate between these strategies by introducing salient distractors and measuring their effect on matching performance.

This task begs a more general question: how do we detect a match when our target is a sequence? This is necessary in many real-world tasks, from recognising music to interpreting echocardiograms. The EEG signatures of detecting a static target have been well studied[266]; what are the EEG signatures of detecting a dynamic target? Do they appear when the first element in the target sequence is witnessed, or once the last element has been matched? What do these signatures look like when the observer sees a partial sequence match, in which the first elements correspond to the first elements of the target but the rest do not? Recent developments in stimulus presentation and high-density EEG allow these basic questions to be investigated.

7.7 Conclusions

In summary, we used a range of image analysis, clip matching and visual search experiments to look for any evidence of high-level (spatially or temporally global) representations of dynamic flame. We found no such computational encodings, suggesting that there is little global structure present in the images. Neither did we find any

psychophysical evidence of long-range spatiotemporal features, gradually increasing expertise, or pop-out. Image-based methods only found similarities which were very temporally local, and playback detection (using observers' category representation for flame) required highly local motion information.

This evidence suggests strongly that flame is a dynamic texture: it be can be easily detected as a category, but individuating particular exemplars requires costly low-level comparison. This makes temporal search extremely difficult, since low-level representations in visual working memory are constantly being overwritten. Flame, despite its antiquity and aesthetic value, is interpreted by the visual system in a similar way to dynamic pink noise. The visual system appears to match flame by taking local spatiotemporal samples as opposed to computing a compressed, high-level representation.

Bibliography

[1] Kalanit Grill-Spector, Zoe Kourtzi, and Nancy Kanwisher. The lateral occipital complex and its role in object recognition. *Vision research*, 41(10):1409–1422, 2001.

[2] Maximilian Riesenhuber and Tomaso Poggio. Are cortical models really bound by the binding problem? *Neuron*, 24(1):87–93, 1999.

[3] Christoph Von der Malsburg. The what and why of binding: the modelers perspective. *Neuron*, 24(1):95–104, 1999.

[4] Christoph Von Der Malsburg. Binding in models of perception and brain function. *Current opinion in neurobiology*, 5(4):520–526, 1995.

[5] Simon Thorpe, Denis Fize, Catherine Marlot, et al. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

[6] Anil K. Jain, Yu Zhong, and Sridhar Lakshmanan. Object matching using deformable templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(3):267–278, 1996.

[7] Roberto Brunelli and Tomaso Poggio. Face recognition: Features versus templates. *IEEE transactions on pattern analysis and machine intelligence*, 15(10):1042–1052, 1993.

[8] Alan L Yuille, Peter W Hallinan, and David S Cohen. Feature extraction from faces using deformable templates. *International journal of computer vision*, 8(2):99–111, 1992.

[9] Shimon Ullman. Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32(3):193–254, 1989.

[10] Oliver G Selfridge. Pandemonium: a paradigm for learning in mechanisation of thought processes. 1958.

[11] 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.

[12] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

[13] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[14] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[15] David H Hubel, Torsten N Wiesel, and Michael P Stryker. Anatomical demonstration of orientation columns in macaque monkey. *Journal of Comparative Neurology*, 177(3):361–379, 1978.

[16] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[17] Jessie J Peissig and Michael J Tarr. Visual object recognition: do we know more now than we did 20 years ago? *Annu. Rev. Psychol.*, 58:75–96, 2007.

[18] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. 1971.

[19] Pierre Jolicoeur. The time to name disoriented natural objects. *Memory & Cognition*, 13(4):289–303, 1985.

[20] Michael J Tarr and Steven Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive psychology*, 21(2):233–282, 1989.

[21] Isabel Gauthier, William G Hayward, Michael J Tarr, Adam W Anderson, Pawel Skudlarski, and John C Gore. Bold activity during mental rotation and viewpoint-dependent object recognition. *Neuron*, 34(1):161–171, 2002.

[22] Heinrich H Bülthoff and Shimon Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, 89(1):60–64, 1992.

[23] Tomaso Poggio and Shimon Edelman. A network that learns to recognize 3d objects. *Nature*, 343(6255):263–266, 1990.

[24] CG Gross, DB Bender, and CE Rocha-Miranda. Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science*, 166(910):1303–1306, 1969.

[25] Nicole C Rust and James J DiCarlo. Selectivity and tolerance (invariance) both increase as visual information propagates from cortical area v4 to it. *The Journal of Neuroscience*, 30(39):12978–12995, 2010.

[26] Keiji Tanaka, Hide-aki Saito, Yoshiro Fukada, and Madoka Moriya. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J Neurophysiol*, 66(1):170–189, 1991.

[27] DI Perrett, JK Hietanen, MW Oram, PJ Benson, and ET Rolls. Organization and functions of cells responsive to faces in the temporal cortex [and discussion]. *Philosophical transactions of the royal society of London. Series B: Biological sciences*, 335(1273):23–30, 1992.

[28] NK Logothetis and J Pauls. Psychophysical and physiological evidence for viewer-centered object representations in the primate. 1995.

[29] MC Booth and Edmund T Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, 8(6):510–523, 1998.

[30] Eucaly Kobatake, Gang Wang, and Keiji Tanaka. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *Journal of Neurophysiology*, 80(1):324–330, 1998.

[31] Keiji Tanaka. Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1):109–139, 1996.

[32] Kazushige Tsunoda, Yukako Yamane, Makoto Nishizaki, and Manabu Tanifuji. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature neuroscience*, 4(8):832–838, 2001.

[33] Martha J Farah. *Visual agnosia*. MIT press, 2004.

[34] Martha J Farah. Is an object an object an object? cognitive and neuropsychological investigations of domain specificity in visual object recognition. *Current Directions in Psychological Science*, pages 164–169, 1992.

[35] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[37] Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

[38] Shimon Ullman, Michel Vidal-Naquet, and Erez Sali. Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, 5(7):682–687, 2002.

[39] Danijela Vukadinovic and Maja Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 2, pages 1692–1698. IEEE, 2005.

[40] Daniel González-Jiménez and José Luis Alba-Castro. Shape-driven gabor jets for face description and authentication. *Information Forensics and Security, IEEE Transactions on*, 2(4):769–780, 2007.

[41] Geraldine Dawson, Leslie Carver, Andrew N Meltzoff, Heracles Panagiotides, James McPartland, and Sara J Webb. Neural correlates of face and object recognition in young children with autism spectrum disorder, developmental delay, and typical development. *Child development*, 73(3):700–717, 2002.

[42] Marijke Brants, Johan Wagemans, and Hans P Op de Beeck. Activation of fusiform face area by greebles is related to face similarity but not expertise. *Journal of cognitive neuroscience*, 23(12):3949–3958, 2011.

[43] Guy Wallis. Toward a unified model of face and object recognition in the human visual system. *Frontiers in psychology*, 4, 2013.

[44] Wulfram Gerstner and Werner M Kistler. Mathematical formulations of hebbian learning. *Biological cybernetics*, 87(5-6):404–415, 2002.

[45] Leon D Harmon and Bela Julesz. Masking in visual recognition: Effects of two-dimensional filtered noise. *Science*, 180(4091):1194–1197, 1973.

[46] Risto Näsänen. Spatial frequency bandwidth used in the recognition of facial images. *Vision research*, 39(23):3824–3833, 1999.

[47] Li Zhao and Charles Chubb. The size-tuning of the face-distortion after-effect. *Vision research*, 41(23):2979–2994, 2001.

[48] Tamara L Watson and Colin WG Clifford. Pulling faces: An investigation of the face-distortion aftereffect. *Perception-London*, 32(9):1109–1116, 2003.

[49] Tim Valentine. Face-space models of face recognition. *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges*, pages 83–113, 2001.

[50] Vicki Bruce. Perceiving and recognising faces. *Mind & Language*, 5(4):342–364, 1990.

[51] Vicki Bruce and Andy Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986.

[52] D Marr. Vision, 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, 1982.

[53] Robert J Baron. Mechanisms of human facial recognition. *International Journal of Man-Machine Studies*, 15(2):137–178, 1981.

[54] Andrew W Young, Deborah Hellawell, and Dennis C Hay. Configurational information in face perception. *Perception*, 16(6):747–759, 1987.

[55] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[56] W Scheirer, S Anthony, Ken Nakayama, and D Cox. Perceptual annotation: Measuring human vision to improve computer vision. 2014.

[57] Beat Fasel. Robust face analysis using convolutional neural networks. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 40–43. IEEE, 2002.

[58] Plinio Moreno, Manuel J Marín-Jiménez, Alexandre Bernardino, José Santos-Victor, and Nicolás Pérez de la Blanca. A comparative study of local descriptors for object category recognition: Sift vs hmax. In *Pattern recognition and image analysis*, pages 515–522. Springer, 2007.

[59] David I. Perrett, Amanda J. Mistlin, and Andrew J. Chitty. Visual neurones responsive to faces, 1987.

[60] L. F. Abbott, Edmund T. Rolls, and Martin J. Tovee. Representational capacity of face coding in monkeys. *Cerebral Cortex*, 6(3):498–505, 1996.

[61] Winrich A Freiwald and Doris Y Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science (New York, N.Y.)*, 330(6005):845–851, 2010.

[62] Daniel W Piepers and Rachel A Robbins. A review and clarification of the terms holistic,configural, and relational in the face perception literature. *Frontiers in psychology*, 3, 2012.

[63] Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 2013.

[64] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

[65] Mark H Johnson, Suzanne Dziurawiec, Hadyn Ellis, and John Morton. Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1):1–19, 1991.

[66] Menaka Rajapakse and Yan Guo. Multiple landmark feature point mapping for robust face recognition. In *Audio-and Video-Based Biometric Person Authentication*, pages 96–101. Springer, 2001.

[67] Elinor McKone and Galit Yovel. Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? toward a new theory of holistic processing. *Psychonomic Bulletin & Review*, 16(5):778–797, 2009.

[68] Valérie Goffaux and Bruno Rossion. Faces are" spatial"–holistic face perception is supported by low spatial frequencies. *Journal of Experimental Psychology: Human Perception and Performance*, 32(4):1023, 2006.

[69] Xiong Jiang, Ezra Rosen, Thomas Zeffiro, John VanMeter, Volker Blanz, and Maximilian Riesenhuber. Evaluation of a shape-based model of human face discrimination using fmri and behavioral techniques. *Neuron*, 50(1):159–172, 2006.

[70] Robert K Yin. Looking at upside-down faces. *Journal of experimental psychology*, 81(1):141, 1969.

[71] Martha J Farah, James W Tanaka, and H Maxwell Drain. What causes the face inversion effect? *Journal of Experimental Psychology: Human perception and performance*, 21(3):628, 1995.

[72] Elan Barenholtz and Michael J Tarr. Reconsidering the role of structure in vision. *Psychology of Learning and Motivation*, 47:157–180, 2006.

[73] James V Haxby, Barry Horwitz, Leslie G Ungerleider, Jose Ma Maisog, Pietro Pietrini, and Cheryi L Grady. The functional organization of human extrastriate cortex: a pet-rcbf study of selective attention to faces and locations. *Journal of Neuroscience*, 14(11):6336–6353, 1994.

[74] James V Haxby, Elizabeth A Hoffman, and M Ida Gobbini. The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233, 2000.

[75] Li Fei-Fei, Rufin VanRullen, Christof Koch, and Pietro Perona. Why does natural scene categorization require little attention? exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, 12(6):893–924, 2005.

[76] Arien Mack and Irvin Rock. Inattentional blindness: Perception without attention. *Visual attention*, 8:55–76, 1998.

[77] Michael A Cohen, George A Alvarez, and Ken Nakayama. Natural-scene perception requires attention. *Psychological science*, 2011.

[78] Dariu M Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[79] John N Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 37(11):2049, 1979.

[80] Viren Jain, H Sebastian Seung, and Srinivas C Turaga. Machines that learn to segment images: a crucial technology for connectomics. *Current opinion in neurobiology*, 20(5):653–666, 2010.

[81] Darren Newtson. Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28(1):28, 1973.

[82] Albert Cardona, Stephan Saalfeld, Johannes Schindelin, Ignacio Arganda-Carreras, Stephan Preibisch, Mark Longair, Pavel Tomancak, Volker Hartenstein, and Rodney J Douglas. Trakem2 software for neural circuit reconstruction. *PloS one*, 7(6):e38011, 2012.

[83] Darren Newtson. Foundations of attribution: The perception of ongoing behavior. *New directions in attribution research*, 1:223–247, 1976.

[84] Aude Oliva. Gist of the scene. *Neurobiology of attention*, 696(64):251–258, 2005.

[85] Mary C Potter and Ellen I Levy. Recognition memory for a rapid sequence of pictures. *Journal of experimental psychology*, 81(1):10, 1969.

[86] Mary C Potter. Understanding sentences and scenes: The role of conceptual short-term memory. *Fleeting memories: Cognition of brief visual stimuli*, pages 13–46, 1999.

[87] Paul Fraisse. The psychology of time. 1963.

[88] David M Eagleman. Human time perception and its illusions. *Current opinion in neurobiology*, 18(2):131–136, 2008.

[89] David M Eagleman, U Tse Peter, Dean Buonomano, Peter Janssen, Anna Christina Nobre, and Alex O Holcombe. Time and the brain: how subjective time relates to neural time. *The Journal of Neuroscience*, 25(45):10369–10371, 2005.

[90] Yongchang Wang and Barrie J Frost. Time to collision is signalled by neurons in the nucleus rotundus of pigeons. 1992.

[91] Matthew I Leon and Michael N Shadlen. Representation of time by neurons in the posterior parietal cortex of the macaque. *Neuron*, 38(2):317–327, 2003.

[92] CR Gallistel. Time has come. *Neuron*, 38(2):149–150, 2003.

[93] Uma R Karmarkar and Dean V Buonomano. Timing in the absence of clocks: encoding time in neural network states. *Neuron*, 53(3):427–438, 2007.

[94] Oliver J Braddick, KH Ruddock, MJ Morgan, and D Marr. Low-level and high-level processes in apparent motion [and discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290(1038):137–151, 1980.

[95] Patrick Cavanagh and George Mather. Motion: the long and short of it. *Spatial vision*, 4(2):103–129, 1989.

[96] Curtis L Baker Jrll and Oliver J Braddick. Temporal properties of the short-range process in apparent motion. *Perception*, 14:181–192, 1985.

[97] George Mather et al. Temporal properties of apparent motion in subjective figures. *Perception*, 17(6):729–736, 1988.

[98] Charles Chubb and George Sperling. Drift-balanced random stimuli: a general basis for studying non-fourier motion perception. *JOSA A*, 5(11):1986–2007, 1988.

[99] Alan Johnston, Christopher P Benton, and Peter W McOwan. Induced motion at texture–defined motion boundaries. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1436):2441–2450, 1999.

[100] Christopher P Benton and Alan Johnston. A new approach to analysing texture-defined motion. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1484):2435–2443, 2001.

[101] Zhong-Lin Lu and George Sperling. The functional architecture of human visual motion perception. *Vision research*, 35(19):2697–2722, 1995.

[102] Johannes M Zanker. Theta motion: a paradoxical stimulus to explore higher order motion extraction. *Vision research*, 33(4):553–569, 1993.

[103] Jamie Carroll Theobald, Brian J Duistermars, Dario L Ringach, and Mark A Frye. Flies see second-order motion. *Current Biology*, 18(11):R464–R465, 2008.

[104] Jamie C Theobald, Patrick A Shoemaker, Dario L Ringach, and Mark A Frye. Theta motion processing in fruit flies. *Frontiers in behavioral neuroscience*, 4, 2010.

[105] Cristóbal Curio, Heinrich H Bülthoff, and Martin A Giese. *Dynamic faces: Insights from experiments and computation*. MIT Press Cambridge, MA, 2011.

[106] John N Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of experimental psychology: human perception and performance*, 4(3):373, 1978.

[107] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[108] Graham E Pike, Richard I Kemp, Nicola A Towell, and Keith C Phillips. Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition*, 4(4):409–438, 1997.

[109] Barbara Knight and Alan Johnston. The role of movement in face recognition. *Visual cognition*, 4(3):265–273, 1997.

[110] Karen Lander, Fiona Christie, and Vicki Bruce. The role of movement in the recognition of famous faces. *Memory & cognition*, 27(6):974–985, 1999.

[111] Karen Lander, Vicki Bruce, and Harry Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology*, 15(1):101–116, 2001.

[112] Karen Lander and Vicki Bruce. Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology*, 12(4):259–272, 2000.

[113] P Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M Blackburn, Elham Tabassi, and Mike Bone. Face recognition vendor test 2002. In *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, page 44. IEEE, 2003.

[114] Fiona Christie and Vicki Bruce. The role of dynamic information in the recognition of unfamiliar faces. *Memory & cognition*, 26(4):780–790, 1998.

[115] Alice J O'Toole, Dana A Roark, and Hervé Abdi. Recognizing moving faces: A psychological and neural synthesis. *Trends in cognitive sciences*, 6(6):261–266, 2002.

[116] Glyn W Humphreys, Nick Donnelly, and M Jane 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.

[117] Andrew J Calder, A Mike Burton, Paul Miller, Andrew W Young, and Shigeru Akamatsu. A principal component analysis of facial expressions. *Vision research*, 41(9):1179–1208, 2001.

[118] Glyn Andrew Cowe. *Example-based computer-generated facial mimicry*. PhD thesis, University College London (University of London), 2003.

[119] Ruth Campbell and DI Perrett. The neuropsychology of lipreading [and discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):39–45, 1992.

[120] Leslie L Steede, Jeremy J Tree, and Graham J Hole. I can't recognize your face but i can recognize its movement. *Journal of Vision*, 6(6):671–671, 2006.

[121] Karen Lander, Glyn Humphreys, and Vickl Bruce. Exploring the role of motion in prosopagnosia: Recognizing, learning and matching faces. *Neurocase*, 10(6):462–470, 2004.

[122] Karen Lander, Harold Hill, Miyuki Kamachi, and Eric Vatikiotis-Bateson. It's not what you say but the way you say it: matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance*, 33(4):905, 2007.

[123] Harold CH Hill, Nikolaus F Troje, and Alan Johnston. Range-and domain-specific exaggeration of facial speech. *Journal of Vision*, 5(10):4, 2005.

[124] Cristóbal Curio, Martin A Giese, Martin Breidt, Mario Kleiner, and Heinrich H Bülthoff. Probing dynamic human facial action recognition from the other side of the mean. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pages 59–66. ACM, 2008.

[125] Alan Johnston. Is dynamic face perception primary in dynamic faces: Insights from experiments and computation, 288.

[126] Patrick Cavanagh. Size and position invariance in the visual system. *Perception*, 7(2):167–177, 1978.

[127] Guy Wallis and Edmund T Rolls. Invariant face and object recognition in the visual system. *Progress in neurobiology*, 51(2):167–194, 1997.

[128] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973.

[129] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5):2, 2002.

[130] JA Beintema and M Lappe. Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences*, 99(8):5661–5663, 2002.

[131] Peter Neri, M Concetta Morrone, and David C Burr. Seeing biological motion. *Nature*, 395(6705):894–896, 1998.

[132] Andrew B Watson. Probability summation over time. *Vision research*, 19(5):515–522, 1979.

[133] DC Burr. Temporal summation of moving images by the human visual system. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 211(1184):321–339, 1981.

[134] Antonino Casile and Martin A Giese. Critical features for the recognition of biological motion. *Journal of vision*, 5(4):6, 2005.

[135] Joachim Lange, Karsten Georg, and Markus Lappe. Visual perception of biological motion by form: A template-matching analysis. *Journal of Vision*, 6(8):6, 2006.

[136] Heinrich H Bulthoff, Christian Wallraven, and Arnulf Graf. View-based dynamic object recognition based on human perception. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 768–776. IEEE, 2002.

[137] James V Stone. Object recognition: View-specificity and motion-specificity. *Vision Research*, 39(24):4032–4044, 1999.

[138] Guy Wallis and Heinrich H Bülthoff. Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences*, 98(8):4800–4804, 2001.

[139] Edmund T Rolls. Invariant visual object and face recognition: neural and computational bases, and a model, visnet. *Frontiers in computational neuroscience*, 6, 2012.

[140] Terrence J Sejnowski and Gerald Tesauro. The hebb rule for synaptic plasticity: algorithms and implementations. *Neural models of plasticity: Experimental and theoretical approaches*, pages 94–103, 1989.

[141] Bruce A Bobier and Michael Wirth. Content-based image retrieval using hierarchical temporal memory. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 925–928. ACM, 2008.

[142] Dan Nie, Xiao-Wei Wang, Li-Chen Shi, and Bao-Liang Lu. Eeg-based emotion recognition during watching movies. In *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*, pages 667–670. IEEE, 2011.

[143] Bruno Rossion and Adriano Boremanse. Robust sensitivity to facial identity in the right human occipito-temporal cortex as revealed by steady-state visual-evoked potentials. *Journal of Vision*, 11(2):16, 2011.

[144] Shimon Edelman and Daphna Weinshall. A self-organizing multiple-view representation of 3d objects. *Biological Cybernetics*, 64(3):209–219, 1991.

[145] Haidong D Lu, Gang Chen, Hisashi Tanigawa, and Anna W Roe. A motion direction map in macaque v2. *Neuron*, 68(5):1002–1013, 2010.

[146] Bruce A Draper, Kyungim Baek, Marian Stewart Bartlett, and J Ross Beveridge. Recognizing faces with pca and ica. *Computer vision and image understanding*, 91(1):115–137, 2003.

[147] T Troscianko, J Fennell, C Benton, and R Baddeley. The perception of sunsets. In *PERCEPTION*, volume 38, pages 62–63. PION LTD 207 BRONDESBURY PARK, LONDON NW2 5JN, ENGLAND, 2009.

[148] Roland W Fleming. Visual perception of materials and their properties. *Vision research*, 94:62–75, 2014.

[149] Roland W Fleming, Shinya Nishida, and Karl Gegenfurtner. Visual perception of materials: the science of stuff. *Vision Research*, 2015.

[150] Takahiro Kawabe, Kazushi Maruya, Roland W Fleming, and Shinya Nishida. Seeing liquids from visual motion. *Vision research*, 2014.

[151] Jacob Beck and Slava Prazdny. Highlights and the perception of glossiness. *Attention, Perception, & Psychophysics*, 30(4):407–410, 1981.

[152] 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.

[153] Mario Kleiner, David Brainard, Denis Pelli, Allen Ingling, Richard Murray, and Christopher Broussard. Whats new in psychtoolbox-3. *Perception*, 36(14):1, 2007.

[154] Miguel P Eckstein. Visual search: A retrospective. *Journal of Vision*, 11(5):14, 2011.

[155] Chao-Ching Ho. Machine vision-based real-time early flame and smoke detection. *Measurement Science and Technology*, 20(4):045502, 2009.

[156] Hideaki Yamagishi and Junichi Yamaguchi. A contour fluctuation data processing method for fire flame detection using a color camera. In *Industrial Electronics Society, 2000. IECON 2000. 26th Annual Confjerence of the IEEE*, volume 2, pages 824–829. IEEE, 2000.

[157] Yigithan Dedeoglu, B Ugur Töreyin, Ugur Güdükbay, and A Enis Cetin. Real-time fire and flame detection in video. In *ICASSP (2)*, pages 669–672, 2005.

[158] Wen-Bing Horng, Jim-Wen Peng, and Chih-Yuan Chen. A new image-based real-time flame detection method using color analysis. In *Networking, Sensing and Control, 2005. Proceedings. 2005 IEEE*, pages 100–105. IEEE, 2005.

[159] Hideaki Yamagishi and Jun'ichi Yamaguchi. Fire flame detection algorithm using a color camera. In *Micromechatronics and Human Science, 1999. MHS'99. Proceedings of 1999 International Symposium on*, pages 255–260. IEEE, 1999.

[160] B Uğur Töreyin, Yiğithan Dedeoğlu, et al. Flame detection in video using hidden markov models. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II–1230. IEEE, 2005.

[161] B Uğur Töreyin, Yiğithan Dedeoğlu, Uğur Güdükbay, and A Enis Cetin. Computer vision based method for real-time fire and flame detection. *Pattern recognition letters*, 27(1):49–58, 2006.

[162] Judith H Langlois and Lori A Roggman. Attractive faces are only average. *Psychological science*, 1(2):115–121, 1990.

[163] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391–412, 2003.

[164] GJ Burton and Ian R Moorhead. Color and spatial structure in natural scenes. *Applied Optics*, 26(1):157–170, 1987.

[165] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 4(12):2379–2394, 1987.

[166] DJ Tolhurst, Y Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992.

[167] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[168] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[169] Richard Dosselmann and Xue Dong Yang. A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, 5(1):81–91, 2011.

[170] Francis Galton. Composite portraits, made by combining those of many different persons into a single resultant figure. *Journal of the Anthropological Institute of Great Britain and Ireland*, pages 132–144, 1879.

[171] Abdul J Jerri. The shannon sampling theorem its various extensions and applications: A tutorial review. *Proceedings of the IEEE*, 65(11):1565–1596, 1977.

[172] Ajay Divakaran, Regunathan Radhakrishnan, Kadir Peker, et al. Motion activity-based extraction of key-frames from video shots. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–932. IEEE, 2002.

[173] Walter S Pritchard. The brain in fractal time: 1/f-like power spectrum scaling of the human electroencephalogram. *International Journal of Neuroscience*, 66(1-2):119–129, 1992.

[174] Claude Bedard, Helmut Kroeger, and Alain Destexhe. Does the 1/f frequency scaling of brain signals reflect self-organized critical states? *Physical review letters*, 97(11):118102, 2006.

[175] David L Gilden, Thomas Thornton, and Mark W Mallon. 1/f noise in human cognition. *Science*, 267(5205):1837–1839, 1995.

[176] Yi-Cheng Zhang. Complexity and 1/f noise. a phase space approach. *Journal de Physique I*, 1(7):971–977, 1991.

[177] Oliver J Braddick, Justin MD O'Brien, John Wattam-Bell, Janette Atkinson, Tom Hartley, and Robert Turner. Brain areas sensitive to coherent visual motion. *Perception-London*, 30(1):61–72, 2001.

[178] Emily Grossman, M Donnelly, R Price, D Pickens, V Morgan, G Neighbor, and R Blake. Brain areas involved in perception of biological motion. *Journal of cognitive neuroscience*, 12(5):711–720, 2000.

[179] Wilson S Geisler. Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.*, 59:167–192, 2008.

[180] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2432–2439. IEEE, 2010.

[181] Alan Johnston, Peter W McOwan, and Christopher P Benton. Robust velocity computation from a biologically motivated model of motion perception. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1418):509–518, 1999.

[182] A Johnston and CWG Clifford. Perceived motion of contrast-modulated gratings: predictions of the multi-channel gradient model and the role of full-wave rectification. *Vision research*, 35(12):1771–1783, 1995.

[183] Alan Johnston, PW McOwan, and H Buxton. A computational model of the analysis of some first-order and second-order motion patterns by simple and complex cells. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 250(1329):297–306, 1992.

[184] Berthold K Horn and Brian G Schunck. Determining optical flow. In *1981 Technical Symposium East*, pages 319–331. International Society for Optics and Photonics, 1981.

[185] Michael J Black and Paul Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996.

[186] Daniel Scharstein and Richard Szeliski. Middlebury stereo vision page. *Online at http://www. middlebury. edu/stereo*, 2002.

[187] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

[188] Stefano Soatto, Gianfranco Doretto, and Ying Nian Wu. Dynamic textures. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 439–446. IEEE, 2001.

[189] Andrew J Calder and Andrew W Young. Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, 6(8):641–651, 2005.

[190] Koray Balci and Volkan Atalay. Pca for gender estimation: which eigenvectors contribute? In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 363–366. IEEE, 2002.

[191] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE, 2006.

[192] Fintan Nagle, Harry Griffin, Alan Johnston, and Peter McOwan. Techniques for mimicry and identity blending using morph space pca. In *Computer Vision-ACCV 2012 Workshops*, pages 296–307. Springer, 2013.

[193] Fatos Berisha, Alan Johnston, and Peter W McOwan. Identifying regions that carry the best information about global facial configurations. *Journal of vision*, 10(11):27, 2010.

[194] Fatos Berisha, Alan Johnston, and Peter McOwan. Spatial location of critical facial motion information for pca-based performance-driven mimicry. *Journal of Vision*, 7(9):495–495, 2007.

[195] DC Pace, M Shi, JE Maggs, GJ Morales, and TA Carter. Exponential frequency spectrum and lorentzian pulses in magnetized plasmas. *Physics of Plasmas (1994-present)*, 15(12):122304, 2008.

[196] DC Pace, M Shi, JE Maggs, GJ Morales, and TA Carter. Exponential frequency spectrum in magnetized plasmas. *Physical review letters*, 101(8):085001, 2008.

[197] Patrik Vuilleumier, Jorge L Armony, Jon Driver, and Raymond J Dolan. Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature neuroscience*, 6(6):624–631, 2003.

[198] Anne Guérin-Dugué and Aude Oliva. Classification of scene photographs from local orientations features. *Pattern Recognition Letters*, 21(13):1135–1140, 2000.

[199] Karla K Evans and Anne Treisman. Perception of objects in natural scenes: is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, 31(6):1476, 2005.

[200] Monica S Castelhano and John M Henderson. The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, 34(3):660, 2008.

[201] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[202] Jeremy M Wolfe and Todd S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501, 2004.

[203] John D Mollon. tho'she kneel'd in that place where they grew the uses and origins of primate colour vision. *Journal of Experimental Biology*, 146(1):21–38, 1989.

[204] Laurent Itti, Carl Gold, and Christof Koch. Visual attention and target detection in cluttered natural scenes. *Optical Engineering*, 40(9):1784–1793, 2001.

[205] Ulric Neisser. Cognitive psychology. 1967.

[206] Anne Treisman and Stephen Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychological review*, 95(1):15, 1988.

[207] Terry Caelli, Bela Julesz, and Edgar Gilbert. On perceptual analyzers underlying visual texture discrimination: Part ii. *Biological Cybernetics*, 29(4):201–214, 1978.

[208] Persi Diaconis and David Freedman. On the statistics of vision: the julesz conjecture. *Journal of Mathematical Psychology*, 24(2):112–138, 1981.

[209] Bela Julesz and James R Bergen. Human factors and behavioral science: Textons, the fundamental elements in preattentive vision and perception of textures. *Bell System Technical Journal, The*, 62(6):1619–1645, 1983.

[210] Henryk Palus. Representations of colour images in different colour spaces. In *The Colour image processing handbook*, pages 67–90. Springer, 1998.

[211] Anthony J Bell and Terrence J Sejnowski. The independent components of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

[212] Vicki Bruce, Steve Langton, et al. The use of pigmentation and shading information in recognising the sex and identities of faces. *PERCEPTION-LONDON-*, 23:803–803, 1994.

[213] Vilayanur S Ramachandran. Perception of shape from shading. 1988.

[214] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *Solid-State Circuits, IEEE Journal of*, 23(2):358–367, 1988.

[215] Timothy J Andrews and Michael P Ewbank. Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *Neuroimage*, 23(3):905–913, 2004.

[216] Byron Bernal, Magno Guillen, and Juan Camilo Marquez. The spinning dancer illusion and spontaneous brain fluctuations: An fmri study. *Neurocase*, 20(6):627–639, 2014.

[217] Shinsuke Shimojo, Gerald H Silverman, and Ken Nakayama. Occlusion and the solution to the aperture problem for motion. *Vision research*, 29(5):619–626, 1989.

[218] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

[219] Qinqin Wang, Patrick Cavanagh, and Marc Green. Familiarity and pop-out in visual search. *Perception & psychophysics*, 56(5):495–500, 1994.

[220] Anne Treisman and Janet Souther. Search asymmetry: a diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114(3):285, 1985.

[221] Jeremy M Wolfe. Visual memory: What do you know about what you saw? *Current Biology*, 8(9):R303–R304, 1998.

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

[223] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[224] Arthur Jersild. Primacy, recency, frequency, and vividness. *Journal of Experimental Psychology*, 12(1):58, 1929.

[225] William D Crano. Primacy versus recency in retention of information and opinion change. *The Journal of Social Psychology*, 101(1):87–96, 1977.

[226] Alan L Yuille. Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, 1991.

[227] NS Sutherland. Object recognition. *Handbook of perception*, 3:157–185, 2012.

[228] John P Lewis. Fast template matching. In *Vision interface*, volume 95, pages 15–19, 1995.

[229] Dariu M Gavrila. Multi-feature hierarchical template matching using distance transforms. In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, volume 1, pages 439–444. IEEE, 1998.

[230] Shihong Lao, Yasushi Sumi, Masato Kawade, and Fumiaki Tomita. 3d template matching for pose invariant face recognition using 3d facial model built with isoluminance line based stereo vision. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 2, pages 911–916. IEEE, 2000.

[231] Michael N Tombu, Christopher L Asplund, Paul E Dux, Douglass Godwin, Justin W Martin, and René Marois. A unified attentional bottleneck in the human brain. *Proceedings of the National Academy of Sciences*, 108(33):13426–13431, 2011.

[232] Steven J Luck and Edward K Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281, 1997.

[233] Daniel Kahneman, Anne Treisman, and Brian J Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219, 1992.

[234] Jeremy M Wolfe and Sara C Bennett. Preattentive object files: Shapeless bundles of basic features. *Vision research*, 37(1):25–43, 1997.

[235] Eduardo Valle and Matthieu Cord. Advanced techniques in cbir: local descriptors, visual dictionaries and bags of features. In *Computer Graphics and Image Processing (SIBGRAPI TUTORIALS), 2009 Tutorials of the XXII Brazilian Symposium on*, pages 72–78. IEEE, 2009.

[236] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[237] Stephen O'Hara and Bruce A Draper. Introduction to the bag of features paradigm for image classification and retrieval. *arXiv preprint arXiv:1101.3354*, 2011.

[238] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[239] Jason Clarke and Arien Mack. Iconic memory for the gist of natural scenes. *Consciousness and cognition*, 30:256–265, 2014.

[240] Graham Byatt and Gillian Rhodes. Identification of own-race and other-race faces: Implications for the representation of race in face space. *Psychonomic Bulletin & Review*, 11(4):735–741, 2004.

[241] Emery Schubert. Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, 51(3):154–165, 1999.

[242] Konstantin Vasily Petrides, Ria Pita, and Flora Kokkinaki. The location of trait emotional intelligence in personality factor space. *British Journal of Psychology*, 98(2):273–289, 2007.

[243] T Druzgal and Mark D'esposito. Dissecting contributions of prefrontal cortex and fusiform face area to face working memory. *Cognitive Neuroscience, Journal of*, 15(6):771–784, 2003.

[244] Susan M Courtney, Laurent Petit, José Ma Maisog, Leslie G Ungerleider, and James V Haxby. An area specialized for spatial working memory in human frontal cortex. *Science*, 279(5355):1347–1351, 1998.

[245] George A Alvarez and Patrick Cavanagh. The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science*, 15(2):106–111, 2004.

[246] Anne M Burrows. The facial expression musculature in primates and its evolutionary significance. *BioEssays*, 30(3):212–225, 2008.

[247] GJ Hitch and AD Baddeley. Verbal reasoning and working memory. *The Quarterly Journal of Experimental Psychology*, 28(4):603–621, 1976.

[248] Akiko Ikkai, Andrew W McCollough, and Edward K Vogel. Contralateral delay activity provides a neural measure of the number of representations in visual working memory. *Journal of neurophysiology*, 103(4):1963–1968, 2010.

[249] Earl Hunt. *Human intelligence*. Cambridge University Press, 2010.

[250] Charles Spearman. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

[251] Marc S Tibber, Gemma SL Manasseh, Richard C Clarke, Galina Gagin, Sonja N Swanbeck, Brian Butterworth, R Beau Lotto, and Steven C Dakin. Sensitivity to numerosity is not a unique visuospatial psychophysical predictor of mathematical ability. *Vision research*, 89:1–9, 2013.

[252] James V Haxby, Leslie G Ungerleider, Vincent P Clark, Jennifer L Schouten, Elizabeth A Hoffman, and Alex Martin. The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, 22(1):189–199, 1999.

[253] Tim Valentine. Upside-down faces: A review of the effect of inversion upon face recognition. *British journal of psychology*, 79(4):471–491, 1988.

[254] Raymond Bruyer. Rôle du langage et de la mémoire visuelle dans la perception des visages: Effet des lésions cérébrales unilatérales. *Psychologie Francaise*, 27:146–157, 1982.

[255] John Shepherd. Social factors in face recognition. *Perceiving and remembering faces*, 1981.

[256] Wilson P Tanner Jr and John A Swets. A decision-making theory of visual detection. *Psychological review*, 61(6):401, 1954.

[257] Joshua I Gold and Michael N Shadlen. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36(2):299–308, 2002.

[258] Jeffrey M Beck, Wei Ji Ma, Roozbeh Kiani, Tim Hanks, Anne K Churchland, Jamie Roitman, Michael N Shadlen, Peter E Latham, and Alexandre Pouget. Probabilistic population codes for bayesian decision making. *Neuron*, 60(6):1142–1152, 2008.

[259] Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1):35–42, 2011.

[260] Alexander D Logvinenko. The anchoring effect in lightness perception in humans. *Neuroscience Letters*, 334(1):5–8, 2002.

[261] Colin WG Clifford and Gillian Rhodes. *Fitting the mind to the world: Adaptation and after-effects in high-level vision*, volume 2. Oxford University Press, 2005.

[262] Ranald R MacDonald. Intertrial dependence in detection and recognition tasks. *Acta Psychologica*, 38(5):357–365, 1974.

[263] Ingo Fründ, Felix A Wichmann, and Jakob H Macke. Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of vision*, 14(7):9, 2014.

[264] Heinz Wässle. Parallel processing in the mammalian retina. *Nature Reviews Neuroscience*, 5(10):747–757, 2004.

[265] Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.

[266] Jeffrey S Johnson and Bruno A Olshausen. The earliest eeg signatures of object recognition in a cued-target task are postsensory. *Journal of Vision*, 5(4):2, 2005.

Appendix A

Video data on CD

Several videos have been included on the accompanying CD.

Successive t-slices of the 3D spatiotemporal power spectrum of dynamic flame:

FFT_tSlices.avi The x axis shows horizontal power, the y axis vertical power. We can see that power is mainly horizontal at high temporal frequencies, but shows a characteristic x-shape at low temporal frequencies.

FFT_ySlices.avi The x axis shows horizontal power, the y axis temporal power. There is a horizontal line at the temporal DC component, which is an artefact of the lack of temporal windowing. We can see that at low vertical frequencies, power is concentrated near low horizontal frequencies.

FFT_xslices.avi The x axis shows vertical power, the y axis temporal power. There is a horizontal line at the temporal DC component, which is an artefact of the lack of temporal windowing. We can see that at low horizontal frequencies, power is concentrated near low vertical frequencies.

Successive t-slices of the Gaussian-windowed 3D spatiotemporal power spectrum of dynamic flame:

FFT_Gaussian_tSlices.avi The x axis shows horizontal power, the y axis vertical power. We can see that power is mainly horizontal at high temporal frequencies, but shows a characteristic X shape at low temporal frequencies.

FFT_Gaussian_ySlices.avi The x axis shows horizontal power, the y axis temporal

power. There is a horizontal line at the temporal DC component, which is an artefact of the lack of temporal windowing.

FFT_Gaussian_xslices.avi The x axis shows vertical power, the y axis temporal power. There is a horizontal line at the temporal DC component, which is an artefact of the lack of temporal windowing.

PixelFFTFrequencyVideo.avi: having access to the power spectra for each pixel, we generated a series of images showing the power of each pixel at a particular frequency, from the DC component to the Nyquist frequency of 25 Hz. This video runs through these images. A selection of still images are shown in Fig. 3.11,

sMcGMDynamicFire.m4v: we applied the sMcGM to a stack of 500 images, using a temporal filter size of 23 frames (contrast to two frames previously). This video shows the results; sample images are shown in Fig. 3.23.