Scuola Internazionale Superiore di Studi Avanzati (SISSA)

Via Bonomea 265, 34136, Trieste, Italy

Representation of natural movies in rat visual cortex

Candidate:

Liviu Soltuzu

Supervisor:

Prof. Davide Zoccolan

Thesis submitted for the degree of

Doctor of Philosophy in Cognitive Neuroscience

Trieste, January 2018

ACKNOWLEDGEMENTS

I want to thank my PhD supervisor, Prof. Davide Zoccolan, for wise guidance, availability, and for his friendly encouragements and support that kept me going these four years.

My gratitude also goes to the following:

To my first teachers, Federica Rosselli and Sina Tafazoli, and our lab magicians, Margherita Riggi and Daniele Bertolini, without whom my work in the lab would have been so much more difficult;

To my colleagues Riccardo Caramellino and Walter Vanzella who wrote beautiful code alongside me; to Alessandro Di Filippo, Nader Nikbakht, Alessio Ansuini, Natalia Grion, Eis Annavini, Francesca Pulecchi, Alessio Isaja, and the Mechatronics Lab (Marco Gigante, Erik Zorzin, Fabio Meneghini) who contributed software, hardware and technical expertise every time I was in trouble; to Sebastian Reinartz, Giulio Matteucci, Rosilari Bellacosa, Emilian Ciupercă, and Ivan Girardi, for their contributions at various stages of the project;

To other people who brought some adventure after my long days at SISSA: Anna Marchenkova, Sergey Antopolsky, Vladimir Djurdjevic, and Carol Coricelli;

To those who took the time to discuss my project and also taught me how a PhD student’s life can be richer: Sofia Osimo and Romain Brasselet;

To my family, who always believed in me more than I did.

CONTENTS

ABSTRACT 7

1. INTRODUCTION 9

VISUAL OBJECT RECOGNITION 9

RODENT VISION 12

Rat visual system 13

Object recognition in rats 16

EFFICIENT CODING UNDER NATURAL STIMULATION 20

Properties of natural images 20

Efficient coding hypothesis 23

Sparseness 24

Slowness 27

2. METHODS 31

SURGICAL PROCEDURE 31

STIMULI 33

Receptive field mapping 34

Movies 36

DATA ANALYSIS 41

Preprocessing 41

Selection of neurons 43

Receptive field estimation 43

Sparseness as a measure of selectivity 44

Measures of timescale of processing 45

Speed of the visual input 47

3. Representation of natural movies in rat visual cortex 49

Statistics of response properties across visual areas 49

Sparseness and slowness 51

Example responses 52

Sparseness 58

Slowness 61

Noise autocorrelations 63

4. Population decoding 67

Stimulus set 68

Selection of the populations of neurons 70

Decoding analysis 71

Results 75

Black objects 76

White objects 79

The effect of bin sizes on the decoding performance 80

Slowness of movie segments 82

5. Discussion 85

References 89

ABSTRACT

The neuroscientific study of mammalian vision has yielded important achievements in the last decades, but a thorough understanding is still lacking at anatomical and functional levels. From an operational perspective, this understanding would amount to being able to create an artificial system that reaches the performance and versatility of human vision. A first step to reach this goal requires understanding what neuroscientists call *core object recognition*, i.e., the rapid and largely feed-forward processing of visual information that mediates the identification and categorization of objects undergoing various identity-preserving transformations (DiCarlo *et al.*, 2012). Electrophysiological experiments on primates have revealed that populations of neurons along the so-called *ventral stream* – a succession of areas running from the occipital to the temporal cortex – support core recognition, thanks to two key properties: along the pathway, neuronal responses become increasingly more *selective* to objects identities and increasingly more *invariant* to their transformations. With similar goals in mind, albeit with an engineering focus, the machine learning field has developed artificial neural networks, with feed-forward multi-layered architectures, that reach human-level performance in various object recognition tasks. Yet these artificial networks are only loosely inspired by the biological ones, they are mainly trained with supervised techniques, and perform on static images, so they fall short at providing a model for the understanding of the mammalian visual system (Kriegeskorte, 2015).

Electrophysiological investigations therefore remain an important aspect of vision research. Primates are the closest species to humans, but conducting research with them has become more and more difficult (for practical and ethical reasons). Thus, over the past decade, rodents have been used as complementary models to monkeys in the study of visual processing,

as they are smaller, reproduce faster, and, importantly, are more suitable for a large batch of experimental techniques. Recent physiological studies in the mouse and the rat have described successions of areas that resemble the visual pathways found in the monkey (Niell, 2011; Vermaercke *et al.*, 2014; Tafazoli *et al.*, 2017), while behavioral studies have shown that the rat visual system is capable of sophisticated object recognition (Zoccolan *et al.*, 2009; Alemi-Neissi *et al.*, 2013).

Until recently, most of the vision research has focused on simple, static and parametric artificial stimuli (e.g. bars), leaving the representations of time-varying natural images (i.e. movies) little explored. Natural images are those we see during every-day life: they are characterized by high spatial and temporal correlations, i.e. they contain well-defined structures that have similar color intensities over extended areas, and that remain present in the scene over long intervals (for example, a tree trunk is all brown and doesn’t disappear from a scene from moment to moment). Some have argued though that natural images are too complex and still poorly understood to allow well-controlled hypothesis-driven experiments (Rust and Movshon, 2005); others have instead stressed the fact that organisms have evolved within a natural environment and they must have adapted to process natural images in the most efficient way (by exploiting their spatiotemporal regularities), hence the requirement to use this type of stimuli in vision studies (Barlow, 1961; Felsen and Dan, 2005).

The theory that formalizes this hypothesis is named “efficient coding”. The aim of this PhD project is to investigate whether we find evidence in support of the theory in the visual cortex of rats. Specifically, we are addressing two important predictions: 1) that neural responses are increasingly persistent in time, which amounts to measuring if neurons across different areas fluctuate at different timescales in response to the same input (which would be a sign of invariance), and 2) that response distributions successively become sparser (a sign of selectivity).

We recorded the neuronal activity in four rat visual areas: in order from the most medial to the most lateral, V1, LM, LI and LL. The results we found are described in two chapters. In the first one, “Representation of natural movies in rat visual cortex”, we observe a tendency towards an increase of slowness estimated with two different measures, and a decrease of sparseness across the four areas. In the second one, “Population decoding”, we are implementing a population decoding technique and show that LL neurons are better than those from other areas at maintaining a self-similar object representation over time. In the last chapter of the thesis we discuss possible implications of our findings.

1. INTRODUCTION

VISUAL OBJECT RECOGNITION

Any individual object in our visual environment can be presented to us in an infinite number of appearances produced by identity-preserving transformations. The images formed on our retinae are therefore also infinite, yet we are able to rapidly and accurately discriminate them. We call this ability "invariant" object recognition because it affords identification (of one named object) and categorization (of a set of objects) over a large range of transformations. When achieved within less than approximately 300 ms, we can refer to it as "core" object recognition, because it constrains the type of feedback involved in the process (DiCarlo and Cox, 2007). Due to the large variability of the world, the encounter of an observer with an object is almost unique. Transformations in position, scale, pose, lighting and context on a single object will all create different retinal patterns, but the visual system needs to label them as belonging to a single object, i.e. to be selective for object identity and invariant to changes that merely alter the appearance. *Selectivity* and *invariance* (or *tolerance*) thus become the central aspects of object recognition.

It is now well established that primate visual cortex is hierarchically organized in two pathways: the *ventral pathway* or *stream* is specialized in "object" vision and runs from the occipital to the temporal lobe, while the *dorsal pathway* is specialized in "spatial" vision and projects from the occipital to the posterio-parietal cortex (Mishkin *et al.*, 1983). In primates object recognition is resolved through the ventral pathway, that consists of the primary visual cortex (V1), V2, V4, and the various stages of the inferotemporal (IT) (Figure 1).

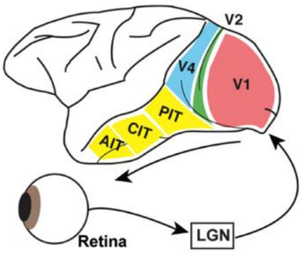

Figure 1. Ventral visual pathway in the primate brain.

LGN, lateral geniculate nucleus; V1, primary visual cortex; AIT, anterior inferotemporal; CIT, central inferotemporal; PIT, posterior inferotemporal.

Taken from DiCarlo *et al.* (2012).

Neural populations along consecutive stages of the ventral stream progressively reformat the visual input in a largely feed-forward manner (DiCarlo *et al.*, 2012). The type of reformatting that is taking place can be intuitively understood with a geometrical perspective (DiCarlo and Cox, 2007). We can consider the response pattern of a population of neurons to a default view of an object as a point in a high-dimensional space in which each dimension is the activity of one neuron in the population (Figure 2A). Any transformation of that object (such as pose or lighting) will result in a different response pattern that will correspond to another point close to the default one. The cloud of points emerging from all possible transformations of the object will define an "object manifold": a continuous, low-dimensional, curved surface inside the space.

In the retina, the first relay of visual information, receptive fields of retinal ganglion cells are relatively small and well driven by simple light patterns, therefore the object manifold will be highly curved. Moreover, manifolds of different objects will be tangled in the retinal representation (Figure 2B), which implies it's impossible to reliably separate them with a linear decision function. So the aim of the ventral processing stream is to reformat the retinal representation of an object, so as to make it flatter and therefore more separable from other object representations (Figure 2B).

In support of the geometrical interpretation, numerous studies have shown that neuronal populations at the highest level of the monkey visual stream – the IT cortex – are capable of supporting object recognition (reviewed by DiCarlo *et al.*, 2012), and that simple linear classifiers can accurately decode object category from the firing rates of IT populations (Hung *et al.*, 2005; Rust and DiCarlo, 2010).

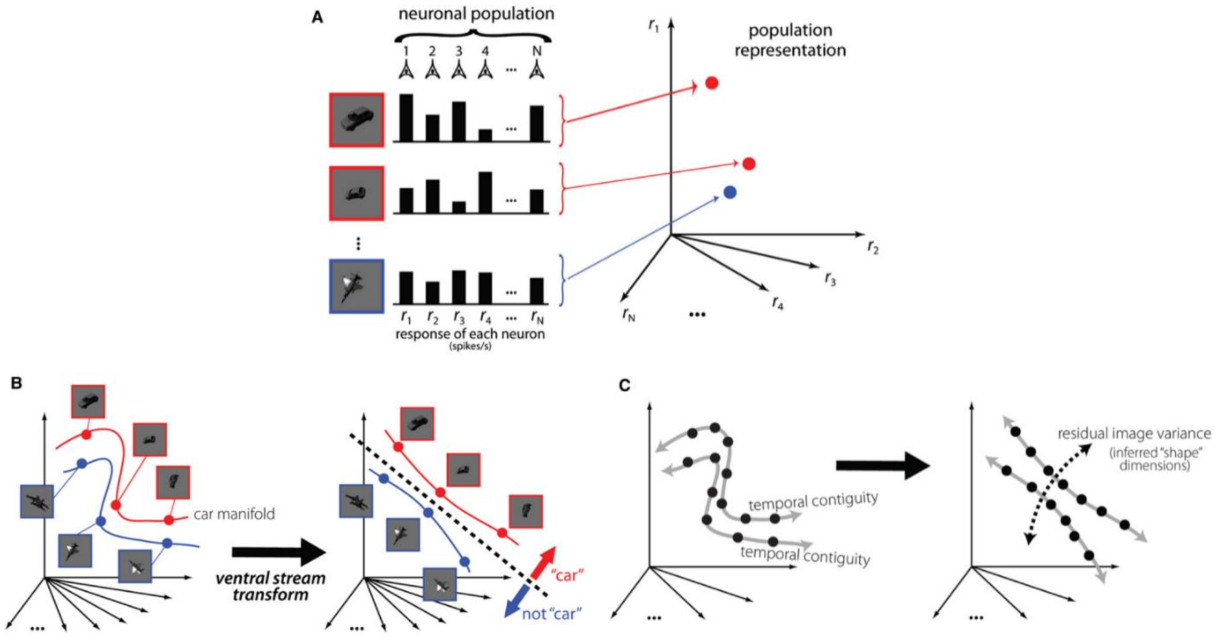

Figure 2. Untangling object representations

A. The response pattern of a population of N neurons to an image is a point in a high-dimensional space where each axis is the response r_i of each neuron. Three images and their corresponding representations are shown: two views of a car (red) and a plane (blue).

B. All possible identity-preserving transformations of an object will form a low-dimensional and highly curved manifold of points in the high-dimensional space. Here, all views of the car and plane lie on their respective manifolds shown as curves for simplicity. The ventral stream is successively transforming each representation until the two manifolds are smoother and separable by a hyperplane (dashed line).

C. An unlabeled object (depicted with black points) undergoing a transformation in time (such as translation) will create views that share a large amount of structure, so the population response patterns will also be similar. Therefore, the manifold produced by an identity-preserving transformation at different time points will be a continuous, curved and low-dimensional surface (gray curves here for simplicity). Cortical circuits can infer shape dimensions by utilizing the temporal continuity of these structures and untangle their representations.

Adapted from DiCarlo *et al.* (2012).

Additional computational ideas that are consistent with physiology enable manifold untangling (DiCarlo and Cox, 2007). First, the flow of information from the retina is projected into an even higher dimensional space in V1 (the number of neurons in V1 is two orders of magnitude larger than the one in the retina), which spreads the data into a larger space (Olshausen and Field, 2004). Second, at each stage, neurons optimize their responses to match the distributions of the visual input: this increases the effective over-completeness, i.e. the number of code elements available to represent visual information (Olshausen and Field, 1996;

Simoncelli and Olshausen, 2001). Third, time implicitly supervises manifold untangling (Figure 2C and details below): theoretical work has formalized the notion that temporal evolution of a retinal image offers clues about object identities in the surrounding world (Földiák, 1991; Wiskott and Sejnowski, 2002).

RODENT VISION

To date, the most common animal model used in object vision studies is the non-human primate, because its visual system closely resembles our own. Unfortunately, the neural mechanisms underlying the formation of explicit object representation in the primate temporal cortex is still not understood. There is probably a combination of factors that make the problem of studying primates so difficult: their visual system is extremely complex (Felleman and Van Essen, 1991); they are relatively big animals and difficult to handle; and genetic, molecular and highly invasive experimental manipulations are often unpractical with this species. To overcome these issues, over the last decade, small mammals, and in particular rodents, have become increasingly popular models for studying visual processing in the brain (reviewed by Huberman and Niell, 2011; Zoccolan, 2015).

Among rodents, mice are the model-of-choice for vision studies. They have poor spatial acuity (0,5 cycles per degree, cpd) and lack some basic visual functions (such as foveation), yet they seem to have a fully functioning visual neural machinery (Prusky *et al.*, 2000; Niell, 2011). Recent advances in imaging techniques allowed extensive charting and functional description of mouse extrastriate areas (Figure 3) (Wang and Burkhalter, 2007; Andermann *et al.*, 2011; Marshel *et al.*, 2011). For example, Andermann and colleagues found that AL (anterolateral) neurons in awake mice were responsive particularly to low spatial frequencies and high temporal frequencies – i.e. large features moving fast, while PM (posteromedial) neurons were responsive to high spatial frequencies and low temporal frequencies – i.e. fine detail moving slowly. The properties of these two areas might suggest disjunctive processing, but Wang and colleagues (2012), by analyzing the cortico-cortical connections, place them along the same processing pathway. In fact, although there are still conflicting results between various groups, an interesting idea that emerges from these studies is that mouse areas form modules that are analogous to the monkey ventral and dorsal pathways. Specifically, the ventral module includes areas LM, LI, P and POR; the dorsal module includes areas AL, RL, PM, AM and A (reviewed by Glickfeld and Olsen, 2017).

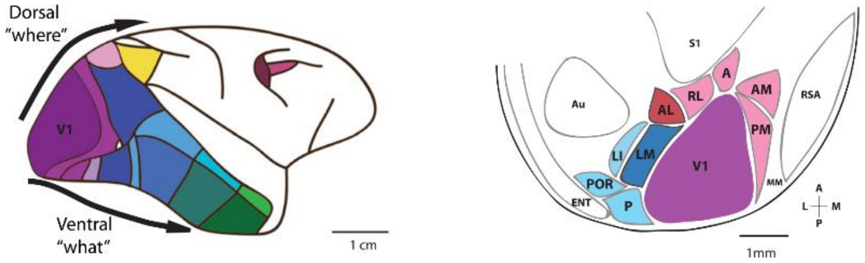

Figure 3. Extrastriate visual areas in primate and mouse.

Ventral (blue shades) and dorsal (pink shades) pathways in the macaque brain (left) and their putative corresponding pathways in mouse visual cortex (right).

Taken from Niell (2011).

Despite these results, high-level vision in mice has been little explored (Cox, 2014). Rats instead were recently shown to be capable of learning complex object discrimination tasks and were put forward as a viable model for the study of invariant object recognition (Zoccolan *et al.*, 2009; Tafazoli *et al.*, 2012; Alemi-Neissi *et al.*, 2013). In the following we will describe rat's visual system and their visual abilities in behavioral and physiological experiments.

RAT VISUAL SYSTEM

Rats' eyes are located on the side of their heads, therefore they have a wider field of view and less binocular overlap (76°) than humans (105°). The dynamics of their movements are complex, regularly disconjugate and are often asymmetrical, as reported by Wallace and colleagues (2013), who tracked eye and pupil movements in freely moving rats. They also show that a large proportion of the eye movements are driven by the vestibulo-ocular reflex: i.e., rotating the head nose-up or nose-down shift the pupils up or down, respectively; and turning the head to the side result in elevation of the lower pupil and declination of the higher pupil.

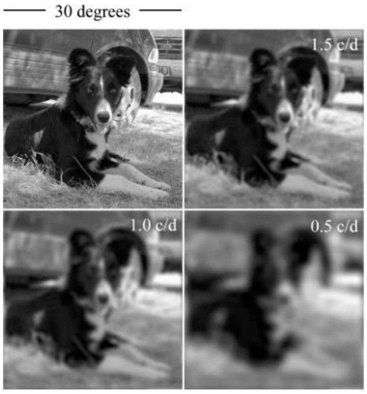

Figure 4. Visual perception of rat strains in the Visual Water Task.

In the visual water task, rats are released into a tank of water that contains two monitors at one end separated by a long divider, and a hidden platform in front of one of the monitors. The rat has to swim towards the monitors, and at the end of the divider to make a choice about which arm of the pool to enter to locate the platform.

The original image has been blurred to model the perception of rats with different acuities of the visual stimuli (width 30°) used in this experiment: Fisher-Norway (top-right), Long Evans (bottom-left), Wistar (bottom-right).

Taken from Prusky *et al.* (2002).

Rat visual acuity ranges from 0,5 cpd to 1,5 in various albino and pigmented strains (Figure 4), but we should note that intense light (e.g. used in the laboratory) can damage the retina and reduce the acuity (Burn, 2008). The Long Evans rat's visual acuity measured in a behavioral task is 1 cpd (Prusky *et al.*, 2002).

Even though the rat's eye cannot accommodate, so light cannot be focused on a discrete point on the retina, it has a considerable depth of focus with its pupil constricted. Rat retina is able to function in scotopic conditions, owing to the preponderance of rods, which account for more than 99% of the photoreceptors. Photopic vision is mediated by two types of cones: one has a peak response at 510 nm (blue-green light), while the other one at about 360 nm (in the UV range) (Sefton *et al.*, 2004). It's worth noting that, due to the color sensitivities of cones in the rat retina, the white emitted by monitors – as a combination of red, green and blue light adjusted for human vision – would probably not appear as white to rats (Burn, 2008).

There is yet no consensus on the critical flicker-fusion frequency (CFF) in rats, which is the frequency at which flickering stimuli appear as continuous. CFF is strongly dependent on the luminous intensity of the stimulus and ranges from approximately 16-19 Hz in pigmented rats (Long Evans) to 39 Hz in albinos, when measured in psychophysical experiments (Williams *et al.*, 1985; Heath and Newland, 2006). Physiological measures that directly look at responses of visual cortical neurons to full-field light pulses place CFF between 10 to 35 Hz in pigmented rats (Wells *et al.*, 2001).

Rat visual cortex consists of at least 13 areas (Figure 5), of which V1 is the largest, accounting for about 15% of the total neocortex (Sereno and Allman, 1991). The topographical organization of V1 is such that the upper visual field (lower retina) is represented caudally, the nasal visual field (temporal retina) is represented laterally, and the vertical meridian is at the lateral border of V1 (Figure 6). The lateral part constitutes the binocular field, whereas in its medial part only the monocular field from the contralateral eye is represented; approximately 80% of V1 cells are binocular (Sefton *et al.*, 2004).

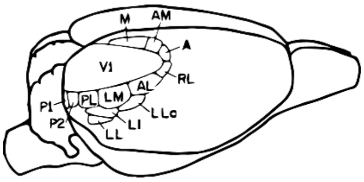

Figure 5. Rat visual areas

Adapted from Sereno and Allman (1991).

Extrastriate areas receive extensive projections that don't originate in V1 (but in thalamus and superior colliculus) and are visually driven in the absence of V1 (Sereno and Allman, 1991). Additionally, asymmetric projections exist within lateral extrastriate cortex and between lateral and medial extrastriate cortex (Coogan and Burkhalter, 1993), and lesions in the most lateral visual areas induce deficits in rats' ability to recognize changes in the visual characteristics of objects (Tees, 1999). The projection patterns and lesioning results, along with the fact that receptive fields are wider in the most lateral extrastriate areas (Espinoza and Thomas, 1983), support the idea that rat visual areas are hierarchically organized on multiple levels, similarly to what has been suggested in mice. Recent functional descriptions of temporal areas seem to confirm that there is a rat visual pathway specialized in object processing (Vermaercke *et al.*, 2014; Tafazoli *et al.*, 2017). These studies pay particular attention to a putative object processing stream: V1 and three extrastriate areas, that run laterally to it, LM (lateromedial), LI (laterointermediate), and LL (laterolateral).

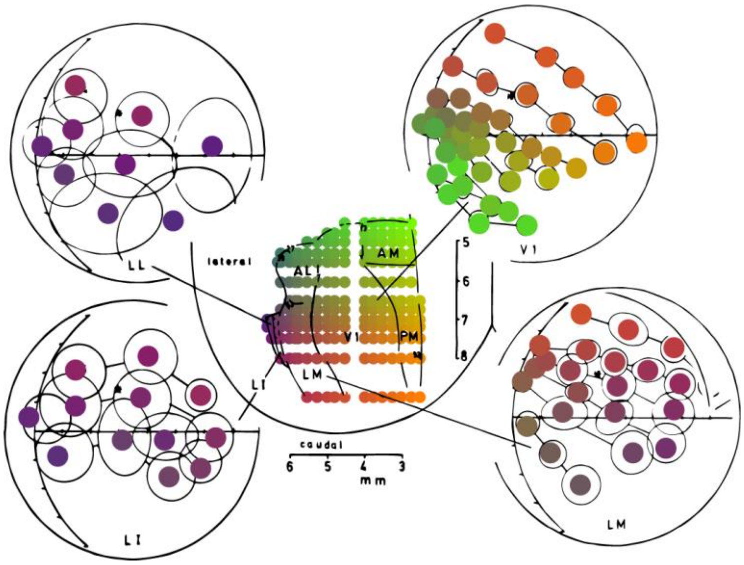

Figure 6. Retinotopic organization of V1, LM, LI and LL

The dorsal view of the left posterior parietal cortex was colored in each of the recording sites to illustrate the representations of the right eye visual field. Color indicates rostrocaudal (green to red) and mediolateral (yellow to blue) positions. Adapted from Espinoza and Thomas (1983).

These areas are retinotopically organized (Espinoza and Thomas, 1983), display receptive field reversal at each border, and each of them encodes a complete representation of the visual field (Figure 6). In LM the upper visual field is represented caudally and the nasal visual field medially, being thus a mirror image of V1. In LI, the upper visual field is medial and the nasal visual field, lateral, being thus a mirror image of LM, or a reduced copy of V1. The mediolateral displacement in LL determines a nasal to temporal progression of receptive fields, and the rostrocaudal displacement, a lower to upper visual field progression of receptive fields (Figure 6).

OBJECT RECOGNITION IN RATS

A first exploration of rat visual abilities started in the 1930s with Karl Lashley's experiments that employed a jumping stand apparatus in two-alternative forced-choice tasks, which required the rat to jump from a platform towards the target stimulus in order to obtain a reward. His results showed that rats use their vision to discriminate patterns and generalize to variations (e.g. due to size or cluttering) of previously learned discriminations (reviewed by Zoccolan, 2015).

Further experiments by other groups confirmed that rats are capable of processing shape information, though without systematically testing their recognition tolerance to transformations in object appearance, nor assessing the perceptual processes underlying this ability, i.e. understanding which visual features were relevant when rats are making a choice. Some of the recent investigations addressed these issues with relatively large sets of natural and artificial stimuli (presented in the following).

Forwood and colleagues (2007) used the so-called "spontaneous oddity preference" task as a means of one-trial testing of cognition. In this task the rat is exposed to three stimuli new for the rat, of which two are identical, while the experimenters record and quantify its exploratory behavior towards the stimuli. Interestingly, when two-dimensional photos were used, rats predominantly explored the different one, showing that rats are capable of spontaneously discriminating (i.e. without training) purely visual stimuli and, moreover, that they can be tested with natural stimuli. Brooks and colleagues (2013) have also used natural images and trained rats to discriminate two pairs of photographic categories (chairs versus flowers and cars versus humans). Rats successfully transferred these discriminations to novel exemplars of each category, partly by relying on global shape-based features of those objects, such as aspect ratio and convexity, but, importantly, not on low-level image attributes, such as brightness or size.

An ecological approach on the study of rat vision was carried out by Vinken and colleagues (2014), who trained rats to discriminate between three classes of 5 sec-long movies – natural movies containing rats, natural movies containing other objects, and artificial (phase-scrambled) movies (Figure 7) – in a visual water maze implementing a two-alternative forced-choice task (described in more detail in the caption of Figure 4). Rats learned to discriminate the stimuli and generalized well to new ones not included in the training set. Better performances were obtained with artificial movies as distractors. Additional controls confirmed that these results were not based on motion energy: specifically, in one experimental manipulation, both target and distractor movies were played at one fourth of the original speed; in another manipulation only one representative frame was shown for each movie. Other controls checked whether rats, in order to solve the task, relied on local luminance features in the lower half of the screen, as proposed by Minini and Jeffery (2006). In this case, the movies and the single frames were altered as to revert the luminance patterns between target and distractor stimuli, which didn't affect the generalization performance.

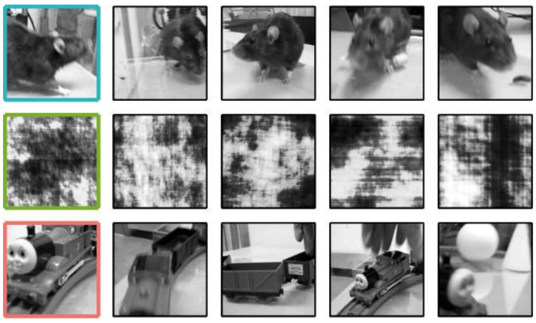

Figure 7. Example frames from the movies used in the study by Vinken *et al.* (2014)

The blue and red rows show snapshots from natural movies that contained rats or other objects (five each). The green row shows snapshots from the five artificial movies generated by phase-scrambling the natural rat movies. This stimulus set was used to test rats' ability to generalize to new, unseen movies.

Adapted from Vinken *et al.* (2014).

The first systematic investigation of rat invariant recognition was carried out by Zoccolan and colleagues, who trained rats to discriminate artificial three-dimensional objects (Zoccolan *et al.*, 2009; Tafazoli *et al.*, 2012; Alemi-Neissi *et al.*, 2013).

In the first study (Zoccolan *et al.*, 2009), rats were trained on the default views of two objects, and afterwards on views that had been transformed along two axes: size and in-depth rotation (Figure 8A and B). After the training phase, rats were tested with all remaining combinations of size and rotation transformations that hadn't been seen by the rats (Figure 8B). Recognition performance was significantly above chance for virtually all pairs of stimuli (Figure 8C), and remained high after the inclusion of new transformations of lighting and in-depth elevation rotation in the generalization part.

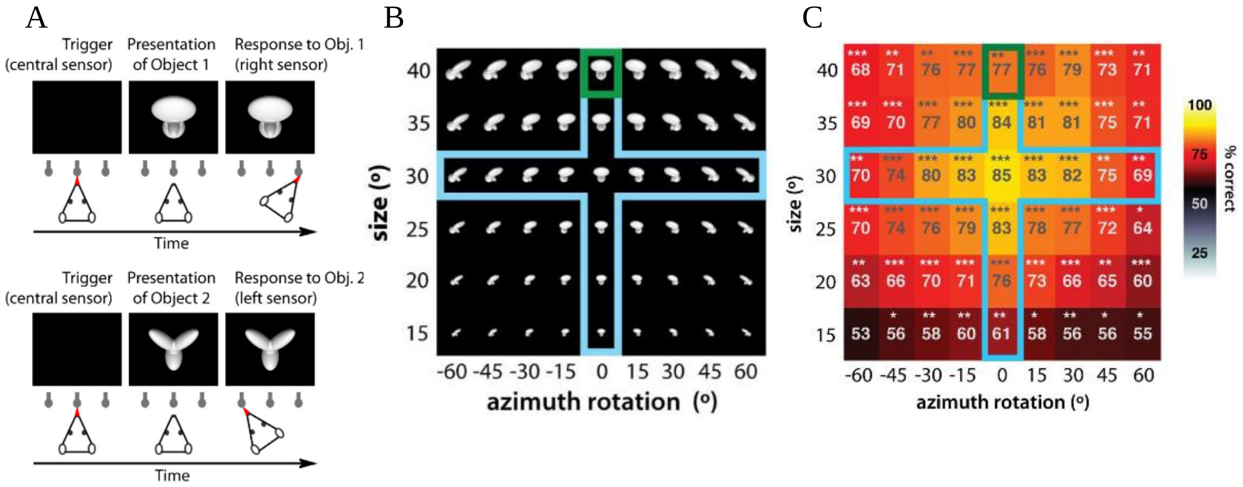

Figure 8. Experimental design in the studies by Zoccolan and colleagues

A. The rat performed the task in a box equipped with one monitor and three sensors placed in front of it. The leftmost and rightmost ports were also providing reward through a licking spout. The rat had to lick the central sensor to trigger stimulus presentation. Afterwards it had to lick either on the right, for one object (top panel), or on the left, for the other object (bottom panel).

B. Transformations of object 1 along the two dimensions: size and azimuth rotation. The green frame highlights the default view used in the first phase of training; the light blue frame highlights the views used in the second phase of training, after the default views had been learned by the rat.

C. The average performance (percentage of correct trials) for each of the object views shown in panel B. Stars indicate the level of statistical significance.

Taken from Zoccolan *et al.* (2009) and Zoccolan (2015).

As already mentioned above, in this experimental paradigm, differently from the two-alternative forced-choice tasks described above, stimuli were presented one at a time on screen. This is important for at least two reasons: first, it is more difficult for rats to rely on low level features because they cannot directly compare the stimuli, but instead they have to remember them; second, at each trial the rat has to compare the stimulus to an internal representation of the object, learned during training. This argues for the claim that rats are capable to form invariant object representations.

In another study, in which the transformed views were only shown as primes of the default views (i.e. briefly flashed before the main stimulus), the same group showed that extensive training on multiple identity-preserving transformations is not necessary for rats to successfully recognize novel views of previously encountered objects (Tafazoli *et al.*, 2012). In other words, assessing invariant recognition using a priming paradigm prevented the invariance endowed by perceptual constancy to be confounded with the invariance resulting from explicitly learning the associative relations among the different views of an object during training (Zoccolan, 2015). Alemi-Neissi and colleagues went even further to show that not only rats can invariantly discriminate objects, but their strategy in doing so is complex, relies on the most informative image features and is close-to-optimal when compared to an ideal observer (Alemi-Neissi *et al.*, 2013).

The behavioral evidence described above is strong enough for a certain level of high-level processing in rats, which justifies physiological investigations. A few studies investigated the functional specialization of the areas along the proposed ventral pathway in rats, i.e. V1, LM, LI and LL.

Vermaercke and colleagues (2014) recorded extracellular signals in awake animals from the four areas and an additional one, TO, that was not described in the literature before (e.g. in the study by Espinoza and Thomas (1983) mentioned above). They reported that response latencies in the two most lateral areas are longer by approximately 20ms than in the medial areas, and that both orientation and direction selectivity increase along the progression, differently from macaques (Vogels and Orban, 1994). By showing simple stimuli at two positions on the screen they were able to measure the position tolerance. More precisely, they displayed 6 moving simple shapes (e.g. H, triangle or +) in the center and on the side of the receptive fields of recorded neurons and trained classifiers to discriminate pairs of shapes from the responses of populations of neurons. Their results seem to suggest that TO is more tolerant than V1, and this is not only due to the increase of receptive field sizes. An interesting confirmation comes from a study by the same group, in which the aim was to find the visual area that is most likely to underlie rats' performance in a behavioral task employing the six shapes. They found that neural population discriminability in V1 correlated well with the low-level properties of the shapes, and neural discriminability in lateral areas correlated well with the behavioral performance. This implies that, in rat temporal areas, visual information is transformed into a representation that is used to drive behavior (Vermaercke *et al.*, 2015).

In another study that explored high-level visual processing of rats, neurons were recorded from V1, LM, LI and LL in anesthetized animals while they were passively presented with a large set of stimuli (Tafazoli *et al.*, 2017). These stimuli were similar to the ones used by Zoccolan and colleagues in the behavioral experiments, and consisted of many transformations along five axes (position, size, in-depth and in-plane rotations, luminance) of 10 three-dimensional objects. By applying information theoretical tools and computational techniques, it was possible to reveal a few neuronal properties along the V1-LL progression: neurons are decreasingly sensitive to stimulus luminance and increasingly so to the shape of the presented objects, and single neurons and population of neurons gain better discriminability of objects in spite of the transformations in their appearance, which amounts to reformatting the visual information in order to make it more accessible for higher brain areas.

The anatomical, behavioral and physiological studies described above indicate rat as a simple enough but potentially good model for studying object recognition.

EFFICIENT CODING UNDER NATURAL STIMULATION

PROPERTIES OF NATURAL IMAGES

Natural scenes are often defined as images of the visual environment that include vegetation, animals and landforms, in which artifacts of civilization do not appear (Olshausen and Field, 2000). In practice, though, this definition is typically relaxed: human artifacts are included or images are stripped of their chromatic or temporal dimensions (Hyvarinen *et al.*, 2009). The naturalistic aspect of the environment can be appreciated by comparison with artificial images. Figure 9 shows example natural and artificial images.

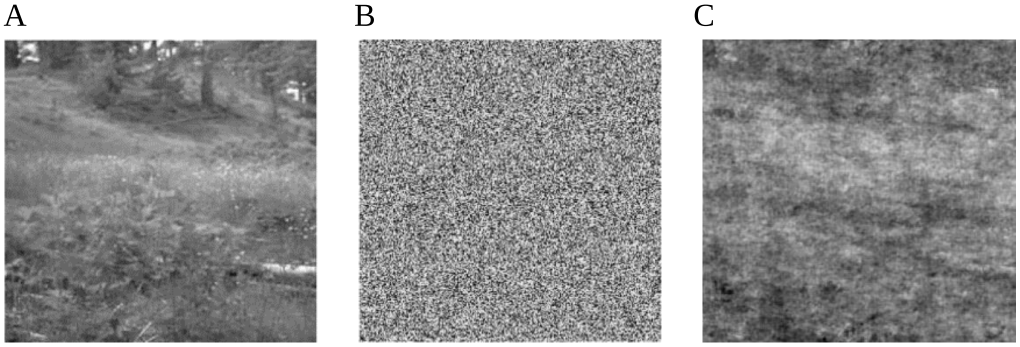

Figure 9. Example natural and artificial images

A. Natural images contains edges and shapes.

B. Some artificial images can have completely random structure: this “white noise” image was generated by assigning to each pixel random intensity values drawn from a uniform distribution.

C. Scrambled images lack consistent edges or sharp features. This image was generated by randomizing the original phase spectrum of the image in A using a 2D fast Fourier transform (more details for the 3D case in Methods).

From a geometrical perspective, images can be thought of as belonging to a high-dimensional state space whose axes are the pixel values of the image. For example, if we're only dealing with images of fixed size of N pixels, one requires an N-dimensional space to represent the set of all possible images, where each of them corresponds to a point in the space. White-noise patterns like the one in Figure 9 would be scattered randomly across the high-dimensional space, whereas natural images, due to the correlations present in them, would occupy a tiny part and lie on a continuous curved surface embedded in the space (Olshausen and Field, 2004).

Static and time-varying images (or movies) have been extensively studied (Dong and Atick, 1995; Simoncelli and Olshausen, 2001; Betsch *et al.*, 2004; Hyvarinen *et al.*, 2009). For simplicity, in the following we will only describe the statistics of time-varying monochromatic images.

Histograms of pixel intensities are the simplest statistical description. By estimating the intensity of light in the original scene for each pixel, one discovers that in most natural images the range of intensity is extremely large, and the distribution of luminance values is typically peaked at the low end with an exponential fall-off toward higher intensities (Olshausen and Field, 2000; Ratliff *et al.*, 2010).

Natural scenes are highly correlated in space and time, which means that the scenes contain patches of relatively constant luminance (indicative of spatial correlations) that are persistent over relatively long intervals (indicative of temporal correlations). At pixel level this translates into groups of pixels having similar values within single frames and across multiple frames (Figure 10A).

The magnitude of correlation between pixel values can be shown by computing the autocorrelation function of an image, which is a function of relative distance between pixels and calculated as the average of products of two pixel values situated at constant distance (Figure 10B). One can see that the strength of the correlations falls with pixel distance within one image or across a set of sequential images (Simoncelli and Olshausen, 2001).

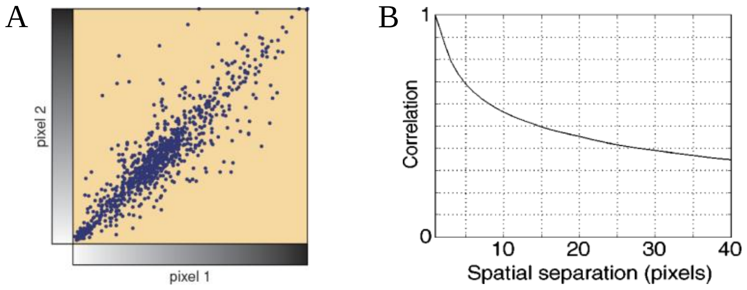

Figure 10. Correlation between pixel intensities

A. The correlation between two adjacent pixels is relatively high in natural images.

B. The autocorrelation function shows the normalized correlation (Y axis) between all pixels placed at a given distance in the image (X axis).

Taken from Olshausen and Field (2000) and Simoncelli and Olshausen (2001)

The properties described above are based on second-order (covariance) statistics. Much of the structure of natural scenes (such as lines, edges, contours) are due to higher-order correlations (Figure 9). This can be shown by computing the two-dimensional power spectrum of an image and then reducing it to a one-dimensional function of spatial frequency by performing a rotational average. Intuitively this means that repetitive patterns along all directions (e.g. edges) can be captured by the Fourier transform, which is then averaged to obtain a single power spectrum plot (Figure 11).

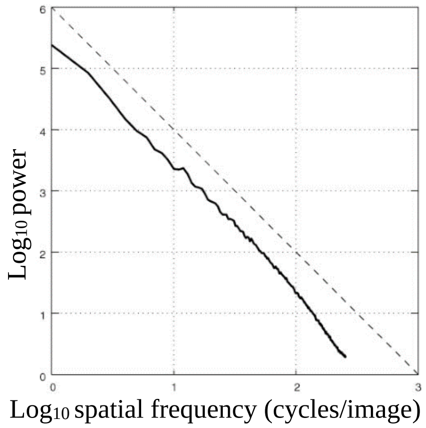

Figure 11. Power spectrum of natural images

The solid line shows the power spectrum of a natural image averaged over all orientations in a log-log plot. The similarity between the power spectrum and the $1/f^2$ function (dashed line) indicates that the power spectrum follows a power law. Taken from Simoncelli and Olshausen (2001).

One can note that, unlike the power of a white noise image which is flat (by definition), the spectral power falls with frequency, f, according to a power law, i.e. $1/f^p$, where p is typically

2. This property suggests that natural images are scale invariant, i.e. their appearance doesn't change when one zooms in or out of it (Olshausen and Field, 2000; Hyvarinen *et al.*, 2009).

Natural scenes have a complex temporal structure that arises from the movement of the observer and of the objects in the world. As in the case of static images, correlations of time-varying images can be best illustrated in the frequency domain (Dong and Atick, 1995). Figure 12 (left panel) shows the spatiotemporal power spectrum averaged across orientations as a function of spatial frequency for different temporal frequencies. We note again a decrease as a reciprocal power of spatial frequency, and that the power spectrum cannot be decoupled into purely spatial and temporal parts (Dong and Atick, 1995).

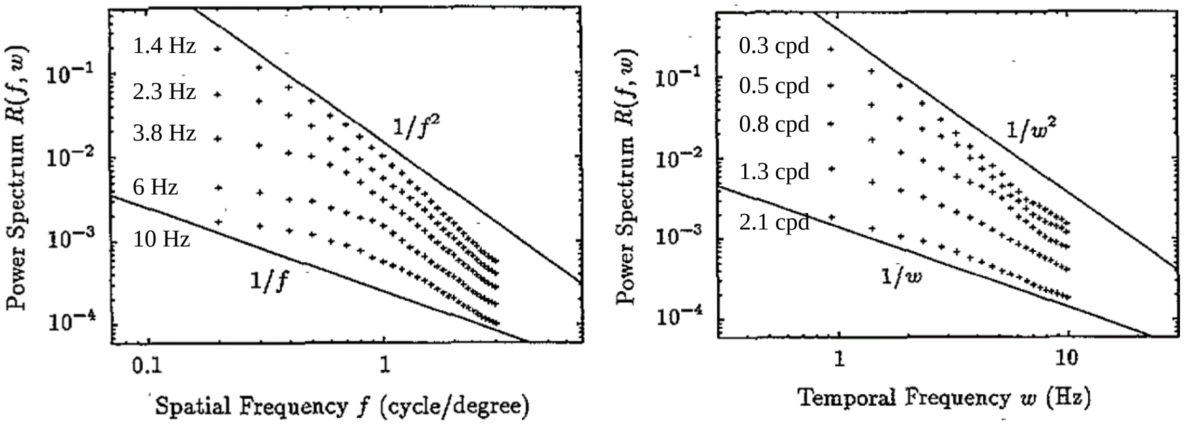

Figure 12. Spatiotemporal power spectrum of natural movies

The spatio-temporal power spectrum was estimated by computing the three-dimensional Fourier transform of short segments from Hollywood and custom-made movies, averaging them together and averaging over orientations. The same power spectrum is plotted in the left panel against spatial frequency (in cycles per degree, cpd) for various temporal frequencies (in Hz), and in the right panel against temporal frequency for various spatial frequencies. Power spectrum slope at high temporal (and spatial) frequencies is shallower than at low temporal (spatial) frequencies.

Adapted from Dong and Atick (1995).

EFFICIENT CODING HYPOTHESIS

Understanding object recognition cannot be achieved by overlooking how the visual system has evolved and developed under the constraints of the environment. The survival of the organism would depend on the ability to quickly identify a potential predator, meal or mate; this ability would in turn be shaped by the neural circuitry at hand. One can then assume that sensorial representations are under strong influence of three fundamental components: the tasks the organism must perform, the metabolic and wiring constraints, and the stimulation it receives (Simoncelli and Olshausen, 2001).

The “efficient coding” hypothesis is a prominent theory in systems neuroscience that states that the visual cortex is optimized to efficiently represent natural scenes. In a seminal paper, Horace Barlow puts forward the ideas that sensory pathways must possess mechanisms for detecting specific relevant stimuli and that “reduction of redundancy is an important principle guiding the organization of sensory messages” (Barlow, 1961; Barlow, 2001). In more concrete terms, this means that neurons are constrained to represent the visual input as compactly as possible, so as to maximize the information they carry about the visual environment: the strategy by which neurons encode sensory information using a small number of active neurons at any given time is often referred to as “sparse coding” (Simoncelli, 2003; Olshausen and Field, 2004). As opposed to “dense codes” in which all neurons simultaneously contribute to represent information, sparse codes are metabolically and computationally efficient (Willmore *et al.*, 2011).

Possible predictions of the sparse coding hypothesis include:

- successively sparser representations should be generated at higher levels of the visual pathway;
- visual cortices should produce sparser representations when presented with natural rather than artificial stimuli;
- neural representations should decorrelate the visual input;
- and neurons are building their tolerances to identity-preserving transformations by exploiting time contiguity of natural scenes.

In the following we will present how these predictions have been approached in the literature, along with a few measures and empirical results.

SPARSENESS

There are many ways in which one can estimate to what extent neural codes are optimal, and these depend on how a constraint on the neuronal code is imposed: on the neural activity of single versus multiple neurons, or alternatively, on the overall activity (i.e. firing rate) versus the shape of the response distribution (Baddeley *et al.*, 1997; Simoncelli and Olshausen, 2001; Willmore and Tolhurst, 2001; Olshausen and Field, 2004; Willmore *et al.*, 2011).

The ***maximization of information*** principle applied on spike trains of single neurons must distinguish between the features that carry the information: the firing rate, i.e. number of spikes over discrete intervals of time, or the interspike intervals. Baddeley and colleagues (Baddeley *et al.*, 1997) discuss three constraints on the neural activity: a maximum firing rate, an average firing rate and the sparseness of the firing distribution. They have shown that when animals are presented with movies of natural scenes, the distributions of firing rates well approximate exponential distributions – which are the maximum entropy distributions for positive values and fixed mean. Mean firing rate constraints are actually likely to exist due to metabolic costs to produce action potentials: the literature estimates an average upper bound of approximately 4 spikes/second in the rodent brain (Baddeley *et al.*, 1997; Attwell and Laughlin, 2001; Willmore *et al.*, 2011).

Laughlin (1981) has shown that neurons in the insect eye, under a maximum firing rate constraint, have response functions that amplify inputs in proportion to their expected frequency so that to achieve a flat response distribution: i.e. they perform a so-called histogram equalization.

The sparseness of the firing distribution is more specifically known as ***lifetime sparseness*** (to indicate that it refers to the stimuli presented to a neuron during its lifetime). This measure quantifies the degree to which a neuron preferentially produces strong responses to certain stimuli, i.e. how peaked its response distribution is. It can be thought of simply as the proportion of stimuli to which neurons respond (Rolls and Tovee, 1995; Vinje and Gallant, 2000): $S = \{1 - [(\sum r_i /n)^2 / \sum (r_i^2/n)]\}/(1 - 1/n)$, where r_i is the firing rate of one neuron to the ith stimulus of in the set of n stimuli. This measure ranges from 0 (responses to all stimuli) to 1 (response to one stimulus). Lifetime sparseness is dependent on the shape of the response distribution, so the mean firing rate doesn't play any role. It results that this is not a measure of metabolic efficiency (Willmore *et al.*, 2011).

One of the prominent application of this metric comes from Vinje and Gallant (2000): in their experiment two monkeys had to fixate the screen for juice reward. During fixation, movies simulating natural vision were played either solely inside the classical receptive fields (CRF) of recorded V1 neurons, or inside the nonclassical receptive fields (nCRF) of the same neurons. Their results show that stimulating the nCRF increases the sparseness from 0.4, when the stimulus is the size of the receptive field, to 0.6, when the stimulus is four times the diameter of the CRF. This suggests that it is the context that sparsifies the responses. Additional evidence for this idea was provided by Froudarakis *et al.* (2014): sparseness was reported to be higher in mice V1 during presentation of full-field natural movies as compared to phase-scrambled artificial movies.

Few studies have systematically compared sparseness in multiple visual areas: Willmore *et al.* (2011) didn't find significant differences of sparseness between V1, V2 and V4 with a stimulus paradigm in which thousands of images were shown to monkeys; similarly, Rust and DiCarlo (2010, 2012) didn't see an increase of sparseness from V4 to IT, and argued that selectivity and tolerance both increase and offset one another to maintain a constant sparse representation in the two areas; Vinken and colleagues (2016) do measure sparseness in three rat visual areas (V1, LI and TO), but don't report any comparison between areas.

Baddeley and colleagues (1997) didn't find evidence for increasing sparseness as they move from the representations in V1 to IT. It's important to note though that the neurons in the two areas were recorded in anaesthetized cats and awake monkeys, respectively, which could make the two datasets incomparable.

Population sparseness takes the formula of the lifetime sparseness, but in this case r_i is the response of neuron i to a given stimulus (Treves and Rolls, 1991; Willmore and Tolhurst, 2001). Due to the similarity to the lifetime sparseness, the properties described before are also valid here. The two measures of sparseness are not necessarily related, i.e. high lifetime sparseness does not imply high population sparseness, unless the corresponding population responses are independent and identically distributed (Willmore *et al.*, 2011). Froudarakis and colleagues (2014) have shown that in large populations of simultaneously recorded neurons, population sparseness was higher for natural movies than for phase-scrambled ones, and additionally that the two measures were highly correlated.

Kurtosis is the fourth standardized moment of a distribution and reflects its "tailedness" (i.e. how heavy the tails are): $k = \frac{1}{n}\sum_{i=1}^{n}\frac{(r_i-\bar{r})^4}{\sigma^4} - 3$, where $\bar{r}$ is the mean firing rate and σ is the standard deviation of the distribution.

In practice this measure is difficult to use with biological neurons, since it requires that the distribution be unimodal, symmetrical (about zero), and with negative tails (Willmore and Tolhurst, 2001; Olshausen and Field, 2004). Vinje and Gallant (2000) measured kurtosis in population response distributions obtained by pooling responses of all neurons to all movie frames (experiment described above). To circumvent the issues mentioned with this metric, they reflected the distributions about the origin and computed kurtosis on the resulting symmetric distribution.

SLOWNESS

The interaction between the visual cortex and the environment happens at multiple timescales, which is an important aspect of the efficient coding hypothesis. It is reasonable to assume that the properties of natural images have remained unchanged for millennia so their effects on the visual system have been hardwired into the organism. Concurrently, alterations at medium and short timescales can be seen during neural development or neural adaptation (Simoncelli and Olshausen, 2001).

As noted before, one of the statistical regularities of natural images is the persistence in time (or slowness) of patterns with common structure. To better illustrate the slowness principle, let's consider the sequence of frames in Figure 13. As a response to the moving object, neurons in the primary visual area will be heavily driven by fluctuations of pixel intensities within their receptive fields. Since the brain needs to extract behaviorally relevant features from the primary sensory signal, such as the identity or location of the object, one can assume that a higher-level cortical area will utilize the persistence and contiguity of objects in the world to form an invariant representation of the natural scene (Hyvarinen *et al.*, 2009; Wiskott *et al.*, 2011). Formally, such an approach belongs to the class of "unsupervised temporal learning" algorithms (DiCarlo *et al.*, 2012).

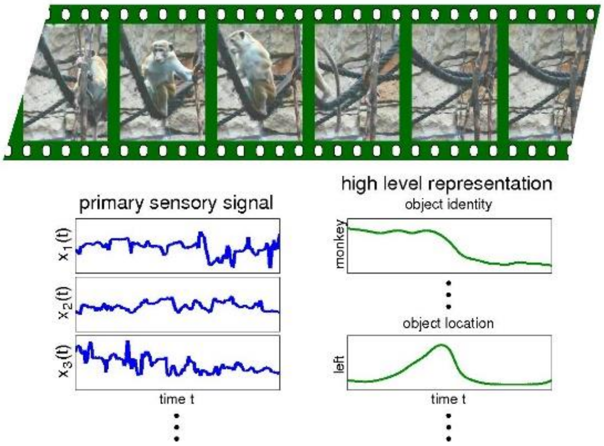

Figure 13. Toy example of the slowness principle

The sequence of frames (top panel) depicts a monkey that stands on a suspended rope and then leaves the field of view to the left. Neurons in the primary visual cortex, with small receptive fields and selective for low-level features (e.g. pixel intensity), will have a highly varying activity in response to the moving monkey (bottom-left panel). Neurons in a higher area, such as IT, will form a high level representation: if they code for object identity, their firing will be maximal and sustained in the first frames while the monkey is present in the scene; location neurons instead will become active when the monkey is in the center (bottom-right panel).
Taken from Wiskott *et al.* (2011).

Computational work has suggested possible implementations of unsupervised temporal learning that result in invariant representations. Földiák (1991) was the first to show that by adding a decay term to a Hebbian rule, a network of simple-cell-like neurons can learn position invariance from temporal patterns undergoing position transformations (shifting oriented bars). This mechanism could explain the emergence of position invariance of complex cells in the primary visual cortex.

Another algorithm that is employing temporal slowness is Slow Feature Analysis, which is able to extract slowly varying components from quickly varying input signals (Wiskott and Sejnowski, 2002; Wiskott *et al.*, 2011). This method had various applications in computational neuroscience: for example it can learn independent representations of object position, angle and identity from quasi-natural stimuli (Franzius *et al.*, 2008), or produce a set of functions that match properties of complex cells in V1 (Berkes and Wiskott, 2005).

There is scarce empirical evidence supporting the efficient coding theory. One of its predictions states that visual areas should recode the visual input into a whitened (decorrelated) form. Dan and colleagues (1996) have addressed this prediction in a straightforward way. They recorded responses of individual neurons in cat lateral geniculate nucleus during presentation of full-field natural and white-noise movies. The autocorrelation of the responses to natural stimuli showed narrow peaks at 0 msec lags and flat amplitudes beyond the peak (which indicates that LGN output was temporally decorrelated), and the corresponding power spectra were predominantly flat between 3 and 15 Hz (consistent with the theoretical prediction that the natural visual signals should be whitened, Figure 14A).

Their controls, the white-noise stimuli, showed autocorrelations with small dips at lags between 10 and 100 msec, which is reflected in the positive slopes between 10 and 15 Hz in the frequency domain. The increase in power at higher frequencies in the case of white noise stimuli shows that LGN neurons are not simply whitening any stimulus, but only those with low high frequency spectrums such as natural images (compare panels A and B in Figure 14).

These results are consistent with the theoretical predictions: i.e., natural images are selectively decorrelated at early processing stages. A similar conclusion was reached by Vinje and Gallant (2000), in the experiment described before: firing activity of pairs of neurons were less correlated during stimulation within their non-classical receptive fields, than within classical receptive fields.

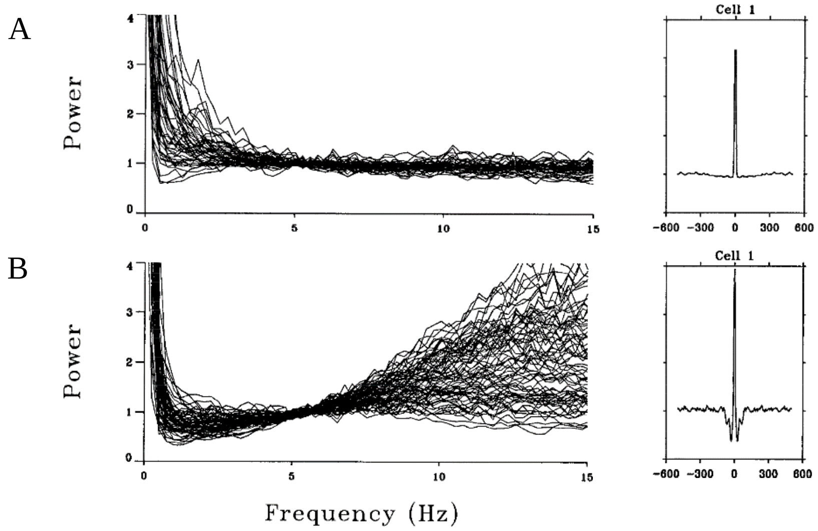

Figure 14. Power spectra of LGN responses

A. Power spectra in response to natural movies. Each trace corresponds to the response of one neuron and, for clarity, is normalized by its value at 5-6 Hz. The box shows the autocorrelation function of the spike trains of one neuron.

B. Power spectra in response to full-field white noise movies. The box shows the autocorrelation function of the same neuron as in panel A.

A series of behavioral and physiological experiments in the group of DiCarlo explored the hypothesis that the ventral visual stream could construct a tolerant object representation by taking advantage of objects' temporal contiguity in the visual input (Cox *et al.*, 2005; Li and DiCarlo, 2008, 2010). In the physiological studies, neural activity was recorded from monkeys involved in simple tasks that allowed them to freely look at a monitor on which two objects appeared intermittently below or above the gaze point. For each recorded neuron objects were selected so that one would induce a strong response (preferred object), and the other one a moderate response (non-preferred object). At each image change, the monkeys reflexively foveated the newly appeared objects. By means of a real-time eye-tracker, the experimenters were able to change the identity of the object at desired trials while the animal was saccading toward the object (swap exposure, Figure 15A).

By consistently swapping the two objects' identities at either location (below or above gaze point), they show that IT neurons changed their position tolerance (by lowering their firing rate) to the initially preferred object in a relatively short interval (Figure 15B). Additional experiments extended these results for size and identity tolerance, and showed that temporal direction of experience is relevant to attain the effect (Li and DiCarlo, 2010).

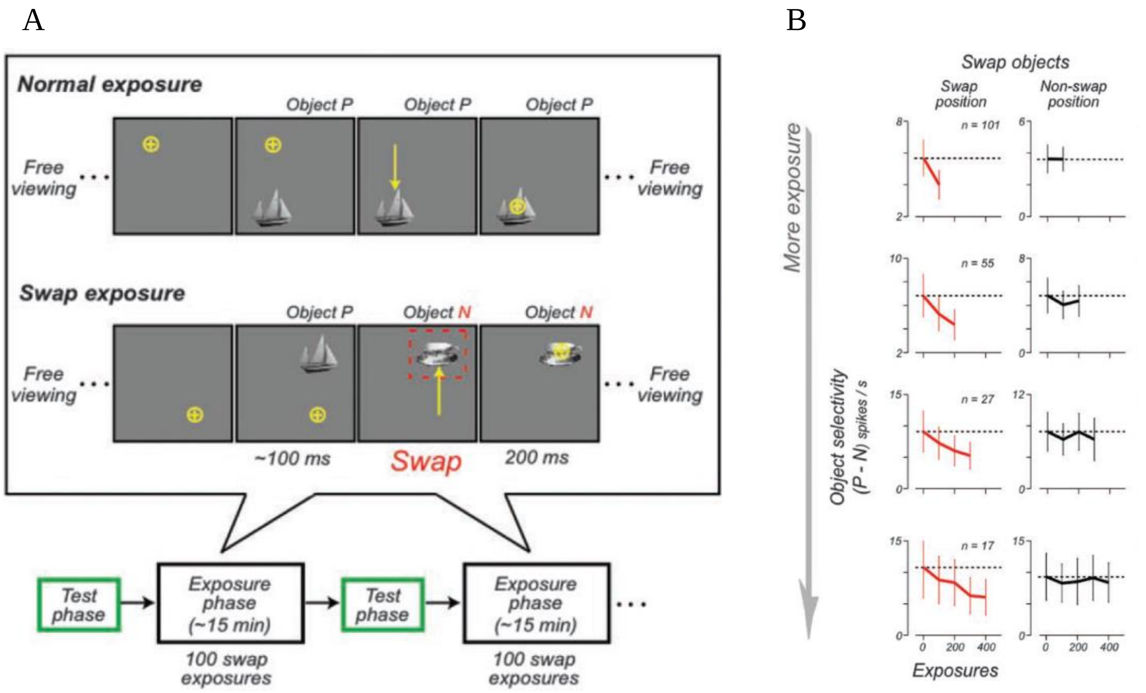

Figure 15. Breaking position tolerance in monkey visual cortex

A. The experiment consisted of a sequence of two alternating phases: test and exposure (bottom row). During these phases one object was presented at either 3° below or above the center of gaze. In the test phase, the position tolerance of the preferred (P) and nonpreferred (P) objects (determined beforehand from a set of stimuli) was measured as the response of the neuron to these objects at the two locations. For the exposure phase, for each neuron a swap position (above or below the center of gaze) is chosen and kept fixed throughout the recording. In the swap position, the object P is maintained while the monkey is saccading (normal exposure, top row), or changed with object N (bottom row, swap exposure).

B. The data reported in these plots were gathered only during test phases. The object selectivity for each neuron is calculated as the difference in its response to object P and N. From top to bottom, each row indicates more exposure phases undergone by surviving neurons. The decreasing trend at the swap position (red traces) indicates that neural responses to object P become smaller as object N replaces P at that position, therefore altering the position tolerance. At the non-swap position objects preserved identity so neurons maintained their selectivity (constant black traces).

Taken from Li and DiCarlo (2008).

These manipulations show that altering the contiguity of visual experience can build artificial neural invariance, therefore confirming the starting hypothesis.

2. METHODS

SURGICAL PROCEDURE

Male Long Evans rats (weighting 350-550 g) were anesthetized with an intraperitoneal injection of a solution of medetomidine (Dormitor: 0,3 mg/kg) and fentanyl (Fentanest: 0,3 mg/kg). Body temperature was kept constant at ~37° with a warming pad. A constant flow of oxygen was delivered to the rat to prevent hypoxia, and oxygen saturation was monitored with a pulse oximeter attached to one of the hind paws. The rat was placed in a stereotaxic apparatus.

After loss of paw and tail reflexes, an incision was performed on the scalp over the left and posterior parietal bones, and then a cranial window (approx. 2×2mm wide) was created at the coordinates intended for recording (left hemisphere). Dura was removed to ease the insertion of the electrode. Throughout the experiment, the exposed brain was irrigated with saline solution to prevent drying. The right eye (contralateral to the hemisphere from which we recorded) was immobilized with an eye ring anchored to the stereotaxic apparatus and covered with ophthalmic solution Epigel (Ceva Vetem) to prevent drying; the left eye was covered with black tape.

Once the surgical procedure was completed, the stereotaxic apparatus was moved on an elevated rotating platform. The rat was maintained under anesthesia by a continuous intraperitoneal flow of the fentanyl and medetomidine solution (0,1 mg/kg/h).

Before penetration, the probe was coated with Vybrant DiI cell-labeling solution (Invitrogen). This dye was used to recover the location of probe insertion through histological procedures.

Recordings were performed with 32-channel silicon probes (Neuronexus Technologies) in various configurations (example Figure 16A). In order to maximize the receptive field coverage, recordings in V1 were performed with 4- or 8-shanks probes, which were inserted perpendicularly on the cortex. For lateral areas, one shank probes were used, which were inserted along the cortex, in order to be able to observe the reversal of retinotopy between successive, adjacent areas (Figure 16B and C). The space between recording sites on each shank ranged from 25 to 200µm; the distance between shanks (when more than one) was 200µm; the surface of recording sites (and as a consequence the site impedance) was either 177 or 413 µm^2.

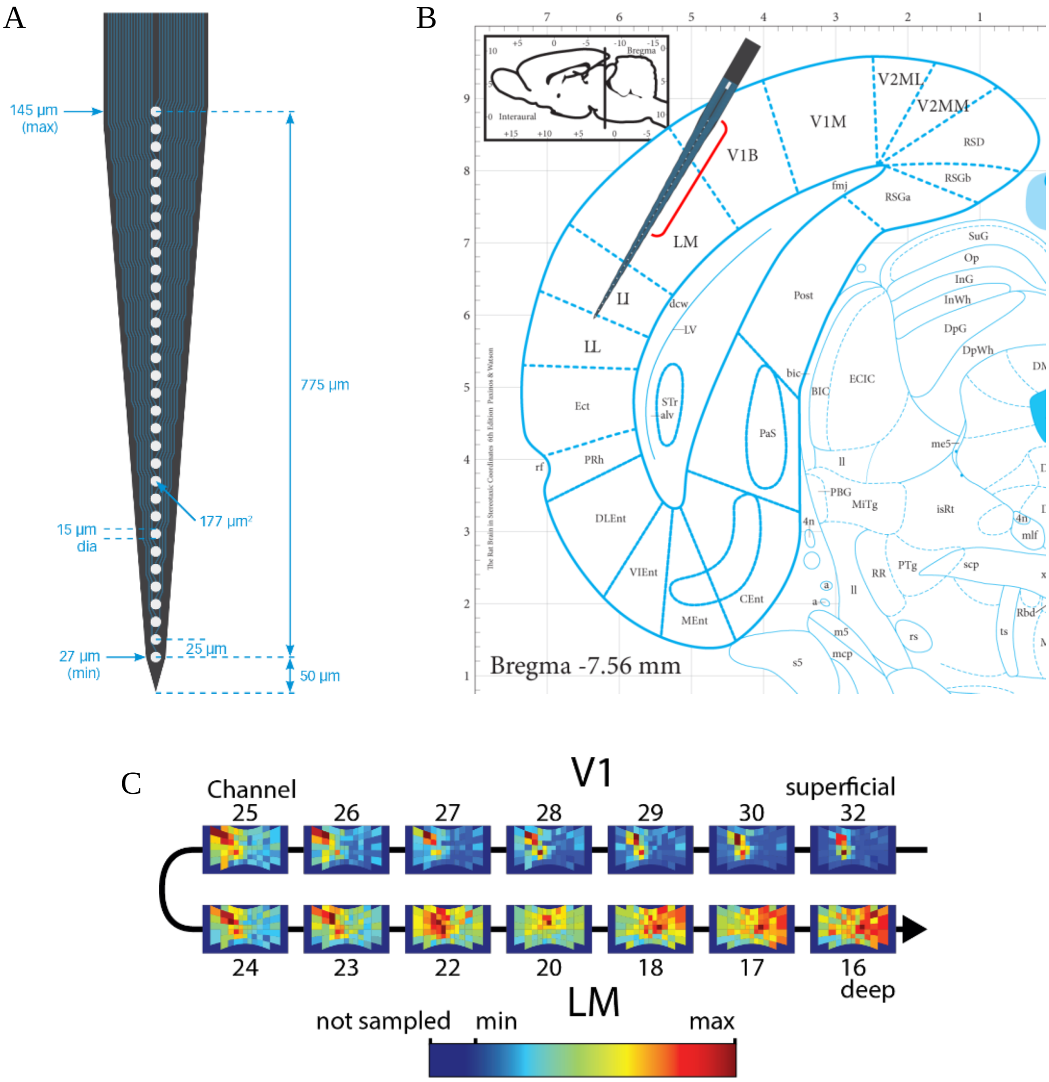

Figure 16. Electrophysiological recording and the reversal of retinotopy

A. Example linear electrode used for lateral penetrations: in this configuration, the recording sites are placed 25µm apart, which facilitates recording from small areas, such as LI and LL.

B. Example penetration for recordings from LM, LI and LL. The probe is inserted at mediolateral position 4 and bregma -7.56 mm with a tilt of 30°. This probe configuration (recording sites at 100 µm apart) allow for simultaneous recordings from multiple areas.

C. The reversal of retinotopy between areas V1 and LM). We note that the receptive fields in channels 32-25 move leftwards, whereas those in channels 24 to 16 move rightwards (recording sites from 16 to 32 are highlighted with red in panel B).
Panel B adapted from Paxinos and Watson (2007).

Extracellular signal was acquired with a TDT system 3 workstation (Tucker-Davis Technologies) at a sampling rate of ~24 kHz. Action potentials (spikes) were detected online by thresholding the bandpass filtered signal (0.3-3 kHz) and were used for data monitoring and the estimation of receptive fields at each recording site (detailed below).

At the end of a recording session, an electrolytic lesion was performed by passing a 5µA current for 2s through 2-4 channels at the tip of the probe.

STIMULI

Stimuli were presented full-field to the right eye at a distance of 30 cm on a 47-inch LCD monitor (SHARP PN-E471R), with 1920 × 1080 resolution, 60 Hz refresh rate, 9 ms response time, 700 cd/m^2 maximum brightness, 1200:1 contrast ratio, spanning a visual field of ~120° azimuth and ~89° elevation.

The rotating platform on which the rat was placed was adjusted so as to center the rat’s right eye over the platform’s rotation axis and align it to the normal from the center of the monitor. The rat was rotated before each recording session in a position that would optimize recorded neurons’ receptive field coverage on the monitor (typically 45° leftwards); its position was adjusted every time the visual fields of the recorded neurons were outside of the screen.

A full recording protocol lasted approximately 2h 15min and consisted of receptive field mapping and movies, which were the main stimuli of the experiment (Figure 17). All stimuli were shown with Psychtoolbox (Kleiner *et al.*, 2007) in a pseudorandom order: movies were repeated 30 times (number of trials) and the receptive field mapping 7 times. Between each stimulus (movie or moving bar) a black screen was shown for 200 ms, except before some movies, because they took up to 4 seconds to load. The activity during these 4 sec-long periods was included in all analyses regarding the spontaneous activity.

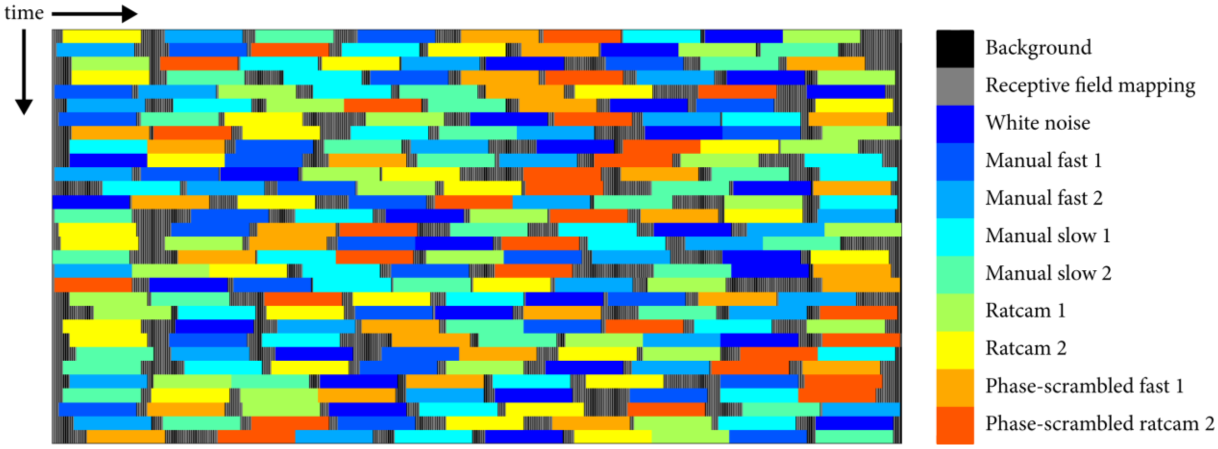

Figure 17. Stimulus protocol

Stimuli were shown in a pseudorandom order during a recording session. This order was frozen across sessions.

Colored horizontal bar represent movies, which were repeated 30 times (rows in the image). Gray stripes indicate short periods of receptive field mapping consisting of drifting moving bars (~0.36 seconds, described below). Black stripes indicate periods when no stimulus was shown, during which the screen was black. The 4 sec intervals in which some movies were loaded are not shown.

RECEPTIVE FIELD MAPPING

Receptive field estimation was obtained by showing moving oriented bars at 73 locations on the screen, covering the central 110° azimuth and central-upper 70° elevation of the total visual field coverage of the monitor (Figure 18A). This configuration was chosen so as to optimize the coverage of the screen and the duration of the receptive field protocol.

Bars measuring 10°×10° were shown translating along four differently oriented axes of movement (horizontal, vertical and two diagonals) from -5° to 5° and back again to -5° with respect to the location point. An example bar at 0° orientation is shown in Figure 18B.

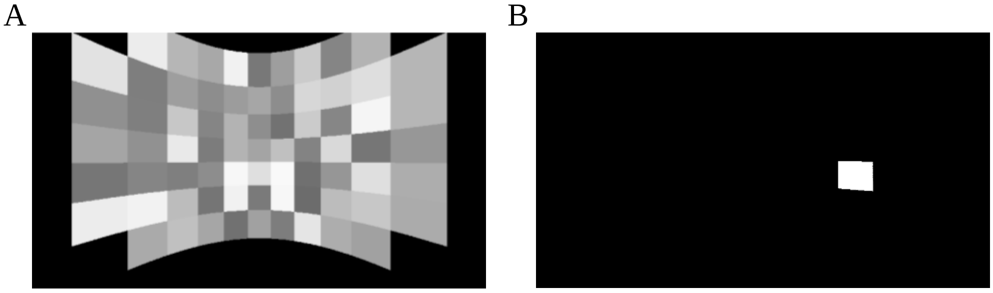

Figure 18. Receptive field mapping

A. All 73 locations covering 110° along the azimuth axis and approximately 70° along the elevation axis. Bar intensity at each location is randomly assigned to illustrate its borders.

B. Example frame of a moving bar.

As mentioned above, the rat's eye was placed at 30 cm from the screen, a choice motivated by behavioral results showing that rats are able to perform complex object recognition at this distance (e.g. Tafazoli *et al.*, 2012). Because of the short viewing distance, stimuli presented at large eccentricities would look distorted compared to stimuli presented directly in front of the eye. To account for this distortion, bars were shown under a spherical projection (described below), so that they appear as if they were displayed at equal distance from the rat's eye (Marshel *et al.*, 2011; labrigger.com, 2012). The procedure basically consists of finding the projection over the monitor of a point belonging to a sphere.

In our three-dimensional coordinate system (Figure 19), we defined the eye as the origin O: the z axis was set to pass through the center of the monitor, and the axes x (width) and y (height) lay on the plane parallel with the monitor, which was at distance R. Therefore, azimuth is the angle θ from the $z-y$ plane, and elevation the angle φ from the $z-x$ plane (Figure 19).

The projection $P(x,y,z)$ on the monitor of any point $Q(R,\theta,\varphi)$ belonging to the bar at location $B(R,\theta_0,\varphi_0)$ on a sphere of radius R centered over the eye was computed as follows:

1) $Q(R,\theta,\varphi)$ was converted from spherical to Cartesian coordinates

$$x_Q = R\cos\varphi\sin\theta$$

$$y_Q = R\sin\varphi$$

$$z_Q = R\cos\varphi\cos\theta$$

2) point Q was rotated about OB by the orientation angle of the bar (0°, 45°, 90°, or 135°, the green angle in Figure 19)(code adapted from Bourke, 1992)

$$Q'(x'_Q, y'_Q, z'_Q) = rot(Q(x_Q, y_Q, z_Q), OB)$$

3) Q' was converted back to spherical coordinates

$$\theta' = \tan^{-1}\frac{x'_Q}{z'_Q}$$

$$\varphi' = \tan^{-1}\frac{y'_Q}{\sqrt{{x'_Q}^2 + {z'_Q}^2}}$$

4) In order to find the position of point P on the plane of the screen, we extended OQ' until it intersected the plane at distance R', which was calculated as:

$$R' = \frac{R}{\cos\theta'\sin\varphi'}$$

5) The Cartesian coordinates of point $P(R',\theta',\varphi')$ were then given by

$$x = R'\cos\varphi'\sin\theta'$$

$$y = R' \sin \varphi'$$
$$z = R' \cos \varphi' \cos \theta' = R$$

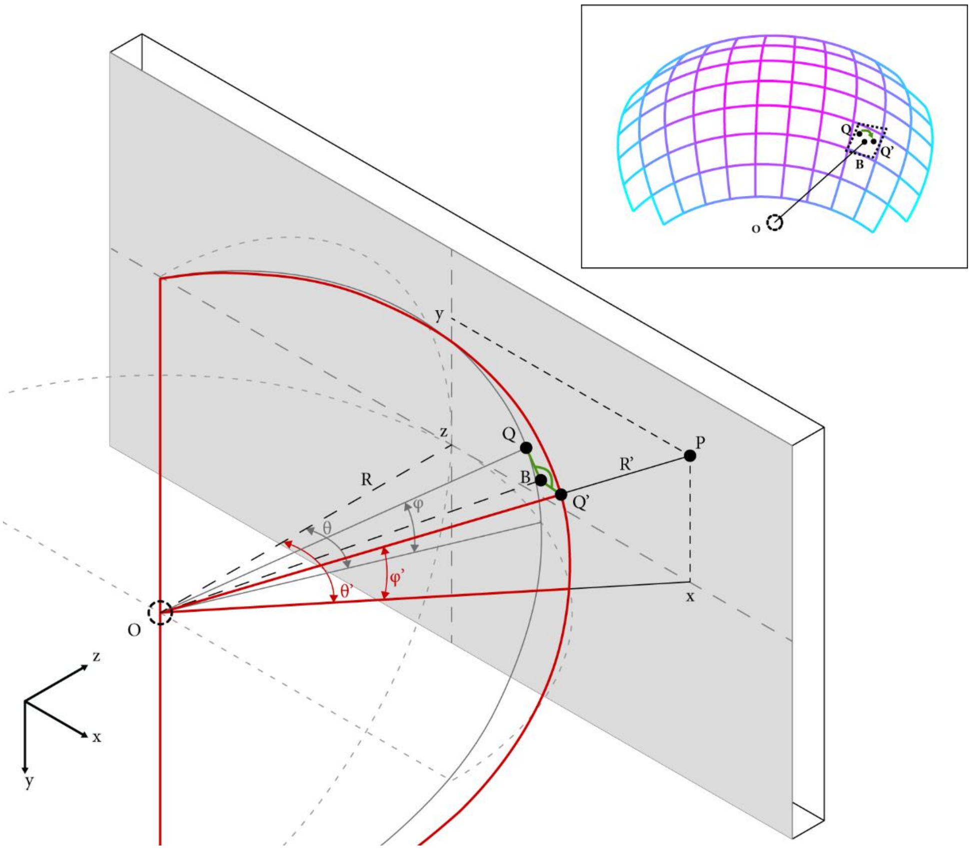

Figure 19. Spherical projection for receptive field mapping

O represents the eye, R is the distance from the eye to center of the screen, B is the location of each bar given as azimuth and elevation angles, Q is a point within the bar B, Q' is the rotation (green angle) of point Q about axis OB, P corresponds to Q on the plane of the screen ($x - y$), R' is the distance between the eye and point P. The **inset** shows again all divisions of Figure 18A and the rotated bar (dashed rectangle).

Note that the reference plane is the one containing the eye (see the axes), and the azimuth and elevation angles are defined differently from the standard convention.

A full receptive field mapping protocol consisted of 292 short movies (73 locations × 4 orientations), 0.363 seconds long (11 frames), which were always followed by a 0.2 seconds interval of black screen.

MOVIES

The main stimulus set consisted of: 2 fast natural movies, 2 slow natural movies, 2 ratcam movies, 1 phase-scrambled version of one fast movie, 1 phase-scrambled version of a ratcam movie (Figure 20). These movies were each 720x1280 pixels, 20 seconds long and were presented at 30 fps (600 frames per movie). An additional white-noise movie 45x80 pixels was shown at 20 fps for 20 seconds (400 frames).

All stimuli were converted to grayscale and were gamma-corrected offline with a lookup table calculated for the monitor used for stimulus presentation.

The videos can be watched at the following link:

youtube.com/playlist?list=PLR5haZHnCJ-gsvtlx-V2nsyhA-89tYiYx

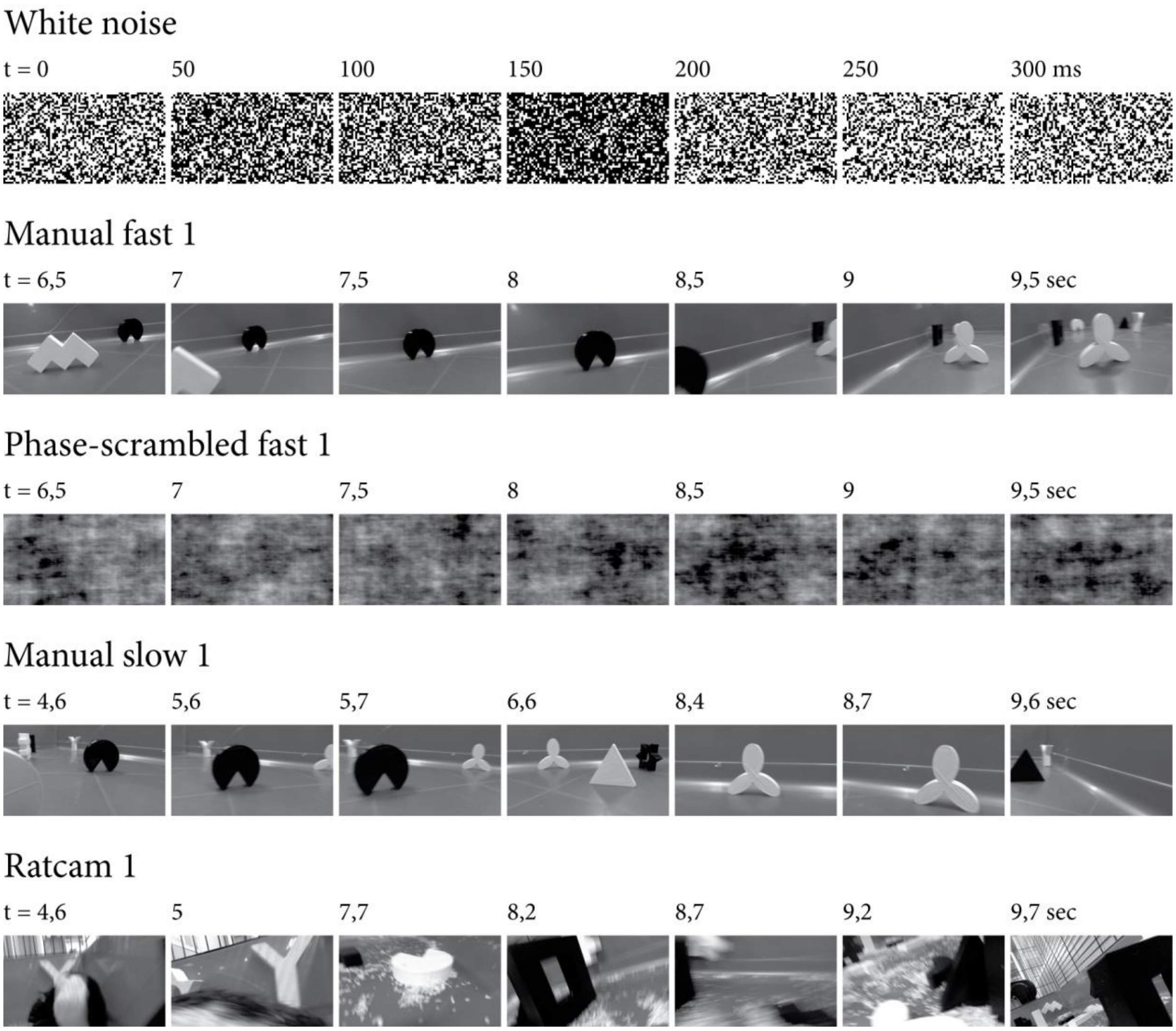

Figure 20. Movies

Example frames for one of each movie type along with the time of each frame within the movie. The frames of the movie Phase-scrambled fast 1 correspond to those of Manual fast 1. Manual movies fast 1 and 2, slow 1 and 2 had similar appearances, but varied in the speed objects moved on screen. Ratcam 1 and 2 perceptually looked indistinguishable. The frames of the movies phase-scrambled fast 1 and phase-scrambled rat 2 looked similar, but varied in their apparent speed (see also Figure 22).

Note that not all frames are equally spaced (for example the first 3 frames of manual slow 1, that illustrate the trajectory of an object on screen).

Natural movies were designed to show solid objects undergoing various transformations (e.g. translation, size and lighting change) in a relatively controlled way. To this aim, we filmed 3D-printed objects painted black and white in an arena (~0.6 m^2) of uniform blue color (which turned into mid gray when the movie was converted to grayscale). The shapes of the objects were selected based on previous reports in monkey (Sereno and Maunsell, 1998) and rat (Vermaercke *et al.*, 2014).

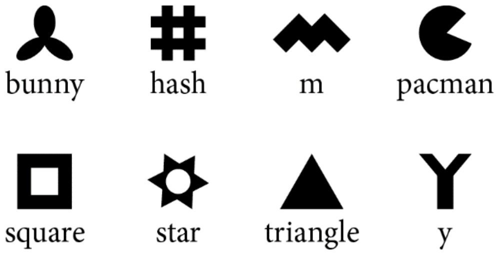

Figure 21. Objects' shapes

The objects used to create the movies. These shapes were 3D-printed with thickness of 3 cm, then spray painted in either black or white to obtain 16 objects in total.

The 3D objects were randomly placed in the arena, but in such a way to allow filming with a hand-held camera a continuous movie that included most objects. The two fast movies included close-up views of all 16 objects; the two slow movies only included a subset of them.

Ratcam movies were intended to simulate the visual input of a rat exploring a natural environment (Froudarakis *et al.*, 2014). They were obtained by placing a small modified web camera (Microsoft Lifecam Cinema HD) on the head of a rat while it was running freely inside the arena that contained some of the printed objects and another rat. To prevent shaky camera effects, we added extra weight on the head. Therefore, the camera captured slow body and head movements, but not eye movements.

Phase-scrambled movies preserve first- and second-order properties of the original movies, but destroy higher-order correlations (see Introduction). Previous reports have described how phase-scrambled images and movies can be obtained from monochromatic or color movies (Fraedrich *et al.*, 2010; Froudarakis *et al.*, 2014; Vinken *et al.*, 2014). Our method was based on the following steps:

1) video frame data (x, y and time) were normalized to mean 0 and standard deviation 1 to obtain M;
2) we computed the spatiotemporal fast Fourier transform (FFT) $Ae^{-i\varphi}$ over M from which we extracted the phase φ and amplitude A spectrums;
3) we saved the indices of identical phases from φ across the three dimensions – so that to preserve the conjugate symmetry of the FFT matrix –, then scrambled the phases and populated the matrix at matching indices to obtain a new phase spectrum φ^*;
4) the original amplitude A and the new phase spectrum φ^* were combined into the matrix M^* by performing an inverse FFT over $Ae^{i\varphi^*}$;
5) the imaginary part of M^* due to round-off errors was discarded, then the original pixel intensity distribution was restored;
6) values outside the 0-255 range were clipped.

In order to establish how similar natural and phase-scrambled movies are at pixel level, we computed four indices from the pixel intensities (I_{xy}) of each image, i.e., across columns $x = 1 \ldots X$ and rows $y = 1 \ldots Y$.

Average intensity contained the average pixel intensity in each frame:

$$\bar{I}(t) = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y} I_{xy}(t)$$

Root-mean-squared contrast (RMS) is defined as the standard deviation of the pixel intensities in each frame:

$$RMS(t) = \sqrt{\frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}\left(I_{xy}(t) - \bar{I}_{xy}(t)\right)^2}$$

Michelson contrast was calculated as follows:

$$MC(t) = \frac{I(t)_{max} - I(t)_{min}}{I(t)_{max} + I(t)_{min}}$$

Time contrast (or pixel change) measures how similar two consecutive frames are, and was defined per frame transition, $t = 1 \ldots T - 1$, as:

$$\overline{PC}(t) = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}\left|I_{xy}(t+1) - I_{xy}(t)\right|$$

The original movies and their corresponding phase-scrambled version had virtually identical average intensity (Figure 22) and RMS contrast when averaged across frames, and slightly different average Michelson contrast.

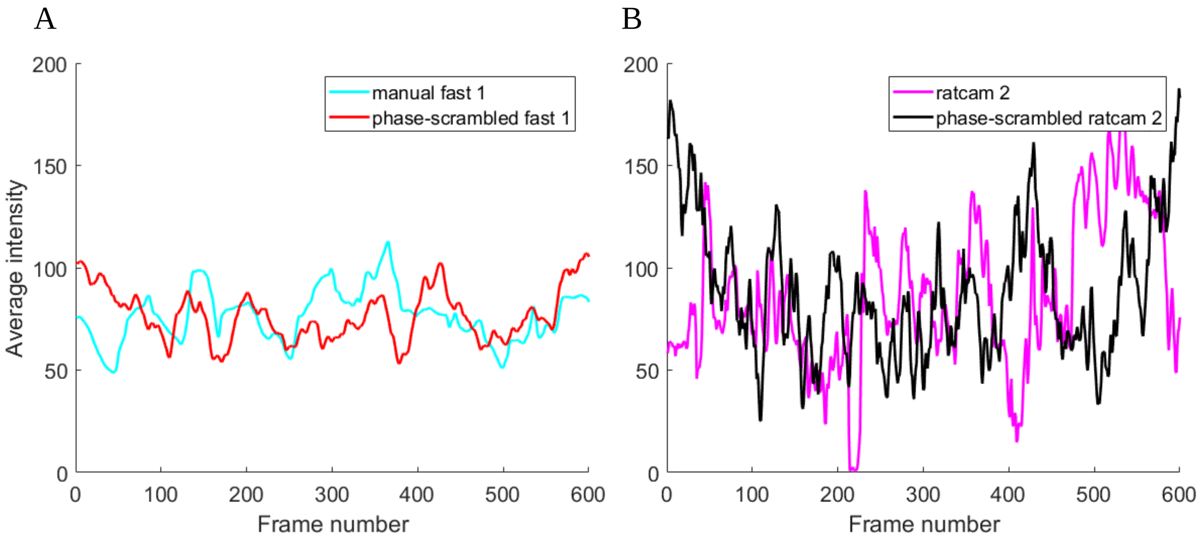

Figure 22. Average intensities of natural and phase-scrambled movies

Average pixel intensity of each frame is shown as a function of frame number. Horizontal lines are grand averages across all frames.

A. One natural fast movie and its phase-scrambled version.

B. One rat movie and its phase-scrambled version.

Average pixel change was also very well preserved (Figure 23), even though this measure varied less from frame to frame in phase-scrambled movies, which is perhaps an indication of their faint flickering appearance.

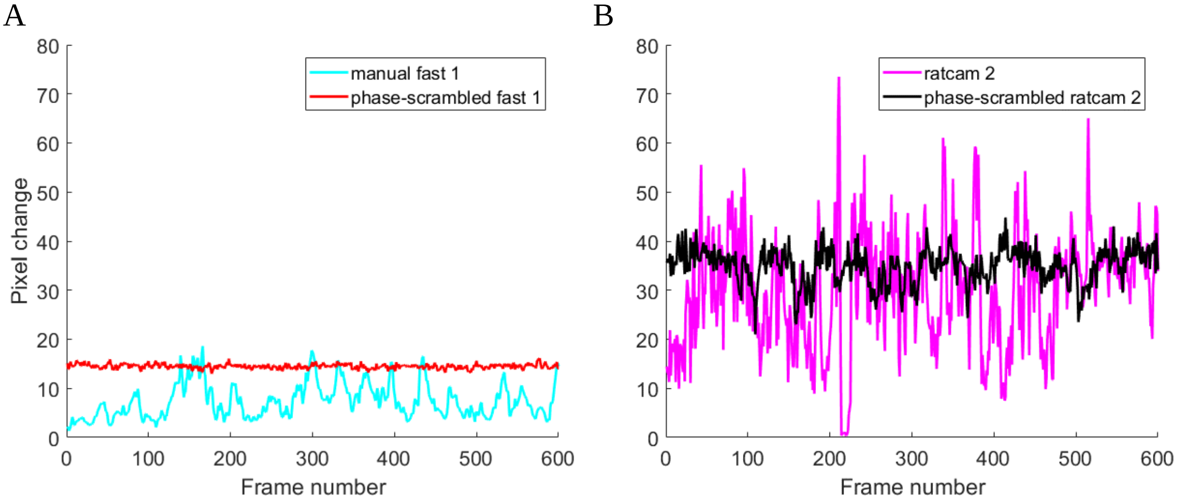

Figure 23. Pixel change

Pixel change shows how similar two consecutive frames are as a function of frame number.

Same notations as in Figure 22.

The **white noise movie** was obtained by drawing values from a standard normal distribution, normalizing and then binarizing them to 0 or 255. The frame size (height 45, width 80 pixels) was chosen to match the rat visual acuity (~1 cpd, see Introduction) at the center of the screen: given that the screen spanned 90×120 degrees of visual angle, one degree covered approximately 2 pixels (i.e. one cycle).

DATA ANALYSIS

After the recording session, raw data were spike-sorted offline with WaveClus (Quiroga *et al.*, 2004) or KlustaKwik (Rossant *et al.*, 2016), in two phases: the first one was an automatic process that separated the spikes in distinctive clusters, the second one consisted of a manual refinement of the clusters generated in the automatic phase. The spike sorting procedure produced for each of the 32 recording sites (or *channels*) a number of *units* (typically 1-4) that represented our pool of available neurons.

All analyses were performed with custom-written MATLAB (Mathworks) code, unless otherwise specified. Some statistical analyses were performed with SPSS Statistics (IBM Corporation).

PREPROCESSING

The duration of each movie (20 sec) was divided in equal time bins of 33 ms for natural and phase-scrambled movies, and of 50 ms for white-noise movies, which resulted in 600 and 400 bins respectively. In the population decoding part, some analyses were performed with bins which were 100 ms long. A time histogram was obtained for each movie repetition (or trial) by counting the number of spikes in each time bin (*spike counts*). These histograms were then averaged across all trials to obtain the peristimulus time histogram (PSTH), which shows the average activity of a neuron during the presentation of one movie (Figure 24A,B).

Because of the sudden and large change of luminance at the onset of each movie, the first ~250ms of the responses (5 bins for white-noise movies, and 7 bins for all other movies) were discarded from all analyses to avoid luminance-driven effects.

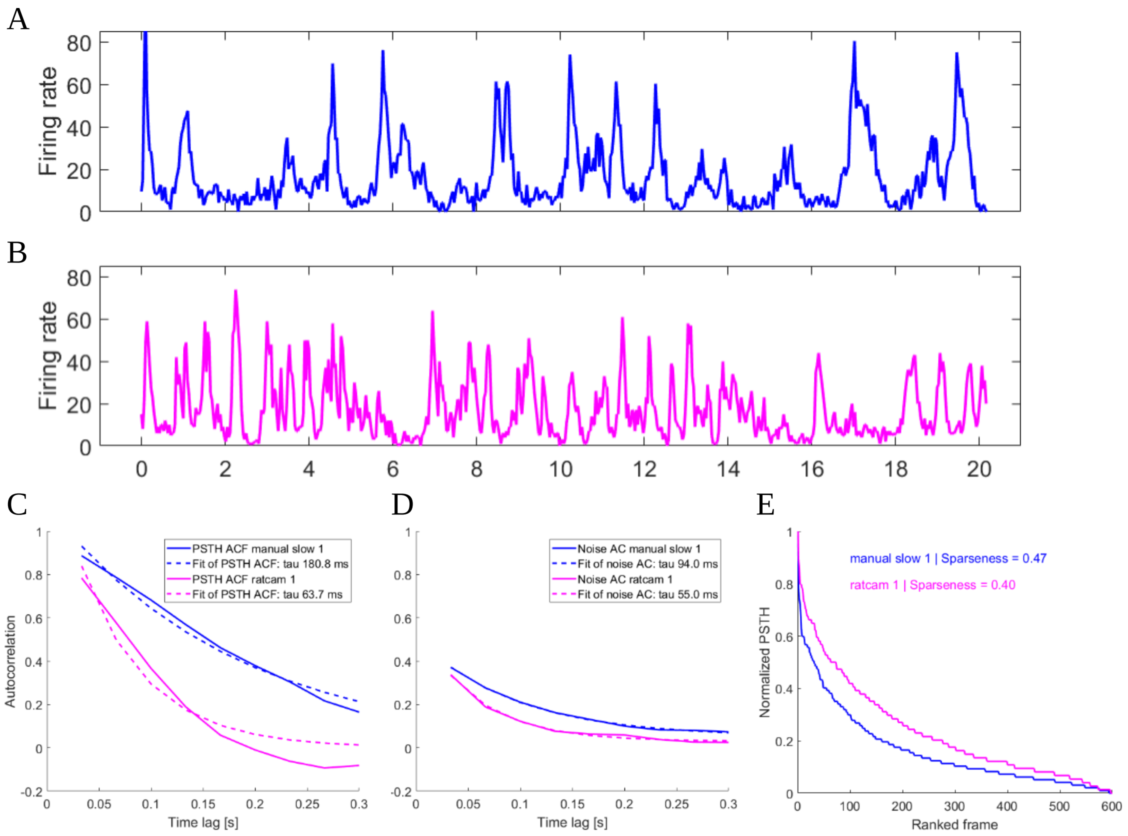

Figure 24. Sparseness, PSTH autocorrelation and noise autocorrelation

This figure refers to the responses of one V1 neuron to two movies. The spike counts from which these plots are obtained are shown in Figure 26 and in section Example responses (Figures 31 and 37).

A. The average firing rate (PSTH) in response to movie manual slow 1.

B. The PSTH of the same neuron in response to movie ratcam 1.

C. The autocorrelation computed from the signals in **A** and **B**, along with their exponential fits and the time constants of the decay (termed slowness). We observe that the autocorrelation for the faster movie (ratcam 1) has a steeper decay than the one for the slower movie, and the time constant are in consequence smaller.

D. The noise autocorrelation computed from the spike counts from which **A** and **B** are obtained, along with their exponential fits and time constants.

E. Sparseness and slowness rely on the same time series so they are strongly interdependent: we can appreciate this by sorting the PSTH and plotting the same values against their rank in the ordered list. The two traces are obtained from the PSTHs in **A** and **B** which have been normalized. High sparseness is associated with narrower ranked PSTHs, and low sparseness with wider ranked PSTHs (the two extremes are pulse-like and step-like functions, respectively).

SELECTION OF NEURONS

Data analysis was restricted to a selection of single and multi units taken from the pool of available neurons, that met our criterion of reproducibility, which is a metric that quantifies how reliable neurons respond to preferred stimuli, i.e. frames that elicit the strongest responses. In more detail, we rank ordered the PSTH for each movie and selected the highest 10% peaks (60 and 40 frames, respectively, for the 33 and 50 ms bins). For each time bin corresponding to the selected peaks, we calculated the coefficient of variation (CV) of the spike counts across trials (i.e. 30 values), and then averaged and normalized the CVs:

$$reproducibility = 1 - \frac{\langle CV_{peaks} \rangle}{maxCV}$$

where $maxCV$ is the maximum value the coefficient of variation can take for a given sample size. It can be shown (whuber, 2011) that the maximum value is reached for a sample that contains only one nonzero and positive element, and all other elements zero, and is equal to the square root of the size of the sample (for a sample of 30 elements, $maxCV = \sqrt{30} = \sim 5.47$).

The resulting metric ranges from 0 to 1, where 1 corresponds to neural responses with perfectly reproducible trials. This metric was chosen over others that rely on the coefficient of correlation between trials (Rikhye and Sur, 2015; Vinken *et al.*, 2016), to avoid the inclusion of silent neurons which would have had high trial-by-trial correlations due to lack of activity.

After visually inspecting the responses of many units in the original dataset (2571 units), we decided to set the inclusion threshold at 0.7, that is to include those neurons which produced for at least one movie a reproducibility index greater or equal than 0.7. By doing so, we restricted our population to 548 responsive and reliable units recorded from 22 animals in 29 sessions (of which 8 sessions included units recorded from two areas):

- 291 in V1 collected over 9 sessions;
- 49 in LM collected over 4 sessions;
- 95 in LI collected over 12 sessions;
- 113 in LL collected over 12 sessions.

In the following, single and multi units will be called neurons.

RECEPTIVE FIELD ESTIMATION

The receptive field position and size for each neuron were calculated from its firing rate in response to drifting bars at 73 positions on the screen and 4 orientations (see Methods, section

Receptive field mapping). For each drifting bar, spikes were counted in windows that matched the duration of the bar on screen (~0.38 sec). At each location, the responses to the four oriented bars were averaged, and the resulting maps were fitted with a two-dimensional Gaussian (Curve Fitting Toolbox, MATLAB). The fitting produced five coefficients: the coordinates of the center, the x and y standard deviations, and the orientation angle θ. The receptive field was defined as the area encompassed by an ellipse whose center coincides with the center of the Gaussian, and whose radii are equal to one standard deviation along the two axes of the Gaussian rotated by angle θ (Figure 25). Each fitting procedure also provided a goodness of fit index ranging from 0 to 1, that was used to establish the reliability of the ellipse in describing the shape and orientation of the actual receptive field.

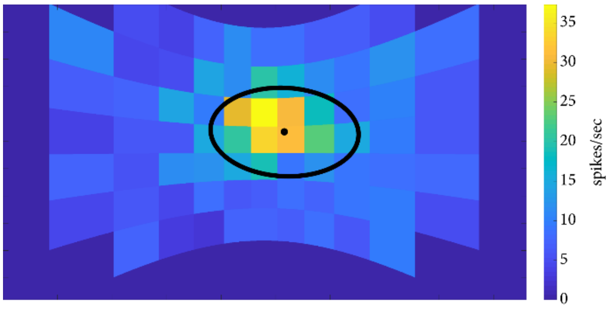

Figure 25. Receptive field estimation

For each neuron, the size and position of its receptive field were obtained by fitting its response to drifting bars at various locations on the screen with a 2D Gaussian. The boundary of the receptive field was defined by the profile of the Gaussian at 1 STD from the center (black ellipse).

SPARSENESS AS A MEASURE OF SELECTIVITY

Sparseness is a measure of neuronal selectivity that shows how sparsely a neuron responds across the frames of the movies. As explained in the Introduction, sparseness S was calculated with the formula

$$S = \frac{1 - \frac{\left(\sum \frac{r_i}{n}\right)^2}{\sum \left(\frac{r_i^2}{n}\right)}}{1 - \frac{1}{n}}$$

where r_i is the value of the PSTH at time bin i and n is the number of frames. S ranges from 0 (no preference for any frame) to 1 (preference for a single frame).

A visual understanding of sparseness can be gathered by plotting together the rank ordered normalized PSTHs of neuronal responses with different sparseness indices, as in Figure

24E. We note that the reordered PSTHs of more selective responses decay faster than those of less selective responses.

MEASURES OF TIMESCALE OF PROCESSING

In our analysis, the timescale of neural processing was assessed in two ways: as the time constant of the autocorrelation function of the PSTH (also referred to as *slowness*), and as the autocorrelation of spike counts, both obtained for single neurons.

The **autocorrelation of the PSTH** signal was computed as follows.

Let $x(t)$ be the PSTH of a neuron in response to a single movie, where $t = 0 \dots N - 1$ represents the time of the binning window of fixed width Δt, obtained as explained in section Preprocessing. Then the autocovariance for lag $\Delta t \geq 0$ is calculated as

$$c(\Delta t) = \sum_{t=0}^{N-\Delta t-1} (x_{t+\Delta t} - \mu_x)\,(x_t - \mu_x)$$

where $\mu_x = \frac{1}{N}\sum_{t=0}^{N-1} x_t$. The autocovariance is then normalized so that the values at zero lag equal 1:

$$r = \frac{c}{c(0)}$$

Next, the autocorrelation r in the lag interval $(0 \dots 0.3]$ seconds was fitted (Curve Fitting Toolbox, MATLAB) with an exponential of the form

$$f(t) = Ae^{\frac{-t}{\tau}} + B$$

where A is the amplitude, B is the offset, and τ is the time constant of the exponential decay, which was our measure of the timescale of the time-averaged neural response. A and B were not included in the analysis.

Figure 24C shows the autocorrelation functions obtained for two PSTHs of an example neuron and their fitted exponentials.

The **noise** (or **spike-count**) **autocorrelation** was computed as described by Murray *et al.* (2014) and explained in the following (Figure 26).

Let matrix C of size $M \times N$ be the spike count data for each movie, where M is the number of trials and N is the number of bins, which contains at each item $c_{i,t}$ the number of spikes that occurred in trial i at time bin t. Then the noise autocorrelation at lag $n\Delta t$ is obtained by averaging all Pearson correlations $R_{t,t+n\Delta t}$ between pairs of column vectors $c_{*,t}$ and $c_{*,t+n\Delta t}$, which include the spike counts across trials at time t and $t + n\Delta t$, respectively.

$$R(\Delta t) = \frac{cov\left(c_{*,t}, c_{*,t+n\Delta t}\right)}{std\left(c_{*,t}\right) \times std\left(c_{*,t+n\Delta t}\right)}$$

$$R(\Delta t) = \frac{\sum_{i=1}^{M}\left(c_{i,t} - \mu_t\right)\left(c_{i,t+\Delta t} - \mu_{t+n\Delta t}\right)}{\sqrt{\sum_{i=1}^{M}\left(c_{i,t} - \mu_t\right)}\sqrt{\sum_{i=1}^{M}\left(c_{i,t+n\Delta t} - \mu_{t+n\Delta t}\right)}}$$

where $\mu_t = \frac{1}{M}\sum_{i=i}^{M} c_{i,t}$.

We note that, because of the subtraction of mean in each time bin, this measure corrects for nonstationarity in the mean firing rate during a movie, therefore we can say that this measure is not directly dependent on the stimulus (Murray *et al.*, 2014).

Next, the autocorrelation $R(\Delta t)$ in the interval $(0 \dots 0.3]$ seconds was fitted with an exponential, as explained above for the autocorrelation of the PSTH.

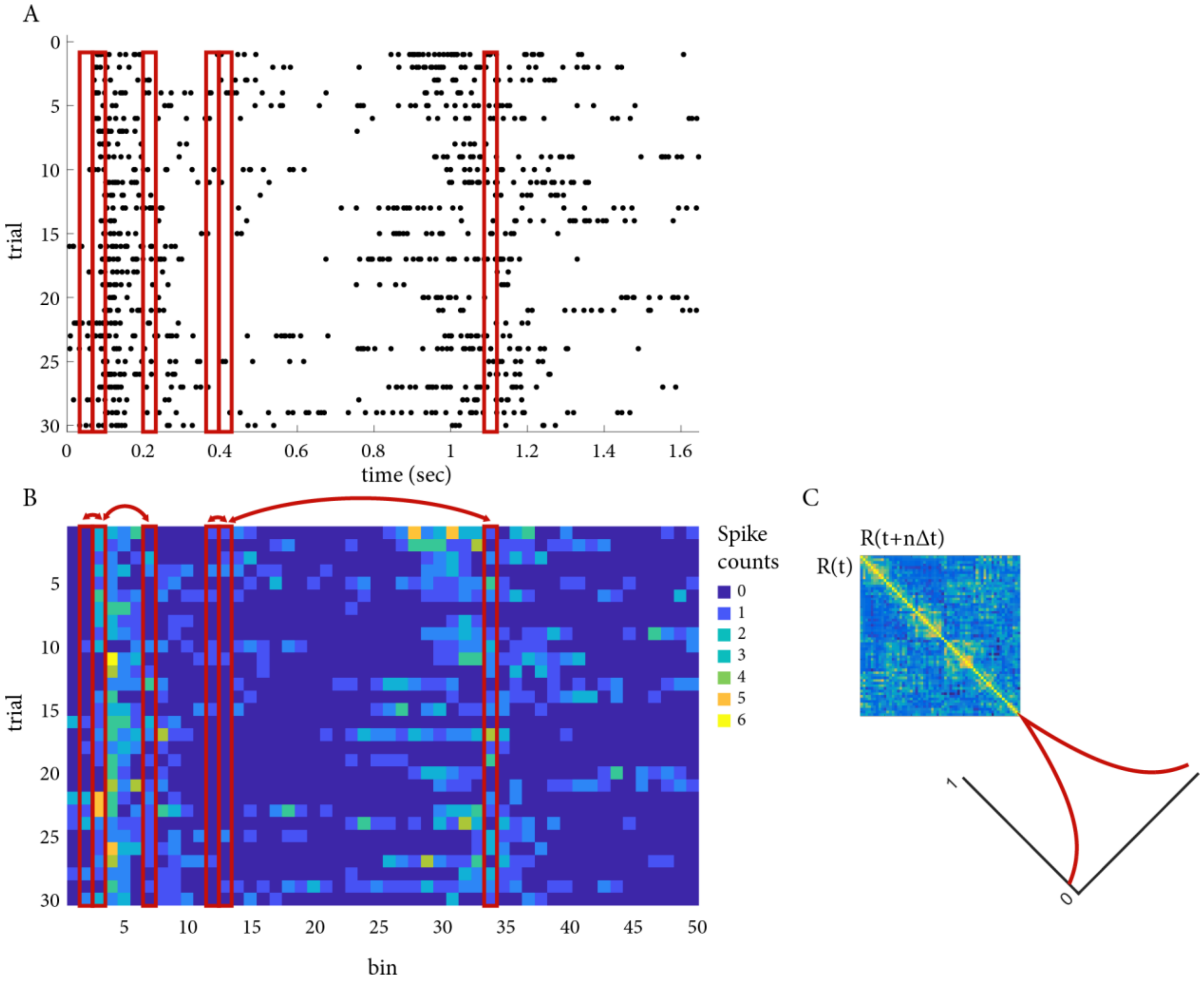

Figure 26. Noise (spike-count) autocorrelation

A. The raster plot of the first 1.65 seconds of the spiking activity across 30 trials of one V1 neuron during the presentation of manual movie fast 1. Each dot represents one spike.

B. The same activity as in A but discretized in time (in total 50 bins). The color of each bin indicates the number of spikes triggered within that bin. The red rectangles in panels A and B indicate, out of the many possible combinations, 2 pairs of bins with a lag of 1 Δt (small arrows), 1 pair with a lag of 4 Δt (medium arrow), and 1 pair with a lag of 21 Δt (long arrow).

C. The noise autocorrelation is obtained as follows. First, the correlation between spike counts in bins separated by a given lag is computed: the matrix shows the correlation coefficients R between bins at all possible lags. Pairs at small lags are more frequent than pairs at longer lags (e.g. there are 599 pairs at lag 1 Δt – on the first diagonal of the matrix – and one pair at lag 600 Δt – the last element on the first line of the matrix). The elements on the diagonal are 1 and represent the correlation between each bin with itself.

Then, all correlations between pairs of bins at a given lag are pooled together and averaged (for example all pairs at lag 1 Δt). This amounts to averaging the matrix along the main diagonal, which results in the red trace.

Figure 24D shows two example noise autocorrelations calculated over the spike counts given in section Example responses (Figures 31 and 37).

Speed of the visual input

The speed of the visual stimulus for a neuron was a scalar defined as the amount of change at pixel-level taking place within that neuron's receptive field during the presentation of one movie (Figure 27A).

First, for each movie, we computed the frame-by-frame difference of pixel intensities, converted to absolute values, and then averaged across time in order to obtain an array S_{xy} the size of a movie frame (720 lines by 1280 columns) that contained at each element the average intensity change of each pixel (Figure 27B-E):

$$S_{xy} = \frac{1}{N}\sum_{t=1}^{N-1}\left|I_{xy}(t+1) - I_{xy}(t)\right|$$

where N is the number of frames in a movie, and $I_{xy}(t)$ is the pixel intensity of the frame at time t. Next, we thresholded and binarized each neuron's receptive field (see section Receptive field estimation) at 0.5, so only the higher half was used in the following calculations. Then, we computed the dot product between the speed map and the receptive field, so pixel change was defined as

$$PC = S_{xy} \cdot RF$$

$$PC = \frac{1}{M_{RF}}\sum_{i=1}^{H}\sum_{j=1}^{W} S_{i,j} \times RF_{i,j}$$

where H and W are the height and width of the movie frame, and M_{RF} the number of nonzero elements of the receptive field mask.

When calculated over two frames, pixel change (speed) ranges from 0, when the frames are identical inside the receptive field, to 255, when one frame is white and the other one black.

Figure 27B-E shows that neurons, depending on the relative position of their receptive fields, can receive very different input. The pixel change measure was used in order to quantify the motion energy inside the receptive field.

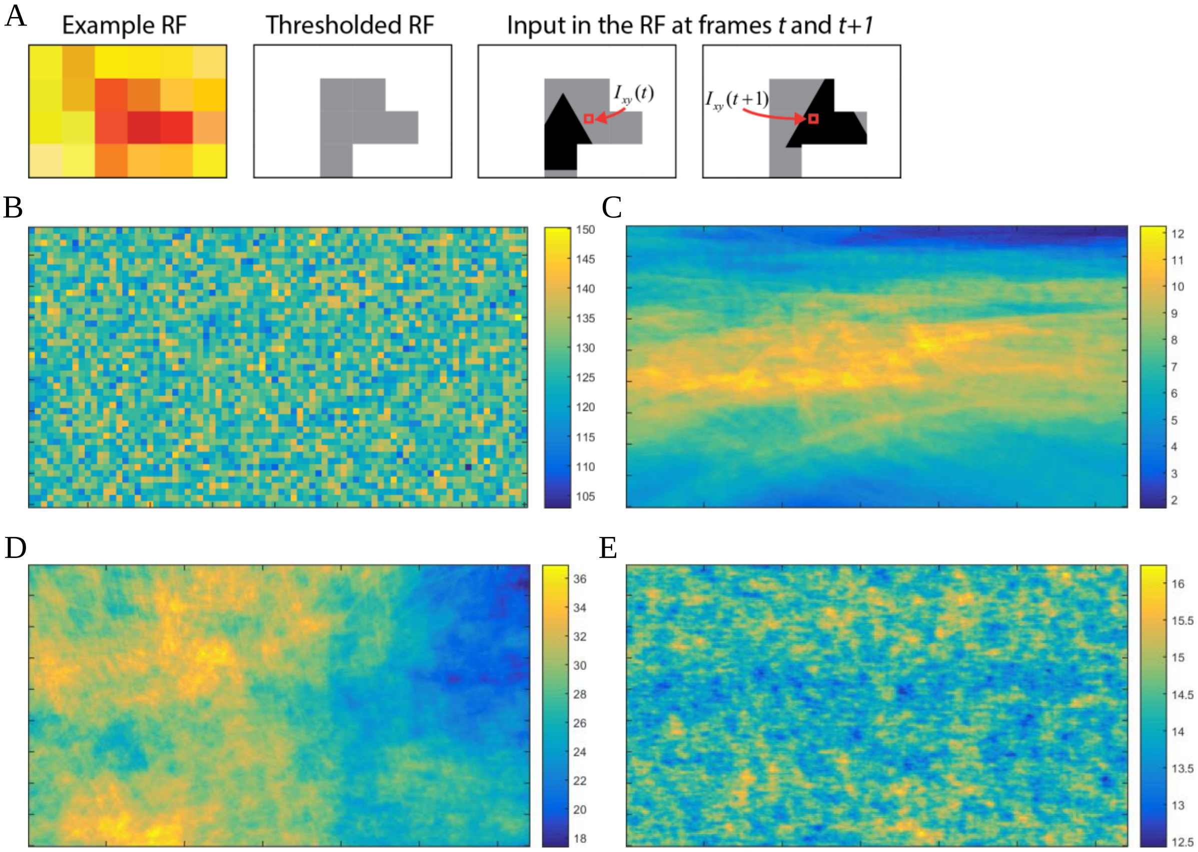

Figure 27. Speed of the visual input

A. Illustration of how the pixel change measure is obtained. For each neuron, receptive fields are normalized to 0-1 and thresholded at 0.5; then each pixel of the receptive field (red square) is multiplied with the average intensity change in the corresponding pixel of a movie (black triangle). **B-E.** Average intensity change for the movies white noise (**B**), manual fast 1 (**C**), ratcam 1 (**D**), and phase-scrambled manual fast 1 (**E**). Please note that the scale bars are different. **C** and **D** show that natural movies had their content localized only to some areas of the image.

In the following we will refer to this measure as *speed* (especially when it is used as a scalar) or *pixel change* (especially when it is a function of time).

3. REPRESENTATION OF NATURAL MOVIES IN RAT VISUAL CORTEX

STATISTICS OF RESPONSE PROPERTIES ACROSS VISUAL AREAS

In this chapter we characterize the responses of neurons recorded in four rat visual areas (V1, LM, LI and LL) while anaesthetized animals were passively exposed to natural and artificial movies.

Natural movies (named *manual fast 1*, *manual fast 2*, *manual slow 1*, *manual slow 2*, *ratcam 1* and *ratcam 2*) consisted of sequences of 3D black and white objects filmed in a rat playground with uniform background. Artificial movies consisted of randomly generated images (*white noise*) and altered versions of two of the natural movies (*phase-scrambled fast 1*, *phase-scrambled rat 2*).

Based on a measure that quantifies how reliably neurons respond from trial to trial, we included in this analysis 548 neurons.

Reproducibility (or reliability, see Methods) was strongly dependent on the average speed of the movie: the same neurons responded differently to the 9 movies (Figure 28). Specifically, the movies ratcam 1 and 2, the white noise and the phase-scrambled version of the ratcam movie evoked the most reproducible responses within each area; the manual movies instead were the least reproducible. This effect is probably due to the sharp contrast changes from frame to frame; in the case of ratcam movies, the changes were induced by the jerky movements of the animal carrying the camera.

By comparing the movies manual fast 1 and ratcam 2 with their phase-scrambled versions, we observe that natural movies evoked more reproducible response than the artificial ones (cyan vs. red bars, and dark pink vs. brown bars, within each area); this has already been observed in mice and rat visual cortices (Froudarakis *et al.*, 2014; Rikhye and Sur, 2015; Vinken *et al.*, 2016).

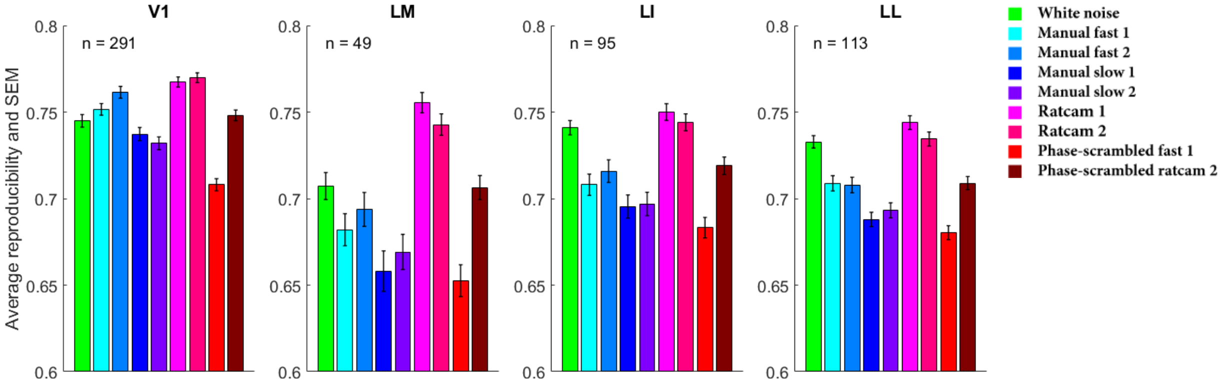

Figure 28. Reproducibility

The four panels show the average reproducibility of the neurons included in the analysis from each area. The inclusion threshold was set at 0.7, but we see smaller values because the threshold was applied on the movie with the highest reproducibility (see Methods).

The average firing rate was calculated as the number of spikes that occurred during all trials for a given movie divided by the duration of the movie and by the number of trials. Spontaneous average activity was estimated from short 4 seconds long periods of black screen before the onset of some movies (see Methods). In confirmation of previous results (Vinken *et al.*, 2016; Tafazoli *et al.*, 2017), the firing rates of V1 neurons were larger than those of the three other areas (Figure 29A). This result was supported by a two-way mixed and unbalanced ANOVA with *area* as the between factor (4 levels) and *movie* as the within factor (9 levels) that yielded a significant main effect for *area*: $\mathrm{F}(3{,}544) = 4{,}36, \mathrm{p} < 0.01$. In the post-hoc analysis only the difference of the pair V1-LL was significant (Dunnett T3 test, $\mathrm{p} < 0.05$).

The movies that elicited higher firing rates (ratcam 1 and 2, phase-scrambled ratcam 2) did so in all four areas: this can be observed in Figure 29B, that shows the coefficients of correlation between the patterns of firing rates measured across the nine movie in the four areas. The correlations are virtually perfect and highly significant (Student’s t-test, $\mathrm{p} < 0.001$, Holm-Bonferroni corrected).

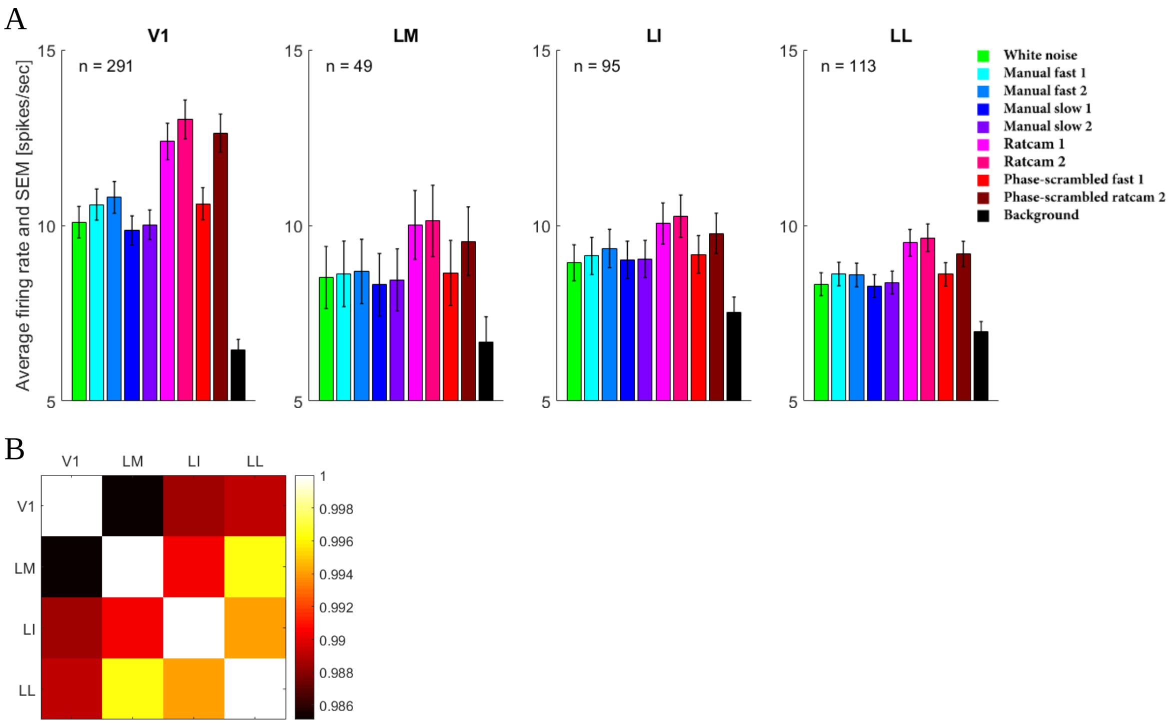

Figure 29. Firing rates

A. Average firing rate per movie and area. Average spontaneous activity is also represented (black).

B. Pairwise linear correlation coefficient between each pair of areas. Each square color codes the Pearson coefficient of correlation between the average firing rates in one area and the average firing rates in other area.

SPARSENESS AND SLOWNESS

The aim of this project was to test some of the predictions of the sparse coding theory such as the selectivity or slowness of the response across visual areas (discussed in the Introduction). To do so, we calculated the lifetime sparseness and the timescale of neural activity for each neuron and movie. Yet, the neurons in the investigated areas had basic response properties that were not matched between areas, thus making difficult to have a fair comparison: as seen above, the firing rates varied between areas and movies, and neurons had receptive fields of different sizes and at various positions on the screen (see examples in Figure 30). While we do expect to see an increase of size across the visual areas (Vermaercke *et al.*, 2014; Tafazoli *et al.*, 2017), the receptive field positions can be adjusted only to some extent during the experiment (Methods and Figure 6). We note in Figure 30, which shows the estimated receptive field of some of our recorded neurons, that V1 neurons mostly cover the center of the screen, whereas LL the top-left and bottom-left corners.

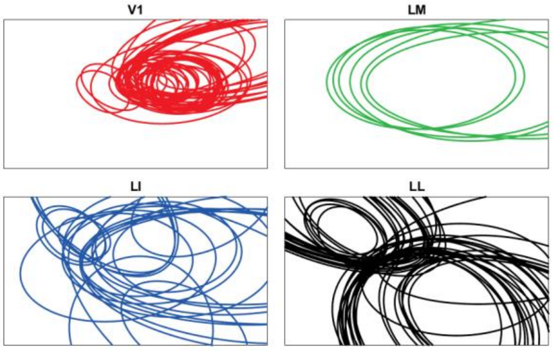

Figure 30. Position and size of receptive fields

The receptive fields with a goodness of fit larger than 0.25 are shown (46 neurons from V1, 6 from LM, 22 from LI and 46 from LL).

Our stimuli add another dimension of variability: movies of the same type (i.e. natural or artificial, slow or fast, manual or ratcam) were rather different between each other. For example, natural manual movies predominantly had their content (i.e. moving objects) localized to the center of the image, while natural ratcam movies to one side (Figure 27C,D).

It follows that recorded neurons in the four visual areas received different amounts of stimulation which depend on the properties of their receptive fields. In order to control for this, we chose to present our selectivity (sparseness) and slowness metrics relative to the speed of the visual input, which shows how much movies change from frame to frame, on average, across the pixels inside the receptive field (described in Methods). To increase the power of the analysis we also pooled together similar movies, i.e. those whose pixel change (speed) clustered together: manual movies formed one group, ratcam movies another one, and the artificial movies where kept separate (from Figure 41 to Figure 47).

Example responses

The neurons included in this analysis had widely different firing activities in response to the nine movies, for what concerns their viability (i.e. for how long we could record them), reproducibility, firing rate, position and size of the receptive field. The next figures show the response of two relatively good neurons, one in V1 and one in LL, to five movies: white noise, manual fast 1, manual slow 1, ratcam 1, and phase-scrambled fast 1.

In our dataset we included only neurons for which we had completed a full recording protocol, i.e. 30 repetitions (e.g. panel C in Figures 31 to 40). The response reproducibility across trials mostly varied across units (Figure 28), but to a large extent even between movies within single neurons: compare, for example, the raster plots in panel C in Figure 31 and Figure

33, which show the responses to ratcam 1 and white noise, respectively. Our reproducibility index measures 0.83 and 0.74 in the two cases.

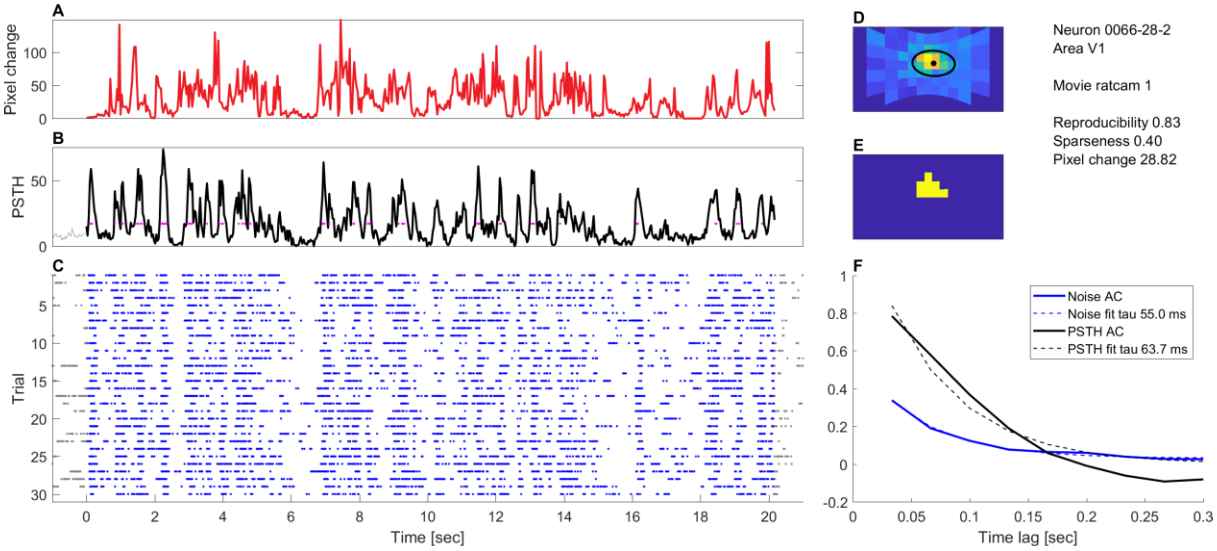

Figure 31. Properties of a V1 neuron with respect to the movie ratcam 1

Neural properties presented in this figure are related to a single neuron and to its response to one movie. The **top-right text** contains, in order:

- a unique ID of the *neuron* (session-channel-unit),

- the *area* where the neuron was recorded,

- the name of the *movie*,

- the *reproducibility* index for the specified movie (as defined in Methods),

- the *sparseness* (as defined in Methods) calculated on the PSTH response to the specified movie (black trace in **B**),

- and the average over time of the *pixel change* measure (red trace in **A**).

A. Pixel change as a function of time (see Methods) in arbitrary units (pixel intensity; it can range from 0 to 255).

B. The PSTH (black trace) is the trial-averaged response of the neuron to the specified movie. The average spontaneous activity is shown before the onset of the movie (gray trace). Magenta points indicate the peaks/bins that have been used to compute the reproducibility index (in total 10% of the number of bins; see Methods); their position on the *y* axis indicates the average firing rate during the movie (i.e. the mean of the black trace).

C. The raster plot shows the spikes produced in response to the 30 repetitions (trials) of the movie (blue dots), and the ones before and after the movie (gray dots).

D. The receptive field and its fitting (see Methods). This panel is identical for all movies.

E. The part of the receptive field that was used to compute the pixel change measure (see Methods). The mask is identical for all movies.

F. The autocorrelation function of the PSTH (black trace), its fit with an exponential (dashed black trace) and the time constant of the decay. The noise (spike-count) autocorrelation based on the spike-count data, its fit and the time constant of the decay (blue traces).

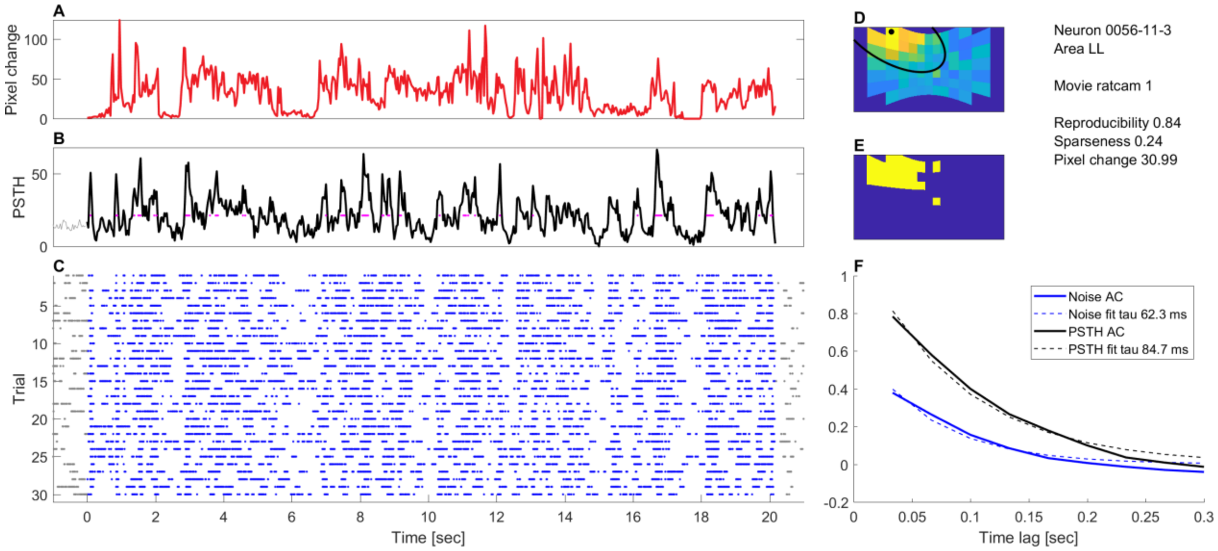

Figure 32. Properties of a LL neuron with respect to the movie ratcam 1

Same content structure as in Figure 31 for a LL neuron in response to the movie ratcam 1.

White noise movies are characterized by high values of the pixel change measure. Panel A in figures 33 and 34 show that it typically reaches values close to 127 (half the maximum range) and is relatively constant. Neurons are poorly triggered by this movie, but some frames seem to strongly drive all neurons (for example those at sec ~7 or sec ~15 in the PSTH, panel B), which results in sparseness being towards the lower end (compare sparseness in Figure 33 and Figure 37).

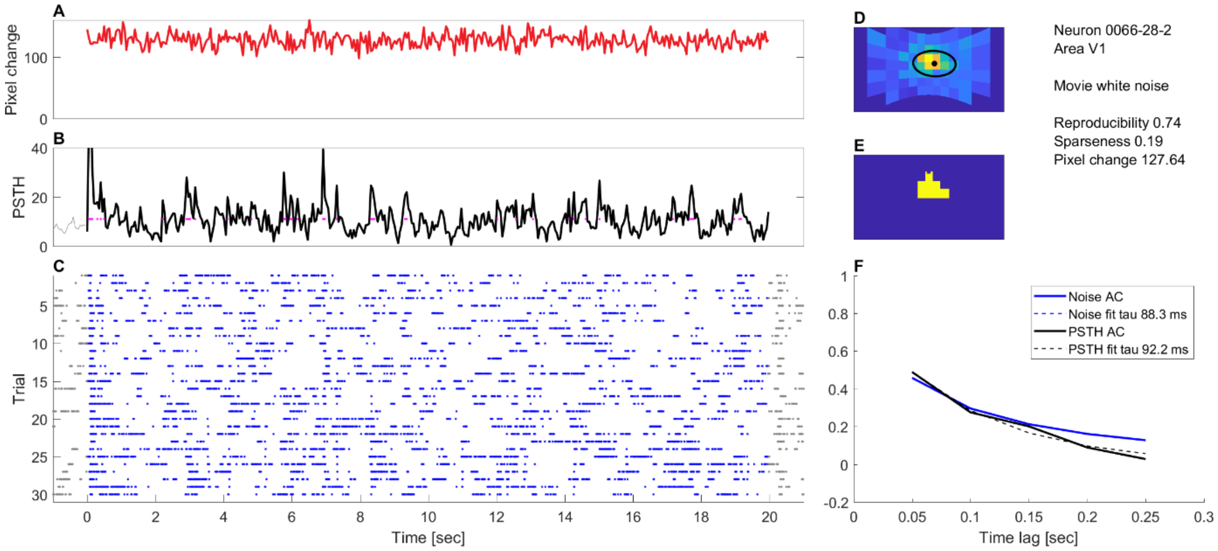

Figure 33. Properties of a V1 neuron with respect to the movie white noise

Same content structure as in Figure 31.

One should also note the difference between receptive fields in the two example neurons: one central, the other one in the top-left corner (panels D and E in all figures in this section). As a consequence, the input received by each neuron had different dynamics and energy, given by the pixel change measure (compare panel A in Figures 35 and 36).

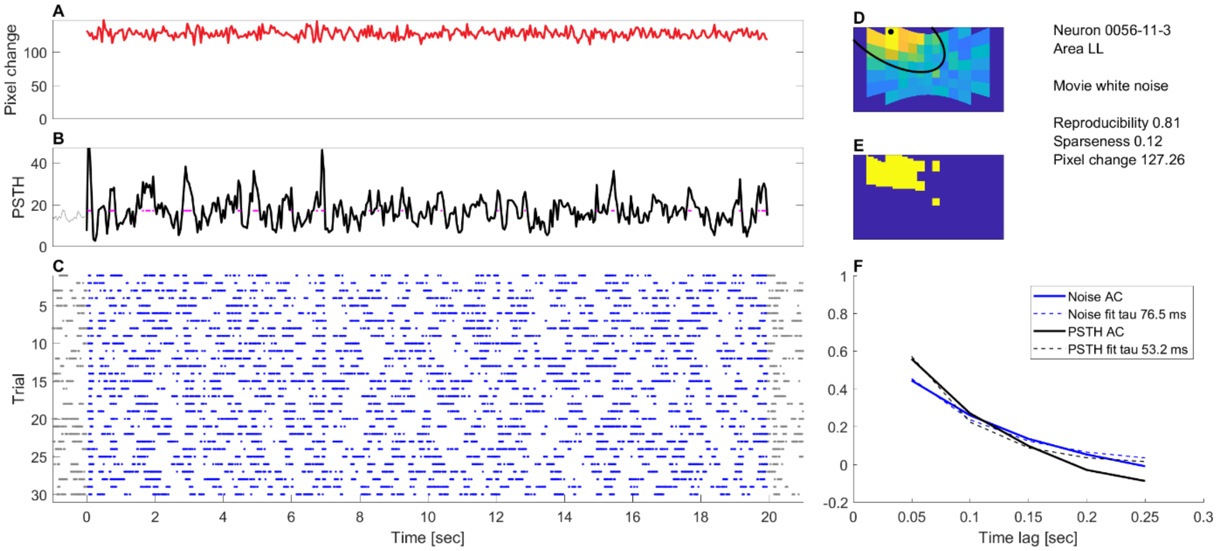

Figure 34. Properties of a LL neuron with respect to the movie white noise

Same content structure as in Figure 31.

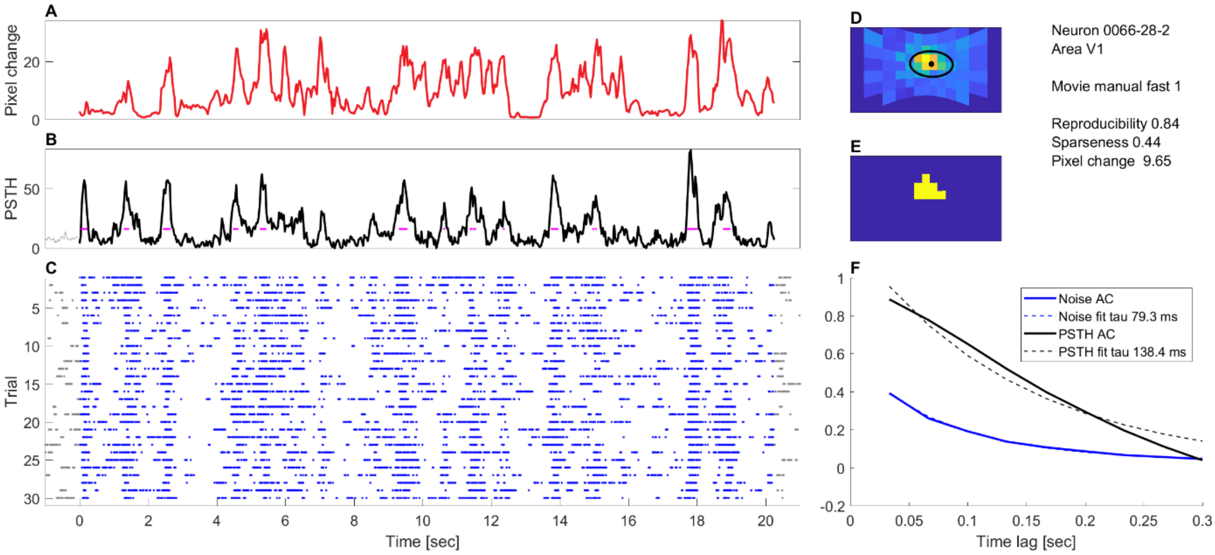

Figure 35. Properties of a V1 neuron with respect to the movie manual fast 1

Same content structure as in Figure 31.

The main focus of this work was to assess whether we can find evidence of activity at different timescales across visual areas. The PSTHs of Figures 35 and 36 show the responses of two neurons (one in V1 and the other in LL) to movie manual fast 1. We can appreciate that the LL neuron has fewer strong responses than the V1 neuron, each with a longer duration, and this can be quantified by computing the autocorrelation of the PSTH ("PSTH AC" in panel E) or the spike-count autocorrelation (“Noise AC” in panel E). The time constants of the decay (“fit tau”) is in both cases larger for the LL neuron, which confirms that the two signals are more self-similar than the corresponding ones of the V1 neuron. However, as discussed in section Sparseness and slowness, the position of the receptive fields in the two cases could confound these effects. In this specific example, the objects contained in the movie manual fast 1 swept more often through the receptive field of the V1 neuron, than through the receptive field of the LL neuron, as can be appreciated by looking at the richer dynamics of the pixel change metric (panel A). A similar pattern was found for the movie manual slow 1 (Figures 37 and 38).

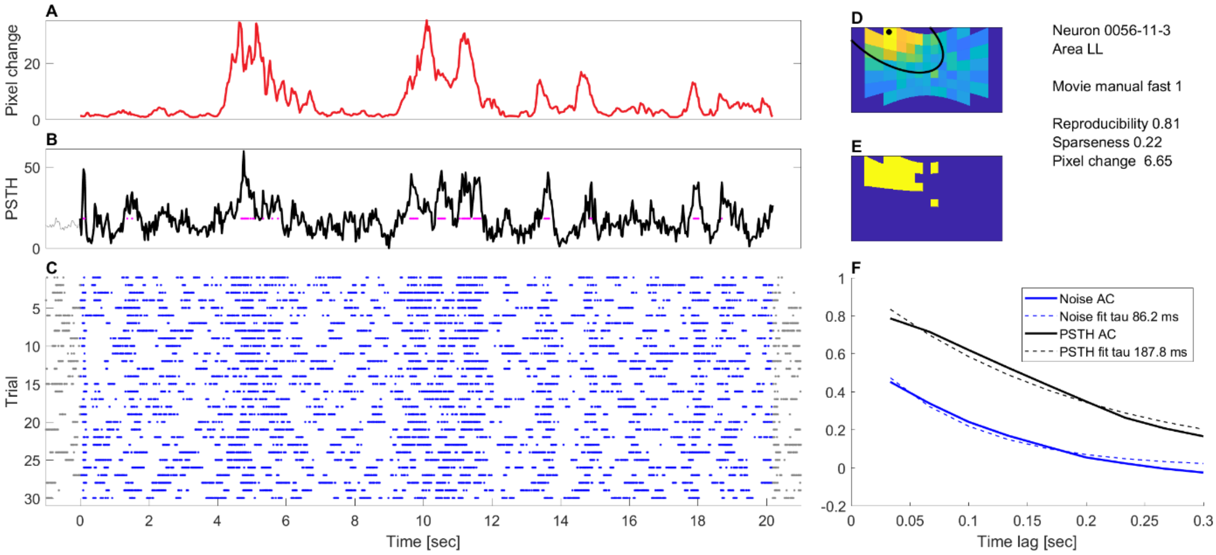

Figure 36. Properties of a LL neuron with respect to the movie manual fast 1

Same content structure as in Figure 31.

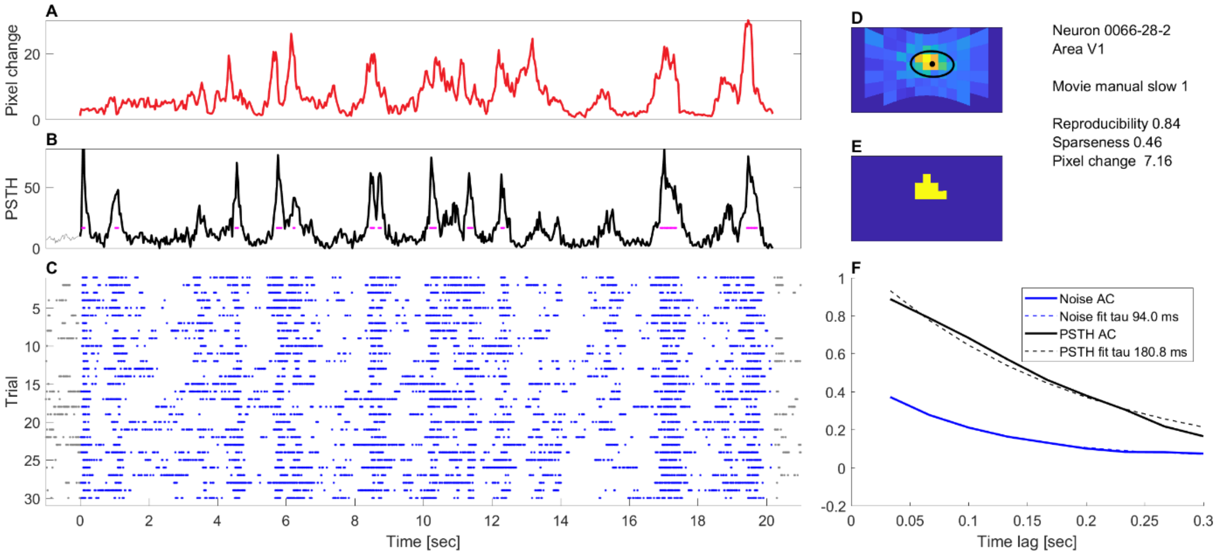

Figure 37. Properties of a V1 neuron with respect to the movie manual slow 1

Same content structure as in Figure 31.

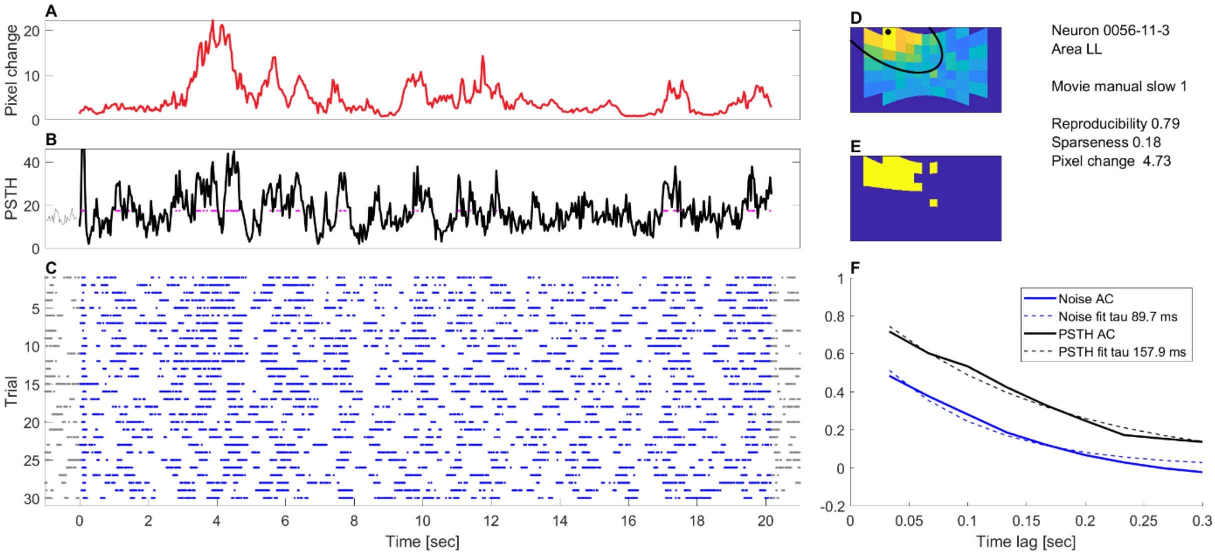

Figure 38. Properties of a LL neuron with respect to the movie manual slow 1

Same content structure as in Figure 31.

The two phase-scrambled movies were similar in their properties and their ability to drive neurons (Figure 39 and Figure 40): they had similar variation of the pixel change (panel A), and, as with the white noise movie, they had less reproducible responses compared to those to natural movies.

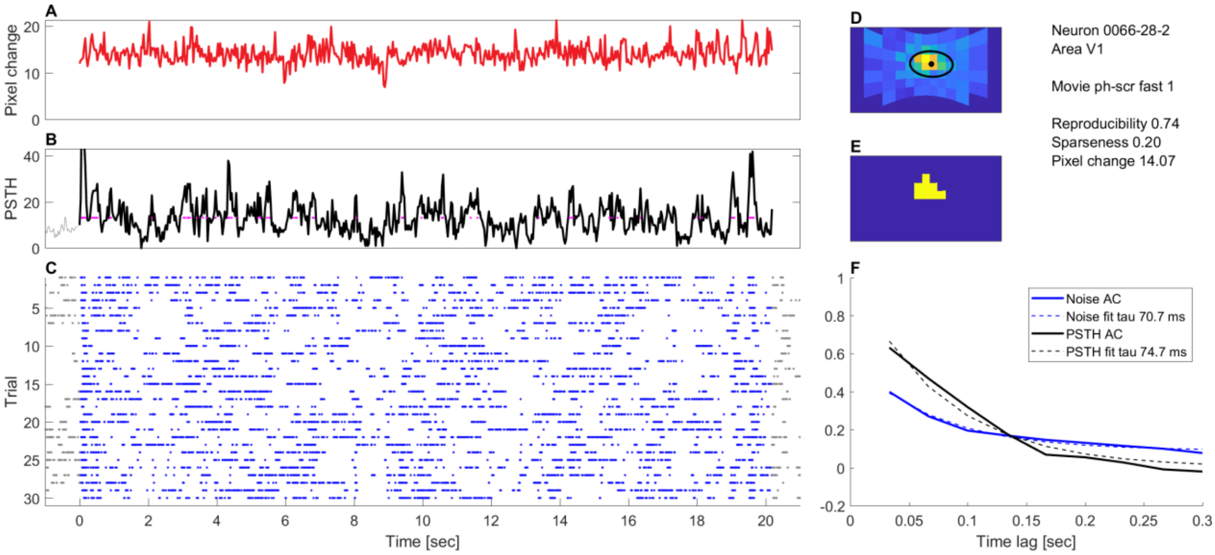

Figure 39. Properties of a V1 neuron with respect to the movie phase-scrambled fast 1

Same content structure as in Figure 31.

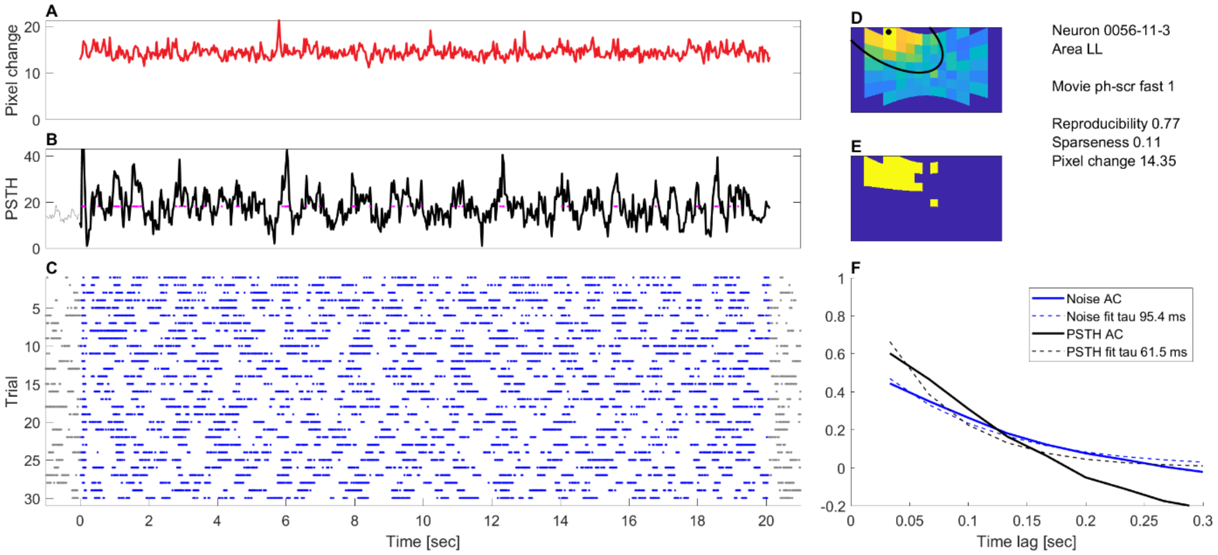

Figure 40. Properties of a LL neuron with respect to the movie phase-scrambled 1

Same content structure as in Figure 31.

In the following sections we will make an analysis of sparseness and slowness of the response of all neurons in our population, relative to the speed of the movie.

Sparseness

To assess how selective individual neurons are when presented with natural and artificial movies, we computed lifetime sparseness over the trial-averaged response (described in Methods). This measure ranges from 0 to 1: it takes small values when neurons fire similarly to all frames and large values when neurons fire only to some frames of the movie.

Sparseness was plotted against the speed of the visual input. As explained above, this allows a fair comparison among the areas, because we could control for differences in terms of receptive field position and size (and, thus, of stimulus luminance changes within the receptive field). In addition, it allowed testing whether sparseness depends on the speed (i.e. on the spatiotemporal energy) of the visual input – a relationship that has not been explored so far in the previous studies of sparseness in visual cortex.

Figure 41 shows the sparseness for manual (speeds between 3 and 11.1) and ratcam movies (speeds from 24 to 35). Figure 42 shows the sparseness for the three artificial movies (speeds from 13.8 to 128.8).

We observe that sparseness generally decreases from V1 to LL in all types of movies, with the exception of the ratcam ones, where LM is the largest (Figure 41 right and Figure 42 middle). We also note that sparseness increases with speed, but only in the case of natural movies. Precisely, mean sparseness is larger for ratcam (faster) than for manual (slower) movies (Figure 41 right vs left), and remains roughly unchanged between the three artificial movies (Figure 42). A two-way ANOVA with *area* and *speed bin* as factors confirms these results and yields a significant main effect of *area* ($F_{3,4896} = 92.73, \mathrm{p} < 0.001$, *speed bin* ($F_{8,4896} = 72.76, \mathrm{p} < 0.001$), and a significant interaction *area*speed bin* ($F_{24,4896} = 4.67, \mathrm{p} < 0.001$).

Finally, sparseness is higher for natural than for artificial movies at similar speeds: this is evident when one compares the bins starting at speed 8.4 with the one starting at 13.8, or bin 31.4 with bin 33.0 in Figure 41 and Figure 42, respectively ($F_{1,4924} = 356.34, \mathrm{p} < 0.001$).

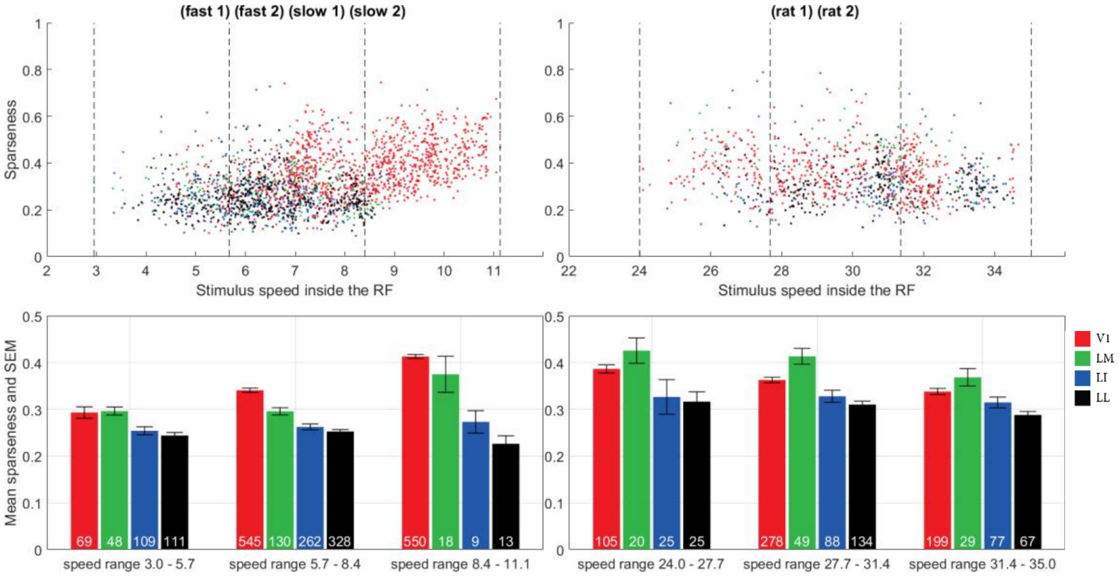

Figure 41. Sparseness of responses to natural movies

Each of the two columns of this figure refer to a group of movies listed at the very top: fast 1, fast 2, slow 1 and slow 2 for the **left column**, and ratcam 1 and ratcam 2 for the **right column**, which together make up the natural movies of our stimulus set.

The **top row** of the figure shows, for each neuron, the sparseness of responses to the listed movies as a function of the movie speed measured inside the receptive field, i.e. each point refers to the response of one neuron to one movie. The *x* axis has been divided in three equal bins for different intervals of speed, from the minimum to the maximum speed evoked in any neuron by the listed movies (dashed lines). The exact margins of each interval are given at the very bottom of the figure (speed range). Also note that speed, on the *x* axis, increases monotonically from left to right in the two columns.

The bars in the **bottom row** of the figure show the average sparseness per area computed from responses with speeds within the aforementioned bins.

Numbers at the base of each bar indicate the number of responses (points in the top panel) used to obtain the value of that bar, so the cloud of dots in each column counts: *the number of movies* × *the size of our population* (4 × 548 for the left column, 2 × 548 for the right column).

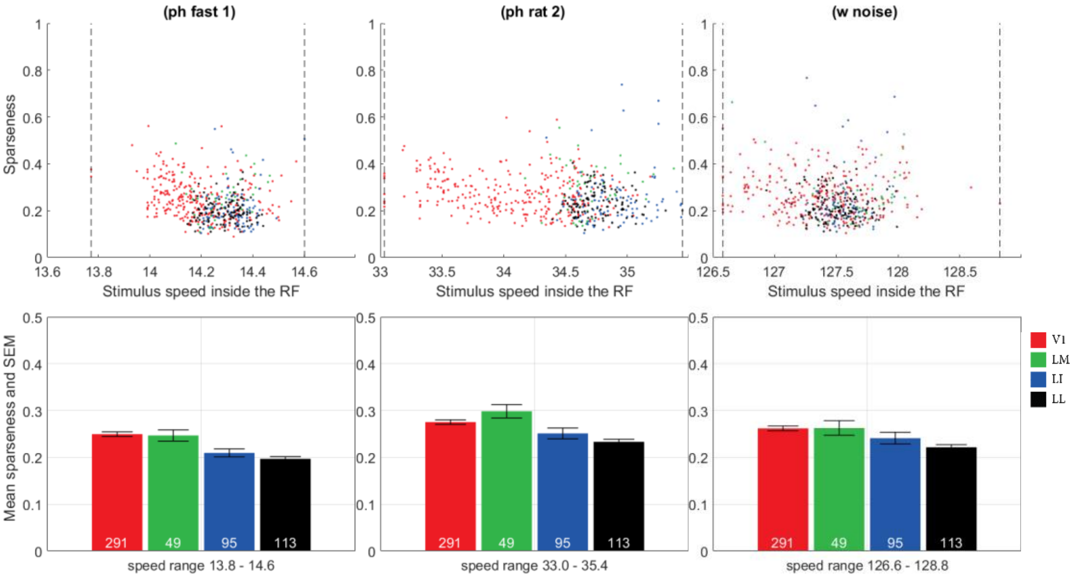

Figure 42. Sparseness of responses to artificial movies

The content structure is similar to the one of Figure 41.

This figure shows the sparseness of responses to three artificial movies: phase-scrambled fast 1, phase-scrambled rat 1 and white noise, in increasing order of speed (*x* axis in both rows).

Note that the movie phase-scrambled rat 2 (bin starting at 33.0) and the third bin of the right column in Figure 41 (bin 31.4) have matching speeds. Phase-scramble fast 1 (bin 13.8) is also very close to bin 8.4 of Figure 41.

Overall, these results suggest that the visual areas we investigated are functionally different in terms of sparseness. Its decrease across areas does not support one of the predictions of the sparse coding theory, i.e., that areas at higher levels of processing should be sparser (see Introduction). However, the reduction of sparseness could result from an increase of invariance of the response to transformation, which could overcome the increase of selectivity (Rust and DiCarlo, 2012). In this sense, our finding could be consistent with the emergence of ventral-like processing across the areas' progression.

The high sparseness values observed in LM for ratcam movies, which are the fastest natural movies in our stimulus set, could be explained by at least one of the following considerations: 1) LM does not belong to the putative V1-LL processing stream, as already suggested by Tafazoli *et al.* (2017); 2) LM neurons are particularly selective to fast-changing stimuli, which argues for a dorsal-like processing, as observed for mice LM neurons, which prefer high temporal and low spatial frequencies (Marshel *et al.*, 2011); 3) our population of LM neurons is too small (49) to allow a proper estimation of sparseness.

Our results also confirm what has already been observed in mice and rats, i.e., that responses to natural movies are sparser than those to artificial ones (Froudarakis *et al.*, 2014; Vinken *et al.*, 2016). Interestingly, in our case, this conclusion is not affected by possible variations of the speed of the phase-scrambled movies, since, for matching speed, the natural movies still surpass the phase-scrambled ones.

SLOWNESS

The slowness of the response of each neuron was assessed by computing the time constant of the autocorrelation function of the trial-averaged response (PSTH) to a given movie (see Methods). The time constant (in ms) indicates the timescale at which neurons respond to various types of movie, i.e., large time constants correspond to slow fluctuations in the stimulus response, whereas small time constants correspond to fast responses.

By construction, this measure is strongly stimulus-dependent, as can be seen in Figure 43, which highlights three emerging groups of similar movies: 1) the four manual ones, 2) the two ratcam and the two phase-scrambled movies, and 3) the white noise movie, which produced the smallest time constants.

We note that qualitatively neurons in various areas respond differently to movies, and that there is an overall increase of slowness from V1 to LL (Figure 43). The group of artificial movies and the ratcam ones (which both are fastest movies in the stimulus set) don’t seem to produce large differences in average time constants in the four areas: $37.4, 32.4, 38.7, 44.7$ ms in V1, LM, LI and LL (Figure 43, red hues). By contrast, for the natural manual movies, which are slower, a sharp increase of slowness can be observed from LM (86.9 ms) to LL (128.3 ms) (Figure 43, blue hues), although V1 time constants are not much smaller than LL ones.

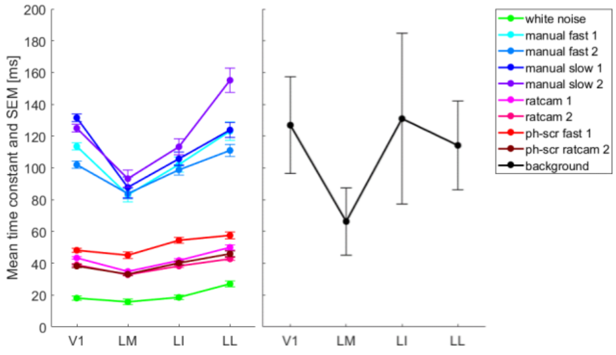

Figure 43. PSTH time constants per movie

For each movie (colored lines) we plot the mean time constants in areas V1, LM, LI and LL. We note the strong dependence of this measure on the movie type, i.e. natural manual movies cluster together (traces of blue hues), and so the fast-moving natural and artificial ones (ratcam and phase-scrambled, traces of red hues).

The black trace (background) represents the time constants computed from the PSTHs relative to short periods of spontaneous activity (see Methods).

The qualitative analysis is confirmed when we plot time constants as a function of speed inside the receptive field: except for the faster manual movies (Figure 44, left column, third bin) which contains few responses, the trends observed in Figure 43 are reproduced at slow (Figure 44), medium and high speeds (Figure 45), irrespective of the movie type (natural or artificial). A two-way ANOVA with *area* and *speed bin* as factors produces a significant main effect of *area* ($F_{3,4896} = 24.8, \mathrm{p} < 0.001$) and *speed bin* ($\mathrm{F}_{8,4896} = 512.44, \mathrm{p} < 0.001$), and a significant interaction *area*speed bin* ($\mathrm{F}_{24,4896} = 13.34, \mathrm{p} < 0.001$). The ANOVA interaction reflects the way V1 neurons respond to slow stimuli, which prevent the trend across areas to be a strict monotonic increase (Figure 44).

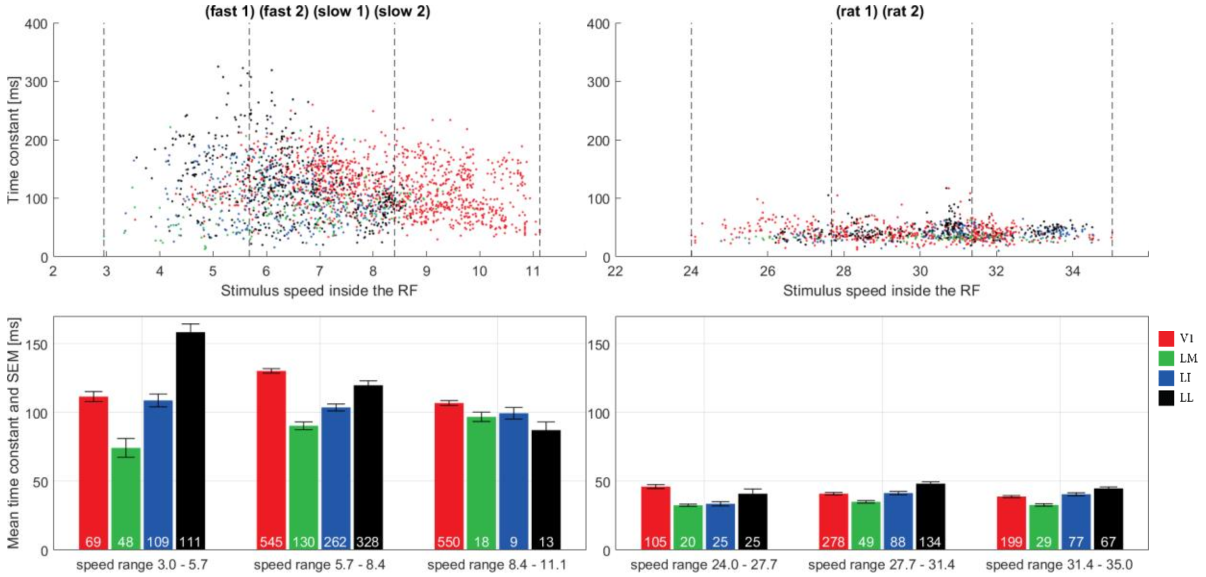

Figure 44. PSTH time constants of responses to natural movies

The content structure is similar to the one of Figure 41.

Overall, by measuring the slowness of the PSTH response we note, as expected, the strong dependence of the visual input, and a slight tendency for lateral areas to fire at slower timescales compared to the medial areas, particularly at very fast stimuli (Figure 45).

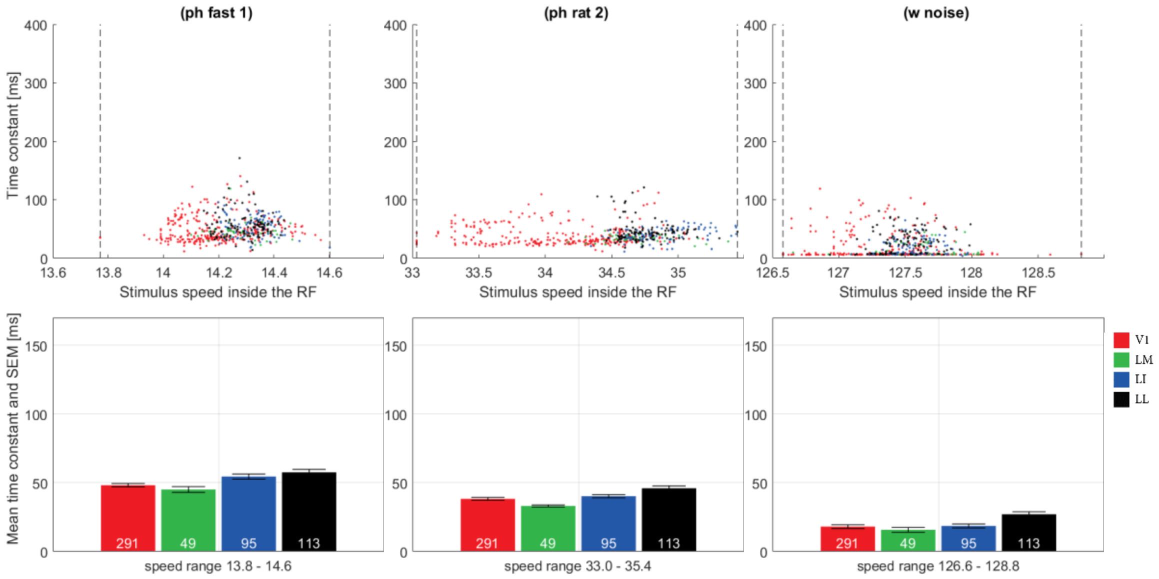

Figure 45. PSTH time constants of responses to artificial movies

The content structure is similar to the one of Figure 41.

Noise autocorrelations

Another way one can analyze neural fluctuations is by looking at the variability at the level of individual trials of a neuron’s response, in relation to the ongoing stimulus. The rationale of this approach is that the PSTH, which is the focus of the previous section, discards the single-trial information by only considering the mean of spike counts in a given time bin, whereas the brain works with single trials.

A possible implementation is to compute, for each neuron and for pairs of time bins separated by a time lag, the correlation of the number of spikes across trials, and then plot the average correlation as a function of time lag to obtain the so-called spike-count or noise autocorrelation (see Methods). Intuitively, the decay of the autocorrelation will be a measure of self-similarity of individual trials, i.e. the time span in which the spiking activity is sustained at a similarly high or low rate.

Murray and colleagues (2014) have employed this method to estimate the timescales of firing activity (i.e. when no stimulus was presented) in various cortical areas of the monkey, and found that sensory areas exhibit shorter timescales while prefrontal areas longer timescales.

Our results show that the activity of neurons in lateral areas is generally at slower timescales: in 8 out of 9 speed bins in Figure 46 and Figure 47, the time constants of LL $(M = 73.5\text{ ms}, \text{SD} = 1.9)$ and LI $(M = 57.4\text{ ms}, \text{SD} = 2.1)$ are larger than those of V1 $(M = 55.6\text{ ms}, \text{SD} = 0.9)$ and LM $(M = 41.2\text{ ms}, \text{SD} = 2.3)$ neurons. A two-way ANOVA with the factors *area* (4 levels) and *speed bin* (9 levels) confirms the effect of *area*: $F_{3,4896} = 40.24, \mathrm{p} < 0.001$.

We also note that, even though the noise autocorrelation doesn't include the mean response (i.e. PSTH), the speed of the stimulus is still reflected in the time constant of the response. The same ANOVA yields a significant main effect of *speed*: $F_{8,4896} = 22.4, \mathrm{p} < 0.001$.

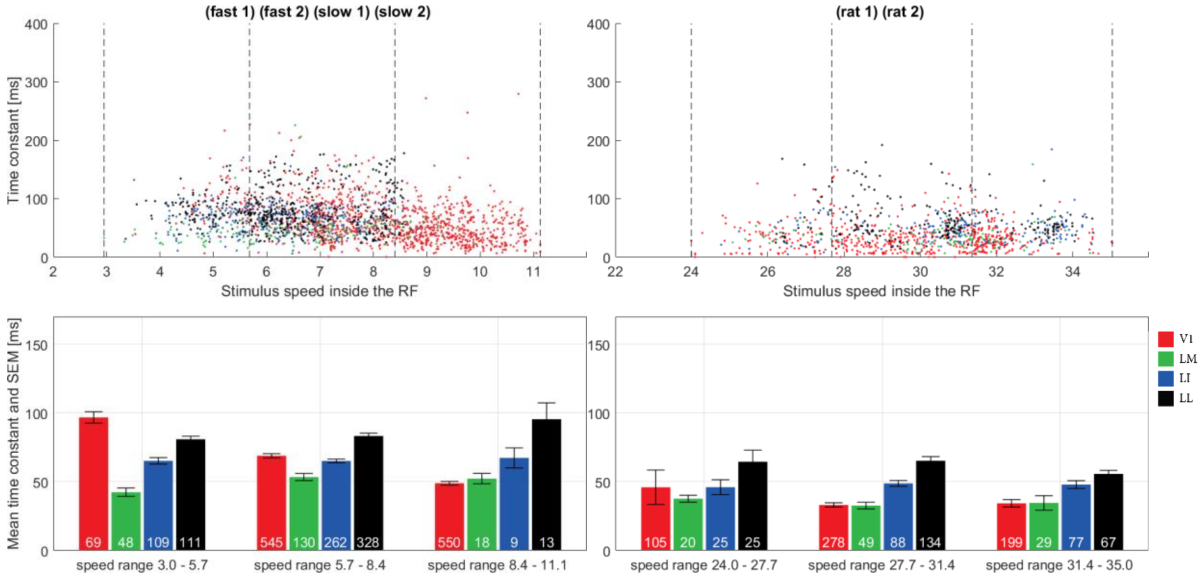

Figure 46. Noise autocorrelations for natural movies

The content structure is similar to the one of Figure 41.

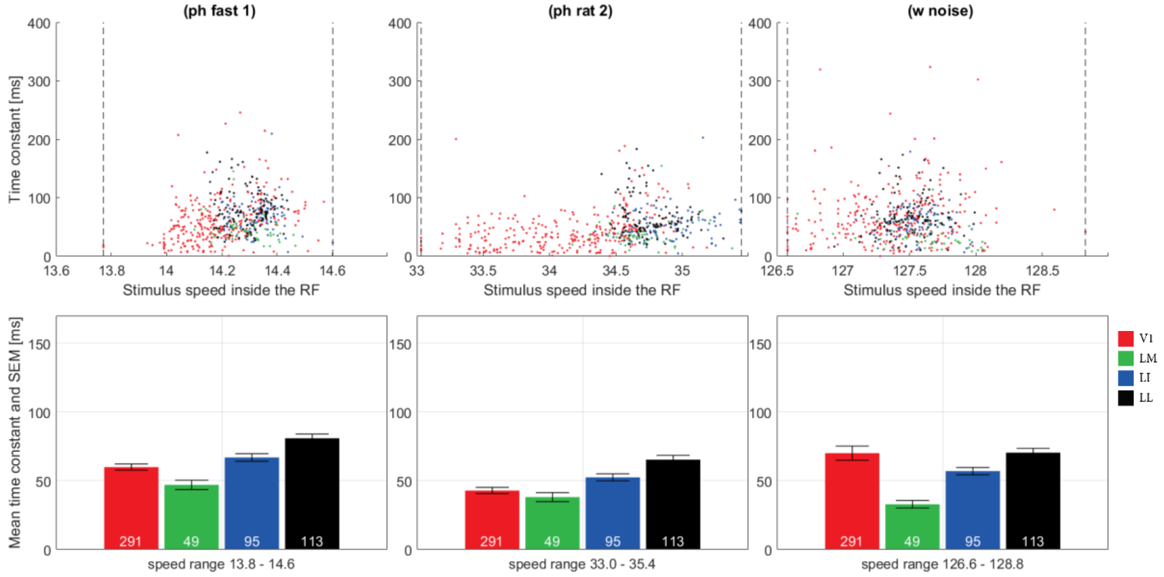

Figure 47. Noise autocorrelations for artificial movies

The content structure is similar to the one of Figure 41.

There is no indication that time constants of responses to natural movies are different from those to artificial movies, as the main effect of *movie type* in a two-way ANOVA with factors *area* and *movie type* (2 levels) is not significant: $F_{1,4924} = 0.76, \mathrm{p} = 0.38$.

In this section we have seen that, when we estimate the trial-by-trial autocorrelation, the emerging timescales of neural activity during stimulus presentation show a progression of increasing slowness from V1 to LL. Noise time constants ($M = 58.4\,ms$) are significantly smaller than PSTH ones ($M = 72.6\,ms$) (Wilcoxon signed-rank test, $p < 0.001$) and overlap with those obtained for the spontaneous activity (Figure 43 and Figure 48). Moreover, we don't see any difference between natural and artificial movies. All of these suggest that noise autocorrelations capture the intrinsic timescales of processing in neural circuits (and less the timescale of the visual input), and our results highlight a putative functional and architectural difference between the four visual areas investigated.

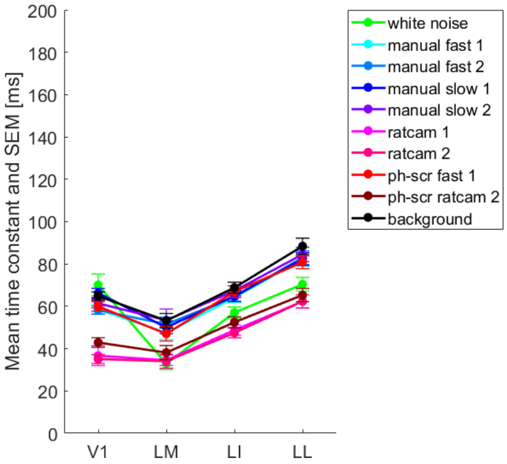

Figure 48. Noise time constants per movie

Mean time constant of noise autocorrelation for each movie and area. To facilitate the comparison, the *y* axis matches the one of Figure 43.

4. POPULATION DECODING

In the previous chapter we have characterized the timescales of single cell responses to movies with various dynamic properties, and assessed to what extent the slowness principle is implemented in the rat visual cortex so as to represent (or encode) time-varying stimuli. A slow and invariant representation is one in which neurons maintain a fairly constant firing over time when presented with stimuli that change their appearance (see Figure 13). If their firing is slower, then one can assume their ability to discriminate between preferred and non-preferred stimuli is also preserved over longer timescales.

A different approach from what has been done in the previous chapter is to assess slowness through single-cell or population decoding, by training classifiers to discriminate between consecutive instances of objects undergoing identity-preserving transformations. The movies we used in this project contain many objects that are translating, rotating, changing size, pose, and lighting. Our working hypothesis is that, if some neuronal responses change slowly over time, we should see it in their ability to assign the same label to consecutive frames of an object changing over time. In practice, in our case, one can address this idea by training a classifier to discriminate between a pair of frames of two objects (taken from two different segments of the same movie or different movies) and testing the ability of the classifier to generalize to pairs of frames containing the same objects that are adjacent, in time, to the training ones: a sharp drop in the performance of the classifier would indicate that the population of neurons is not maintaining a stable (invariant) representation of the two objects, whereas a shallow decrease in performance – as tested by looking at frames that are farther apart in time – would suggest that those representations are slow (Figure 52C).

In the following, we are presenting an application of the population decoding technique used before for the discrimination of static objects (Hung *et al.*, 2005; Rust and DiCarlo, 2010; Vermaercke *et al.*, 2015; Tafazoli *et al.*, 2017). Noticeably, previous studies have applied decoding approaches to time-varying stimuli, but they have not tried to decode the identity of visual objects that are continuously transforming over time: they either decoded the identity of static visual shapes presented in rapid sequence (Nikolic *et al.*, 2006) or random frames of natural movies (Froudarakis *et al.*, 2014). The read-out scheme we apply here is one possible way to estimate the amount of information that neurons convey to downstream visual areas.

STIMULUS SET

In order to limit possible confounds due to the complexity of natural images, for this analysis we only used the four manual movies: fast 1, fast 2, slow 1 and slow 2. Moreover, within each movie, we considered only sequences of frames that contained only one object per frame.

To obtain the objects' masks of each movie we performed a semi-automatic image segmentation process on each frame: the automatic part consisted of selecting the continuous pixels whose colors were within a certain tolerance relative to two manually-defined reference pixels, by using a magic wand tool applied on the color movies (Phung, 2004, updated 2 April 2004); the manual part consisted of refining the masks with a brush-like tool and labeling each object with pre-defined names (16 objects in total, see Methods and Figure 49 below).

Manual movies contain on average 2.55 ($std = 1.58$) objects per frame across the 4 × 600 frames, from which we extracted 25 segments of black and white objects (exactly one per frame) of mean duration 30.3 frames ($std = 13$) (Figure 49).

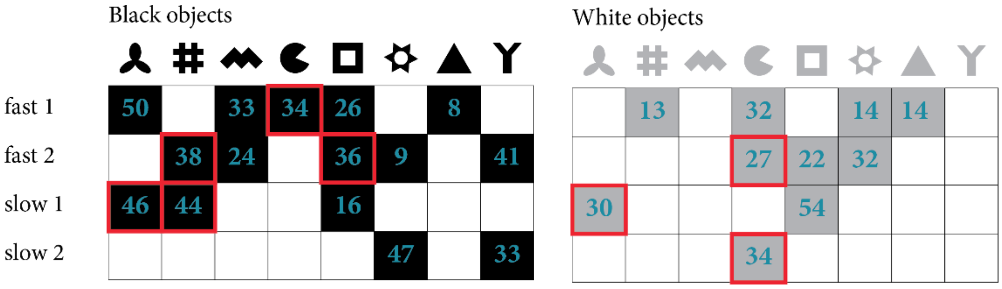

Figure 49. Objects used for population decoding

The two panels show the source of movie segments from the four manual movies. Each filled square indicates whether the object drawn at the top of the table has been found in the movie named on the left side. The numbers inside each square represents the length (i.e. number of frames) of the movie segment. The red contours highlight the objects that were included in the stimulus set (Figure 50). From left to right, the names of the 8 objects are: bunny, hash, m, pacman, square, star, triangle, y, to which their color (black or white) are added, to obtain the total number (16) of objects used in this experiment.

The 25 objects had trajectories (or transformations) that could roughly be classified based on a few qualitative criteria, as follows: start position (center, left or right off-center, extreme right, extreme left, top right of the image), direction (towards left, right, bottom left, bottom right), size (zoom-in, constant size), and magnitude of translation (half screen, full screen). The objects that were eventually included in the analysis are shown in Figure 50A and were matched for all these criteria. Figure 50B shows some objects that were not included in the analysis: either because their trajectories were not similar to other objects' trajectories (to allow grouping them together), or because they appeared for too short durations.

Since for this analysis we also used a bin size the duration of three frames (100 ms), we created resampled movies that contained the original frames at positions $3 \times k - 1$ for $k = 1 \dots 200$.

In the following, with the locution *movie segment*, we refer to a consecutive sequence of frames that contained a specific object (one of those selected above) from the four manual movies, binned at either 33 or 100 ms.

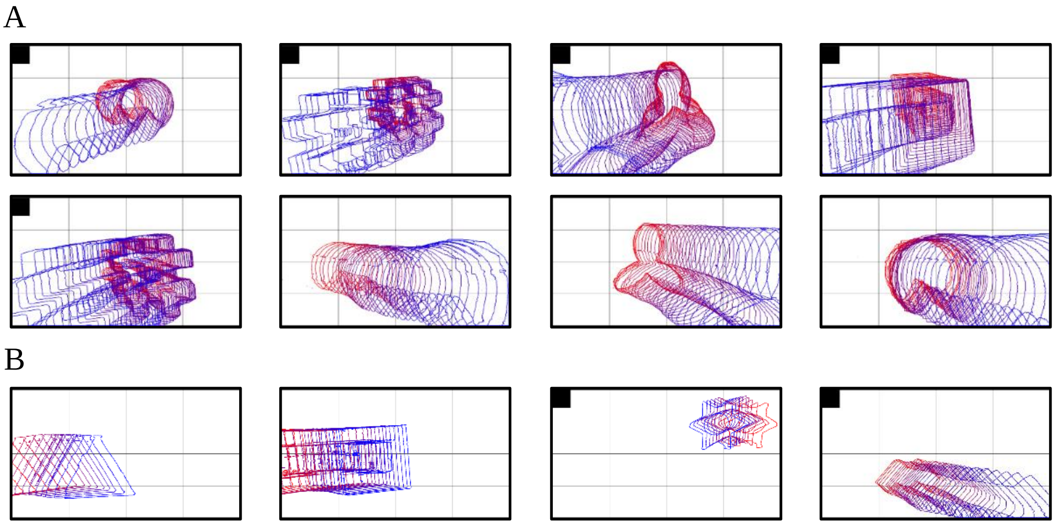

Figure 50. Trajectories of selected objects

Individual panels show objects' trajectories along movie segments. The contours indicate the positions of an object in a given frame, at times coded by color (red represents the starting point, and blue the end). The number of frames of each object is given in Figure 49.

A. We included in this analysis 5 black objects (marked with small black squares) and 3 white objects. We note that objects started being alone in the frame when they were approximately in the center of the image and then moved towards the left or right corner. The two groups of objects (black and white) were analyzed separately because of the difference in color and direction of the trajectory. **B.** Some example objects that were not included in the analysis, for the following reasons: not enough direction matches, not enough start position matches (panels 1, 2, 4), and too few frames (panel 1 and 3).

SELECTION OF THE POPULATIONS OF NEURONS

In this analysis we included only the most reliable neurons in our dataset, i.e. those with a lifetime average firing rate of at least 2 Hz and a reproducibility index larger than 0.75. With these criteria we selected 216 neurons recorded in V1, 28 in LM, 60 in LI and 58 in LL.

Because most objects were present in the center of the image (Figure 50A), we aimed at having populations of neurons whose receptive fields would maximally cover the same area. To this end, we did the following steps: within each area, we pseudorandomly sampled 1000 populations of 15 neurons, and for each one of them we generated the binary union of receptive fields masks of constituent neurons (see Methods). We then assigned to each population a score that was computed as the ratio between the screen coverage of the union U – defined as the number of nonzero pixels –, and the median distance from each pixel of the union to the center of the image $\tilde{d}$.

$$U = RF_1 \cup RF_2 \dots \cup RF_{15}$$

$$score = \frac{\sum U}{\tilde{d}}$$

Finally, from the 1000 populations we selected the 10 with the highest scores for each area, which were used in all the decoding analyses described in the following. Figure 51 shows a few examples of the resulting population receptive fields.

Not all neurons that met our criteria were included in the 40 populations. The unique neurons that made up our populations numbered 90 from V1, 28 from LM, 32 from LI, and 28 from LL.

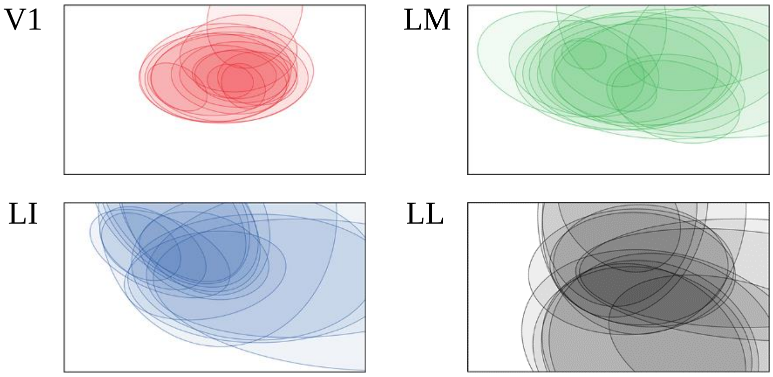

Figure 51. Example population receptive fields

Each panel is showing one of the ten populations in the four areas we investigated. Note that most receptive fields cover at least partially the central area of the image.

DECODING ANALYSIS

The decoding analysis consisted of training classifiers to discriminate between frames i and j belonging to two different movie segments containing, respectively, object O_1 (frames $i = 1 \ldots M_1$) and object O_2 (frames $j = 1 \ldots M_2$), respectively, where M_1 and M_2 are the lengths of the two segments. For every pair of frames i and j, a classifier was built.

The data fed to each of the $M_1 \times M_2$ classifiers comprised the spike count responses of a population of N neurons to P presentations of the frames i and j (see Figure 26B for the spike counts of one neuron). The binning windows Δt, in which spikes were counted, were the duration of one frame (33 ms) or three frames (100 ms), which will be called in the following *small bins* or *large bins*, respectively (Figure 26B). Each presentation of a frame resulted in an array with a dimensionality of $N \times 1$ that corresponds to one point in a high-dimensional space, whose axes represent the responses of the neurons of the population; all P repetitions of a frame formed a cloud of points in this space (Figure 2 and Figure 52A). In our analysis $N = 15$ and $P = 30$, therefore each i, j pair produced two clouds of 30 points in a 15-dimensional space (Figure 52A).

Our procedure for generating populations did not specifically consider whether neurons had been recorded simultaneously or in separate sessions. Obviously, in the case of simultaneously recorded neurons, noise (i.e., trial-by-trial) correlations can affect the representation. To avoid inhomogeneities, given that we randomly mixed neurons recorded in the same and different sessions, we destroyed all noise cross-correlations by shuffling the responses of each neuron in a population across trials (sampling trials with replacement for each neuron), thus obtaining neuronal pseudo-population vectors (Figure 52B). This approach is typically applied when population decoding analyses are utilizing data that were not all recorded simultaneously (Hung *et al.*, 2005; Tafazoli *et al.*, 2017).

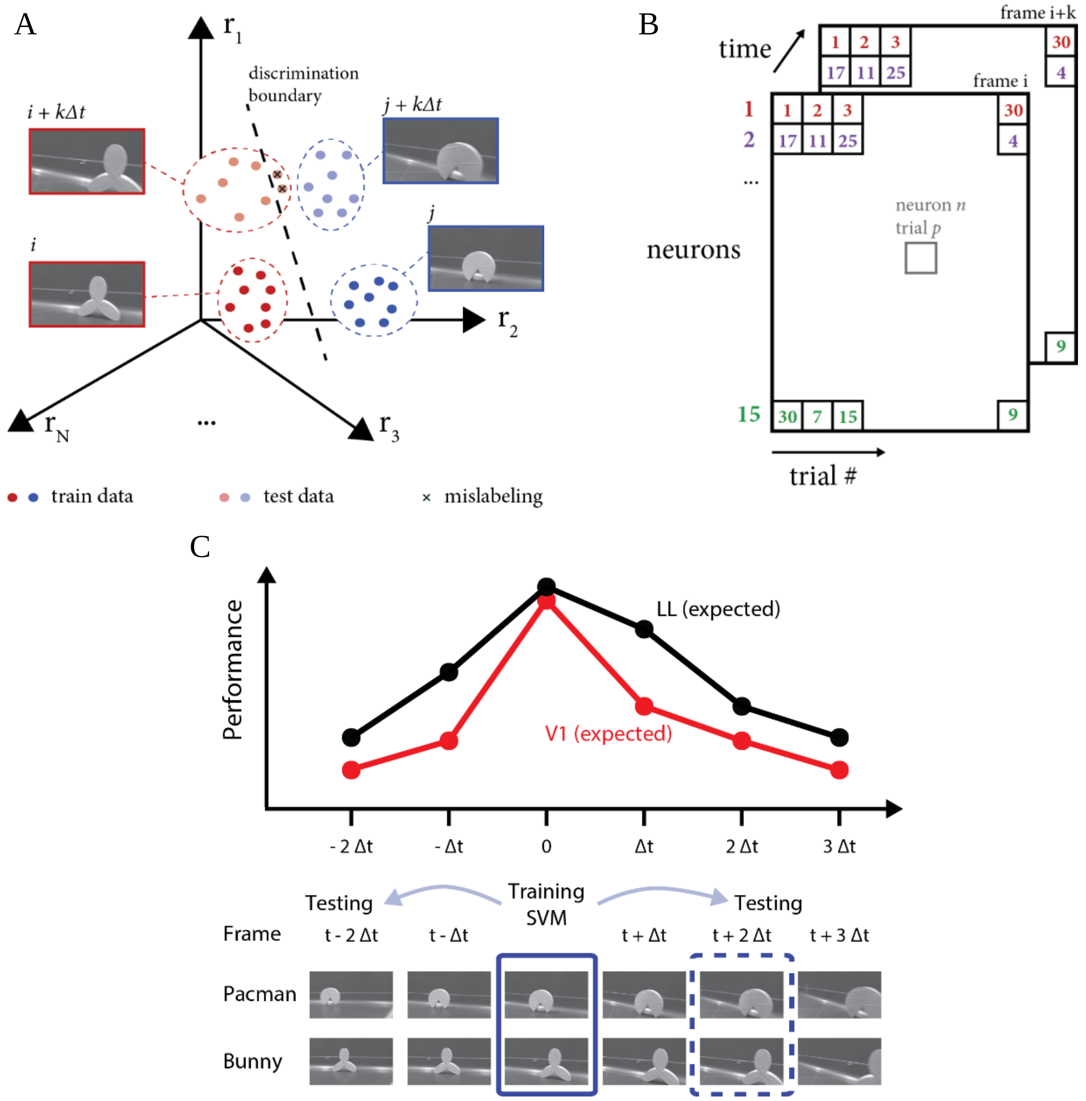

Figure 52. Population decoding

A. The schematic representation of the training and testing procedures in our decoding analysis. The response patterns of a population of N neurons (axes r_n) to the repeated presentation of a frame can be represented in a high-dimensional space as a cloud of points (dots). Training a classifier to discriminate between the responses to frames i and j (red and blue dots) amounts to finding a hyperplane that separates the two classes of responses (dashed line for simplicity). Testing the classifier on the adjacent frames $i + k\Delta t$ and $j + k\Delta t$ amounts to calculating the proportion of correct classifications over the total number of decisions (light red and blue dots). In this case the performance is high, at ~0.88, as only 2/16 datapoints have been mislabeled.

B. The data fed to the classifiers consisted of arrays that stored the responses of one population of neurons (rows, from 1 to 15) to multiple trials (columns, from 1 to 30). The order of trials was shuffled for each neuron within a single frame. Note that neurons 1, 2 and 15 have different orders of the 30 trials which are maintained across frames.

C. The main prediction of this analysis is that training a classifier to discriminate a given pair of frames (solid rectangle) and then testing to discriminate an adjacent pair of frames (dashed rectangle) will result in a decrease of performance (toy red and black traces), which will be different between areas. More specifically, we expect neuronal populations from LL (black trace) to maintain high classification performances for pairs of frames placed at longer time intervals $k\Delta t$ with respect to the training pair, compared to other areas (such as V1, red trace): compare red and black traces at positive and negative lags Δt.

For each i, j pair, a Support Vector Machine (SVM, with a linear kernel and the penalty parameter C set to 1, implemented in the Statistics and Machine Learning Toolbox, MATLAB) was trained to discriminate the two frames with a 10-fold cross-validation procedure, that consists in splitting the data in 10 subsets and using, in each cross-validation loop, nine of them for training a decoder – which amounts to finding a hyperplane that would best separate the two clouds of data points –, and one for testing the decoder – i.e. testing its ability to correctly label the data points in the remaining subset. Performance was measured as the proportion of correct classification decisions in the testing set and it ranges from 0 to 1, where 1 is perfect classification and 0,5 is chance. The average performance across the cross-validation loops for an i, j pair will be called in the following the *cross-validated performance* (Figure 53B). This performance is indicative of how well the individual frames i and j are decodable in the first place and sets an upper bound on the ability of the population to generalize to different frames containing the same objects.

In a similar way, for each i, j pair, an SVM was trained on the full data (i.e. all trials) and then tested on the adjacent frames before or after the training pair, i.e. on pairs $i - k\Delta t, j - k\Delta t$ and $i + k\Delta t, j + k\Delta t$, respectively, where Δt is the bin size, and k is a counter ranging $k = 1 \ldots 10$ for the 33 ms binning, and $k = 1 \ldots 3$ for the 100 ms binning. The performance on this test will be called in the following the *generalization performance* and will be dependent on the time lag $k\Delta t$ between training and testing pairs of frames (Figure 52C and Figure 53). In order to obtain an estimate of the population discrimination for a given object pair O_1 and O_2, we averaged the generalization performance curves obtained from various choices of the training frames i, j at matching lags (Figure 53). This averaging procedure will be extensively used throughout the analysis, as it's giving an estimate of the decoding performance of one neuronal population, tested on one pair of objects, within one single area.

Not all pairs of frames allowed testing bins both before and after the training pair (such as pairs 1,1, or $1, M_1$, or M_1, M_2), because this would have meant including frames outside of the movie segments that contained one object per image (Figure 53A).

Generalization performance at a given time lag $i \pm k\Delta t, j \pm k\Delta t$ is providing an absolute measure of how well the classifier i, j is performing when tested on unseen data (i.e., from adjacent frames); cross-validated performance instead shows what is the maximal discriminability at the very same lag. In order to quantify how well the classifier i, j is performing with respect to the available discriminability at the given lag, one needs to divide the two performances. In the following, the proportion of performance achieved by classifier i, j at time lag $\pm k\Delta t$, calculated as the ratio between generalization and cross-validated performances, will be referred to as *normalized performance* (Figure 53B).

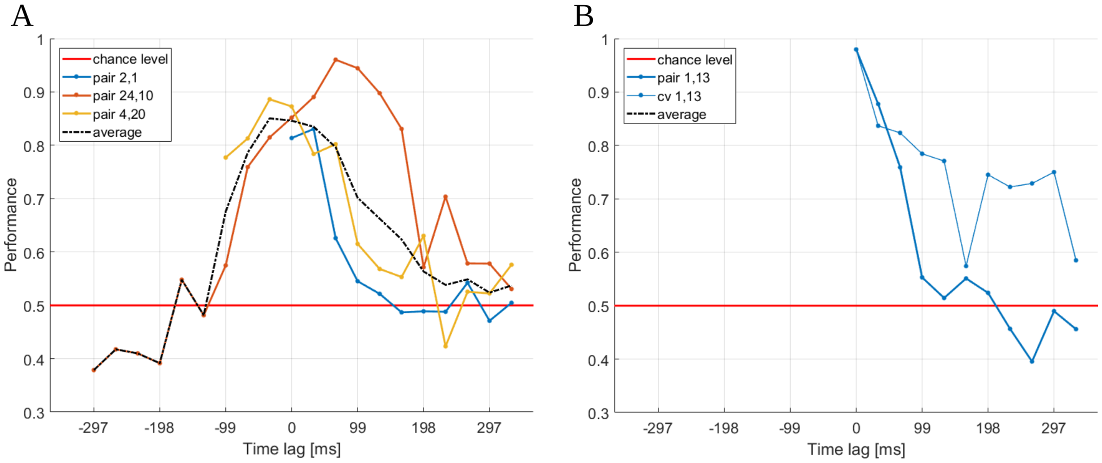

Figure 53. Example performance traces

This figure refers to the pair of black objects bunny – hash (made of 46 and 44 frames, respectively).

A. Each of the 3 (out of $46 \times 44 = 2024$) solid traces represents the generalization performance of the classifier trained with the data corresponding to the frames given in the legend. The value at lag 0 is the cross-validated performance for the same pair of frames.

Note that for some pairs (e.g. 2,1) only the pairs *after* the training pair can be tested (10 positive lags). In the case of the training pair 4,20, the following pairs of frames were tested: from the leftmost dot before lag 0: (1,17)(2,18)(3,19), and after lag 0 (5,21)(6,22)… (14,30).

The mean is computed from the available values at each lag (dashed line), which makes the average trace overlap with the trace for pair 24,10 at negative lags.

The red horizontal line is the chance level for the binary classifier.

B. Similarly to panel A, each value of the thick trace shows the generalization performance for adjacent pairs. The thin trace instead shows the cross-validated performance for adjacent pairs: note that the two traces are actually identical at lag 0, because they both refer to pair 1,13.

The normalized performance is defined as the ratio between the values of the thick trace and the values of the thin trace, individually for each lag. In order, from the leftmost value, the pairs are: (1,13)(2,14)(3,15)…(11,23).

The chance classification performance was estimated by training the SVMs as explained above, with the only difference that the association between the object identity and the spike count data has been shuffled (Figure 55B). The testing procedure did not change. In Figure 52A this would be equivalent to shuffling the color labels between dark red and blue dots.

RESULTS

In this section we will present the decoding performances of populations of neurons from the four visual areas investigated in response to two classes of stimuli extracted from the manual movies. The first group includes five black objects translating from the center of the image to the bottom left corner of it, that can be combined in $\binom{5}{2} = 10$ pairs of objects (Figure 50). The second group includes three white objects translating approximately rightwards, which can be combined in $\binom{3}{2} = 3$ pairs (Figure 50).

Our decoding analysis is based on a large number of combinations of different variables: pairs of objects (13), bin sizes (one small and one large), and areas (4). Moreover, each pair of objects can be tested starting from roughly 900 pairs of different training frames (i.e. ~30×30 frames, Figure 49), and, for each of such pairs, 20 or 6 (depending on the bin size) additional values are obtained in the generalization test (i.e., as points of the generalization performance curves as a function of the time lag from the training frames, see Figure 53A). It becomes apparent that substantial averaging at multiple levels is needed in order to check for possible differences among the visual areas (Figure 54).

As explained in section Decoding analysis, the cross-validated performance provides an upper bound (or reference) of the performance that a population of neurons can achieve when discriminating between any two frames. At the other end, randomizing the identities of the frames during the training procedure (i.e., label-shuffling) is diagnostic of the extent to which the decoder’s outcome is due to chance. In this analysis, we considered these margins on the expected performance sufficient to make our results relevant. Additional possible controls (such as the overlap between objects, or the overlap of objects with the population receptive field) were not implemented for the present report because their effects on the final results were moderate.

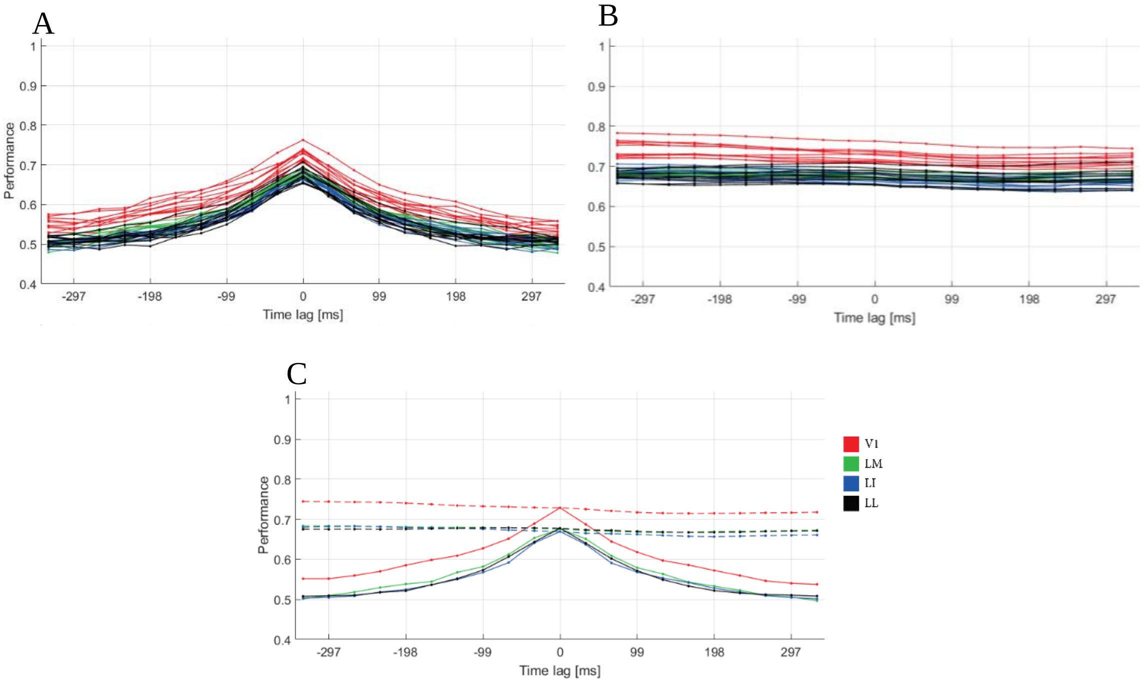

Figure 54. Decoding performance for one object pair

This figure refers to the pair of black objects bunny – hash.

A. Each trace represents the average generalization performance (i.e., the average of 1848 pairs of frames, of which 3 are shown in Figure 53A), for one population of neurons, and one pair (black bunny – black hash). There are 10 traces per area as explained in section Selection of the populations of neurons.

B. Each trace is the average cross-validated performance (thin trace in Figure 53B) for one population (average of the same number of pairs, i.e. 1848, as in panel A).

C. Average per area (across populations) for the combined traces in panel A and B.

Black objects

The generalization performance for the 10 pairs of black objects (solid lines), along with the cross-validated performances (dashed lines), and their averages per area, are shown in Figure 55A. Additionally, Figure 55B shows the performance after training with shuffled label, which is virtually equal to the chance performance; since in all combinations the shuffled labels traces are very similar (i.e. overlapping at 0.5), they will not be plotted anymore. The fact that this type of training is producing chance performance during testing is a clear indication that the effects we are seeing are not artificially created by our SVM implementation.

To assess the slowness (i.e., the persistence) of the representation over time (i.e., at time lags increasingly farther from the training bin), the generalization performance was normalized by the cross-validated performance. The resulting normalized generalization curves (Figures 55C and 56B) show how self-similar the representations of the two objects remain over time. Critically, this normalization equates the visual areas in terms of the decoding power (as measured by the cross-validated performance) that the different populations afford over the time axis. In addition, the normalization discounts the absolute magnitude of the generalization performance, which is not relevant in order to assess the self-similarity of object representation over time. As shown in Figure 55B, V1 affords better generalization in absolute terms, but this is due to the fact that its discriminatory power starts higher in the first place, as compared to the other areas (compare the peaks of the four curves). However, the decay of the discrimination performance from the peak is faster for V1: we see that the performance is surpassed by the other areas at time lags that are very far from the training bin (i.e. the peak). Normalizing by the cross-validated performance allows assessing the variation of the generalization performance, relative to the maximal discriminability of the two objects (which is virtually flat over time).

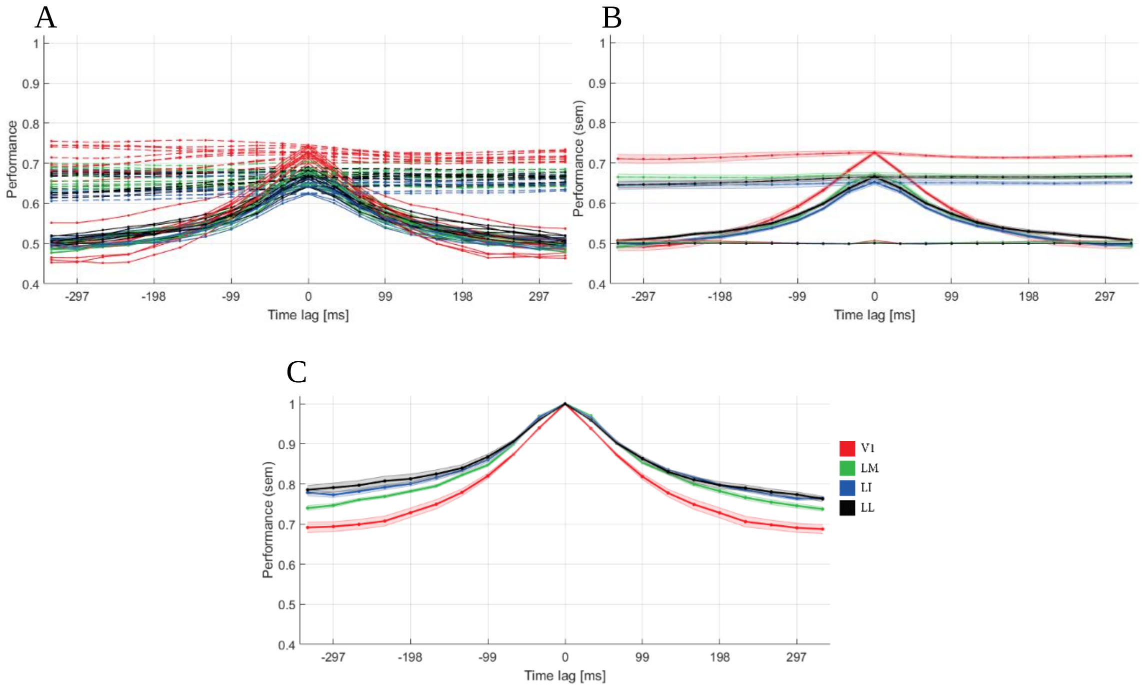

Figure 55. Decoding performance for black objects

A. Solid traces are the generalization performance for each pair of objects (in total 10, including the one in Figure 54C). Dashed traces are the cross-validated performances (Figure 53B), for each pair of objects (as an average across populations).

B. Average performance and SEM of the traces in panel A, obtained for each area. There are four traces that overlap at 0,5, and they represent the averages for each area of the shuffled-labels performances, which are obtained by testing classifiers that had been trained with data with random label assignment.

C. Average normalized performance obtained from the traces in panel A by dividing, for each bin, the generalization performances (solid traces) by the cross-validated performances (dashed traces) (see also Figure 53B).

V1 populations produced significantly higher cross-validated performances than all other areas, at both bin sizes of 33 ms (Figure 55B, values at lag 0) and 100 ms (Figure 56A). This result is supported, in each case, by the main effect of a one-way ANOVA with *area* as a factor and the cross-validated performance of each population as the dependent variable (n = 400 values, i.e. 10 object pairs × 10 populations × 4 areas): $F_{3,396} = 252, \mathrm{p} < 0.001$ for the 33ms bin, and $F_{3,396} = 262, \mathrm{p} < 0.001$ for the 100ms bin. The post-hoc analyses indicate indeed that V1 performances are higher than those of all other areas (Student's t-test, p < 0.001, Bonferroni corrected).

Normalized performances in Figures 55C and 56B, for the two bin sizes (33 ms and 100 ms), show that generalization performance of the four visual areas gradually decrease from LL to V1, especially at large positive and negative lags (> 99 ms in absolute values), which means that object representations in lateral areas remain more self-similar, over time, compared to those in V1.

There are at least two ways to establish whether these observations are statistically meaningful: 1) by comparing the time constants of the decay, or 2) by comparing the values at each lag. For these tests, data at population level have been used, 400 traces in total: 10 object pairs × 10 populations × 4 areas (Figure 54A, where one pair of objects is shown); within each generalization trace, only the values at strictly positive lags were included (10 lags).

For the first approach, the normalized generalization traces were fitted with exponential curves, as described in Methods. One-way ANOVA with *area* as a factor yields a non-significant main effect at both bin sizes.

For the second approach, we compared the generalization traces by considering the values at each lag individually, in which case we used a balanced two-way mixed ANOVA with *area* as a between factor (4 levels) and *lag* as a within factor (10 levels). This analysis yielded, for small bins, significant main effects of *area* ($F_{3,396} = 155.6, \mathrm{p} < 0.001$) and *lag* ($F_{3,1080} = 8035, \mathrm{p} < 0.001$), and a significant interaction *area*lag* ($F_{8,1080} = 40.9, \mathrm{p} < 0.001$); the degrees of freedom were Greenhouse-Geisser corrected. Similar results were obtained for large bins (Figure 56).

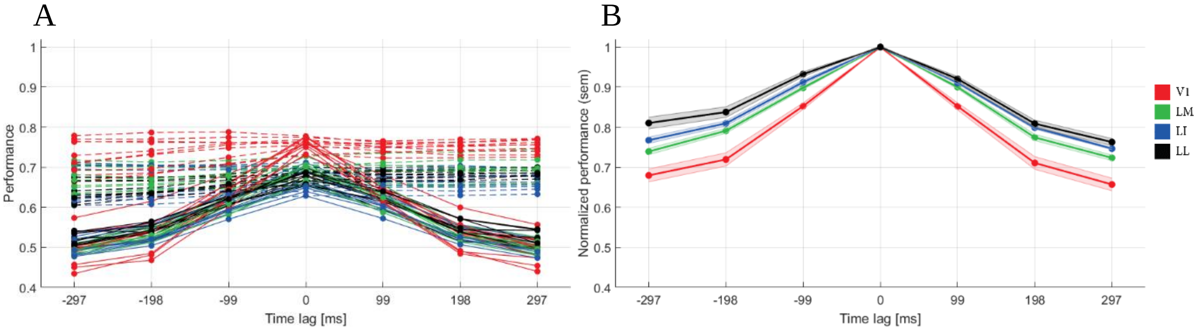

Figure 56. Decoding performance for pairs of black objects at large bins

A. Same contents as in 55A but for 100 ms bins.

B. Same contents as in 55C but for 100 ms bins.

WHITE OBJECTS

The decoding analysis for the white objects was similar to the one for black objects. In this case though, three white objects were used and they could be combined in three pairs (Figure 50).

The average generalization performances for each pair and across populations are shown in Figures 57A and 57C, for the two bin sizes, respectively. As in the case of black objects, V1 cross-validated performances (i.e., at lag 0) are higher than those of all other areas. A one-way ANOVA, similar to the one described above, with *area* as a factor and the cross-validated performance of each population as the dependent variable ($\mathrm{n} = 120$ values, i.e. 3 object pairs × 10 populations × 4 areas), produces a significant main effect of *area* ($F_{3,116} = 222, \mathrm{p} < 0.001$ for the small bin, and $F_{3,116} = 171, \mathrm{p} < 0.001$ for the large bins); the post-hoc analyses confirm that the main effect is driven by V1 performances being significantly higher than those of other three areas (Student's t-test, $\mathrm{p} < 0.001$, Bonferroni corrected).

The normalized performances of Figures 57B and 57D show the same effect as in the case of black objects, i.e. generalization performances of LL populations decrease at shallower slopes, compared to V1. For these pairs, this observation is confirmed by two one-way ANOVAs (one for each bin size) performed on the time constants of the exponential fits of each population trace ($\mathrm{n} = 120$): $F_{3,116} = 5.44, \mathrm{p} = 0.001$ for the small bins, and $F_{3,116} = 3.24, \mathrm{p} = 0.04$ for the large bins.

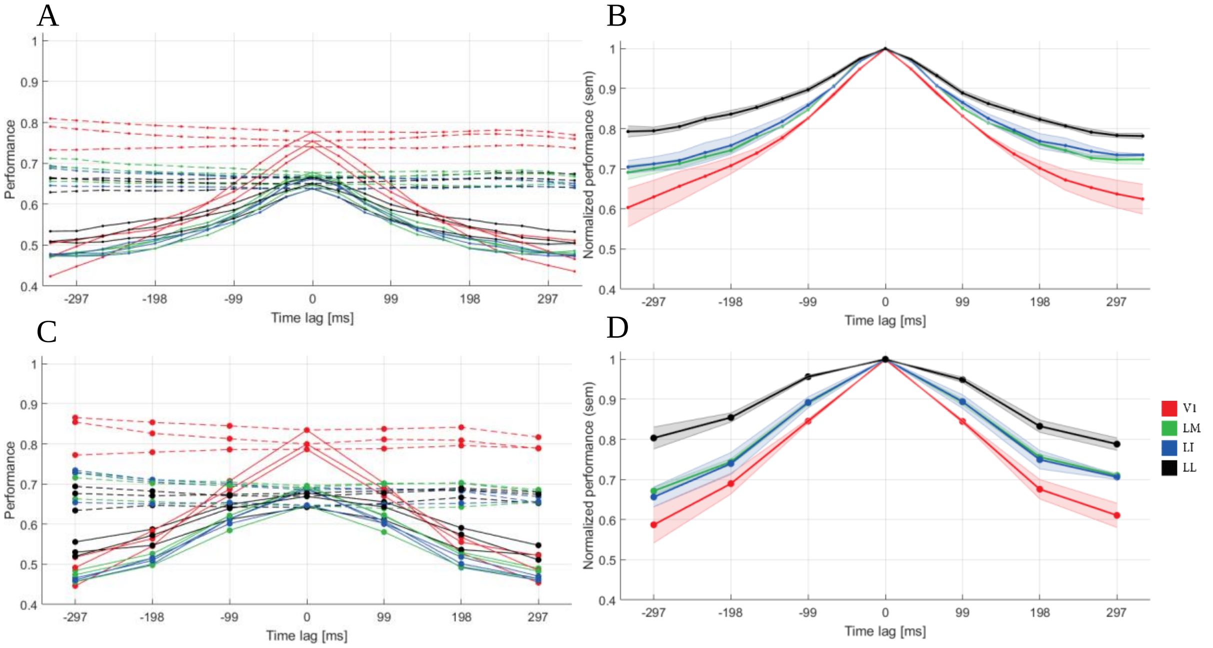

Figure 57. Decoding performance for pairs of white objects

A, C. Generalization (solid traces) and cross-validated performances (dashed traces) for 3 pairs of white objects for small (panel A) and large (panel C) bin sizes.

The contents are similar to those in Figures 55A and 56A.

B, D. Normalized performances obtained from the traces of panels A and C. Similar to Figures 55C and 56B.

A two-way mixed ANOVA (identical to the one described for the black objects) yielded, for small bins, significant main effects of *area* ($F_{3,116} = 97.3, \mathrm{p} < 0.001$) and *lag* ($F_{2,239} = 2467, \mathrm{p} < 0.001$), and a significant interaction *area*lag* ($F_{6,239} = 33.5, \mathrm{p} < 0.001$); the degrees of freedom were Greenhouse-Geisser corrected. Similar significant results were obtained for the large bins.

THE EFFECT OF BIN SIZES ON THE DECODING PERFORMANCE

Until now, we showed results separately for each bin size: 33 ms and 100 ms, which are the widths of the spike counting windows used in our analyses (section Decoding analysis). By directly comparing the performances obtained in each case, we can assess whether the bin size itself is important for how well classifiers generalize. Figure 58A shows the normalized generalization curves obtained for the white objects at both small and large bins. Since the large bin is three times larger than the small bin, the positive and negative lags 99, 198 and 297 ms are common between the curves corresponding to the two bin sizes. This allows comparing the performances obtained at these three time lags, within each area, with the small and large time bins. As it can be appreciated by looking at Figure 58A, the differences between such performances (e.g., between curves of the same color at lags 99, 198, and 297) diminishes at longer lags (for example, compare black traces at negative lags 99 and 297). More importantly, such differences are not the same across the four areas: they are much larger in LL and LI, than in V1 (especially at the shortest lag of 99 ms).

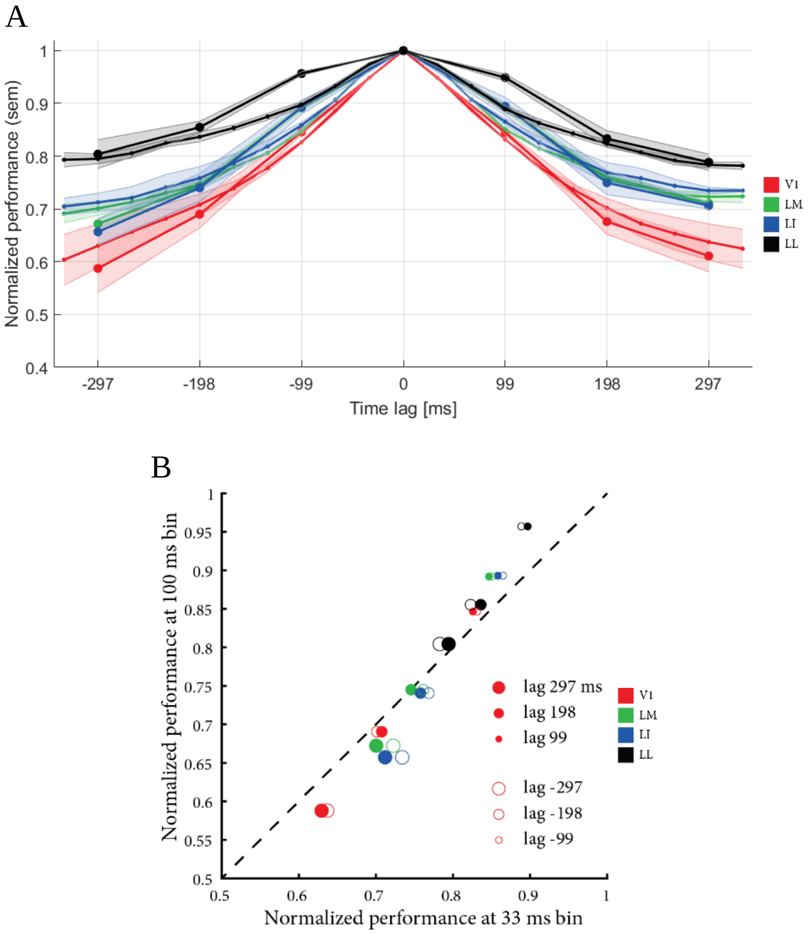

Figure 58. Performance relative to the size of spike count window

A. Normalized performance for pairs of white objects and two bin sizes: 33 and 100 ms. Identical to those in Figures 57B and D.

B. The scatter plot shows the normalized performance at large bins relative to the normalized performance at small bins, for 4 areas and 6 matching positive and negative lags (±99, ±198, ±99, empty and filled dots, respectively). Larger dots indicate longer lags. Values above the identity line (dashed line) show lags for which the performance at 100ms bin is larger than that at 33ms.

Figure 58B shows the comprehensive comparison between the performances at different bin sizes. The dots corresponding to short lags (i.e. 99 ms) are all above the identity line, which means that generalization performance is better when decoders are trained with data from larger bins. Importantly, this effect is particularly strong for LL (black dots) and, to a lesser extent, for LI and LM, than for V1. The dots corresponding to long lags (i.e. 198 or 297 ms) are slightly above the identity line only for LL (thus showing slight improvement in performance when information is integrated over longer time windows). The longer lags for other areas are either on the identity line (no improvement) or below it (and close to the chance level: far left and right dots in Figures 57A and 57C, that show the non-normalized performances).

Similar results were obtained for black objects (not shown).

Overall, the results observed in the previous two sections for the pairs of black and white objects show that there is a gradual increase in the stability of decoding performances along the four visual areas. This indicates a more persistent (i.e. more self-similar) representation over time in lateral areas (especially LL), as compared to V1. We also note that V1 populations are always better, in absolute values, than all other areas at discriminating pairs of objects. This effect could be due to the high reproducibility of V1 neurons (Figure 28) – which increases the performance at lag 0 obtained through a cross-validation procedure –, and to their strong preference for low-level features (such as luminance), that facilitate the discrimination (Tafazoli *et al.*, 2017). At a qualitative level, large spike counting windows seem to especially benefit populations in the lateral areas and at relatively short lags (99 ms), as their performances increase with bin size (Figure 58). All of these results indicate that the timescales of the representations of visual objects become gradually slower from V1 to LL.

Slowness of movie segments

As explained in section Decoding analysis, the results described in this chapter are based on movie segments that contain only one specific object, whereas the previous chapter characterizes the dynamics of full-length movies, i.e., including empty frames or frames with multiple objects.

By repeating the analysis of the first chapter of results (section Slowness), but for the short movie segments, we can assess to what extent the effects we observe in this chapter are due to the self-similarity of the trial-averaged signal (PSTH) or to noise autocorrelations. In this analysis, only the neurons included in the 40 populations were used. Figure 59 shows that the average time constants of the decay of the PSTH (panel A) are approximately equal and around 70 ms. A one-way unbalanced ANOVA with *area* as a factor yields a non-significant main effect ($p = 0.22$).

In the case of noise autocorrelations (panel B), the time constants obtained from LL responses are significantly higher than all other three areas. A similar one-way ANOVA produces a significant main effect of *area* ($F_{3,1420} = 21.4, \mathrm{p} < 0.001$), and the post-hoc analyses confirm that all pairwise comparisons with LL are statistically significant (Student's *t* test, $\mathrm{p} < 0.01$, Holm-Bonferroni corrected) and those between V1, LM and LI are not.

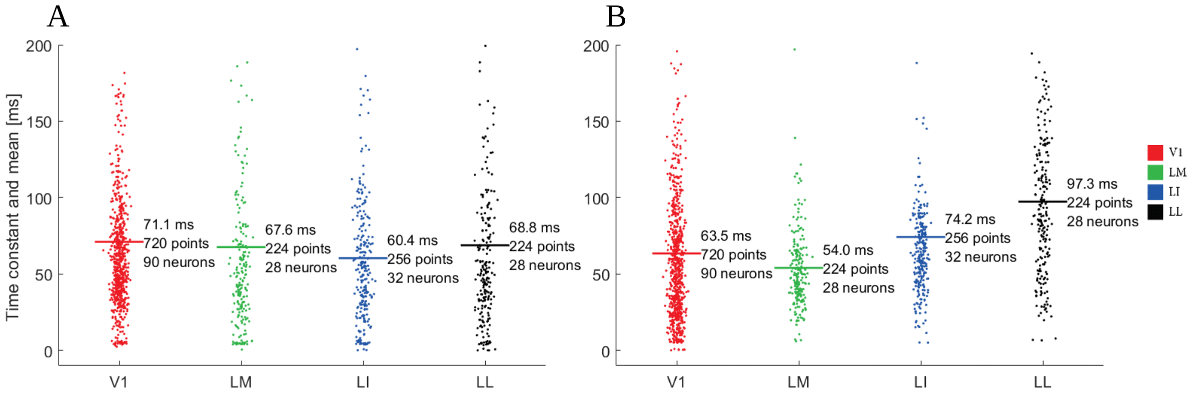

Figure 59. Time constants of the PSTH and noise autocorrelation of movie segments

The two panels of this figure refer to the same data, i.e. neural responses during the movie segments for the 8 objects included in this analysis (Figure 50A). The responses belong to the unique neurons that make up the 10 populations within each area, without repetition (i.e., 90 for V1, 28 for LM, 32 for LI, and 28 for LL; described in section Selection of the populations of neurons).

Dots represent the time constants for a given object and a given neuron (note that the number of points = number of objects (8) × number of neurons in each area). Horizontal lines represent mean time constants.

A. Time constants of the exponential fits of the PSTHs computed from short movie segments (the duration in frames is given in Figure 49).

B. Time constants of the exponential fits of the noise autocorrelations computed from the spike counts of all trials of the movie segments.

These results suggest that stimulus-independent properties of neurons (i.e. noise autocorrelations), not the stimulus-dependent one, are responsible for the pattern of decoding performances we saw in this chapter. We should note again that, even though the neurons in the two analyses are the same, the spike counts fed to the SVM classifiers were shuffled across trials within each neuron (Figure 52B): in practice, this means that the decoders have been trained with pseudo-data obtained by mixing responses from multiple neurons.

5. DISCUSSION

In this study we have investigated the representation of movies in rat visual cortex. We recorded neurons from four visual cortical areas, while animals were passively exposed to stimuli with various spatio-temporal properties (objects undergoing various transformations at different speeds, or noisy content). To our knowledge, this is the first study that compares how neuronal populations in multiple cortical areas represent movies with very different dynamical properties. Other studies have focused on one area (Kayser *et al.*, 2003; Froudarakis *et al.*, 2014; Rikhye and Sur, 2015), on areas from different species (Baddeley *et al.*, 1997), or have used natural and artificial movies with similar properties (Vinken *et al.*, 2016).

In the chapter “Representation of natural movies in rat visual cortex”, section “Statistics of response properties across visual areas”, we looked at two basic properties of the neural activity of each neuron: the trial-by-trial reproducibility (or reliability), and the firing rate.

We found that **reproducibility** is strongly influenced by the speed of the content of each movie (e.g. ratcam or white noise movies, Figure 28). This effect has been described in the literature in the context of temporal precision: it was reported that precision depends on the frequency content of the stimulus and it was suggested that spiking precision at timescales shorter than those of the stimulus are necessary to accurately represent natural stimuli (Butts *et al.*, 2007; Kayser *et al.*, 2010).

Similarly to previous reports (Froudarakis *et al.*, 2014; Vinken *et al.*, 2016), we see a separation between natural and artificial movies: i.e., responses to natural movies are more reliable than those to artificial ones. Two possible interpretations can explain this result: first, neurons are preferably activated by natural movies, so that the stimulus-driven component of their firing is larger than the internal noise; second, neurons reduce variability of their spiking so as to ensure efficient coding (Simoncelli and Olshausen, 2001).

Neurons were always significantly driven by the visual input, with respect to their baseline activity, and their **firing rates** were modulated by the speed of the movies: in particular, the ratcam ones evoked the strongest activations (Figure 29). V1 had larger firing rates than the other areas, but from our data we cannot ascertain whether this was due to unbalanced distributions of single and multi units between areas, or V1 neurons fire more in general.

In the chapter "Representation of natural movies in rat visual cortex", section "Sparseness", we calculated **sparseness** for each neuron and movie and plotted it against the speed of the visual input neurons experienced in their receptive fields. We found that, for natural movies, sparseness decreased significantly along the areas' progression while, at the same time, globally increasing as a function of the speed (Figure 41). Area and speed also showed a significant interaction, driven by the large increase of sparseness in LM at high speeds. Finally, sparseness was lower for the phase-scrambled and white noise movies, despite their large speed values (Figure 42). As already explained (page 60), the reduction in sparseness that we observe for the recorded neurons is not consistent with the sparse coding theory – that predicts an increase along the ventral stream (Olshausen and Field, 2004) –, but alternatively it could be accounted for by an increase of invariance (slowness) from V1 to LL (Rust and DiCarlo, 2012).

Indeed, **slowness** increased along the areas' progression, for most speed bins (Figures 44 and 45). Because of the way it is calculated (as the decay of the autocorrelation function of the PSTH), slowness reflects the dynamics of the stimulus that evokes the responses: hence, as expected, we observe that time constants decreased with the speed of the movie. Surprisingly, the pattern across areas was maintained for both natural and artificial movies, whereas in the efficient coding paradigm we expect invariance to be instantiated only for the stimuli to which the visual cortex is adapted.

Our second slowness metric, **noise autocorrelation**, was calculated on the detrended signal, so it is less dependent on the dynamics or content of the stimuli: we note that noise time constants vary much less than PSTH time constants with the speed of the movies (Figures 46 and 47). At each speed level there is a strong trend of increase of noise time constants across areas' progression. This pattern reflects the intrinsic timescales of neural variability in each visual area, and suggests the existence of a hierarchical ordering and specialization between areas (Murray *et al.*, 2014).

In the chapter “Population decoding” we implemented a **population decoding** technique to assess whether object representations are more persistent over time in some areas than others. To this end, we trained classifiers to discriminate between instances of two objects undergoing transformations and tested on farther apart instances of the same objects: the decay of the generalization performance (normalized by the decoding power, explained on page 77) was then indicative of the slowness of the representation.

For both pairs of black and white objects (Figures 55-57), at small and large bins, we found that LL affords high **generalization performance** for longer durations than the other areas, and much higher than V1, in particular. This effect was stronger in the case of white objects (Figure 57): the difference between black and white objects (for example, Figures 56B and 57D for large bins), was due both to shallower decay of LL generalization power and to sharper decay of V1 generalization power. Although we are not showing it here, from preliminary investigations, it results that the stronger effect for white objects could be ascribed to luminance differences between objects within each pair. Specifically, the two objects that created a pair on which the classifiers were build (e.g., Figure 52C) had different amounts of brightness at pixel level, which was enough to evoke different responses in the neurons making up the populations, therefore facilitating the discriminability (white objects’ pixel intensities ranged between ~100-200, whereas black objects ranged between ~0-15, for a maximum range of 0-255). This could also explain why **V1 populations** outperformed the other areas, in absolute value (lag 0 in Figures 55A, 56A, 57A,C): Tafazoli and colleagues (2017) have shown that V1 neurons are more sensitive to low-level stimulus features, in particular luminance, than other visual areas.

We compared the performances obtained with two **different spike counting bins** (33 and 100 ms) and found that LM, LI, and LL performances increase with the bin size (Figure 58), but only at small lags (±99 ms). The fact that larger bin sizes boost generalization performance in some areas suggest that larger integration windows are more relevant for those specific populations. This result could be further confirmed by expanding the analysis to additional bin sizes (such as 66 or 132 ms).

Finally, we linked the slowness of single-cell responses to the population decoding analysis and showed that a possible source for the generalization performance is the intrinsic timescale of the response, as measured by the decay of noise autocorrelations (Figure 59).

In the following, we describe possible **issues** of this study.

The data on which our results are based are collected from **anesthetized animals**: one of the reasons for this choice is that our experimental design requires lengthy and stable recording sessions; another reason is the need to avoid top-down and learning effects, which are detrimental to our goal to probe the feed-forward visual computations at the first stages of visual processing. While we don't exclude anesthesia influences on our measures (sparseness and slowness), we discount a complete corruption of the effects we presented, as some of the results in another study with identical experimental paradigm were no different from those obtained in awake animals (discussed in Tafazoli *et al.*, 2017).

We explained in section "Sparseness and slowness" that the **position of receptive fields** of our neurons varied between areas (Figure 30). It's important to mention that controlling for the speed of the input might not be enough for a fair comparison. Various studies have shown that contextual effects are modulating the firing rate distributions of individual neurons (Vinje and Gallant, 2000) and affecting the developmental refinement of visual processing (Pecka *et al.*, 2014); therefore, neurons from different areas should ideally be matched for their receptive field position, so as to ensure they receive the same amount of surround stimulation and as similar as possible visual content.

Figure 29 shows that **firing rates** were larger in V1 than in other three areas. We didn't quantify whether and how this affected our sparseness and slowness results. Selecting sub-populations of neurons with matched firing rates could reduce possible confounds.

In the **population decoding analysis**, the cross-validated performance is providing an upper bound of the decoding power a population can afford. The analysis could be further improved by making sure that, within each movie segment, objects do overlap between themselves and with the population receptive field. Additionally, to assess how specific our results are to the objects we selected (Figure 50), we can train and test classifiers on random movie segments (that either contain no objects or multiple objects), or on segments extracted from phase-scrambled movies.

Taken together, our results confirm the specialization of visual neurons for the sparse coding of natural scenes, not only in V1 but also in higher-level areas. They also support the attribution of LM, LI and LL to an object-processing pathway, given the gradual decrease of sparseness and increase of slowness that were found across these areas, which are consistent with the growth of invariance expected for ventral stream representations.

REFERENCES

Alemi-Neissi A, Rosselli FB, Zoccolan D (2013). Multifeatural shape processing in rats engaged in invariant visual object recognition. *J Neurosci* **33**(14): 5939-5956.

Andermann ML, Kerlin AM, Roumis DK, Glickfeld LL, Reid RC (2011). Functional specialization of mouse higher visual cortical areas. *Neuron* **72**(6): 1025-1039.

Attwell D, Laughlin SB (2001). An energy budget for signaling in the grey matter of the brain. *J Cereb Blood Flow Metab* **21**(10): 1133-1145.

Baddeley R, Abbott LF, Booth MC, Sengpiel F, Freeman T, Wakeman EA, Rolls ET (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc Biol Sci* **264**(1389): 1775-1783.

Barlow H (2001). Redundancy reduction revisited. *Network* **12**(3): 241-253.

Barlow HB (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*: 217-234.

Berkes P, Wiskott L (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *J Vis* **5**(6): 579-602.

Betsch BY, Einhauser W, Kording KP, Konig P (2004). The world from a cat's perspective--statistics of natural videos. *Biol Cybern* **90**(1): 41-50.

Bourke P (1992, updated August 2002). "Rotate a point about an arbitrary axis (3 dimensions)." from paulbourke.net/geometry/rotate/.

Brooks DI, Ng KH, Buss EW, Marshall AT, Freeman JH, Wasserman EA (2013). Categorization of photographic images by rats using shape-based image dimensions. *J Exp Psychol Anim Behav Process* **39**(1): 85-92.

Burn CC (2008). What is it like to be a rat? Rat sensory perception and its implications for experimental design and rat welfare. *Applied Animal Behaviour Science* **112**(1-2): 1-32.

Butts DA, Weng C, Jin J, Yeh CI, Lesica NA, Alonso JM, Stanley GB (2007). Temporal precision in the neural code and the timescales of natural vision. *Nature* **449**(7158): 92-95.

Coogan TA, Burkhalter A (1993). Hierarchical organization of areas in rat visual cortex. *J Neurosci* **13**(9): 3749-3772.

Cox DD, Meier P, Oertelt N, DiCarlo JJ (2005). 'Breaking' position-invariant object recognition. *Nat Neurosci* **8**(9): 1145-1147.

Cox DD (2014). Do we understand high-level vision? *Curr Opin Neurobiol* **25**: 187-193.

Dan Y, Atick JJ, Reid RC (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J Neurosci* **16**(10): 3351-3362.

DiCarlo JJ, Cox DD (2007). Untangling invariant object recognition. *Trends Cogn Sci* **11**(8): 333-341.

DiCarlo JJ, Zoccolan D, Rust NC (2012). How does the brain solve visual object recognition? *Neuron* **73**(3): 415-434.

Dong DW, Atick JJ (1995). Statistics of Natural Time-Varying Images. *Network-Computation in Neural Systems* **6**(3): 345-358.

Espinoza SG, Thomas HC (1983). Retinotopic organization of striate and extrastriate visual cortex in the hooded rat. *Brain research* **272**(1): 137-144.

Felleman DJ, Van Essen DC (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex* **1**(1): 1-47.

Felsen G, Dan Y (2005). A natural approach to studying vision. *Nat Neurosci* **8**(12): 1643-1646.

Földiák P (1991). Learning invariance from transformation sequences. *Neural Computation* **3**(2): 194-200.

Forwood SE, Bartko SJ, Saksida LM, Bussey TJ (2007). Rats spontaneously discriminate purely visual, two-dimensional stimuli in tests of recognition memory and perceptual oddity. *Behav Neurosci* **121**(5): 1032-1042.

Fraedrich EM, Glasauer S, Flanagin VL (2010). Spatiotemporal phase-scrambling increases visual cortex activity. *Neuroreport* **21**(8): 596-600.

Franzius M, Wilbert N, Wiskott L (2008). Invariant Object Recognition with Slow Feature Analysis. *Artificial Neural Networks - Icann 2008, Pt I* **5163**: 961-970.

Froudarakis E, Berens P, Ecker AS, Cotton RJ, Sinz FH, Yatsenko D, Saggau P, Bethge M, Tolias AS (2014). Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nature Neuroscience* **17**(6): 851-857.

Glickfeld LL, Olsen SR (2017). Higher-Order Areas of the Mouse Visual Cortex. *Annu Rev Vis Sci* **3**: 251-273.

Heath JC, Newland M (2006). Effects of methylmercury on the critical fusion frequency of rats. NEUROTOXICOLOGY, Elsevier Science BV PO BOX 211, 1000 AE Amsterdam, Netherlands.

Huberman AD, Niell CM (2011). What can mice tell us about how vision works? *Trends Neurosci* **34**(9): 464-473.

Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**(5749): 863-866.

Hyvarinen A, Hurri J, Hoyer PO (2009). *Natural Image Statistics - A probabilistic approach to early computational vision*, Springer.

Kayser C, Salazar RF, Konig P (2003). Responses to natural scenes in cat V1. *J Neurophysiol* **90**(3): 1910-1920.

Kayser C, Logothetis NK, Panzeri S (2010). Millisecond encoding precision of auditory cortex neurons. *Proc Natl Acad Sci U S A* **107**(39): 16976-16981.

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C (2007). What's new in Psychtoolbox-3. *Perception* **36**(14): 1.1-16.

Kriegeskorte N (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science* **1**: 417-446.

labrigger.com (2012, updated 29 August 2012). "Visual stimuli for mice." from labrigger.com/blog/2012/03/06/mouse-visual-stim/.

Laughlin S (1981). A simple coding procedure enhances a neuron's information capacity. *Z Naturforsch C* **36**(9-10): 910-912.

Li N, DiCarlo JJ (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* **321**(5895): 1502-1507.

Li N, DiCarlo JJ (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* **67**(6): 1062-1075.

Marshel JH, Garrett ME, Nauhaus I, Callaway EM (2011). Functional specialization of seven mouse visual cortical areas. *Neuron* **72**(6): 1040-1054.

Minini L, Jeffery KJ (2006). Do rats use shape to solve "shape discriminations"? *Learn Mem* **13**(3): 287-297.

Mishkin M, Ungerleider LG, Macko KA (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences* **6**: 414-417.

Murray JD, Bernacchia A, Freedman DJ, Romo R, Wallis JD, Cai X, Padoa-Schioppa C, Pasternak T, Seo H, Lee D, Wang XJ (2014). A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci* **17**(12): 1661-1663.

Niell CM (2011). Exploring the next frontier of mouse vision. *Neuron* **72**(6): 889-892.

Nikolic D, Haeusler S, Singer W, Maass W (2006). Temporal dynamics of information content carried by neurons in the primary visual cortex. *Trial* **20**.

Olshausen BA, Field DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**(6583): 607-609.

Olshausen BA, Field DJ (2000). Vision and the coding of natural images. *American Scientist* **88**(3): 238-245.

Olshausen BA, Field DJ (2004). Sparse coding of sensory inputs. *Curr Opin Neurobiol* **14**(4): 481-487.

Paxinos G, Watson C (2007). *The rat brain in stereotaxic coordinates*, Academic Press.

Pecka M, Han Y, Sader E, Mrsic-Flogel TD (2014). Experience-Dependent Specialization of Receptive Field Surround for Selective Coding of Natural Scenes. *Neuron*.

Phung SL (2004, updated 2 April 2004). Simulating Photoshop's magic wand tool. mathworks.com/matlabcentral/fileexchange/4698-simulating-photoshop-s-magic-wand-tool, MathWorks File Exchange.

Prusky GT, West PW, Douglas RM (2000). Behavioral assessment of visual acuity in mice and rats. *Vision Res* **40**(16): 2201-2209.

Prusky GT, Harker KT, Douglas RM, Whishaw IQ (2002). Variation in visual acuity within pigmented, and between pigmented and albino rat strains. *Behavioural Brain Research* **136**(2): 339-348.

Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput* **16**(8): 1661-1687.

Ratliff CP, Borghuis BG, Kao YH, Sterling P, Balasubramanian V (2010). Retina is structured to process an excess of darkness in natural scenes. *Proc Natl Acad Sci U S A* **107**(40): 17368-17373.

Rikhye RV, Sur M (2015). Spatial Correlations in Natural Scenes Modulate Response Reliability in Mouse Visual Cortex. *J Neurosci* **35**(43): 14661-14680.

Rolls ET, Tovee MJ (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J Neurophysiol* **73**(2): 713-726.

Rossant C, Kadir SN, Goodman DF, Schulman J, Hunter ML, Saleem AB, Grosmark A, Belluscio M, Denfield GH, Ecker AS, Tolias AS, Solomon S, Buzsaki G, Carandini M, Harris KD (2016). Spike sorting for large, dense electrode arrays. *Nat Neurosci* **19**(4): 634-641.

Rust NC, Movshon JA (2005). In praise of artifice. *Nat Neurosci* **8**(12): 1647-1650.

Rust NC, DiCarlo JJ (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J Neurosci* **30**(39): 12978-12995.

Rust NC, DiCarlo JJ (2012). Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. *J Neurosci* **32**(30): 10170-10182.

Sefton AJ, Dreher B, Harvey A (2004). Visual system. The Rat Nervous System. Paxinos.

Sereno AB, Maunsell JH (1998). Shape selectivity in primate lateral intraparietal cortex. *Nature* **395**(6701): 500-503.

Sereno MI, Allman JM (1991). Cortical visual areas in mammals. The Neural Basis of Visual Function (Vision and Visual Dysfunction). J. Cronly-Dillon, A. Leventhal.

Simoncelli EP, Olshausen BA (2001). Natural image statistics and neural representation. *Annu Rev Neurosci* **24**: 1193-1216.

Simoncelli EP (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology* **13**(2): 144-149.

Tafazoli S, Di Filippo A, Zoccolan D (2012). Transformation-tolerant object recognition in rats revealed by visual priming. *J Neurosci* **32**(1): 21-34.

Tafazoli S, Safaai H, De Franceschi G, Rosselli FB, Vanzella W, Riggi M, Buffolo F, Panzeri S, Zoccolan D (2017). Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex. *Elife* **6**.

Tees R (1999). The effects of posterior parietal and posterior temporal cortical lesions on multimodal spatial and nonspatial competencies in rats. *Behavioural Brain Research* **106**(1-2): 55-73.

Treves A, Rolls ET (1991). What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems* **2**(4): 371-397.

Vermaercke B, Gerich FJ, Ytebrouck E, Arckens L, Op de Beeck HP, Van den Bergh G (2014). Functional specialization in rat occipital and temporal visual cortex. *J Neurophysiol*.

Vermaercke B, Van den Bergh G, Gerich F, Op de Beeck H (2015). Neural discriminability in rat lateral extrastriate cortex and deep but not superficial primary visual cortex correlates with shape discriminability. *Front Neural Circuits* **9**: 24.

Vinje WE, Gallant JL (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**(5456): 1273-1276.

Vinken K, Vermaercke B, Op de Beeck HP (2014). Visual Categorization of Natural Movies by Rats. *Journal of Neuroscience* **34**(32): 10645-10658.

Vinken K, Van den Bergh G, Vermaercke B, Op de Beeck HP (2016). Neural Representations of Natural and Scrambled Movies Progressively Change from Rat Striate to Temporal Cortex. *Cereb Cortex*.

Vogels R, Orban GA (1994). Activity of inferior temporal neurons during orientation discrimination with successively presented gratings. *Journal of Neurophysiology* **71**(4): 1428-1451.

Wallace DJ, Greenberg DS, Sawinski J, Rulla S, Notaro G, Kerr JN (2013). Rats maintain an overhead binocular field at the expense of constant fusion. *Nature* **498**(7452): 65-69.

Wang Q, Burkhalter A (2007). Area map of mouse visual cortex. *J Comp Neurol* **502**(3): 339-357.

Wang Q, Sporns O, Burkhalter A (2012). Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex. *J Neurosci* **32**(13): 4386-4399.

Wells E, Bernstein G, Scott B, Bennett P, Mendelson J (2001). Critical flicker frequency responses in visual cortex. *Experimental Brain Research* **139**(1): 106-110.

whuber (2011, version 20 November 2011). "Maximum value of coefficient of variation for bounded data set." Cross Validated, from stats.stackexchange.com/q/18679.

Williams R, Pollitz C, Smith J, Williams T (1985). Flicker detection in the albino rat following light-induced retinal damage. *Physiology & Behavior* **34**(2): 259-266.

Willmore B, Tolhurst DJ (2001). Characterizing the sparseness of neural codes. *Network* **12**(3): 255-270.

Willmore BD, Mazer JA, Gallant JL (2011). Sparse coding in striate and extrastriate visual cortex. *J Neurophysiol* **105**(6): 2907-2919.

Wiskott L, Sejnowski TJ (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural computation* **14**(4): 715-770.

Wiskott L, Berkes P, Franzius M, Sprekeler H, Wilbert N (2011). Slow feature analysis. *Scholarpedia* **6**(4): 5282.

Zoccolan D, Oertelt N, DiCarlo JJ, Cox DD (2009). A rodent model for the study of invariant visual object recognition. *Proc Natl Acad Sci U S A* **106**(21): 8748-8753.

Zoccolan D (2015). Invariant visual object recognition and shape processing in rats. *Behav Brain Res* **285**: 10-33.