Scuola Internazionale Superiore di Studi Avanzati - Trieste

Shape Processing Strategies Underlying Visual Object Recognition

Candidate:

Alireza Alemi-Neissi

Supervisor:

Davide Zoccolan

Thesis submitted for the degree of Doctor of Philosophy in Neuroscience Area

Trieste, 2013

SISSA - Via Bonomea 265 - 34136 Trieste, Italy.

Shape Processing Strategies Underlying Visual Object Recognition

by

Alireza Alemi-Neissi

Submitted to Neuroscience Area
for the degree of

Doctor of Philosophy

at

Scuola Internazionale Superiore di Studi Avanzati (SISSA)

January 2013

Author . . . . . .
Neuroscience Area
31 Jan 2013

Certified by . . . . . .
Dr. Davide Zoccolan
Assistant Professor
Thesis Supervisor

Accepted by . . . . . .
Mathew Diamond
Director, Neuroscience Area

Shape Processing Strategies Underlying Visual Object Recognition

by

Alireza Alemi-Neissi

Submitted to Neuroscience Area
on 31 Jan 2013 for the degree of
Doctor of Philosophy

Abstract

Large variations in object appearance such as changes in pose or size, do not affect much recognition of the objects in human and some non-human species, yet the computational mechanisms underlying such an invariant shape processing are not full understood. Most studies aimed at understanding higher-level vision and shape/object-processing use monkeys to investigate the neural underpinning of these phenomena, in spite of the limited range of experimental approaches that are available in this species. Rodents, on the other hand, are the animals model-of-choice in many neuroscience sub-disciplines, because of the growing number of powerful experimental tools that have become available in the recent years. However, the evidence for invariant object recognition, and, more in general, higher-level visual processing in rodents, is still limited and under debate.

In this thesis, in order to show how the complexity of visual shape processing in rats compared with that of human, we uncovered the perceptual strategies underlying invariant visual object recognition task in these two species. To characterize the visual recognition strategies, we applied an image masking method, known as the Bubbles method, that revealed the diagnostic features used by the observers to discriminate two objects across a range of sizes, positions, in-depth and in-plane rotations. An ideal observer analysis was also carried out in order to compare the diagnostic features obtained for humans and rats with those obtained for a simulated observer that has full access to the pictorial discriminatory information contained in the different views of the two target objects. We also investigated to what extent, for both rats and humans, the diagnostic object features were preserved across views, thus touching on a long-standing debate between view-dependent and view-invariant models of object recognition.

Based on the diagnostic features obtained using the standard analysis of the Bubbles method, we found that rat recognition relied on combinations of multiple features that were mostly preserved across the transformations the objects underwent, and largely overlapped with the features that a simulated ideal observer deemed optimal to accomplish the discrimination task. These results indicate that rats are able to process and efficiently use shape information, in a way that is largely tolerant to variation in object appearance. The diagnostic features found for humans partially overlapped with those found for rats for one object. However, human and rat recognition strategies substantially differed in the case of the other object. Moreover, human recognition strategy was not correlated with that of the

ideal observer, but, rather, for many object views, was significantly anti-correlated. This can be attributed to the fact that human observers relied on the boundaries and coterminations of the objects as diagnostic features, while rats and the ideal observer mainly relied the bulk of the objects structural parts. Finally, human strategy remained largely stable across transformations, with the exception of extreme rotations in-depth—in that case, due to drastic changes in the appearance of the main objects diagnostic features, the observers changed their perceptual strategy to recognize the object.

The standard analysis for finding diagnostic shape features using the Bubbles method is based on the assumption of an underlying linear observer model—i.e., an observer that performs a weighted sum of the evidences provided, independently, by each pixel in an image to detect the presence of a given object in that image. One problem with this assumption is that it does not allow understanding how multiple, distinct object features may interact to drive the recognition behavior of an observer—i.e., whether such features need to be simultaneously present/visible for the observer to correctly identify the object. As a computational contribution of this thesis, we present a novel, information-theoretic analysis to uncover non-linear interactions between diagnostic features extracted with the Bubbles method. After formulating the problem in a mathematical framework, we carried out two simulations of Bubbles experiments, in which the simulated observers performed either an AND-like or an OR-like interaction between two predefined features of an object in order to recognize it. These simulations showed that our information-theoretic analysis successfully retrieved the simulated non-linear interaction between pairs of diagnostic object features.

To summarize, our experimental results provide the most compelling evidence, to date, that rats process visual objects through rather sophisticated, shape-based, transformation-tolerant mechanisms. As such, given the powerful array of experimental approaches that are available in rats, this model system will likely become a valuable tool in the study of the neuronal mechanisms underlying object vision. We also concluded that human observers generalize their recognition to novel views of visual objects largely by relying on the same patterns of diagnostic features that they used to discriminate previously learned views of the same objects. Our results are consistent with both view-based and view-invariant theories of object recognition, because we found the same diagnostic features to be preserved in most object views, but we also observed the emergence of new features for particularly challenging views. Noticeably, our novel feature-interaction, information-theoretic analysis has the potential to test these theories even further by measuring whether the relationship between diagnostic features will be preserved across transformations. Finally, by uncovering human invariant recognition strategy under a variety of viewing conditions, our work provides new constraints for visual cortex models and neuromorphic machine vision systems, in terms of human-like shape processing mechanisms.

Thesis Supervisor: Dr. Davide Zoccolan
Title: Assistant Professor

Acknowledgments

The thesis would have not been made possible without constant supervision of my supervisor Davide Zoccolan. I would like to sincerely thank him for his ceaseless support throughout my PhD and also giving me freedom to pursue my own theoretical interest towards the end of my PhD. I learned from him to change my perspective from engineering to scientific: effectively "using my eyes" in scientific observation and having a long-term goal in research. I learned from him that it is possible to do world-class science and being a nice person, simultaneously. It is my honor to graduate as his first PhD student.

I would like to express my gratitude to Methew Diamond for all I learned by being a member of his lab in the first year. He was the first neuroscientist whom I personally got aquatinted with. He allowed me to come to SISSA as a visitor and to use my computer engineering skills to extract information from neural populations. His attitude taught me much more than just science.

John Nicholls was a caring and motivating teacher for me. I cannot fully appreciate with words his style of provocative lectures. He gave me different perspective on neuroscience. He voluntarily participated in informal meetings with us, students, to foster and guide us, from how to give a presentation to how neurons work.

I would like to thank my colleague and friend Federica Rosseili. Together with her, we accomplished the rat experiment. She was an example of a hard-working and nice colleague. I also sincerely appreciate Laura Riontinos help in running the human experiment and hiring the human participants.

Many thanks go to Riccardo Zechinna for hosting me during my visit in Torino in the collaboration we have had. The fruitful discussions that I had in his lab with Carlo Baldasi, Andrea Pagnani, Alfredo Braunstein were invaluable to me. Discussions with Stefano Panzeri led me to sharpen my ideas and see my computational problem from different perspectives.

During my stay in Trieste, I had a unique opportunity to be acquainted with many scientists specially in the statistical physics sector and mathematical physics. Courses by Christian Micheleti, Matteo Marsili, and Andrea Gambassi gave me good understanding of

statistical physics approaches that are related to natural sciences. I got many pleasure from discussions I had with my friend, Giacomo Dossena, on mathematics of invariance.

My lab members Sahitya Chetan Pandanaboina, Alessandro Di Filippo, Sina Tafazoli, Francesca Pulecchi helped me with lab-related work and I appreciate their helps. I thank Alessio Isaja for technical support, and Andrea Sciarrone, Riccardo Iancer, Federica Tuniz for administrative assistance at SISSA.

I would like to thank friends with whom we shared good memories in our life in Trieste. Fahimeh Baftizade, Zhaleh Ghaemi, Sam Azadi, Maryam Tavakoli, Mahdi Torabian, Shima Seyed-Allaee, Shima Talehy, Mahmood Safari, Mohammed Zhian Asadzadeh, Mohammad Ali Rajabpour, Ladan Amin, Houman Safaai, Giacomo Dossena, Stefano Giaccari, Georgette Argiris, Aurora Meroni, Michele Burrello, Tonya Blowers, Yabebal Fantaye, Chethan Krishnan, Pietro Giavedoni, Francesco Mancarella, Jacopo Viti, Andrea Trombettoni and many that I may have forgotten to name here.

There is a special person in my life, my wife, Sahar Pirmoradian. In the hardship of lifeacademic or generalshe has constantly been beside me. Words come short to appreciate her. I am also deeply thankful to Sahars parents for their warm emotional support. And finally I thank my mother and father for all they have done for me in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Understanding invariant recognition is one of the greatest challenges in the study of how the brain processes and make sense of sensory inputs. In our daily life, without invariant recognition, we would not be able to recognize objects using any of our sensory systems, whether visual, auditory, or tactile. We recognize a familiar face or object from a distance or in a dark room or from different viewpoints; we recognize the voice of familiar people when they speak over the phone; or we find the keys in our pocket even when only some specific part of our hand makes contact with some specific part of the keys or keychain. Visual object recognition, in particular, needs to cope with an extraordinary array of variation axes—such as contrast, lighting, size, orientation, position, in-depth rotation, etc. Such a richness and variety of appearances that each visual object can take is perhaps one of the reasons why the visual system has been evolved to be the most refined sensory system in the human brain. This makes the visual system the best sensory system to study invariance problem.

Not only to modern scientists has the invariant recognition problem been intriguing, but to ancient thinkers as well. Plato perhaps is among the first philosophers who formalized the problem, arguing through the main character of his treatise—Socrates—about a theory of invariant recognition. His theory of Forms, dating back to 23 centuries ago, proposed that non-material abstract forms of objects remain the same even though the objects' appearance may change under different conditions. Though Plato's theory had metaphysical elements in it, yet it is clear that it captures the gist of the invariance problem.

In modern era, the invariant visual object recognition problem has been tackled from different disciplines such as neurophysiology, psychophysics/psychology, and computer vision. From electrophysiological studies in different mammals, we have learned neurons in primary visual cortex (V1) are mainly selective for oriented edges (Hubel and Wiesel, 1959; Hubel and Wiesel, 1968) and, as we consider higher-order visual cortical areas along the pathway, V1 $\rightarrow$ V2 $\rightarrow$ V4 $\rightarrow$ inferotemporal (IT) cortex, the complexity of the visual features that neurons are selective for increases (Felleman and Van Essen, 1991; Tanaka, 1996). In particular, at the culmination of this pathway (known as the ventral visual pathway/stream) IT neurons exhibit the highest featural tuning complexity and the highest degree of tolerance to variation in object appearance, such as size, clutter, and position (Logothetis and Sheinberg, 1996; Tanaka, 1996; Orban, 2008; DiCarlo et al., 2012), besides view-selective responses to specific views of objects (Logothetis et al., 1995). At the population level, identity of objects can be read out from a population of IT neurons in monkeys regardless of object position or size (Hung et al., 2005). Moreover, some IT neurons have been shown to be selectively tuned to specific, non-accidental features of visual objects (Op de Beeck et al., 2001).

While electrophysiological recordings provide direct information about visual processing mechanisms in the brain, behavioral/psychophysics studies aim at understanding the visual processing when the access to neurons is not easy, especially in humans, given the inability of performing invasive experiments (with the exception of therapeutic approaches). Psychophysical studies have used different paradigms including priming (Biederman and Cooper, 1991), adaptation-aftereffect (Afraz and Cavanagh, 2008; Afraz and Cavanagh, 2009), and the Bubbles method (Nielsen et al., 2008; Vermaercke and OpdeBeeck, 2012) to study invariant object recognition in humans. These methods not only reveal the overall capabilities and limitations of human invariant recognition but also lay the ground to compare human visual processing abilities with those of other species.

While some studies reached to the conclusion that human representation of visual objects is fully invariant with respect to size (Biederman and Cooper, 1992) and position (Biederman and Cooper, 1991) changes, some recent studies concluded that tolerance of human object recognition to position, size, and rotation changes is more limited than previously

thought (Afraz and Cavanagh, 2008; Afraz and Cavanagh, 2009). The issue becomes even more debated when it comes to tolerance/invariance for non-affine transformations, such as in-depth rotations. There are two main theories that have emerged from psychophysical studies about how we are able to recognize 3D objects from different viewpoints. One theory argues in favor of a view-invariant representation of objects, based on the object features that remain largely recognizable across rotations (often named non-accidental features) and their relations (such as cotermination of edges) and their relations (Biederman, 1987; Biederman and Cooper, 1991; Biederman and Gerhardstein, 1993; Biederman, 2000). The other theory maintains that the brain achieves view invariance by learning and storing multiple view-dependent representations of objects at different viewpoints, and recognizing novel views by interpolating between such stored, previously learned representations (Poggio and Edelman, 1990; Bülthoff and Edelman, 1992; Logothetis et al., 1994; Blthoff et al., 1995; Tarr and Blthoff, 1998; Edelman, 1999; Riesenhuber and Poggio, 2000; Freedman et al., 2005). More recent developments suggest that the mechanisms proposed by both theories may be at work in the brain, depending on the amplitude of the in-depth rotation an object is undergoing (Lawson, 1999; Hayward, 2003; Foster and Gilson, 2002). One issue in all these previous psychophysical studies is that they did not directly retrieve the visual features underlying recognition of an object across multiple views. As such, they did not provide direct evidence about specific features being preserved (or not being preserved) across viewpoints.

All the electrophysiological and psychophysical studies should eventually provide evidence about how the visual system works so as to enable us to build computer vision systems that have capabilities similar to primates in terms of invariant object recognition. To date, in spite of many attempts, no artificial vision system has been developed that fully matches the ability of the human visual system (Pinto et al., 2008). This is another motivation to study the neural underpinning of visual object recognition in animal models, so as to reverse engineer the brain machinery underlying object vision. To achieve this goal, the choice of animal model is really important. The main animal model that has been used, so far, in the investigation of the visual system, is the non-human primate, since its visual system closely mirrors the human one (Felleman and Van Essen, 1991;

Nassi and Callaway, 2009). This model, however, does not provide access to the most powerful experimental approaches (e.g., genetics, molecular, optical, etc.) that have been developed in recent years in rodents (Huberman and Niell, 2011). The visual system of rodents, on the other hand, has not been extensively studied, because of a widespread belief that it is not capable of advanced, higher-level processing. For instance, some behavioral studies have suggested that rats use low-level strategies to process visual shape (Minini and Jeffery, 2006; Vermaercke and OpdeBeeck, 2012). In contrast, other studies have argued in favor of advanced object recognition abilities in rats (Zoccolan et al., 2009; Tafazoli et al., 2012). No study, however, has investigated what perceptual strategy underlies rat recognition behavior, and whether such a strategy is consistent with some form of higher-level shape processing.

The goal of this thesis was to verify more directly how complex rat object recognition strategy is, when the animals are engaged in a demanding, invariant object recognition task. We made use of a method for randomly sampling of the image space, known as the Bubbles method (Gosselin and Schyns, 2001), that enabled us to find the diagnostic features underlying rat discrimination of two visual objects across a range of transformations. We also applied this same paradigm to human participants to compare their recognition strategy with that of rats. This also gave us the opportunity to investigate the long-standing issue of view-invariant versus view-dependent representations (see above) by comparing the diagnostic features obtained for different views of an object. Finally, we also developed a computational method to extract non-linear interactions between the diagnostic features of an object.

The outline of the thesis is as the following. In Chapter 2, I describe the behavioral study investigating the shape-processing strategy underlying rat invariant object recognition. In Chapter 3, I describe the psychophysics study investigating the shape processing strategy underlying human invariant recognition (using the same visual stimuli and experimental/analytical approach used in the rat study). In Chapter 4, I propose a new approach to reveal the presence of non-linear interactions between the diagnostic features of an object.

Chapter 2

Shape Processing in Rats

2.1 Introduction

Visual object recognition is an extremely challenging computational task, because of the virtually infinite number of different images that any given object can cast on the retina. While we know that the visual systems of humans and other primates successfully deal with such a tremendous variation in object appearance, thus providing a robust and efficient solution to the problem of object recognition (Logothetis and Sheinberg, 1996; Tanaka, 1996; Rolls, 2000; Orban, 2008; DiCarlo et al., 2012), the underlying neuronal computations are poorly understood, and transformation-tolerant (a.k.a. "invariant") recognition remains a major obstacle in the development of artificial vision systems (Pinto et al., 2008). This is not surprising, given the formidable complexity of the primate visual system (Felleman and Van Essen, 1991; Nassi and Callaway, 2009) and the limited understanding of neuronal mechanisms that primate studies allow at the molecular, synaptic and circuitry level. In recent years, the powerful array of experimental approaches that has become available in rodents has re-ignited the interest for rodent models of visual functions (Huberman and Niell, 2011; Niell, 2011), including visual object recognition. However, it remains controversial whether rodents possess higher-order visual processing abilities, such as the capability of processing shape information and extracting object-defining visual features in a way that is comparable with primates.

The studies that have explicitly addressed this issue have reached opposite conclusions.

Minini and Jeffery (2006) concluded that rats lack advanced shape-processing abilities and rely, instead, on low-level image cues to discriminate shapes. By contrast, two recent studies have shown that rats can recognize objects despite remarkable variation in their appearance (e.g., changes in size, position, lighting, in-depth and in-plane rotation), thus arguing in favor of a sophisticated recognition strategy in this species (Zoccolan et al., 2009; Tafazoli et al., 2012). However, studies based on pure assessment of recognition performance cannot reveal the complexity of rat recognition strategy, i.e., they cannot tell: 1) whether shape features are truly extracted from the test images; 2) what these features are and how many; and 3) whether they remain stable across the object views the animals face. In spite of a recent attempt at addressing these issues by Vermaercke and Op de Beeck (2012), who used a version of the same image classification technique we have applied in our study in this chapter, these questions remain largely unanswered. In fact, the authors' conclusion that rats are capable of using a flexible mid-level recognition strategy is affected by several limitations of their experimental design, which prevented a true assessment of shape-based, transformation-tolerant recognition (see Discussion for details).

In the study discussed in this chapter, we trained six rats to discriminate two visual objects across a range of sizes, positions, in-depth rotations and in-plane rotations. Then, we applied to a subset of such transformed object views the Bubbles masking method (Gosselin and Schyns, 2001; Gibson et al., 2005; Gibson et al., 2007), which allowed extracting the diagnostic features used by the rats to successfully recognize each view. Our results show that rats are capable of a sophisticated, shaped-based, invariant recognition strategy, which relies on extracting the most informative combination of object features across the variety of object views the animals face.

2.2 Materials and Methods

2.2.1 Subjects

Six adult male Long Evans rats (Charles River Laboratories) were used for behavioral testing. Animals were 8 weeks old at their arrival, weighted approximately 250 g at the onset of training and grew to over 600 g. Rats had free access to food but were water-deprived throughout the experiments, that is they were dispensed with 1 hour of water pro die after each experimental session, and received an amount of 4-8 ml of pear juice as reward during the training. All animal procedures were conducted in accordance with the National Institutes of Health, International, and Institutional Standards for the Care and Use of Animals in Research and after consulting with a veterinarian.

2.2.2 Experimental Rig

The training apparatus consisted of six operant boxes. Each box hosted one rat, so that the whole group could be trained simultaneously, every day, for up to two hours. Each box was equipped with: 1) a 21.5" LCD monitor (Samsung 2243SN) for presentation of visual stimuli, with a mean luminance of 43 cd/mm^2 and an approximately linear luminance response curve; 2) an array of three stainless steel feeding needles (Cadence Science) ~10 mm apart from each other, connected to three capacitive touch sensors (Phidgets 1110) for initiation of behavioral trials and collection of responses; and 3) two computer-controlled syringe pumps (New Era Pump Systems NE-500), connected to the left-side and right-side feeding needles, for automatic liquid reward delivery.

A 4-cm diameter viewing hole in the front wall of each box allowed each tested animal to extend its head out of the box, so to frontally face the monitor (placed at ~30 cm in front of the rat's eyes) and interact with the sensors' array (located at 3 cm from the opening). The location and size of the hole was such that the animal had to reproducibly place its head in the same position with respect to the monitor to trigger stimulus presentation. As a result, head position was remarkably reproducible across behavioral trials and very stable during stimulus presentation. Video recordings obtained for one example rat showed that

the standard deviation of head position, measured at the onset of stimulus presentation across 50 consecutive trials, was ±3.6 mm and ±2.3 mm along the dimensions that were, respectively, parallel (x axis) and orthogonal (y axis) to the stimulus display (with the former corresponding to a jitter of stimulus position on the display of ±0.69º of visual angle). For the same example rat, the average variation of head position over 500 ms of stimulus exposure was $\Delta x = 2.5 \pm 0.5$ mm (mean ± SD) and $\Delta y = 1.0 \pm 0.2$ mm (n = 50 trials), with the former corresponding to a jitter of stimulus position on the display of ~0.48º of visual angle. Therefore, the stability of rat head during stimulus presentation was close to what achieved in head-fixed animals. This guaranteed a very precise control over stimulus retinal size and prevented head movements from substantially altering stimulus retinal position (see Results and Discussion for further comments about stability of stimulus retinal position).

2.2.3 Visual stimuli

Each rat was trained to discriminate the same pair of 3-lobed visual objects used in (Zoccolan et al., 2009). These objects were renderings of three-dimensional models that were built using the ray tracer POV-Ray (http://www.povray.org/). Both objects were illuminated from the same light source location and, when rendered at the same in-depth rotation, their views were approximately equal in height, width and area (see Fig. 1A). Objects were rendered in a white, bright (see below for a quantification), opaque hue against a black background. Each object's default size was 35° of visual angle, and their default position was the center of the monitor (their default view was the one shown in Fig. 1A).

As explained below, during the course of the experiment, transformed views of the objects were also shown to the animals (i.e., scaled, shifted, in-plane and in-depth rotated object views; see Fig. 1C). The mean luminance of the two objects (measured in their center) across all the transformed views was 108 ± 3 cd/mm^2 (mean ± SD) and 107 ± 4 cd/mm^2, respectively, for Object 1 and 2 (thus, approximately matching the display maximal luminance). Therefore, according to both behavioral and electrophysiological evidence (Naarendorp et al., 2001; Jacobs et al., 2001; Fortune et al., 2004; Thomas et al.,

2007), rats used their photopic vision to discriminate the visual objects.

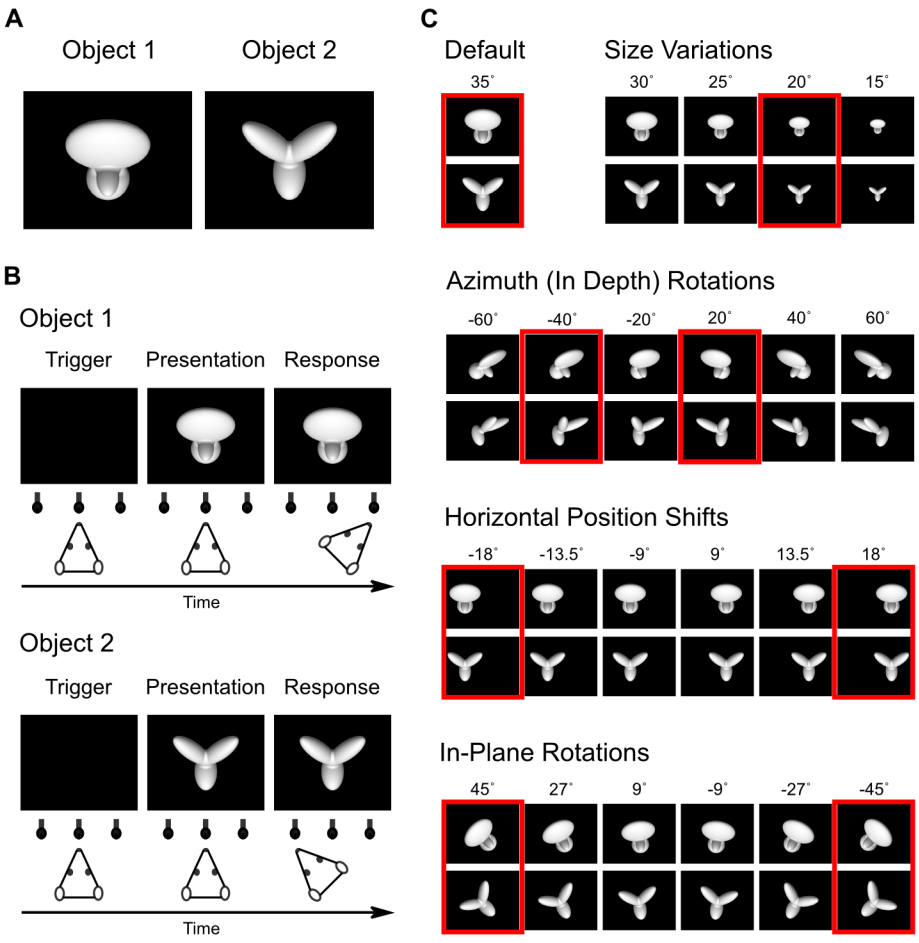

Figure 2-1: Visual stimuli and behavioral task. **A.** Default views (0º in-depth and in-plane rotation) of the two objects that rats were trained to discriminate during Phase I of the study (each object default size was 35º of visual angle). **B.** Schematic of the object discrimination task. Rats were trained in an operant box that was equipped with an LCD monitor for stimulus presentation and an array of three sensors. The animals learned to trigger stimulus presentation by licking the central sensor, and to associate each object identity to a specific reward port/sensor (right port for Object 1 and left port for Object 2). **C.** Some of the transformed views of the two objects that rats were required to recognize during Phase II of the study. Transformations included: 1) size changes; 2) azimuth in-depth rotations; 3) horizontal position shifts; and 4) in-plane rotations. Azimuth rotated and horizontally shifted objects were also scaled down to a size of 30º of visual angle; in-plane rotated objects were scaled down to a size of 32.5º of visual angle and shifted downward of 3.5º . Note that each variation axis was sampled more densely than shown in the figure – sizes were sampled in 2.5º steps; azimuth rotations in 5º steps; position shifts in 4.5º steps; and in-plane rotations in 9º steps. This yielded a total of 78 object views. The red frames highlight the subsets of object views that were tested in bubbles trials (see Fig. 2).

2.2.4 Experimental design

Phase I: critical shape features underlying recognition of the default object views

Rats were initially trained to discriminate the two default views of the target objects (see Fig. 1A). Animals initiated each behavioral trial by inserting their heads through the viewing hole in the front wall of the training box and licking the central sensor. This prompted presentation of one of the target objects on the monitor placed in front of the box. Rats learned to associate each object identity with a specific reward port (see Fig. 1B). In case of correct response, reward was delivered through the port and a reinforcement tone was played. An incorrect choice yielded no reward and a 1-3 s time out (during which a failure tone sounded and the monitor flickered from black to middle gray at a rate of 15 Hz). The default stimulus presentation time (in the event that the animal made no response after initiating a trial) was 2.5 s. However, if the animal responded correctly before the 2.5 s period expired, the stimulus was kept on the monitor for an additional 4 s from the time of the response (i.e., during the time the animal collected his reward). In the event of an incorrect response, the stimulus was removed immediately and the time-out sequence started. If the animal did not make any response during the default presentation time of 2.5 s, it still had 1 s, after the offset of the stimulus presentation and before the end of the trial, to make a response.

Once a rat achieved $\geq$70% correct discrimination of the default object views (this typically required 3-12 weeks of training), an image masking method, known as the *Bubbles* (Gosselin and Schyns, 2001), was applied to identify the visual features the animal relied upon to successfully accomplish the task. This method consists in superimposing on a visual stimulus an opaque mask punctured by a number of circular, semi-transparent windows (or *bubbles*; see Fig. 2A). When one of such masks is applied to a visual stimulus, only those parts of the stimulus that are revealed through the bubbles are visible. Hence, this method allows isolating the image patches that determine the behavioral outcome, for whether a subject (e.g., a rat) can identify the stimulus depends on whether the uncovered portions of the image are critical for the accomplishment of the recognition task.

In our implementation of the Bubbles method, any given bubble was defined by shaping

the transparency (or alpha) channel profile of the image according to a circularly symmetrical, two-dimensional Gaussian (with the peak of the Gaussian corresponding to full transparency). Multiple such Gaussian bubbles were randomly located over the image plane. Overlapping of two or more Gaussians lead to summation of the corresponding transparency profiles, thresholded with the maximal level corresponding to full transparency. The size of the bubbles (i.e., the standard deviation of the Gaussian-shaped transparency profiles) was fixed at 2º of visual angle, while the number of bubbles was chosen so as to bring each rat's performance to be ~10% lower than in unmasked trials (this typically brought the performance down from ~70-80% correct obtained in unmasked trials to 60-70% correct obtained in bubbles masked trials; see Figs. 3A and 4). This was achieved by randomly choosing the number of bubbles in each trial among a set of values that was specific for each rat. These values ranged between 10 and 50 (in steps of 20) for top performing rats, and between 50 and 90 (again, in steps of 20) for average performing rats (examples of objects occluded by masks with a different number of bubbles are shown in Fig. 2B).

Trials in which the default object views were shown unmasked (named *regular trials*) were randomly interleaved to trials in which they were masked (named *bubbles trials*; see Fig. 2C). The fraction of bubbles trials presented to a rat in any given daily session varied between 0.4 and 0.75. In order to obtain enough statistical power to extract the critical features underlying rat recognition, at least 3,000 bubbles trials for each object were collected over the course of 16.3 ± 4.4 sessions (rat group average ± SD, $n = 6$).

Phase II: critical shape features underlying recognition of the transformed object views

After having collected sufficient data to infer the critical features used by rats to discriminate the default object views, the animals were trained to tolerate variation in the appearance of the target objects along a variety of transformation axes. The goal of this training was to engage those high-level visual processing mechanisms that, at least in primates, allow preserving the selectivity for visual objects in the face of identity-preserving object transformations (Zoccolan et al., 2005; Zoccolan et al., 2007; Li et al., 2009; Rust and DiCarlo, 2010; DiCarlo et al., 2012). Four different transformations were tested (see Fig. 1C), in the following order: 1) *size* variations, ranging from 35° to 15° visual angle;

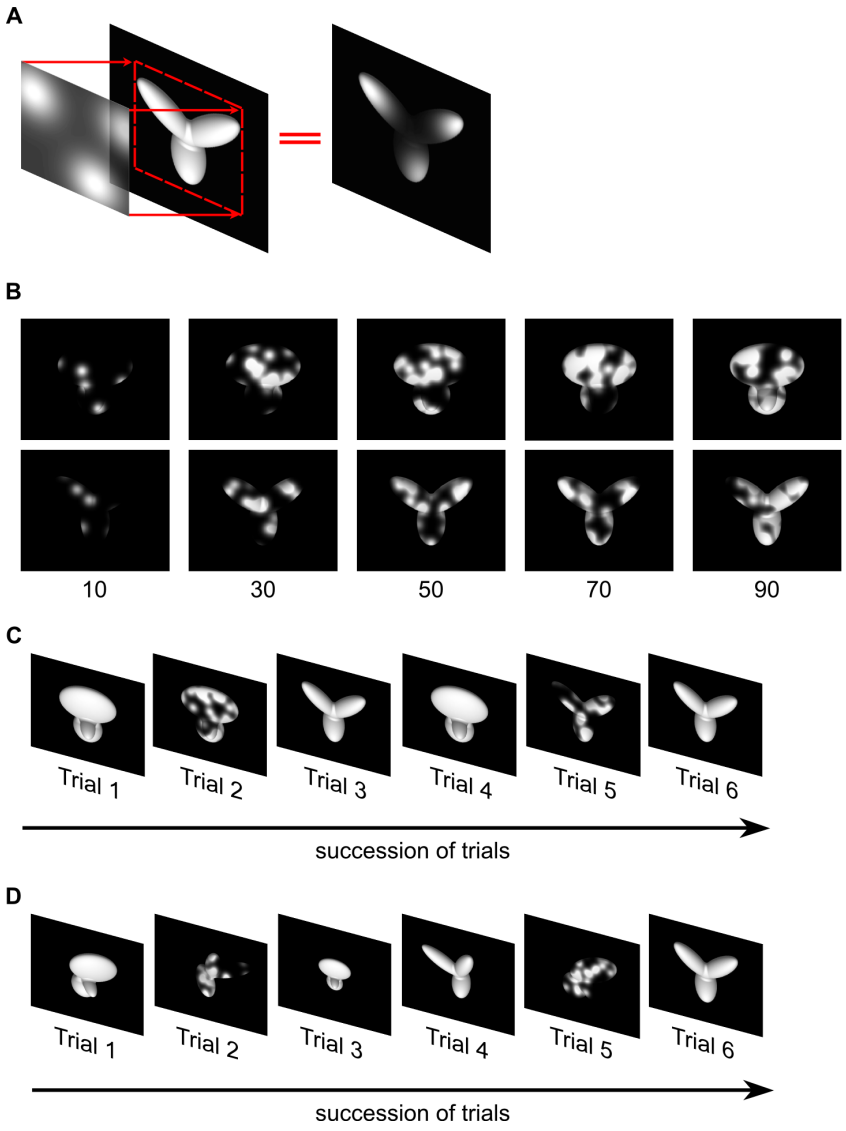

Figure 2-2: The Bubbles method. **A.** Illustration of the Bubbles method, which consists in generating an opaque mask (fully black area) punctured by a number of randomly located transparent windows (i.e., the *bubbles*; shown as white, circular clouds) and then overlapping the mask to the image of a visual object, so that only parts of the object remain visible. **B.** Examples of the different degrees of occlusion of the default object views that were produced by varying the number of bubbles in the masks. **C.** Examples of trial types shown to the rats at the end of Experimental Phase I. The object default views were presented both unmasked and masked in randomly interleaved trials (named, respectively, *regular* and *bubbles* trials). **D.** Examples of trial types shown during Experimental Phase II, after the rats had learned to tolerate size and azimuth rotations. The animals were presented with interleaved regular and bubbles trials. The former included all possible unmasked object views the rats had been exposed to up to that point (i.e., size and azimuth changes), while the latter included masked views of the -40º azimuth rotated objects.

2) *azimuth* rotations (i.e., in-depth rotations about the objects' vertical axis), ranging from -60° to 60° ; 3) horizontal *position* changes, ranging from -18º to +18º visual angle; and 4) *in-plane* rotations, ranging from -45° to +45° .

Size transformations were trained first, using an adaptive staircase procedure that, based on the animal performance, updated the lower bound of the range from which the object size was sampled (the upper bound was fixed to the default value of 35° visual angle). Once the sizes' lower bound reached a stable (asymptotic) value across consecutive training sessions (i.e., 15º of visual angle), a specific size (i.e., 20º of visual angle; see red frame in the top row of Fig. 1C) was chosen so that: 1) its value was different (lower) enough from the default one; and 2) most rats achieved about 70% correct recognition for that value. Rats were then presented with randomly interleaved regular trials (in which unmasked objects could be shown across the full 15º–35º size range) and bubbles trials (in which bubbles masks were superimposed on the 20º-scaled objects).

This same procedure was repeated for each of the other tested object transformations. For instance, after having trained size variations and having applied the Bubbles method to the 20º-scaled objects, a staircase procedure was used to train the rats to tolerate the azimuth rotations. After reaching asymptotic azimuth values (i.e., ±60º), two azimuth rotations were chosen (using the same criteria outlined above) for application of the Bubbles method: -40º and +20º (see red frames in the second row of Fig. 1C). Again, regular trials (in which unmasked objects could be shown across the full 15º–35º size range and the full -60º /+60º azimuth range) were then presented randomly interleaved with bubbles trials (in which bubbles masks were superimposed to either the -40º or the +20º azimuth rotated objects; see Fig. 2D).

After the azimuth rotations, position changes were trained (with bubbles masks applied to objects that were horizontally translated of ±18º of visual angle; see red frames in the third row of Fig. 1C) and then in-plane rotations (with bubbles masks applied to objects that were rotated of ±45º ; see red frames in the forth row of Fig. 1C). Noticeably, as explained above, while a new transformation was introduced, the full range of variation of the previously trained transformations was still shown to the animal, with the result that the task became increasingly demanding in terms of tolerance to variation in object appearance.

The staircase training along each transformation axis typically progressed very rapidly. On average, rats reached: i) the asymptotic size value in 1.2 ±0.4 sessions (mean± SD; $n = 6$); ii) the asymptotic azimuth rotation values in 5.5 ± 1.0 sessions ($n = 6$); iii) the asymptotic position values in 2.8 ± 1.9 sessions ($n = 5$); and iv) the asymptotic in-plane rotation values in 2.0 ± 0.0 sessions ($n = 2$). As for the default object views, at least 3,000 bubbles trials were collected for each of the transformed views that were tested with the Bubbles method. This required, on average across rats: i) 34.0 ± 13.7 sessions (mean ± SD; $n = 6$) for the 20º scaled view; ii) 40.1 ± 19.5 sessions ($n = 6$) and 15.6 ± 9.3 sessions ($n = 5$) for, respectively, the -40º and the $+20$º azimuth rotated views; iii) 27.6 ± 15.3 sessions ($n = 5$) and 14.0 ± 5.2 sessions ($n = 3$) for, respectively, the -18º and the $+18$º horizontally shifted views; and iv) 21 sessions ($n = 1$) and 20.5 ± 14.9 sessions ($n = 2$) for, respectively, the -45º and the $+45$º in-plane rotated views. In general, for each rat, bubbles trials could be collected only for a fraction of the seven transformed views we planned to test (see red frames in Fig. 1C). This was because the overall duration of Experimental Phase II varied substantially among rats, depending on: 1) how many trials each animal performed per session (this number roughly varied between 250 and 500); 2) what fraction of trials were bubbles trials (this number, which ranged between 0.4 to 0.75, had to be adjusted in a rat-dependent way, so to avoid the performance in bubbles trials to drop below ~10% less of the performance in regular trials); and 3) the longevity of each animal (some rats fell ill during the course of the experiment and had to be euthanized before being able to complete the whole experimental phase).

All experimental protocols (from presentation of visual stimuli to collection of behavioral responses) were implemented using the freeware, open-source software package MWorks (http://mworks-project.org). An ad-hoc plugin was developed in C++ to allow MWorks building bubbles masks and presenting them superimposed on the images of the visual objects.

2.3 Data Analysis

2.3.1 Computation of the saliency maps

The critical visual features underlying rat recognition of a given object view were extracted by properly sorting all the correct and incorrect bubbles trials obtained for that view. More specifically, saliency maps were obtained by measuring the correlation between the transparency values of each pixel in the bubbles masks and the behavioral responses. That is, saliency map values c^i for each pixel $\mathbf{x}^i$ were defined as:

$$c^i = \frac{\mathbf{x}^i.\mathbf{B}}{\|\mathbf{x}^i\|_{L1}}\ , \tag{2.1}$$

where $\mathbf{x}^i$ is a vector containing the transparency values of pixel i across all collected bubbles trials for a given object view; $\mathbf{B}$ is a binary vector coding the behavioral outcomes on such trials (i.e., a vector with elements equal to either 1 or 0, depending on whether the object view was correctly identified or not); and $\|\mathbf{x}^i\|_{L1}$ is the L_1 norm of $\mathbf{x}^i$, i.e.:

$$\|\mathbf{x}^i\|_{L1} = \sum_{n=1}^{N} \|\mathbf{x}^i\|, \tag{2.2}$$

where N is the total number of collected bubbles trials. Throughout the thesis, saliency maps are shown as grayscale masks superimposed on the images of the corresponding object views, with bright/dark pixels indicating regions that are salient/anti-salient, i.e., likely/unlikely to lead to correct identification of an object view when visible through the bubbles masks (e.g., see Figs. 3 and 6). For the sake of providing a clearer visualization, the saliency values in each map are normalized by subtracting their minimum value and dividing by their maximum value, so that all saliency values are bounded between zero and one.

To show which pixels, in the image of a given object view, had a statistically significant correlation with the behavior, the following permutation test was performed. All the bubbles trials that had been collected for that object view were divided in subsets, according to the number of bubbles that were used in a trial (e.g., 10, 30 or 50 for top performing

rats; see previous section). Within each subset of trials with the same number of bubbles, the behavioral outcomes (i.e., the elements of vector **B**) were randomly shuffled. Chance saliency map values c^i were then computed according to eq. 1, but using the shuffled vector **B**. Among all the chance saliency values, only those corresponding to pixels within the image of the object view were considered (i.e., those corresponding to background pixels were discarded). This yielded an average of 28,605 chance saliency values per object view. This procedure was repeated 10 times, so to obtain a null distribution of saliency values for each object view.

Based on this null distribution, a one-tailed statistical test was carried out to find what values, in each saliency map, were significantly higher than what obtained by chance ($p < 0.05$), and, therefore, what pixels, in the image, could be considered as significantly salient. Similarly, significant anti-salient pixels were found by looking for corresponding saliency values that were significantly lower than what expected by chance ($p < 0.05$). Throughout the thesis, significantly *salient regions* of an object view are shown in red, whereas *anti-salient regions* are shown in cyan (e.g., see Figs. 3 and 6).

Group average saliency maps and significant salient and anti-salient regions were obtained using the same approach described above, but after pooling the bubbles trials obtained for a given object view across all available rats (see Fig. 10).

2.3.2 Ideal observer analysis

Rats' average saliency maps were compared with the saliency maps obtained by simulating a linear ideal observer (Gosselin and Schyns, 2001; Gibson et al., 2005; Vermaercke and OpdeBeeck, 2012). Given a bubble-masked input image, the simulated observer classified it as being either Object 1 or 2, based on which of the eight views of each object, to which the mask could have been applied (shown by red frames in Fig. 1C), matched more closely (i.e., was more similar to) the input image. In other words, the simulated ideal observer performed a template matching operation between each bubble-masked input image and the 16 templates (i.e., eight views for each object) it had stored in memory. The ideal observer was linear in the template matching operation consisted in computing a normalized dot

product between each input image and each template. For better consistency with the experiment, we used the bubbles masks that were presented to one rat that was tested with all the eight object views (i.e., rat 3; see Fig. 6B). Also, to better match the actual retinal input to the rats, each input image was low pass-filtered so that its spatial frequency content did not exceeded 1 cycle per degree [i.e., the maximal retinal resolution of Long-Evans rats (Keller et al., 2000; Prusky et al., 2002)]. Finally, to lower the performance of the ideal observer and bring it close to the performance of the rats, Gaussian noise (std $= 0.5$ of the image grayscale) was independently added to each pixel of the input images. This assured that potential differences between rats' and ideal observer's saliency maps were not merely due to performance differences. Crucially, this constraint did not force the recognition strategy of the ideal observer to be similar to the one used by rats (the ideal observer had no knowledge of how rats responded to the bubble-masked input images). This was empirically assessed by running the ideal observer analysis with different levels of noise added to the input images, and verifying that the resulting saliency maps did not substantially change as a function of noise level (i.e., as a function of the ideal observer's performance). Saliency maps and significant salient and anti-salient regions for the ideal observer were obtained as described above for the rats (see the previous section).

Each rat group average saliency map was compared with the corresponding map obtained for the ideal observer by computing their Pearson correlation coefficient. The significance of the correlation was assessed by running a permutation test. In the permutation test, the behavioral outcomes of the bubbles trials, within each subset of trials with the same number of bubbles, were randomly shuffled 100 times for both the average rat and the ideal observer, yielding 100 pairs of random rat-ideal saliency maps. Computation of the Pearson correlation coefficient between each pair of random maps yielded a null distribution of 100 correlation values, against which the statistical test was carried out at $p = 0.05$ significance level.

All data analyses were performed in MATLAB (2011a, MathWorks Co, http://www.mathworks.com).

2.4 Results

The goal of this study was to understand the visual processing strategy underlying rat ability to recognize visual objects in spite of substantial variation in their appearance [e.g., see (Zoccolan et al., 2009; Tafazoli et al., 2012)]. To this aim, rats were trained in a visual object recognition task that required them to discriminate two visual objects under a variety of viewing conditions. An image masking method [known as the *Bubbles* (Gosselin and Schyns, 2001)] was then applied to a subset of the trained object views to infer what object features rats relied upon to successfully recognize these views. This approach not only revealed the complexity of rat recognition strategy, but also allowed tracking if and how such a strategy varied across the different viewing conditions the animals were exposed to.

2.4.1 Critical features underlying recognition of the default object views

During the initial experimental phase, 6 Long-Evans rats were trained to discriminate the default views (or appearances) of a pair of visual objects (shown in Fig. 1A). Details about the training/testing apparatus and the behavioral task are provided in Material and Methods and Fig. 1B. The animals were trained for 3-12 weeks until they achieved $\geq$70% correct discrimination. Once this criterion was reached, *regular* trials (i.e., trials in which the default object views were shown unoccluded) started to be randomly interleaved with *bubbles* trials (i.e., trials in which the default object views were partially occluded by opaque masks punctured by a number of circular, randomly located, semi-transparent windows; see Material and Methods and Fig. 2C).

The rationale behind the application of the bubbles masks was to make it harder for the rats to correctly identify an object, by revealing only parts of it (Gosselin and Schyns, 2001; Gibson et al., 2005; Gibson et al., 2007) (e.g., see Fig. 2A). Obviously, the effectiveness of a bubbles mask at impairing recognition of an object depended on the position of the semi-transparent bubbles (thus revealing what object features a rat relied upon to successfully recognize the object), but also on their size and number. Following previous applications of the Bubbles method (Gosselin and Schyns, 2001; Gibson et al., 2005), in our experiments the bubbles' size was kept fixed (i.e., set to 2º of visual angle), while their number

was adjusted so to bring each rat performance in bubbles trials to be ~10% lower than in regular trials (this was achieved by randomly sampling the number of bubbles in each trial either from a 10-50 or from a 50-90 range, according to the fluency of each animal in the recognition task; see examples of bubbles masked objects in Fig. 2B and Material and Methods for details). In the case of the default object views tested during this initial experimental phase, rat average recognition performance dropped from ~75% correct in regular trials to ~65% correct in bubbles trials (see Fig. 3A).

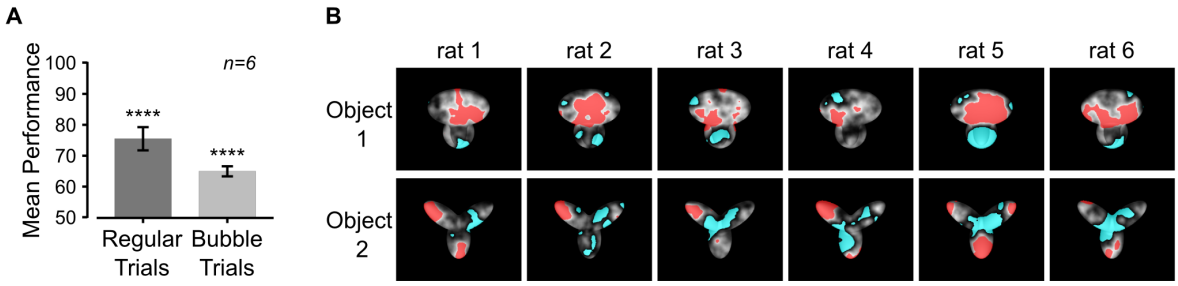

Figure 2-3: Critical features underlying recognition of the default object views. **A.** Rat group average performance at discriminating the default object views was significantly lower in bubbles trials (light gray bar) than in regular trials (dark gray bar; $p < 0.001$; one-tailed, paired t-test). Both performances were significantly higher than chance (****$p < 0.0001$; one-tailed, unpaired t-test). Error bars are SEM. **B.** For each rat, the saliency maps resulting from processing the bubbles trials collected for the default object views are shown as grayscale masks superimposed on the images of the objects. The brightness of each pixel indicates the likelihood, for an object view, to be correctly identified when that pixel was visible through the bubbles masks. Significantly salient and anti-salient object regions (i.e., regions that were significantly positively or negatively correlated with correct identification of an object; $p < 0.05$; permutation test) are shown, respectively, in red and cyan.

The critical visual features underlying rat recognition of the default object views were extracted by computing saliency maps that measured the correlation between bubbles masks' transparency values and rat behavioral responses, as done in (Gosselin and Schyns, 2001; Gibson et al., 2005) (see Materials and Methods for details). For each rat, the resulting saliency maps are shown, in Figure 3B, as grayscale masks superimposed on the images of the corresponding object views (with the brightness of each pixel corresponding to the correlation between the pixel transparency values and the correctness of responses of an

observer identifying an object view through the bubbles masks). Whether a saliency map value was significantly higher or lower than expected by chance was assessed through a permutation test at $p < 0.05$ (see Materials and Methods). This led to the identification of significantly *salient* and *anti-salient regions* in the images of the default object views (shown, respectively, as red and cyan patches in Fig. 3B). These regions correspond to those objects' parts that, when visible through the masks, likely led, respectively, to correct identification and misidentification of the object views.

Visual inspection of the patterns of salient and anti-salient regions obtained for the different rats revealed several key aspects of rat object recognition strategy. In the case of Object 1, salient and anti-salient regions were systematically located within the same structural parts of the object (Fig. 3B, top row). Namely, for all rats, salient regions were contained within the larger, top lobe, while anti-salient-regions lay within the smaller, bottom lobes. Therefore, in spite of some variability in the size and compactness of the salient and anti-salient regions (e.g., compare the large, single salient region found for rats 2 and 5 with the smaller, scattered salient patches observed for rats 3 and 4.), the perceptual strategy underlying recognition of Object 1 was highly preserved across subjects.

In contrast, a substantial inter-subject variability was observed in the saliency patterns obtained for Object 2 (Fig. 3B, bottom row). Although the central part of the object (at the intersection of the three lobes) tended to be consistently anti-salient across rats, and the salient regions were always located within the peripheral part (the tip) of one or more lobes, the combination and the number of salient lobes varied considerably from rat to rat. For instance, rats 2 and 3 relied on a single lobe (the upper-left one), while rats 1, 4 and 6 relied on the combination of the upper-left and bottom lobes, and rat 5 relied on all three lobes. Moreover, some lobe (e.g., the bottom one) was salient for some animal, but fully (rat 2) or partially (rat 4) anti-salient for some other.

The larger inter-subject diversity in the pattern of salient and anti-salient features that was found for Object 2, as compared with Object 1, is not surprising, given the different structural complexity of the two objects. In fact, Object 2 is made of three fully visible, clearly distinct and roughly equally sized lobes, while the three lobes of Object 1 are highly varied in size, with the smaller, bottom lobes that are partially overlapping and, therefore,

harder to distinguish. As a consequence, Object 2 affords a larger number of distinct structural parts, as compared with Object 1, hence a larger number of "perceptual alternatives" to be used for its correct identification. As such, the saliency patterns obtained for Object 2 are more revealing of the complexity and diversity of rat recognition strategies.

Specifically, our data suggest that rat recognition may typically rely on a combination of multiple object features, as long as those features are, structure-wise, distinct enough to be parsed by the rat visual system. This is demonstrated by the fact that four out of six rats recognized Object 2 based on a combination of at least two significantly salient lobes. In addition, for the two remaining rats, saliency map values were high (i.e., bright) not only in the significantly salient upper-left lobe, but also in one (rat 2) or more (rat 3) additional lobes (although they did not reach significance in these lobes, except in a very small, point-like patch of the upper-right lobe, in the case of rat 2, and of the bottom lobe, in the case of rat 3). Overall, this suggests that rats naturally tend to adopt a shape-based, multi-featural processing strategy, rather than a lower-level strategy, based on detection of a single image patch. In particular, the fact that salient features were found in both the upper and lower half of Object 2 (together with the observation that salient and anti-salient features were typically found in the upper lobes of both objects; e.g., see rat 5) rules out the possibility that rat recognition was based on detection of very low-level stimulus properties, such as the amount of luminance in the lower or upper half of the stimulus display (Minini and Jeffery, 2006).

2.4.2 Recognition of the transformed object views

After the critical features underlying recognition of the default object views were uncovered (see Fig. 3B), each rat was further trained to recognize the target objects in spite of substantial variation in their appearance. Namely, objects were transformed along four different variation axes: size, in-depth azimuth rotation, horizontal position and in-plane rotation (the trained ranges of variation are shown in Fig. 1C). These transformations were introduced sequentially (i.e., size variation was trained first, followed by azimuth, then by position and finally by in-plane variation) and each of them was trained gradually, using

a staircase procedure (see Materials and Methods). Rats reached asymptotic values along these variation axes very quickly (i.e., in about 1-5 training sessions, depending on the axis; see Materials and Methods for details), which is consistent with their recently demonstrated ability to spontaneously generalize their recognition to novel object views (Zoccolan et al., 2009; Tafazoli et al., 2012) (see Discussion). Once the animals reached a stable, asymptotic value along a given transformation axis, one or two pairs of transformed object views along that axis were chosen for further testing with the bubbles masks (these pairs are marked by red frames in Fig. 2C). Such views were chosen so to be different enough from the objects' default views, yet still recognized with a performance close to 70% correct by most animals. Rats were then presented with randomly interleaved bubbles trials (in which these transformed views were shown with superimposed bubbles masks) and regular trials (in which unmasked objects were randomly sampled across all the variation axes tested up to that point; e.g., see Fig. 2D). A total of seven different pairs of transformed object views were chosen for testing with bubbles masks (see Fig. 1C), although, due to across-rat variation in longevity and fluency in the invariant recognition task (see Materials and Methods for details), all animals but one (rat 3) were tested only with a fraction of them.

Rat average recognition performance was significantly higher than chance for almost all tested object transformations (see legend of Fig. 4), typically ranging from ~70% to $\geq$80% correct, and dropping below 70% correct only for some extreme transformations (Fig. 4, gray lines). This confirmed that rat recognition is remarkably robust against variation in object appearance, as recently reported by two studies (Zoccolan et al., 2009; Tafazoli et al., 2012). As previously observed in the case of the default views (see Fig. 3A), rat performance at recognizing the transformed object views was generally 5-15% lower in bubbles trials than in regular trials (see black diamonds in Fig. 4).

Rat average reaction time (RT) was ~850 ms for the default object views. RT increased gradually (almost linearly) as object size became smaller (reaching ~950 ms for the smallest size; see Fig. 5A). A gradual (albeit not as steep) increase of RT was also observed along the other variation axes, with RT reaching ~900 ms for the most extreme azimuth rotations (see Fig. 5B), position shifts (see Fig. 5C) and in-plane rotations (see Fig. 5D). Overall, the smooth increase of RT as a function of the transformation magnitude and the fact that such

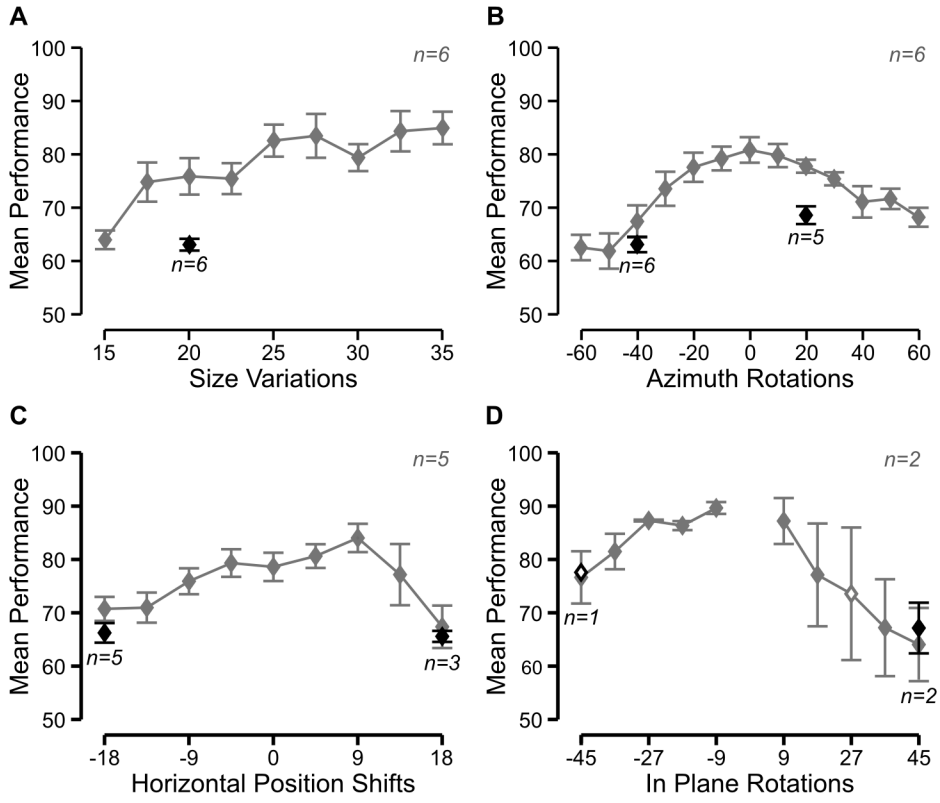

Figure 2-4: Recognition performance of the transformed object views. Rat group average recognition performance over the four variation axes along which the objects were transformed. Gray and black symbols show performances in, respectively, regular and bubbles trials that were collected over the course of the same testing sessions of Experimental Phase II (see Materials and Methods). Solid and open symbols indicate performances that were, respectively, significantly and non-significantly higher than chance (p textless 0.0001 in **A**, **B** and **C**; $p < 0.05$ in **D**; one-tailed, unpaired t-test). Error bars are SEM.

an increase was at most ~50 ms (with the exception of the smallest sizes) strongly suggest that rats did not make stimulus-triggered saccades (or head movements) to compensate for the retinal transformations the objects underwent. In fact, it is well established that, in primates, target-oriented saccades have a latency of at least 200 ms [a.k.a. *saccadic latency*, i.e., the interval between the time when the decision is made to move the eyes and the moment when the eye muscles are activated (Melcher and Colby, 2008)]. Therefore, it is reasonable to assume that in rats [which move their eyes much less frequently than primates do (Chelazzi et al., 1989; Zoccolan et al., 2010)] the saccadic latency should have at least the same magnitude (no data are available in the literature, since no evidence of

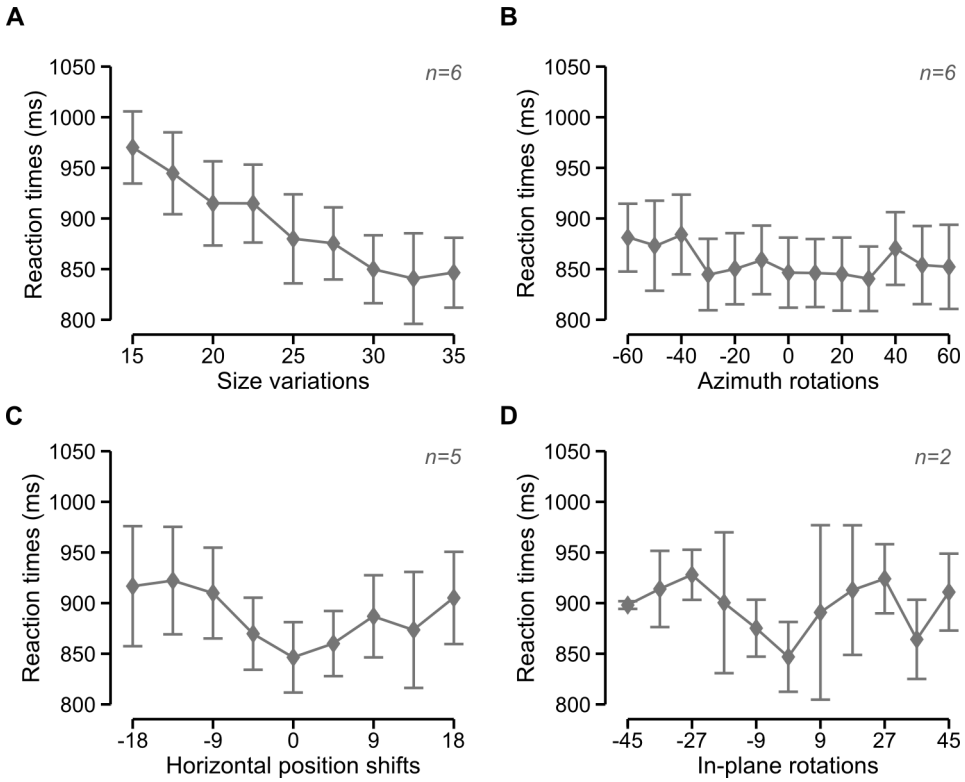

Figure 2-5: Reaction times along the variation axes. Rat group average reaction times (RTs) over the four variation axes along which the objects were transformed. RTs were measured across all the sessions performed by each rat during Experimental Phase II (see Materials and Methods) and then averaged across rats. Error bars are SEM.

target-oriented saccades has ever been reported in rodents). As a consequence, if rats made target-oriented saccades to (e.g.) adjust their gaze to the visual field locations of the target objects, a much larger and abrupt increase of the RT would have been observed for the horizontally shifted views (relative to the default views), as compared with what is shown in Figure 5C. Rather, the gradual increase of RT as a function of the transformation magnitude is consistent with the recognition task becoming gradually harder and, therefore, requiring an increasingly longer processing time (as revealed by the overall agreement between the trends shown in Figs. 4 and 5). Among the tested transformations, size reductions had the strongest impact on RT, likely because substantially longer processing times were required to compensate for the loss of shape details in the smallest object views (given rat poor visual acuity).

2.4.3 Critical features underlying recognition of the transformed object views

The critical features underlying rat recognition of a transformed view were extracted by properly processing all the correct and incorrect bubbles trials obtained for that view (see previous section and Materials and Methods). This yielded saliency maps with highlighted significantly salient and anti-salient regions that revealed if and how each animal recognition strategy varied across the different viewing conditions he was exposed to (see Fig. 6).

As previously reported for the default object views (see Fig. 3B), a larger inter-subject variability was observed in the patterns of critical features obtained for Object 2, as compared with Object 1. Namely, while for most rats a single and compact salient region was consistently found in the larger, top lobe of Object 1, regardless of the transformation the object underwent (see Figs. 6A-C, odd rows), in the case of Object 2 not only different rats relied on different combinations of salient lobes, but, for some rats, such combinations varied across the transformed object views (see Figs. 6A-C, even rows). Therefore, the saliency patterns obtained for Object 2 were more revealing of the diversity and stability of rat recognition strategies in the face of variation in object appearance.

For some rats, all the lobes used to discriminate the default view of Object 2 remained salient across the whole set of transformations the object underwent (see yellow arrows in Fig. 6A). This was particularly striking in the case of rat 5, which consistently relied on all three lobes of Object 2 as salient features across all tested transformations. Rat 6 showed a similarly consistent recognition strategy, although he relied only on two salient lobes (the upper-left and bottom ones). Also in the case of rat 3, the single salient lobe that was used for recognition of the default object view (the upper-left one) remained salient for all the subsequently tested transformed views (see yellow arrows in Fig. 6B). In this case, however, the bottom lobe, which only contained a point-like hint of a salient patch in the default view, emerged as a prominent salient feature when the animal had to face size variations, and remained consistently salient for all the ensuing transformations (see white arrows in Fig. 6B). In still other cases, lobes that were originally used by a rat to

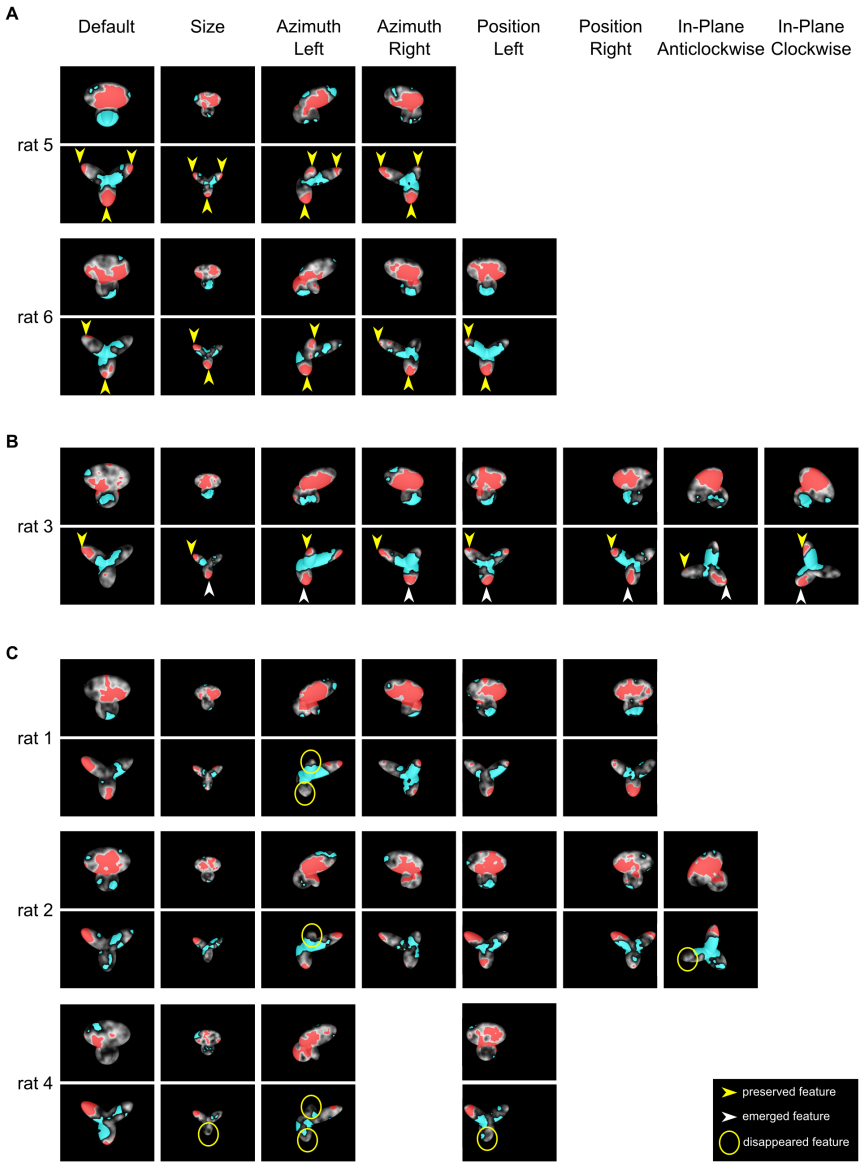

Figure 2-6: Critical features underlying recognition of the transformed object views. For each rat, the saliency maps (with highlighted significantly salient and anti-salient regions; same color code as in Fig. 3B) that were obtained for each transformed object view are shown. Maps obtained for different rats are grouped in different panels according to their stability across the tested views. **A.** For rats 5 and 6, the same pattern of salient features (i.e., lobes' tips) underlay recognition of all the views of Object 2 (see yellow arrows). **B.** For rat 3, one salient feature (i.e., the tip of the upper-left lobe) was preserved across all tested views of Object 2 (see yellow arrows), while a second feature (i.e., the tip of the bottom lobe) became salient after the animal started facing variation in object appearance (see white arrows). **C.** For rats 1, 2 and 4, features that underlay recognition of Object 2's default view became no longer salient for some of the transformed views (see yellow circles) and were replaced by other salient features.

discriminate Object 2's default view became no longer salient for some of the transformed views (see yellow circles in Fig. 6C) and were replaced by other salient lobes.

In summary, half of the rats showed a remarkably stable recognition strategy (Figs 6A, B) in the face of variation in object appearance, with the same combination of the salient object parts (i.e., lobes) being relied upon across all (Fig. 6A) or most (Fig. 6B) object views. The other half of the rats showed a more variable recognition strategy, based on view-specific salient features' patterns (Fig. 6C). Such a difference in the stability of the recognition strategies may reflect a different ability of rats to spontaneously generalize their recognition to novel object appearances (thus consistently relying on the same salient features), without the need to explicitly learn the associative relations between different views of an object (Tafazoli et al., 2012) (see Discussion).

Crucially, regardless of its stability across transformations, rat recognition strategy relied on a combination of at least two different salient features for most tested views of Object 2 (i.e., in 26 out of 34 cases). Since these features are located in structurally distinct parts of the object (i.e., distinct lobes), and, in all cases, in both its lower and upper half, this strongly suggests that rats are able to process global shape information and extract multiple structural features that are diagnostic of object identity. Such a shape-based, multi-featural processing strategy not only rules out a previously proposed low-level account of rat visual recognition in terms of luminance detection in the lower half of the stimulus display (Minini and Jeffery, 2006), but also suggests that rats are able to integrate shape information over much larger portions of visual objects (virtually, over a whole object) than reported by a recent study (Vermaercke and OpdeBeeck, 2012).

However, having assessed that rats are able to process global shape information does not imply, *per se*, that they are also capable of an advanced transformation-tolerant (or invariant) recognition strategy. In fact, having excluded one very low-level account of rat object vision (Minini and Jeffery, 2006) does not automatically rule out that some other low-level recognition strategies may be at work when rats have to cope with variation in object appearance. For instance, rats could rely on detection of some object feature(s) that is (are) largely preserved (in terms of position, size and orientation) across the tested object transformations. This would result in higher-than-chance recognition of the transformed

object views, without the need, for rats, to form and rely upon higher-level, transformation-tolerant object features' representations. This is not a remote possibility, since recent computational work has shown that even large databases of pictures of natural objects (commonly used by vision and computer vision scientists to probe invariant recognition) often do not contain enough variation in each object appearance to require engagement of higher-level, truly invariant recognition mechanisms (Pinto et al., 2008). This is especially true if objects do not undergo large enough changes in position over the image (e.g., retinal) plane (Pinto et al., 2008).

Assuring that the transformations we applied to the objects produced enough variation in the appearance of the objects' diagnostic features is particularly crucial in this study, since many transformations (i.e., size changes, azimuth rotations and in-plane rotations) did not alter the position of the objects over the stimulus display and, therefore, the images of many object views substantially overlapped (see examples in Fig. 7). As a consequence, the possibility that a given object feature partially retained its position/size/orientation cannot be excluded (Pinto et al., 2008). One could argue that, in spite of object position being unchanged over the stimulus display, some amount of trial-by-trial variation in the *retinal* position of the object views was likely present. In fact, stimulus presentation was not conditional upon fixation of a dot (given that rats, differently from primates, cannot be trained to make target-oriented saccades towards a fixation dot) and, therefore, rat gaze direction was not necessarily reproducible from trial to trial. However, the few data available in the literature show that rat gaze tends to be very stable over long periods of time, and, typically, comes back to a default, "resting" position after the rarely occurring saccades (Chelazzi et al., 1989; Zoccolan et al., 2010). Therefore, it is not safe to assume that across-trial variations in the retinal position of the object views would prevent rat recognition to be based on some transformation-preserved diagnostic features. Rather than relying on such an assumption, we explicitly tested if diagnostic features existed that remained largely unchanged across the transformations the object underwent (see next section). We tested this, assuming the "worst case scenario" that rat gaze was fully stable across trials and, therefore, no trial-by-trial variation in the retinal position of the object views diminished their overlap in the visual field of the animals.

2.4.4 No transformation-preserved diagnostic features can explain rat invariant recognition

As a way to inspect if any diagnostic object feature existed that was consistently preserved across the tested object transformations, the saliency maps obtained for different pairs of object views were superimposed and the overlap between pairs of significantly salient regions assessed. In fact, the existence of transformation-preserved features that are diagnostic of object identity would result in a large, systematic overlap between the significantly salient regions obtained for different object views. Examples of superimposed patterns of salient regions obtained for two or more object views are shown in Figure 7 (for clarity, the salient regions of different object views are depicted in different colors and the corresponding views are shown in the background).

In the case of Object 1, the lobe in which the salient region was located (the top one) was so large that, in spite of the transformations the object underwent, a substantial portion of it always occupied the same area within the image plane (i.e., the stimulus display). As a consequence, a substantial overlap was typically observed between the salient regions obtained for different views of the object (see orange patches in the top rows of Figs. 7A-C). However, not only such an overlap was not complete, but, for same pairs of object views, was minimal (e.g., see default vs. azimuth left-rotated view, or default vs. in-plane rotated views for rat 3, in the top row of Fig. 7C) and, in the case of the horizontally shifted views, it was null (see leftward- vs. rightward-shifted views in the top rows of Figs. 7B, C).

As previously observed, in the case of Object 2, the larger number of smaller and distinct structural parts (as compared with Object 1) resulted in a richer variety of patterns of multiple salient features, each located at the tip of a lobe. Such tips typically occupied different, non-overlapping portions of the image plane when the object underwent size, rotation and position changes. As a consequence, the overlapping between salient features obtained for different views was null or minimal (see bottom rows in Figs. 7A-C). Noticeably, the lack of overlap was observed not only when a salient lobe was replaced by a different one, following a given transformation (e.g., see default vs. azimuth left-rotated

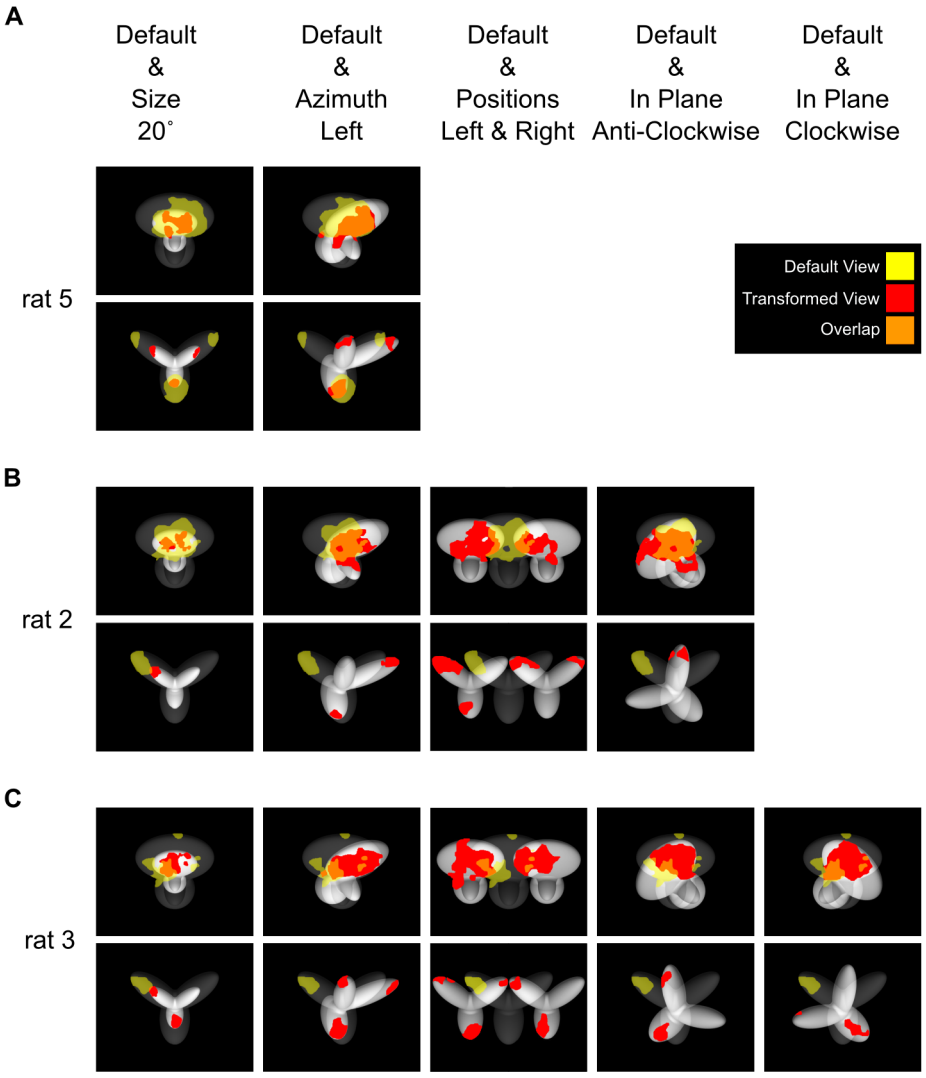

Figure 2-7: Overlap between the salient features obtained for different views of an object. Several pairs/triplets of views of Object 1 and 2 are shown superimposed, together with their salient features, to allow appreciating whether, and to what extent, such features overlapped. The salient features are the same as those shown in Figure 6, only here are shown in either yellow or red, to distinguish the features obtained, respectively, for the default and the transformed views. The orange patches indicate the overlap between the salient features of two different views in a pair/triplet. Panels **A**, **B** and **C** refer, respectively, to rats 5, 2 and 3.

view, or default vs. in-plane rotated view for rat 2 in Fig. 7B), but also when the same lobe remained salient across multiple object views (e.g., see the non-overlapping red and yellow patches in the object's left lobe for rats 5 and 3 in Figs. 7A and C). These examples exclude that rat recognition of Object 2 may have relied on some transformation-preserved feature

that was diagnostic of the object's identity across all or most the tested views.

To further assess whether rat recognition strategy was more consistent with a high-level, transformation-tolerant representation of diagnostic features or, rather, with low-level detection of some transformation-preserved image patches, we measured the overlap between the salient features obtained for all possible pairs of object views produced by affine transformations (i.e., all tested object views with the exclusion of in-depth azimuth rotations). The overlap was computed for both: 1) *raw* salient features' patterns, in which the image planes containing the salient features of the views to compare were simply superimposed (see second row of Fig. 8A, left plot); and 2) *aligned* salient features' patterns, in which the transformations that produced the two object views were "undone" (or reversed), so to perfectly align one view on top of the other (e.g., in the case of the comparison between the default and the horizontally translated views shown in Fig. 8A, the latter was shifted back to the center of the screen and scaled back to 35° , so to perfectly overlap with the default view; see second row of Fig. 8A, right plot). The overlap was quantified as the ratio between overlapping area and overall area of the significantly salient regions of the two object views (Nielsen et al., 2006) (e.g., as the ratio between the orange area and the sum of the red, yellow and orange areas in Fig. 8A, second row).

The resulting pairs of raw and aligned overlap values obtained for all tested combinations of object views are shown in Figure 8B (circles and diamonds refer, respectively, to pairs of views of Object 1 and 2). Similarly to what done by (Nielsen et al., 2006), the significance of each individual raw and aligned overlap was assessed through a permutation test, in which the salient regions of each object view in a pair were randomly shifted within the minimum bounding box enclosing each view. As illustrated by the example shown in Figure 8A, in the case of the raw overlap, the bounding boxes enclosing the two views partially overlapped (compare the white frames in the third row of Fig. 8A, left plot), while, in the case of the aligned overlap, by construction, the bounding boxes enclosing the two views were coincident (see the single white frame in the third row of Fig. 8A, right plot). Null distributions of raw and aligned overlap values were obtained by running 1,000 permutation loops, and the significance of the measured raw and aligned overlaps was assessed at $p = 0.05$ (significance is coded by the shades of gray filling the symbols in Fig.

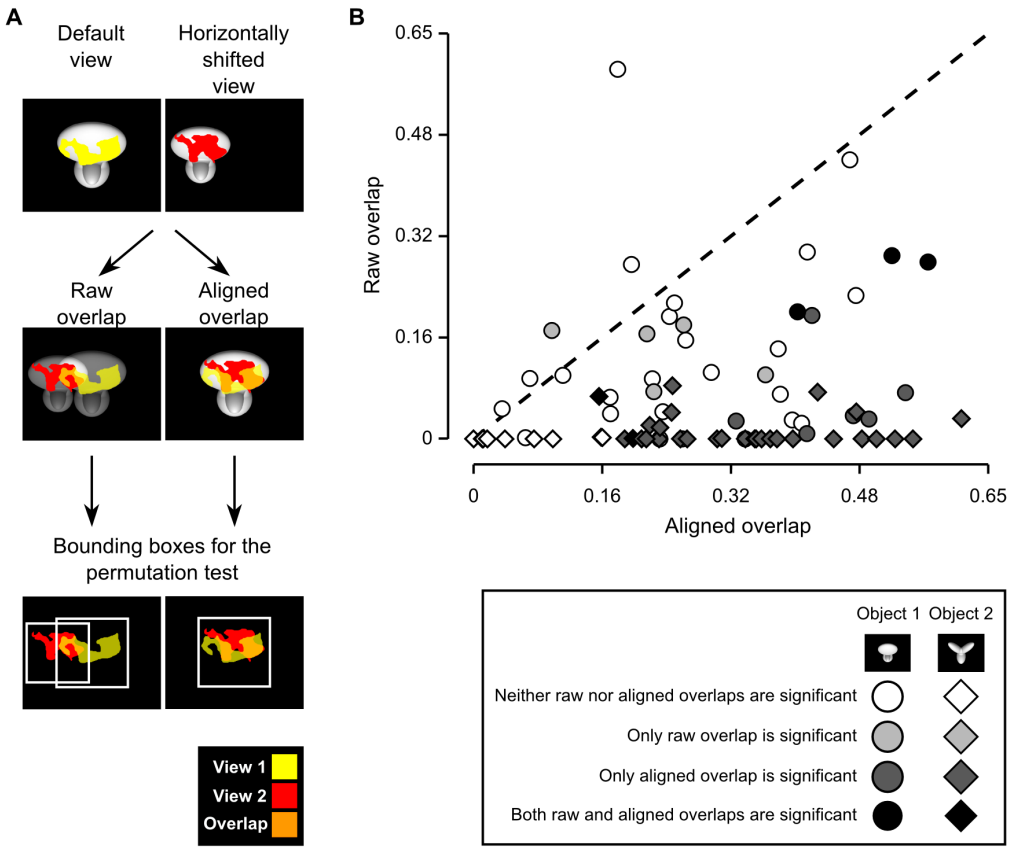

Figure 2-8: *Raw* vs. *aligned* features' overlap for all pairs of object views. **A.** Illustration of the procedure to compute the *raw* and *aligned* overlap between the salient features' patterns obtained for two different views of an object. The default and the leftward horizontally shifted views of Object 1 are used as examples (first row). To compute the raw features' overlap, these two object views (and the corresponding features' patterns) were simply superimposed (second row, left plot), as previously done in Figure 7. To compute the aligned features' overlap, the transformation that produced the leftward horizontally shifted view was reversed. That is, the object was shifted to the right of 18º and scaled back to 35º , so to perfectly overlap with the default view of the object itself (second row, right plot). In both cases, the overlap was computed as the ratio between the orange area and the sum of the red, yellow and orange areas. The significance of the overlap was assessed by randomly shifting the salient regions of each object view within the minimum bounding box enclosing each view (see Results for details). Such bounding boxes are shown as white frames in the third row of the figure, for both the raw and aligned views. **B.** The raw features' overlap is plotted against the aligned features' overlap for each pair of views of Object 1 (circles) and Object 2 (diamonds) resulting from affine transformations (i.e., position/size changes and in-plane rotations). The shades of gray indicate whether the raw or/and the aligned overlap values were significantly larger than expected by chance ($p < 0.05$; see caption).

8B; see caption).

For most pairs of object views (71 out of 76), the overlap between salient features was higher in the aligned than in the raw case (Fig. 8B). Namely, the average overlap between aligned views was 0.30 ± 0.01 (mean ± SEM), while the average overlap between raw views was 0.07 ± 0.01, with the former being significantly higher than the latter ($p < 0.0001$; significance was assessed through a paired permutation test, in which the sign of the difference between aligned and raw overlap for each pair of views was randomly assigned in 10,000 permutation loops).

The larger overlap found for aligned vs. raw views was particularly striking in the case of Object 2, with most raw overlap values being zero, and the average overlap being one order of magnitude larger for aligned than raw views (i.e., 0.30 ± 0.02 vs. 0.01 ± 0.00; such a difference was statistically significant at $p < 0.0001$, according to the paired permutation test). Moreover, in the large majority of cases (30/38), the aligned overlap was significantly higher than expected by chance (see black and dark gray diamonds in Fig. 8B), while the raw overlap was significantly higher than chance only for a few pairs of object views (2/38; see black diamonds in Fig. 8B). This confirmed that the transformations Object 2 underwent were large enough to displace its diagnostic features in non-overlapping regions of the stimulus display (hence, the zero or close-to-zero salient features' overlap observed for the raw views), thus preventing rats from relying on any transformation-preserved feature to succeed in the invariant recognition task (see also Fig. 7). At the same time, the large and significant salient features' overlap found for the aligned views of Object 2 indicates that the same structural parts were deemed salient for most of the object's views the rats had to face, thus suggesting that rats truly had to rely on some transformation-tolerant representation of these diagnostic structural features.

In the case of Object 1, in agreement with the examples shown in Figure 7, the overlap between raw views was considerably larger, as compared with what obtained for Object 2 (compare circles and diamonds in Fig. 8B). However, in most cases, the overlap between aligned views was higher than the corresponding overlap between raw views (i.e., 33 out of the 38 circles are in the lower quadrant in Fig. 8B) and, in several cases, the raw overlap was zero or close-to-zero. As a result, the average overlap was significantly larger for

aligned than raw views (i.e., 0.30 ± 0.02 vs. 0.13 ± 0.02; $p < 0.0001$, paired permutation test). This suggests that, although a salient feature existed that was partially preserved across many tested views, rat strategy was nevertheless more consistent with "tracking" that feature (i.e., its position, size, orientation) across the transformations Object 1 underwent, rather than merely relying on the portion of that feature that remained unchanged across such transformations. This observation, together with the fact that the same feature was relied upon also when shifted in non-overlapping locations of the stimulus display (as in the case of the horizontally translated views), indicates that also recognition of Object 1 was more consistent with a high-level, transformation-tolerant representation of diagnostic features, rather than with low-level detection of some transformation-preserved luminance patch. Finally, the fraction of overlap values that were significantly higher than expected by chance was similarly small for both the raw (8/38) and the aligned (9/38) views of Object 1 (see gray and black circles in Fig. 8B). This reflects the fact that relatively large overlaps were produced by chance in the permutation test (given the large area occupied by Object 1's salient regions), thus making the threshold to reach significance higher than in the case of Object 2. This confirms that the saliency regions/maps obtained for Object 2 were, in general, more powerful to understand the complexity of rat recognition strategy, as compared with the ones obtained for Object 1.

Noticeably, the overlap analysis described above was carried out under the "worst case scenario" assumption that rat gaze was stable across trials and, therefore, the relative position of two raw views on the retina matched their relative position on the stimulus display (see previous section). Any deviation from this assumption (i.e., any possible across-trial variability in rat gaze direction) could only reduce the overlap between a pair of raw views on the retina. Therefore, for any given pair of views, the raw overlap reported in Figure 8B actually represents an *upper bound* of the possible raw retinal overlap. Finding that such an upper bound was, in general, lower than the overlap between aligned views, guarantees that, regardless of the stability of rat gaze direction, no trivial recognition strategy (based on detection of some transformation-preserved diagnostic features) underlay rat recognition behavior.

As a further assessment of the complexity of rat recognition strategy in the face of vari-

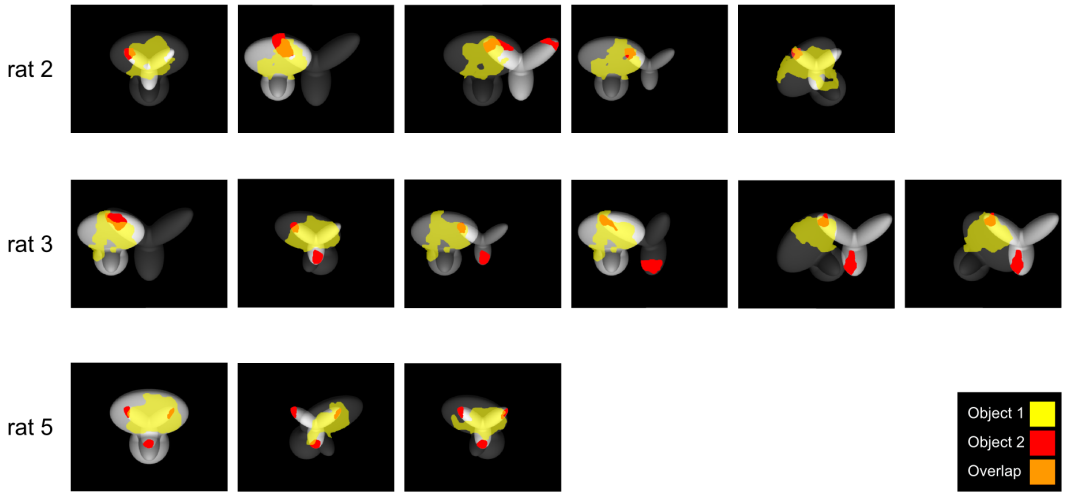

Figure 2-9: Overlap between the salient features obtained for exemplar views of Object 1 and 2. Several examples in which one or more salient features of Object 2 (red patches located at the tips of the upper and bottom lobes) overlapped with the salient feature located in the upper lobe of Object 1 (yellow patch). Overlapping regions are shown in orange.

ation in object appearance, we checked whether any salient feature underlying recognition of Object 1 overlapped with a salient feature underlying recognition of Object 2. The rationale behind this analysis is that, if the features that are salient for a view of Object 1 and a view of Object 2 do overlap, then the area of the overlap within the stimulus display (i.e., the animal's visual field) cannot, *per se*, be diagnostic of the identity of any object. This, in turn, implies that rats must definitely adopt a strategy that goes beyond associating high luminance in a given local region of the display with a given object identity – they need to take into account the shape (e.g., size, orientation, aspect ratio, etc) of that high luminance region and, possibly, rely on the presence of additional diagnostic regions (i.e., salient features) to successfully identify what object that region belongs to. As shown in Figure 9, several cases could indeed be found, in which one of the salient features (red patches) located in the top lobes of Object 2 overlapped with the salient feature (yellow patch) located in the top lobe of Object 1 (the overlapping area is shown in orange).

Overall, the overlap analyses shown in Figures 7-9 indicate that rat invariant recognition of visual objects does not trivially rely on detection of some transformation-preserved object features that are diagnostic of object identity across multiple object views. Instead, rat recognition appears to be consistent with the existence of higher level neuronal representa-

tions of visual objects that are largely tolerant to substantial variation in the appearance of the objects' diagnostic features.

2.4.5 Comparison between the critical features' patterns obtained for the average rat and a simulated ideal observer

Having found that rats are capable of an advanced, shape-based and transformation-tolerant recognition strategy raises the question of just how optimal such strategy is, given the amount of discriminatory information a pair of visual objects (each presented under many different viewing conditions) affords. To address this issue, we built a linear ideal observer and we extracted the critical features underlying its recognition of the same bubble-masked images that had been presented to one of the rats. The simulated observer was *ideal*, since it had stored in memory, as templates, the eight views each object could take (i.e., those marked by red frames in Fig. 1C), and was *linear*, since it classified each bubble-masked input image as being either Object 1 or 2, based on which of these templates had the highest correlation with the image itself (see Materials and Methods for details). Given its full access to all possible appearances the objects could take, the ideal observer, by construction, was able to perform optimally in the invariant recognition task and, as such, its recognition strategy represents an upper, optimal bound. Note that the ideal observer considers the object in a pixel-based representation, and the relations between the pixels are not considered.

The saliency maps obtained for the ideal observer were compared with rat group average saliency maps, i.e., the maps obtained by pooling the bubbles trials collected for a given object view across all available rats. Such group average maps summarized rat invariant recognition strategy in a way that was more robust to noise (given the larger number of trials they were based on) and more suitable for comparison with the ideal observer, since idiosyncratic aspects of individual rat strategies were averaged out, while the features that were more consistently relied upon across subjects emerged more clearly. As a result, the patterns of critical features extracted from the average saliency maps (see red and cyan patches in Figs. 10A, B, top rows) were a cleaner version of what observed at the level of individual rats (see Fig. 6). For Object 1, a large salient region (covering most of the upper

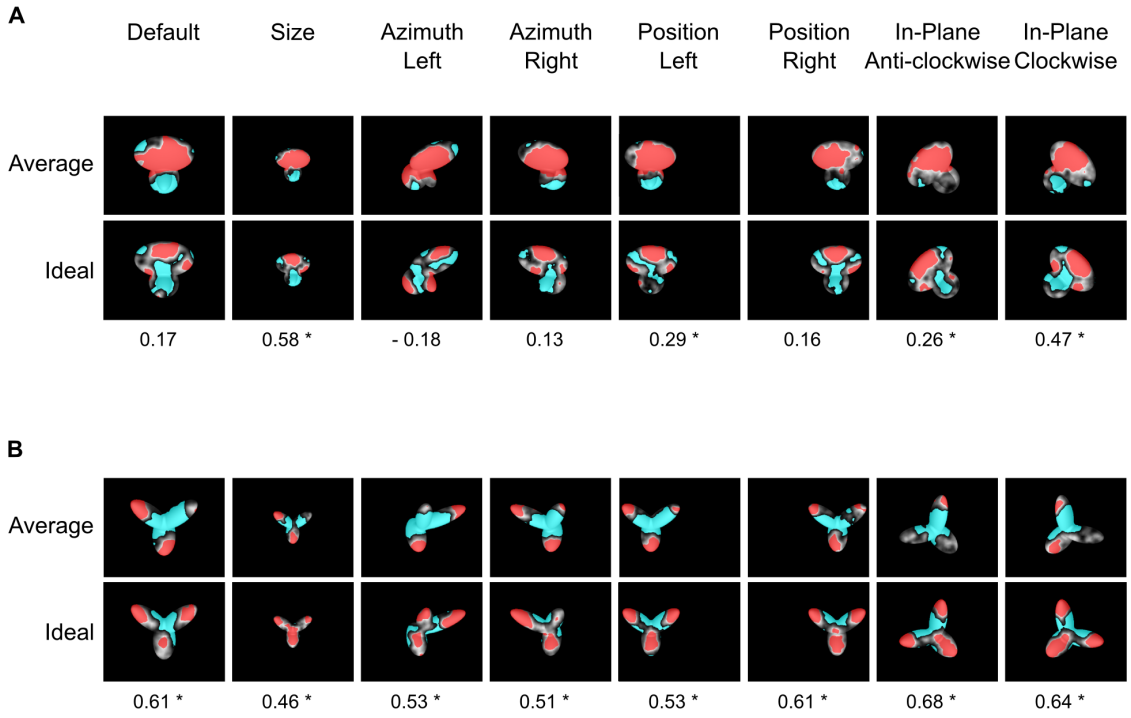

Figure 2-10: Critical features' patterns obtained for the average rat and a simulated ideal observer. Rat group average saliency maps, with highlighted significantly salient (red) and anti-salient (cyan) features (top rows in **A** and **B**), are compared with the saliency maps obtained for a linear ideal observer (bottom rows in **A** and **B**). For each object view, the Pearson correlation coefficient between the saliency maps obtained for the average rat and the ideal observer is reported below the corresponding maps (* indicates a significant correlation at $p < 0.05$; permutation test).

lobe) and a smaller anti-salient region (covering the bottom part of the lower lobes) were found (Fig. 10A, top row). For Object 2, different combinations of salient features (located at the tips of the lobes) and a large anti-salient area (located at the lobes' intersection) were found (Fig. 10B, top row).

These patterns of critical features bore many similarities, but also some key differences, with those obtained for the ideal observer (Figs. 10A and B, bottom rows). The structural parts in which the salient and anti-salient features were located were largely the same for the rats and the ideal observer. However, in the case of Object 1, the salient region in the upper lobe was fragmented and smaller for the ideal observer, as compared with the average rat, while the anti-salient region was larger, extending from the bottom lobes to the

upper one, and branching in two arms that resembled an outline of Object 2 (compare top and bottom rows in Fig. 10A). In the case of Object 2, the location and size of the salient features found for the ideal observer and the average rat closely matched, although the salient region in the bottom lobe was larger for the ideal observer and typically extended over the lobes' intersection, at the expenses of the central anti-salient region, which was smaller and restricted to the base of the upper lobes (compare top and bottom rows in Fig. 10B). For any given object view, the similarity between the saliency maps obtained for the average rat and the ideal observer was quantified by computing the Pearson correlation coefficient (r values are reported under each pair of saliency maps in Fig. 10). Such a correlation was significantly higher than expected by chance for four out of eight views of Object 1 and for all the views of Object 2 ($p < 0.05$; permutation test; see Materials and Methods).

Overall, this comparison shows that rat recognition strategy was highly consistent with that of the ideal observer and, as such, relied on close-to-optimal use of the discriminatory information afforded by the two objects across their various appearances. At the same time, it is interesting to note that where the rats' strategy departed from the ideal one was mainly because it better parsed the structure of the objects. That is, objects' structural parts, such as the upper lobe of Object 1, were considered salient as a whole by rats, while the ideal observer carved the negative image of Object 2 out of Object 1, even if this operation resulted in a critical features' pattern that did not match the natural boundary of Object 1's upper lobe (see Discussion for possible implications).

In general, the pattern of critical features found for the view of a given object closely resembled the negative image of the pattern of visual features found for the matching view of the other object. This was more apparent for the ideal observer (because of the above-mentioned carving of the silhouette of Object 2 out of Object 1) but it was true also in the case of rat recognition strategy. In fact, all the saliency maps obtained for matching views of the two objects in Figures 6 and 10 shows a clear phase opposition. This was quantified, in the case of the ideal observer and the average rat, by computing the Pearson correlation coefficient between saliency maps of matching object views. Most correlation coefficients ranged between -0.6 and -0.8 (see Table 2-1) and were all significantly lower than expected

Table 2.1: Phase opposition of the saliency maps obtained for matching views of Object 1 and 2. Pearson correlation coefficients between the saliency maps obtained for matching views of Object 1 and 2 (i.e., the same maps shown in Figure 10). For both the average rat (top row) and the ideal observer (bottom row), the correlation coefficients were all negative and significantly lower than expected by chance (* indicates significance at $p < 0.05$; permutation test).

	Default	Size	Azimuth Left	Azimuth Right	Position Left	Position Right	In-Plane Anticlockwise	In-Plane Clockwise
Average rat	-0.80^*	-0.66^*	-0.82^*	-0.90^*	-0.76^*	-0.67^*	-0.76^*	-0.75^*
Ideal observer	-0.43^*	-0.61^*	-0.63^*	-0.60^*	-0.68^*	-0.76^*	-0.77^*	-0.78^*

by chance ($p < 0.05$; permutation test; see Materials and Methods), thus showing that the saliency maps of matching object views were strongly anti-correlated, for both the average rat and the ideal observer. Although the average correlation coefficient was larger for the average rat than for the ideal observer (-0.76 ± 0.03 vs. -0.66 ± 0.04), such a difference was not significantly larger than expected by chance ($p = 0.5$, paired permutation test). Overall, this suggests that the phase opposition of the saliency maps obtained for matching views of the two objects is a property of rat recognition strategy that is fully consistent with extraction of the optimal discriminatory information afforded by the tested objects' views.

2.5 Discussion

In this study, we investigated the perceptual strategy underlying rat invariant recognition of visual objects, by exploiting an image classification technique known as the Bubbles method (Gosselin and Schyns, 2001; Gibson et al., 2005; Gibson et al., 2007; Nielsen et al., 2006; Nielsen et al., 2008; Vermaercke and OpdeBeeck, 2012). This approach uncovered four key aspects of rat recognition strategy.

First, when it comes to recognize a given object view, rats appear to rely on most or all the distinct structural parts that object view affords (see Fig. 6). Second, for many rats, the recognition strategy was remarkably stable in the face of variation in object appearance. That is, in many cases, the combination of diagnostic structural parts a rat relied upon was the same across all or most the object views the animal faced (see Figs. 6A,B). Third, no trivial low-level strategies (e.g., relying on transformation-preserved diagnostic features; see Figs. 7-9) could account for rat invariant recognition behavior. Forth, the critical features' patterns underlying rat recognition strategy closely (although not fully; see Results) matched those obtained for a simulated ideal observer engaged in the same invariant recognition task (see Fig. 10). Overall, these findings imply that rats: 1) do process global shape information and make close-to-optimal use of the array of diagnostic features an object is made of; and 2) do so, in a way that is largely tolerant to variation in the appearance of diagnostic object features across a variety of transformation axes.

2.5.1 Comparison with previous studies

Our findings directly compare with those of two recent studies (Minini and Jeffery, 2006; Vermaercke and OpdeBeeck, 2012). Minini and Jeffery (2006) found that rats did not process global shape information and relied, instead, on luminance in the lower half of the stimulus display to discriminate two geometrical shapes (a square and a triangle). Vermaercke and Op de Beeck (2012) also concluded that rats discriminate squares and triangles by relying on their bottom part, unless this part is largely occluded (the authors too used the Bubbles method) – hence, the authors' conclusion that rats are capable of a mid-level, context-dependent recognition strategy.

Our findings not only contradict the low-level account of rat visual processing provided by Minini and Jeffery, but also argue in favor of an invariant, shape-based, multi-featural recognition strategy that is way more advanced than postulated by Vermaercke and Op de Beeck. Such a discrepancy can be explained by several key methodological differences between these previous studies and ours.

First, in our study we used renderings of three-dimensional object models that were made of several structural parts, as opposed to the simple, planar geometrical shapes used in these previous studies. Our findings show that the complexity of rat recognition strategy closely matches the structural complexity of the object to process (see Results and Fig. 6). Therefore, the squares and triangles used in these previous studies simply lacked the structural complexity to properly probe advanced visual shape processing.

Second, in our experiments, rats were required to tolerate large variation in object appearance along a variety of transformation axes, while, in these previous studies, a much smaller number of simpler transformations was tested [e.g., only position changes were tested in Vermaercke and Op de Beeck (2012)]. The extended training in a challenging, invariant recognition task was likely crucial to engage rat more advanced shape processing abilities in our study.

Finally, these previous studies used a two-alternative forced-choice procedure that requires the animals to compare two simultaneously presented visual objects, which may consistently differ in some low-level visual feature even across transformed views, especially if both objects are equally transformed. Noticeably, this is the case of Vermaercke and Op de Beeck (2012), in which, when position tolerance was tested, not only both target shapes were simultaneously presented to the rats and shifted of the same amount, but they were also covered by the same pattern of occluding bubbles. Therefore, the shapes could have been easily discriminated by adopting such a low-level strategy as looking for the stimulus that was brighter in its lower part. As a consequence, the conclusions of Vermaercke and Op de Beeck, while useful to explain the findings of Minini and Jeffery, cannot be taken as a general assessment of rat invariant recognition abilities. By contrast, in our study, each transformed object view was presented in isolation, and a rat had to implicitly compare it with all other possible transformed views of the other object to succeed in the

task. This forced the subjects to perform a truly transformation-tolerant recognition task, which could hardly be solved by relying on low-level image cues.

2.5.2 Validity and limitations of our findings

One limitation of our study is that we did not probe pure generalization of rat recognition to novel object views. This requires withholding feedback (e.g., reward) to the rats about the correctness of their response, which can only be done in a small fraction of trials (Zoccolan et al., 2009). However, very likely, rat recognition of the transformed views mainly resulted from generalizing rather than learning/memorizing each individual view. This speculation is based on three arguments. First, two previous studies (Zoccolan et al., 2009; Tafazoli et al., 2012) have rigorously established that rats do spontaneously generalize their recognition to novel appearances of visual objects across many different transformations axes/ranges (including those tested in this study). Second, rat progression along each transformation axis during the staircase training was very quick, in some case requiring a single training session (see Materials and Methods). Third, bubbles trials were randomly interleaved with regular trials, in which objects were sampled from any of the variation axes a rat had been exposed to. For instance, when bubbles were applied to in-plane rotations, any of 78 possible different views could be presented in a regular trial, which makes highly unlikely that rats memorized each of them. Moreover, the bubbles masks themselves produced large changes in object appearance. Crucially, masks varied randomly across trials, thus making even more unlikely for rats to memorize each object appearance.

Another potential limitation of our study is that rat gaze was not monitored. However, our study was appositely designed so that the lack of precise eye control did not affect our conclusions. First, rat recognition strategy was recovered by a method that, by its very nature, is not only alternative, but superior to eye-tracking in revealing the visual information correlated with behavioral outcomes [see (Schyns et al., 2002; Murray, 2011; Jack et al., 2012)] – none of the behavioral studies that have previously applied the Bubbles method in humans, avian and rodents has made use of eye-tracking or fixation dots [e.g., see (Gosselin and Schyns, 2001; Gibson et al., 2005; Vermaercke and OpdeBeeck, 2012)].

Second, with the exception of position shifts, none of the transformations the objects underwent in our study could be *undone* by compensatory eye movements (i.e., saccades). Although size changes and in-plane rotations could, in principle, be compensated by head movements, rat head position was highly reproducible/stable across trials (see Materials and Methods). This guaranteed full control over stimulus size and in-plane rotation. Azimuth rotation was equally well controlled, given that view-point changes were virtual. Finally, position shifts too were unlikely to be compensated by saccades. In fact, the increase of reaction time as a function of stimulus position was so gradual and small (see Fig. 5C) to exclude that rats, following the trial onsets, made target-oriented saccades to bring the horizontally shifted stimuli always in the same retinal position [this, by itself, is a very remote possibility, since rats do not have a fovea (Paxinos, 2004), saccade much less frequently than primates do (Chelazzi et al., 1989; Zoccolan et al., 2010), and no evidence of target-oriented saccades has even been reported in rodents]. Finally, although we cannot exclude possible (unmeasured) trial-by-trial variations in the retinal position of the object views (caused by across-trial variations in rat gaze direction), such variations would make the recognition task even *more* (not *less*) demanding in terms of tolerance to variation in object appearance. This, in turn, would make even less likely for rats to rely on some low-level strategy to succeed in the recognition task. Therefore, the lack of gaze monitoring in our study could only result in an *underestimation* (not an *overestimation*) of rat invariant shape processing abilities (e.g., see the overlap analysis shown in Fig. 8).

2.6 Conclusions

Over the past five years, a new tide of studies, encompassing behavior (Zoccolan et al., 2009; Meier et al., 2011; Busse et al., 2011; Tafazoli et al., 2012; Vermaercke and Op de Beeck, 2012; Histed et al., 2012), imaging (Greenberg et al., 2008; Sawinski et al., 2009; Bonin et al., 2011; Andermann et al., 2011; Marshel et al., 2011) and electrophysiology/anatomy (Van Hooser, 2006; Niell and Stryker, 2008; Niell and Stryker, 2010; Gao et al., 2010; Wang et al., 2011; Kerr and Nimmerjahn, 2012) has reignited the interest for the use of rodent models in vision research (Van Hooser, 2006; Huberman and Niell, 2011;

Niell, 2011). However, it remains unclear to what extent the rodent visual system can support advanced, shape-based processing of visual objects. To our knowledge, our study provides the most compelling evidence, to date, that rats process visual objects through rather sophisticated, shape-based, transformation-tolerant mechanisms. As such, given the powerful array of experimental approaches that are available in rats (Ohki et al., 2005; Lee et al., 2006; Sawinski et al., 2009), this model system will likely become a valuable tool in the study of the neuronal mechanisms underlying object vision.

Chapter 3

Critical Shape Features Underlying Object Recognition in Human

3.1 Introduction

In the study described in Chapter 2, we applied an image classification technique (i.e., the Bubbles method) to uncover the perceptual strategy underlying rat invariant recognition of visual objects. Our primary goal was to understand whether rats are capable of processing visual shape information, given that a previous study has concluded that they are not (Minini and Jeffery, 2006). Our secondary goal was to understand how advanced rat shape processing is, given that a previous study has concluded that rat recognition relies on a low-level strategy (Vermaercke and OpdeBeeck, 2012). Our final goal was to understand how stable rat recognition strategy is across the different views afforded by an object, given that previous studies have shown that rats are capable of invariant recognition (Zoccolan et al., 2009; Tafazoli et al., 2012), but they have not investigated what strategy underlies this ability. Our comparison with the recognition strategy of a simulated ideal observer served as a benchmark to quantitatively understand how efficient is rat extraction/processing of visual discriminatory information in the face of variation in object appearance.

A natural extension of this study is to compare rat recognition strategy with the strategy of human observers engaged in the same invariant object recognition task. In fact, the human visual system is the most powerful invariant recognition system that is known

to exist in nature. Therefore, the human invariant recognition strategy would provide the ultimate benchmark against which to compare the complexity of rat recognition strategy. More importantly, unveiling the pattern of diagnostic features underlying human recognition of visual objects across different appearances would provide interesting clues about the mechanisms governing human invariant recognition. In fact, while human invariant recognition abilities have been investigated by several performance-based (Blthoff et al., 1995; Tarr and Blthoff, 1998), priming-based (Biederman and Cooper, 1992; Kravitz et al., 2010), or adaptation-based (Afraz and Cavanagh, 2008; Afraz and Cavanagh, 2009) studies, very little is known about the perceptual strategy underlying such abilities. The Bubbles method, as well as other image classification approaches (Murray, 2011), is well fit to address this question. However, only a few studies have applied it to map the patterns of diagnostic features underlying recognition of an object across different views (Nielsen et al., 2008; Vermaercke and OpdeBeeck, 2012). Moreover, in these studies, only views along a single variation axis (e.g., in-plane rotation or position) were tested. As a consequence, many crucial aspects of human perceptual strategy in the face of variation in object appearance remain unanswered. Given a set of objects to discriminate, how consistent the recognition strategy will be among different human observers? Will humans rely on the same set of diagnostic features across object views or, rather, will they adopt a view-dependent recognition strategy? How will human perceptual strategy compare with the strategy of a simulated ideal observer that has stored in memory, as templates, all the views an object can take?

In the study discussed in this chapter, we addressed these issues by testing four human participants in the same invariant recognition task, applying the same image classification approach described in Chapter 2. Our results show that human recognition strategy departs from rat and ideal observer recognition strategy in interesting ways that underscore the peculiarly advanced shape processing abilities of our species.

3.2 Materials and Methods

3.2.1 Participants

Four participants, three females and one male, with normal or corrected to normal vision volunteered for the experiment. Their ages were between 26 and 29 and all were right-handed. They were all naive with respect to the experiment and the visual stimuli. Participants signed a written informed consent and received 10 euro for each hour of testing they underwent. All the procedure of this study was approved by SISSA ethics committee.

3.2.2 Setup

The participants were placed in a sound-proof chamber. A chin-rest was used to fixate the head in front of a 22-inch widescreen Samsung LCD monitor (300 cd/m2; 1680 x 1050 resolution pixel at 60 Hz refresh rate), keeping the distance of the head from the screen fixed at 61cm. The participants were asked to touch the relevant keys of a three-button touch-pad in front of them, initiating trials with the central button and reporting the identity of the object by touching the left-hand side or the right-hand side buttons (each button was associated to a specific object identity). All of the experimental protocol was implemented using, the freeware, open-source package *MWorks* (http://mworks-project.org) with a custom-written plugin to build trial-specific bubbles masks in real time. The experiment was run on a Mac mini (2.53GHz intel Core 2 Duo processer; NVIDIA GeForce 9400 graphic card; Mac OS X 10.6.7).

3.2.3 Visual Stimuli

Each participant was trained to discriminate the default views of the same objects used in the rat experiment (see Chapter 2 and Figure 3-1A). The objects default size was 17.5º of visual angle and their default position was the center of the monitor. The viewing distance (i.e., the distance between the eyes of the participants and the monitor) was 61 cm. Therefore, as compared with the rat experiment, the viewing distance was larger and the default size was smaller by a factor 0.5. This scaling factor was applied to all the transformed

object views that were presented to the human observers in later phases of the experiment, so as to keep the relative size and distance between object views the same as in the rat experiment.

3.2.4 Experimental Design

In the preliminary phase, the participants were familiarized with the default views of the two objects and instantly they learned to associate Object 1 to the left-hand side button of the touch-pad and Object 2 to the right-hand side button. The structure of each trial was as following. At the onset of each trial, a fixation dot appeared on the stimulus display. Participants were instructed to fixate on the dot and then press the central button on the touch-pad to trigger stimulus presentation. One of the two target objects (see Figure 3-1A) was shown on the stimulus display for 100 ms, followed by presentation of a mask consisting of a grid of small squares with random brightness and spanning the whole display. The participants were instructed to report the identity of the objects by touching the appropriate left/right buttons. The objects appeared intact, and the performance was close to 100% since the beginning of the training because the task was really easy for human observers.

Following this familiarization phase, participants were tested in two consecutive test phases, in which the objects were partially occluded by bubbles masks of the kind defined in Chapter 2, and participants received no feedback about the correctness of their response (see Figure 3-2A and 3-2B) . Briefly, a bubbles mask is an opaque frame that is punctured by a number of circular transparent apertures, called bubbles. Bubbles masks were obtained by manipulating the transparency channel of the image as described in Chapter 2. Using this method, we were able to randomly sample a subset of shape features in each trial, testing whether participants were able to recognize the objects given the available shape information in each trial. The number of bubbles was adaptively adjusted on a trial-by-trial basis so that the performance of the participants remained ~75% correct responses – performance was computed in consecutive blocks of 10 trials and the number of bubbles was increased/decreased if performance was above/below 70/80% correct responses. The number of bubbles changed between 3 and 25. Examples of trials with different number

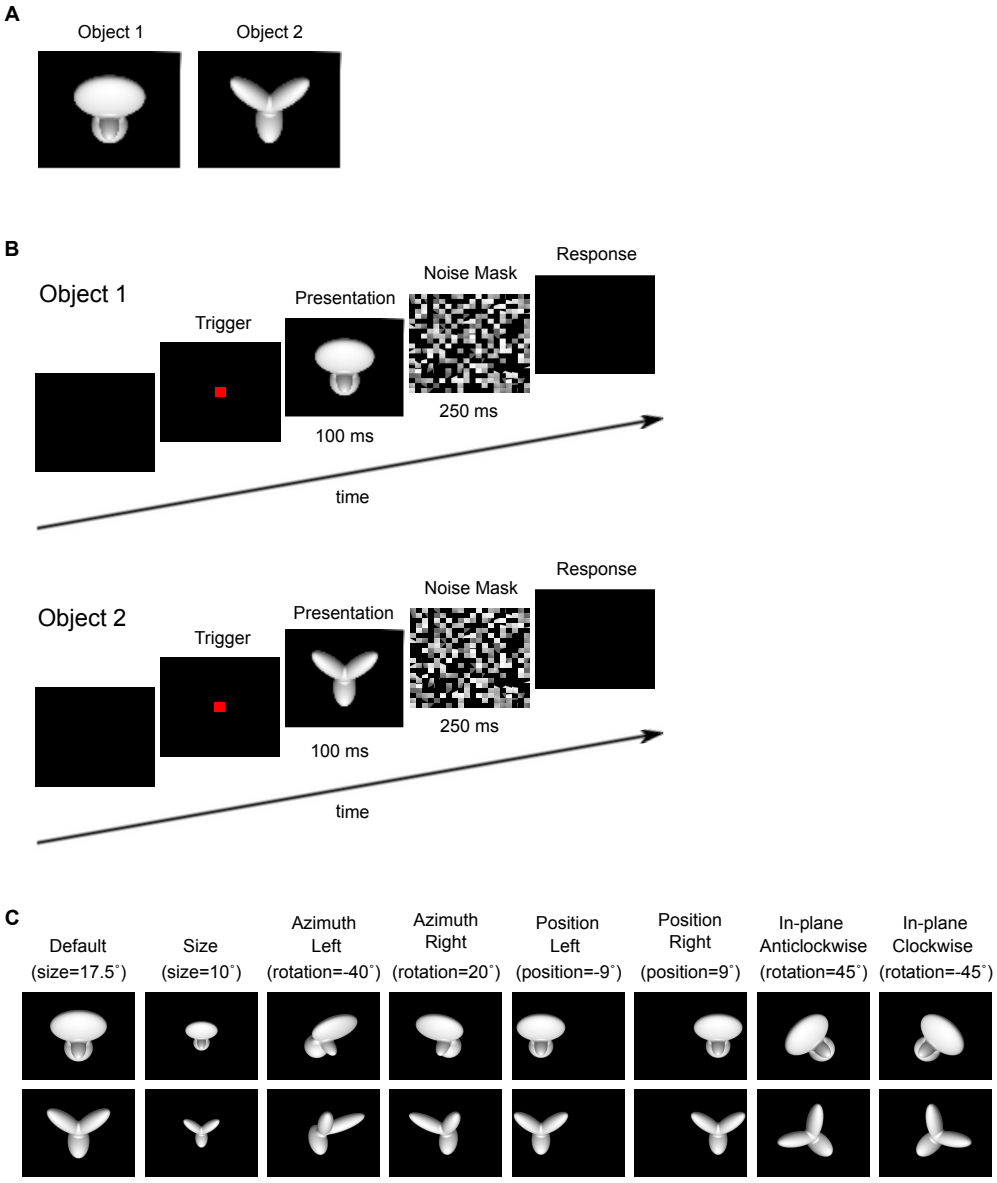

Figure 3-1: Visual stimuli and behavioral task in the human experiment. **A.** Default views (0º in-depth and in-plane rotation) of the two objects that humans were trained to discriminate during the training phase (each object default size was 17.5º of visual angle). **B.** Schematic of the object discrimination task. Once a fixation dot appeared on the stimulus display, human participants had to fixate and then press a button to trigger stimulus presentation. Stimuli (i.e., one of the object views shown in **A.**) were immediately followed by a noise mask. The stimuli were presented on an LCD monitor humans during the training phase (each object default size was 17.5º of visual angle). **C.** The unoccluded transformed views of the two objects that humans were required to recognize during the test phase. Transformation axes included: size changes; azimuth in-depth rotations; horizontal position shifts; and in-plane rotations. Azimuth rotated and horizontally shifted objects were also scaled down to a size of 15.5º of visual angle; in-plane rotated objects were scaled down to a size of 16.25º of visual angle and shifted downward of 1.75º.

of bubbles are shown in Figure 3-2C. A small fraction of *regular trials* (<5%) with intact default views of the objects were randomly interleaved with the trials in which bubbles masks were superimposed on the objects (named *bubbles trials*).

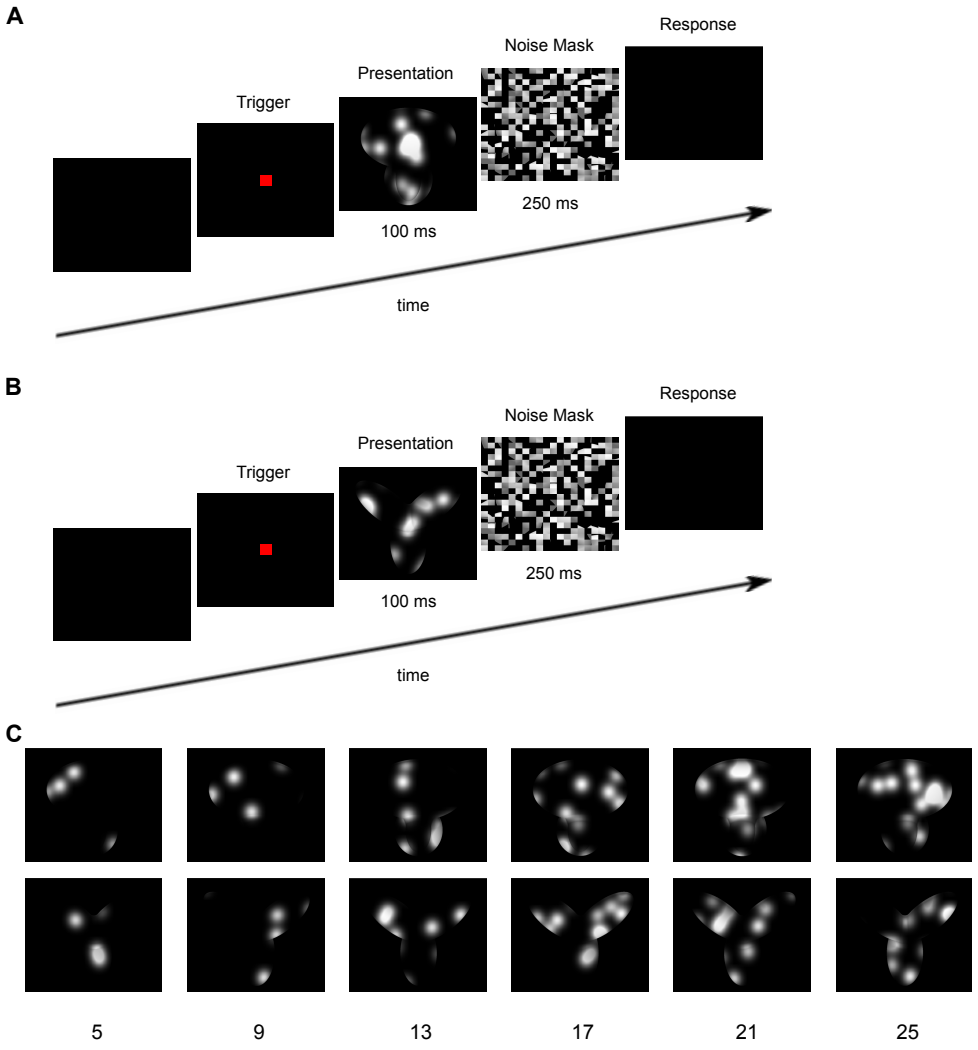

Figure 3-2: The Bubbles method. **A.** Example of bubbles trial in which the default view of Object 1 was occluded by a bubbles mask in test phase I of the human experiment. **B.** Example of bubbles trial in which Object 2 was occluded. **C.** Examples of the different degrees of occlusion of the default object views that were produced by varying the number of bubbles in the bubbles masks in the human experiment.

In test phase I, only the default views of the objects (shown Figure 3-1A) were tested (by randomly interleaving regular and bubbles trials, as explained above). In test phase II, along with the default view, seven additional views of each object were introduced on which bubbles masks were superimposed (see Figure 3-1C). Object transformations included: 1)

size reduction (down to 10º of visual angle); 2) left and right rotations about the objects vertical axes (i.e., -40º and 20º in-depth azimuth rotations); 3) horizontal leftward and rightward shifts (i.e., ±9º horizontal shifts); and 4) clockwise and anticlockwise in-plane rotations (±45º). The transformed object views were always shown occluded by bubbles masks in randomly interleaved trials. For a meaningful comparison with the results shown in Chapter 2, the eight views used in test phase II matched those that were tested in the bubbles trials of the rat experiment (see Figure 2-1C, red frames), although sizes/distances were scaled down by a factor 0.5 (see next section).

The experiment was conducted in blocks of 15~20 minutes in which ~500 trials were collected on average (in some cases the blocks lasted up to 30 min). Overall, about 39 blocks were carried out across 1~3 weeks. Depending on the participant, 1-6 blocks were run in a day with at least 5 min break time between blocks. The average number of collected trials per person was 3,000 in the first phase and 18,000 in the second phase.

3.2.5 Comparison with the rat experiment in Chapter 2

One of the goals of this study was to compare the recognition strategies of humans with those obtained for the rats in Chapter 2. For this reason, we designed a task that was as close as possible to the one used in the rat experiment. However, some adjustments to the design used in the rat study were necessary to make the visual task more suitable to human superior visual recognition and cognitive abilities.

- In the test phases, we did not provide feedback to the human participants, while with rats we were forced to provide them feedback, because the task was already exceptionally demanding and the bubbles trials were more than half of the total trial count (therefore, without receiving reward in bubbles trials, the animals would have simply stopped working). Obviously, this difference has an important implication: in the human study described in this chapter, we could test pure generalization of recognition to novel object views, while, in the rat study, we cannot exclude that rat recognition performance/strategy was at least partially based on learning the association between each novel view and the appropriate reward port (although we believe that this is

unlikely; see Discussion in Chapter 2).

- The participants were not exposed to the intact transformed object views. Moreover, unlike the experiment with rats, we did not gradually expose the participants to increasingly large transformations along a given variation axis. For example, in the case of the azimuth rotation axis, we trained the rats to tolerate increasingly wider rotations, starting from 0º up to from ±60º. After this training/exposure phase took place, we selected two rotated views within this range (namely -40º and 20º) and we presented these views occluded by masks in bubbles trials. Instead, in the experiment with humans, we just selected these two views for bubbles trials

- Differently from the rat experiment, we used an adaptive procedure to automatically update the number of bubbles on a trial-by-trial base, so as to keep the performance of the participants at ~75% correct responses.

- In the human experiment, all object views (with superimposed bubbles masks) were shown in randomly interleaved trials during the course of all testing sessions (in phase II). Instead, in the rat experiment, we tested in bubble trials one object view at the time, and only after collecting enough bubbles trials for a given view we switched to the next one.

- In the human experiment, the size of the objects and their distance from the center of the monitor were scaled down to the half of the size (and distance) used in the rat experiment.

- To avoid ceiling of human performance and to make the task hard enough so that human participants make some mistakes (which are essential to uncover the diagnostic object features through the Bubbles method), we presented the masked objects very briefly (~100 ms) and we used a backward-masking protocol (Serre et al., 2007).

3.2.6 Data Analysis

Data were analyzed using the same approach described in Chapter 2. All the data analysis code was developed in MATLAB (2011a, MathWorks Co, `http://www.mathworks.`

com) and the data was stored in SQLite (http://www.sqlite.com) databases. All the data analysis was run on an Apple Mac Pro computer (with 2 CPUs at 2.93GHz Quad-Core Intel Xeon; Mac OS 10.6.8).

3.3 Results

We tested four human participants in the experimental phases described in Materials and Methods. The first phase was just meant to familiarize the participants with the default views of the visual objects they had to discriminate and teach them the correct association between objects and response buttons. The following test phases were meant to uncover the patterns of diagnostic features underlying recognition of the default (test phase I) and transformed views (phase II) of the objects. This was achieved by occluding the visual objects with bubbles masks, as explained in Materials and Methods and in the previous chapter. On average, ~3,150 bubbles trials per participant were collected in phase I and ~2,500 in phase II per object view per participant.

3.3.1 Critical Shape Features Underlying Recognition of the Default Object Views

During test phase I of the study, the bubbles masks were applied to the default views of the visual objects (see Figure 3-2). The rationale behind the application of the bubbles masks was that, by applying a multiplicative noise field to the object stimuli, the task became harder, since the objects became only partially visible through the masks. As explained in Chapter 2, the level of occlusion produced by a bubbles mask depended on two parameters, both affecting recognition performance: the number of bubbles and their size (i.e., standard deviation of the Gaussian aperture of a semi-transparent bubble; see Chapter 2). The ratio of bubbles size to objects size in the default view was set to the same value as in the rat experiment, so as to make the saliency maps comparable in the two experiments (i.e., the bubbles size of 1° and the object size of 17.5° were used in the default view of the human experiment). The number of bubbles was adjusted in real-time, through an adaptive

procedure, to keep the performance of each participant in bubbles trial close to 75% correct. The number of bubbles ranged between 3 and 25, thus producing a much larger occlusion than in the rat experiment (compare Figure 3-2C to Figure 2-2B). On average, across trials, the number of bubbles (±SD) for the four participants in phase I was: 4.5 ± 1.6 (H1), 3.7 ± 1.1 (H2), 6.6 ± 2.3 (H3) and 3.8 ± 1.3 (H4). In the phase II, since the task was harder, these averages increased to the following: 5.3 ± 1.2 (H1), 5.6 ± 1.9 (H2), 8.4 ± 3.2 (H3) and 6.1 ± 2.3 (H4).

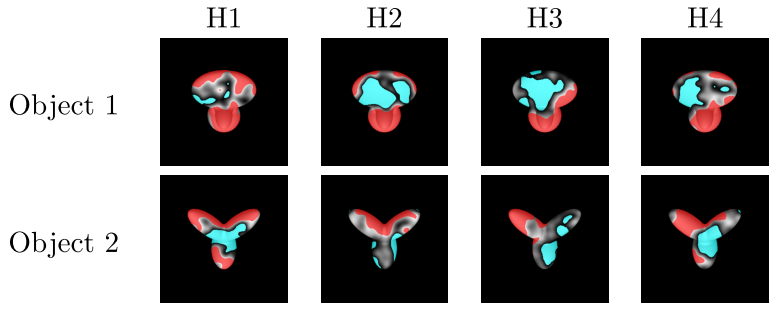

Figure 3-3: Critical features underlying recognition of the default object views in test phase I of the human experiment. For each participant, the saliency maps resulting from processing the bubbles trials collected for the default object views are shown as grayscale masks superimposed on the images of the objects. The brightness of each pixel in an object view is proportional to the correlation of the pixel brightness with the behavioral outcome. Significantly salient and anti-salient object regions (i.e., regions that were significantly positively or negatively correlated with correct identification of an object; $p \leq 0.05$; permutation test) are shown, respectively, in red and cyan.

As done in the rat experiment (see Chapter 2), for each participant, we computed saliency maps by measuring the correlation between bubbles masks transparency values and behavioral responses of the participant, for each object separately. The saliency maps are shown as gray-scale masks overlaid on the images of the corresponding objects (see Figure 3-3).The statistical significance of each pixel value in a saliency map was assessed by comparing that value against the 5^{th} and 95^{th} percentiles of a null distribution of saliency map values. Such a null distribution was obtained by randomly shuffling 100 times the labels reporting the trials' outcomes (i.e., correct or wrong), within the set of trials with the same number of bubbles (see Chapter 2 for details). The values above the 95^{th} percentile were taken as *significantly salient* regions (or features) and those below the 5^{th} percentile as *significantly anti-salient* regions. These regions correspond to those objects'

parts, whose visibility (through the bubbles masks) was more correlated/anti-correlated with the responses of the participant, and, therefore, was more likely to lead, respectively, to correct identification and misidentification of an object view. The pattern of salient and anti-salient regions obtained for a given object view revealed the strategy underlying recognition of that view, in the context of the task that the participant had to perform.

The patterns of salient features obtained for Object 1 for the four participants (Figure 3-3) show that the most prominent salient feature of the object (and the most preserved across participants) was its bottom part (made of two overlapping lobes). In addition, for some participants (i.e., H1 and H2) the top boundary of the top lobe was almost entirely salient, whereas for one participant (H4) saliency was limited to the top-right side edge of the lobe. Yet another participant (H3) did not rely on the top boundary of the top lobe, rather preferring its bottom right-side edge. Finally, the central part of the top lobe was fully anti-salient for two participants (H2 and H3) and at least partially anti-salient for another participant (H4). This is in striking contrast with the results obtained for the rats – in that case, the central part of the top lobe was the only salient feature of Object 1 (see Figure 2-3B), while its bottom part was its main anti-salient feature (see below for further discussion). In particular, it should be noticed that all human participants relied on multiple salient features to recognize Object 1 and that such features were distributed across the whole object (i.e., both in the top and bottom lobes).

The saliency patterns obtained for Object 2 in the human experiment showed some similarities with those obtained in the rat experiment. As observed for the rats (see Figure 2-3B), any of the three lobes of Object 2 could be used by the human participants as salient diagnostic features. Two participants used all three lobes (i.e., H1 and H4), although in different proportions—H1 relied on equally large portions of the lobes, while H4 mainly relied on the left one. Another participant relied on both top lobes (H2), while the forth participant (H3) relied on just one lobe (the left one). Therefore, most human participants relied on a combination of at least two distinct structural parts (i.e., lobes) to recognize Object 2, similarly to what most rats did. However, in spite of this similarity, rat and human strategies differed in a subtle but crucial way. Rats only used the tip of the lobes (see Figure 2-3B), while most human participants (H1, H2 and H4) relied on the edges of

the top lobes, often not including their tips (see H2 and H4), but including, instead, their v-shaped conjunction (or cotermination). This is an indication of superiority of human visual system in shape processing.

Overall, these data show that humans, even more than rats, use multiple shape features to discriminate visual objects and, differently from rats, tend to rely on the boundary of the objects and the cotermination of their structural parts, when available (e.g., in the case of Object 2). This suggests a superior ability of our species to process edges and complex shape features (such as the v-shaped conjunction of Object 2's top lobes).

3.3.2 Critical Features Underlying Recognition of the Transformed Object Views

In test phase II, we introduced seven novel views of the target objects. The full set of new and default views of the two objects yielded a total of 16 stimulus conditions that were randomly interleaved during the course of each experimental session. Critically, all object views were always shown in bubbles trials (i.e., partially occluded by bubbles masks) and, contrary to the rat experiment (see Chapter 2), no feedback was ever provided to the participants about whether they correctly recognized them. This allowed assessing how the recognition strategy of human observers generalizes to previously unseen appearances of previously learned objects.

The saliency maps obtained for all the participants and all the object views tested with bubbles masks in test phase II are shown in Figure 3-4. The first interesting aspect to take into account in examining the figure is that the saliency maps obtained for the default views (first column) can be compared with those shown, for the same views, in Figure 3-3. The difference between the two figures is that the participants, in one case, were exposed only to the default views themselves (Figure 3-3), while, in the other case, were exposed to both the default and the transformed views in randomly interleaved trials (Figure 3-4). For most participants, dealing with the transformed object views had an impact on the saliency patterns obtained for the default object views. In the case of Object 1, while the patterns of salient features obtained for participants H1 and H2 were not substantially altered, partici-

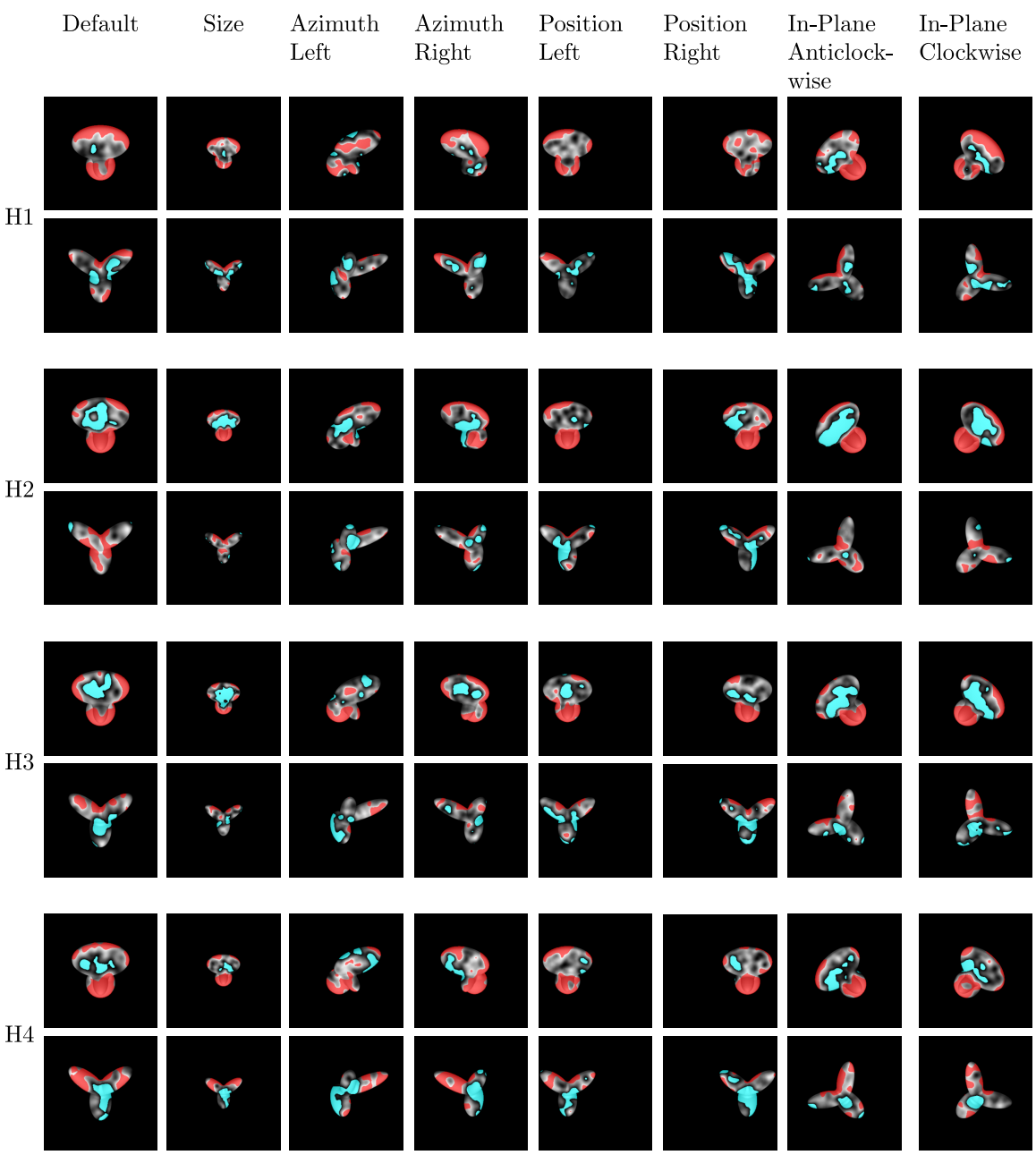

Figure 3-4: Critical features underlying recognition of the transformed views in test phase II of the human experiment. For each participant, the saliency maps (with highlighted significantly salient and anti-salient regions; same color code as in Figure 3-3) that were obtained for each transformed object view are shown. The top-edge and the bottom part of Object 1 and the v-shaped cotermination of the top lobes in Object 2 were consistently used as a diagnostic feature across almost all the views by the human observers.

pants H3 and H4 started relying more robustly on the top boundary of the object's top lobe.

As a result, the saliency patterns obtained for Object 1 became more symmetric and more

consistent/similar across participants. An increased symmetry was also observed for the saliency patterns obtained for Object 2—for all participants, significantly salient features were almost evenly distributed between the top left and right lobes, although their extent was reduced, as compared with what observed in Figure 3-3. Most noticeably, a narrow region marking the v-shaped cotermination of the top lobes became the most prominent and most consistent salient feature across participants, while the bulk of the top lobes failed to reach significance in many cases (e.g., see participant H1). Overall, the introduction of the new object views seems to have brought the participants to rely more consistently, and more evenly, on the edges and cooterminations of the default object views.

As discussed in Chapter 2, Object 2 affords a larger number of perceptual alternatives to be used for its correct identification, as compared to Object 1. This is because Object 2 is made of three lobes that are approximately equal in size and, therefore, equally easy to parse. Therefore, an observer could, in principle, use any combination of lobes to recognize Object 2, and such a combination could change as a function of the specific object view to process. Observation of Figure 3-4 shows that, for some participants, the saliency of a lobe varied considerably across transformations. For instance, in the case of H1, the right lobe was salient for the default, the scaled, the rightward-shifted and the anticlockwise in-plane rotated views, but not for the other views. Other participants showed a more stable pattern of salient features. For example, for H4, portions of the left lobe remained salient across all tested transformations, with the exception of the azimuth left-rotated and the anticlockwise in-plane views. More in general, for all participants, the left lobe stopped being salient for the azimuth left-rotated view, likely reflecting the fact that such a transformation dramatically reduced the size and changed the orientation/position of the left lobe. However, the most remarkable finding emerging from observation of Figure 3-4 is that the v-shaped cotermination of the top lobes of Object 2 was the salient feature all participants consistently relied upon across virtually all the object views. The only view in which the top lobes' cotermination was consistently (across participants) not salient was, again, the azimuth left-rotated view, likely because such transformation brought the left lobe to partially overlap (and occlude) the right one, thus making the angle of their cotermination narrower and harder to parse. Interestingly, however, for most participants (i.e., H1, H2 and H4), the

edge of the right lobe ending at the v-shaped cotermination was still significantly salient.

When it comes to Object 1, unlike rats, who mostly relied on a single salient feature to recognize the object (i.e., the central part of the top lobe; see Figure 2-6), all human observers relied on two distinct object features: the boundary of the top lobe and the bottom part of the object (made of two distinct, but overlapping, lobes). Such a pattern of salient features was mostly preserved across participants and object views. Again, the exception was the azimuth left-rotated view, which was recognized by most participants (with the exception of H2) by relying on the central part of the top lobe, rather than on its boundary. This was probably because the azimuth rotation made the top lobe of Object 1 substantially thinner and oriented along the same axis of Object 2's top-right lobe, thus making harder to use its edge as discriminatory feature of the object. Another interesting observation is that, since the azimuth left rotation removed the overlap between the two bottom lobes, for most participants (i.e., H1, H3 and H4), both lobes emerged as distinct diagnostic salient features.

When compared with rat strategy, human recognition strategy of the transformed object views showed some similarities and some crucial differences. Recognition of Object 1 was based on almost orthogonal sets of diagnostic salient features in rats and humans—the central area of the top lobe in rats versus the boundary of the top lobe and the bottom part of the object in humans (compare Figure 3-4 to Figure 2-6). Interestingly, the recognition strategies of the two species, although very different from each other, were similarly consistent across participants and object views. This implies that the shape of Object 1 was processed in fundamentally different, species-specific ways by rats and humans. The patterns of salient features obtained for Object 2 in rats and humans were more similar, since both species tended to rely on some combination of the three object lobes. However, while the rats exclusively relied on the tips of the lobes (with the lobes' intersection being anti-salient), the human observers consistently relied on a thinner, but more extended, region of the top lobes, which defined their boundary and was centered at their v-shaped intersection/cotermination. This suggests that the rats and the humans relied on partially overlapping strategies to recognize Object 2, with the humans preferentially processing the edges of the object and more complex features' patterns such as the top lobes' cotermina-

tion. Interestingly, rats and humans showed a similar trend with respect to specific object views. For instance, all human observers, as well as three rats (see Figure 2-6C), stopped relying upon Object 2's top-left lobe, when the object was azimuth rotated of 40º to the left (thus making the lobe a less prominent feature). For other transformations rats and humans developed opposite trends. For instance, horizontal translations brought rats to rely more on the part of Object 1's top lobe that was closer to the center of the stimulus display, while most humans tended to rely more on the boundary of Object 1's top lobe that was further apart from the center of the stimulus display.

3.3.3 Transformation-tolerant diagnostic features underlying human invariant recognition

In the rat study, one of our goals was to understand to what extent this species is capable of processing visual shape as compared with relying on lower-level perceptual strategies (such as comparing the overall brightness produced by the two objects in the lower half of the stimulus display). To quantitatively assess this, we ran an overlap analysis, in which the saliency maps obtained for different pairs of views of a given object were superimposed and the overlap values between pairs of significantly salient regions were assessed (see Figure 2-8). This was done for both raw and aligned views (in the latter case, the transformations that produced the two object views were "undone", so as to perfectly align one view on top of the other; see Figure 2-8A). The resulting aligned overlap values were much larger, on average, than the raw overlap values (and more often significant at the level of individual pairs of views), thus showing that: 1) rat recognition strategy could not be trivially accounted by the existence of some transformation-preserved object feature; and 2) the set of object features that were considered diagnostic across views substantially overlapped. This, in turn, suggests that rats likely relied on some transformation-tolerant representation of the objects' diagnostic features.

In the case of humans, there is no doubt about the capability of our species to process visual shape, yet it is interesting to check how much the salient features' patterns obtained for different views of an object overlapped, in the raw and aligned case. In fact, this would

quantitatively estimate how preserved the recognition strategy of the participants was in the face of transformation in object appearance. It would also be an important sanity check to make sure, as in the rat experiment, that the invariant recognition task we designed engaged the participants' higher-level visual processing abilities. To assess this and better compare rat and human recognition strategies, we carried out the same overlap analysis we performed in the rat study (see Figure 2-8), by plotting the raw overlap values versus the aligned overlap values for each pair of views of an object that were produced by affine transformations (see Figure 3-5).

Figures 3-5, and Figure 2-8B show a similar trend. Similarly to what was found in the rat experiment (see Chapter 2), for all pairs of object views, the overlap values between salient features were higher in the aligned than in the raw case, with the average overlap values between aligned views (0.36 ± 0.20; mean ± SEM) being significantly larger than the average overlap values between raw views (0.05 ± 0.06; $p < 10^{-6}$, one-tailed paired t-test). As in the rat experiment, this was true also when overlap values corresponding to the two objects (compare circles vs. diamonds in Figure 3-5) were considered separately—for both Object 1 and 2 the average aligned overlap value was significantly higher than the average raw overlap value (0.52 ± 0.13 vs. 0.06 ± 0.07, in the case of Object 1 and 0.21 ± 0.10 vs. 0.03 ± 0.04, in the case of Object 2; $p < 10^{-6}$, one-tailed paired t-test). Differently from the rat experiment, however, the aligned overlap value was twice as large for Object 1 as for Object 2 (and this difference was significant; $p < 10^{-6}$, one-tailed paired t-test). This is not surprising, since, in the case of Object 2, the salient region that was consistently used across views was the small v-shaped intersection of the top lobes, while the saliency of the lobes themselves showed a larger view-dependence (see Figure 3-4). For Object 1, on the other hand, the same saliency patterns were, in general, consistently were used across views (see Figure 3-4). Finally, as in the case of rats (see Figure 2-8), the large majority of the aligned overlap values were significantly higher than expected by chance ($p < 0.05$; significance was assessed through the permutation test described in Chapter 2; see also Figure 2-8A), while the raw overlap, differently from what observed in the rat experiment, was significantly higher than expected by chance only for one pair of views of Object 2 (see the black diamond in Figure 3-5).

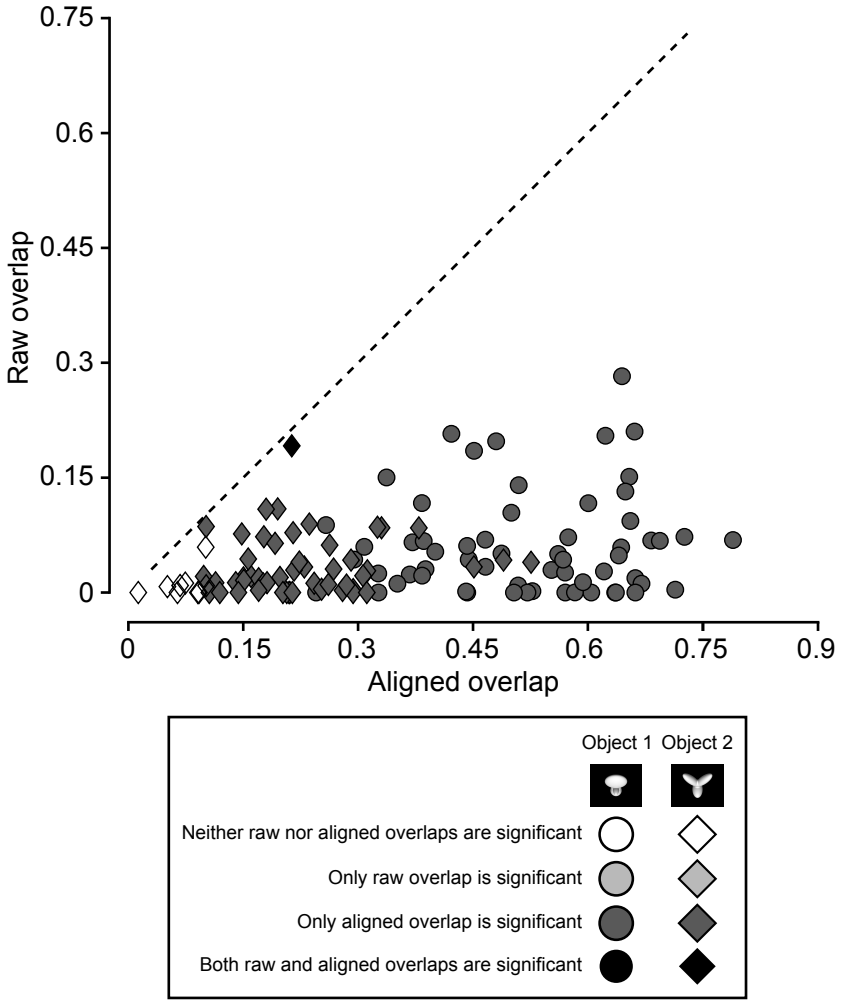

Figure 3-5: *Raw* vs. *aligned* features overlap for all pairs of object views in the human experiment. The raw features's overlap values are plotted against the aligned features overlap values for each pair of views of Object 1 (circles) and Object 2 (diamonds) resulting from affine transformations (i.e., position/size changes and in-plane rotations). The shades of gray indicate whether the raw or/and the aligned overlap values were significantly larger than expected by chance (p ¡ 0.05; permutation test). The raw and aligned overlap values are computed the same way as in the rat experiment (illustrated in Figure 2A).

Overall, the overlap analysis showed that human recognition strategy did not rely on any transformation-preserved diagnostic features. Rather, the large overlap values obtained for the aligned view pairs imply that humans' strategy tracked the objects' diagnostic features across the transformations the objects underwent, even more that rats did (especially for Object 1). Since our participants never received feedback about the correctness of their responses, this strongly suggests that human recognition was based on transformation-

tolerant representations of the objects' discriminatory features.

3.3.4 Comparison between critical features' patterns obtained for the average human and a simulated ideal observer

We performed the same ideal observer analysis that we carried out in the rat experiment, with the difference that here we did not low-pass filter the images of the object views (this was done, in the rat experiment, to match rat retinal spatial resolution). Rather the full-resolution images used in the human experiment were fed to the simulated ideal observer. The ideal observer had stored in memory, as templates, the eight views each object could take (see Figure 3-1C), and classified each bubble-masked input image as being either Object 1 or 2, based on which of these templates had the highest linear correlation with the image itself. Therefore, the saliency maps obtained for the ideal observer provide a benchmark, against which the strategies used by different species can be compared, to understand to what extent they make optimal use of the discriminatory information afforded by the various object views at the level of individual pixels.

The saliency maps obtained for the ideal observer (Figure 3-6A, B, middle rows) were virtually identical to those previously obtained in the rat study (see Figure 2-10), thus showing that image resolution did not affect much the outcome of the template-matching operation performed by the simulated observer. Briefly, for the ideal observer, the top lobe of Object 1 was mostly salient, while the central part and the edges of the bottom lobes were, respectively, anti-salient and salient. The tips of all the lobes of Object 2 were salient for the ideal observer, whereas their intersection/cotermination was consistently anti-salient. As previously explained (see Chapter 2), such saliency patterns bore many similarities with those obtained for the average rat (see Figure 2-10), which, for easier comparison, are shown again in Figure 3-6A, B (bottom rows).

These ideal observer and average rat saliency maps were compared with the equivalent average saliency maps obtained for the human participants (i.e., the maps obtained by pooling the bubbles trials collected for any given object view across all human observers; top rows of Figure 3-6A, B). As previously discussed in the rat study (see Chapter 2), these

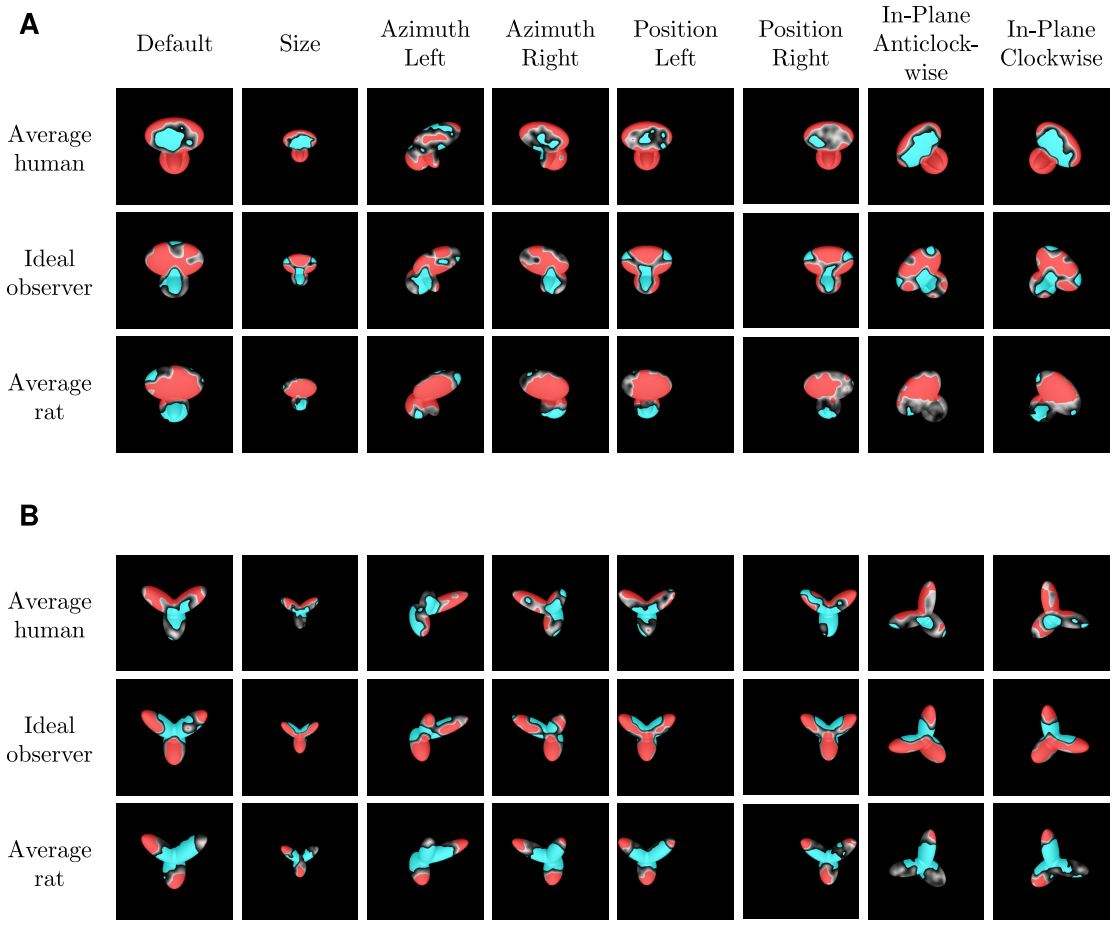

Figure 3-6: Critical features' patterns obtained for the average human, a simulated ideal observer and the average rat. Human group average saliency maps, with highlighted significantly salient (red) and anti-salient (cyan) features (top rows in **A** and **B**), are compared with the saliency maps obtained for a linear ideal observer (central rows in **A** and **B**). The saliency maps of the average rat from the previous chapter is replotted in the bottom rows of **A** and **B** for easier comparison.

group average saliency maps (with highlighted significantly salient and anti-salient object features) summarize the invariant recognition strategy of a group of observers in a way that is more robust to noise and less depending on idiosyncratic aspects of individual participants' strategies. In other words, these maps show the diagnostic object features that most observers relied upon. In the case of Object 1, the average human saliency maps reaffirmed the finding that, for the human observers, the boundary of the object's top lobe and its bottom part were consistently salient, while the central part of the top lobe was consistently anti-salient. Only for the azimuth left-rotated view the human observers did not fully use the boundary of the top lobe, instead partially relying on the central part of the lobe. In

Table 3.1: Pearson correlation coefficient between saliency maps obtained for the average human the ideal observer for the two objects. For each object view, Pearson correlation coefficients between the saliency maps obtained for average human and the ideal observer (shown in Figure 3-5) in reported in the table. The * indicates a significant at $p < 0.05$ by a permutation test. For Object 1 in almost all the views the correlation coefficients were significantly negative, whereas for the Object 2 this is true only for three cases out of eight. This confirms, as it is observed in Figure 3-5, that the human strategy clearly departed from the ideal observer strategy.

	Default	Size	Azimuth Left	Azimuth Right	Position Left	Position Right	In-Plane Anticlock-wise	In-Plane Clockwise
Object 1	-0.42^*	-0.24^*	-0.41^*	-0.30^*	-0.24^*	-0.42^*	-0.13	-0.36^*
Object 2	-0.09	-0.45^*	0.16	0.08	-0.25	-0.18	-0.60^*	-0.58^*

the case of Object 2, the v-shaped cotermination of the top lobes and their boundaries were consistently used across virtually all the object views, again with the only exception of azimuth left-rotated view. As previously observed, this is likely due to the dramatic change in the size/orientation of the object's left lobe and in the angle subtended by the top lobes resulting from the azimuth in-depth rotation.

Observation of Figure 3-6 shows that the saliency maps obtained for the average human observer were quite different form those obtained for both the ideal observer and the average rat. In the case of Object 1, the major difference was that the center of the object's top lobe was anti-salient for the humans, while it was salient for the rats and the ideal observer. Similarly, the bottom part of the object was fully salient for the humans, while it was mostly anti-salient for the rats and the ideal observer. This resulted in negative Pearson correlation coefficients between the saliency maps obtained for the average human observer and the ideal observer (see Table 3-1). Such correlations were significantly lower than expected by chance for seven out of eight views of Object 1 (permutation test; $p < 0.05$; see Table 3-1). In the case of Object 2, while the bulk of the object's three lobes (including their tips but not their intersection) was salient for both the rats and the ideal observer, only a narrow region defining the boundary of the object's top lobes (and including their cotermination) was salient for the humans. This resulted only in a minimal overlap between the salient regions obtained for the average human and the ideal observer. Given that the cotermination of the object's top lobes was anti-salient for the ideal observer, the saliency maps obtained for

Table 3.2: Pearson correlation coefficients between the saliency maps obtained for matching views of Object 1 and 2 (i.e., the same maps shown in Figure 3-5). For the ideal observer (bottom row), the correlation coefficients were all negative and significantly lower than expected by chance, but for the average human with the exception of one case, none of the correlation values was significant (the * indicates a significant at p ¡ 0.05; permutation test).

	Default	Size	Azimuth Left	Azimuth Right	Position Left	Position Right	In-Plane Anticlockwise	In-Plane Clockwise
Average human	-0.05	-0.21^*	-0.26	-0.10	-0.06	-0.18	-0.08	-0.02
Ideal observer	-0.66^*	-0.92^*	-0.82^*	-0.85^*	-0.91^*	-0.94^*	-0.77^*	-0.83^*

the average human observer and the ideal observer were expected to be anti-correlated for some of the object views. In fact in three out of eight views, such negative correlations were significantly lower than expected by chance (permutation test; $p < 0.05$; see Table 3-1).

In summary, the recognition strategy of the human observers, far from matching the optimal strategy of the ideal observer (as rat strategy did), was rather orthogonal to it. Moreover, human recognition strategy also departed from the ideal one in another important way. In the case of ideal observer, the pattern of critical features found for a view of a given object closely resembled the negative image of the pattern of critical features found for the matching view of the other object. Although a similar trend was found for the average rat (see Table 2-1), such a phase opposition was not barely present for the average human observer (see Table 3-2). The Pearson correlation coefficients between the saliency maps of matching views of the two objects were mostly close to zero for the average human observer with only one case out of eight reached to significance level, whereas they were ranged between -0.66 and -0.94 for the ideal observer (all of them were significant according to a permutation test; $p < 0.05$). Overall, this indicates that humans' recognition strategy was not consistent with a linear, template-matching-based extraction of the optimal discriminatory visual information across the object views. This raises the question of what shape processing mechanisms may be consistent with the observed human recognition strategy (see Discussion below).

3.4 Discussion

The aim of this study was to uncover the perceptual strategy underlying human invariant recognition of visual objects and make a comparison with rat strategy. To fulfill aim, we ran an invariant object recognition protocol that was almost identical to the one run with rats (see the previous chapter), and we used the same visual objects and the same image classification technique (the Bubbles method) used in the rat experiment. This approach revealed several key aspects of human object recognition strategy, showing how such a strategy departed from the one used by rats and by a simulated ideal observer.

3.4.1 Implications of our findings and comparison with previous studies

Our major conclusion is that humans generalize their recognition to novel views of visual objects largely by relying on the same patterns of diagnostic features they used to discriminate the previously learned, default views of the objects. Several studies have tested human capability of generalizing recognition to previously unseen object views, often reaching opposite conclusions, especially when non-affine, in-depth rotations are involved (Biederman, 1987; Biederman and Cooper, 1991; Tarr and Blthoff, 1998; Poggio and Edelman, 1990; Bülthoff and Edelman, 1992; Biederman and Gerhardstein, 1993; Logothetis et al., 1994; Blthoff et al., 1995; Lawson, 1999; Biederman, 2000). However, no study so far has systematically investigated what perceptual strategy humans use to succeed in such a generalization. The only two studies addressing this question have tested only a single transformation axis—i.e., in-plane rotations (Nielsen et al., 2008) or translations (Vermaercke and OpdeBeeck, 2012)—and no in-depth rotations.

Our experiments revealed that human observers use combinations of diagnostic features they had previously learned, while discriminating the trained views of two visual objects, to recognize unseen transformed views of those objects across a variety of axes: size, position, in-plane rotation and azimuth in-depth rotation. In general, this was true for all tested transformations, with the exception of those (such as extreme azimuth rotations) that radically altered the shape of the diagnostic features (e.g., the v-shaped cotermination of

Object 2's top lobes). In such cases, human observers relied on a subset of the diagnostic features that were less altered by the transformation (e.g., Object 2's right lobe) or started relying on new discriminatory object features (i.e., the central part of Object 1's top lobe or its smaller, bottom lobe). Since the participants did not receive feedback about the correctness of their responses, it is unclear how such alternative diagnostic object features may have emerged. One possibility is that humans implicitly teach themselves to rely on new features (without the need of explicit feedback), based on their own judgment/perception of their performance under particularly challenging viewing conditions.

Overall, these findings suggest that humans cope with variation in object appearance, largely by relying on representation of diagnostic object features that are tolerant to a wide range of image-level transformations. Interesting, these features, mainly correspond to the boundaries of the objects' more prominent structural parts (such as the curved boundary of Object 1' top lobe or the straight edges of Object 2's top lobs) and their cotermination (such as the v-shaped intersection of Object 2's op lobes). This conclusion is in general agreement with theories stating that human invariant recognition relies on extraction of objects' nonaccidental features and their relations (Biederman, 1987; Biederman, 2000; Biederman and Gerhardstein, 1993). These nonaccidental features are features that are primarily unaffected by rotation in depth such as whether a contour is straight or curved or whether a pair of lines is parallel or, rather, joints in a vertex (Biederman, 1987; Biederman, 2000; Lowe, 1987). Interestingly, a comparative human-pigeon study based on the Bubbles method reached a similar conclusion (i.e., both species were found to mainly rely on cotermination and edges to discriminate four objects) (Gibson et al., 2005). Our study extends this conclusion to a real-world scenario, in which objects are subject to variation in their appearance (which were not tested by Gibson et al, 2007), showing that boundaries and coterminations are diagnostic features that humans consistently rely upon to recognize objects across the variety of transformations they undergo. At the same time, our study shows that view-specific diagnostic features can emerge as a result of extreme in-depth rotations (that dramatically alter the shape of nonaccidental object features), thus supporting view-based theories of object recognition (Poggio and Edelman, 1990; Bülthoff and Edelman, 1992; Logothetis et al., 1994; Blthoff et al., 1995; Tarr and Blthoff, 1998; Edelman, 1999;

Riesenhuber and Poggio, 2000; Freedman et al., 2005).

Another major contribution of this study is the possibility of comparing invariant recognition strategies in humans, rats and a simulated ideal observer. As observed in the results, human and rat strategies had in common the fact of being largely stable across the tested object views. This was quantitatively assessed for the affine transformations, by measuring the overlap between the saliency patterns obtained for all possible pairs of views of an object, after each view in a pair was aligned back to its default appearance. In most cases, for both species, the overlap between aligned views was larger than expected by chance and virtually always larger than the overlap between raw views (see Figs. 2-8 and 3-5). As explained in the previous paragraphs, this means that, for both species, many of the discriminatory features afforded by the two objects were consistently relied upon across many different object views. However, the specific patterns of diagnostic features each species relied upon were radically different in the case of Object 1 and partially different in the case of Object 2. The main difference was that rats relied on the bulk of the objects' lobes, while humans mainly relied on their boundaries and cotermination (see Figure 3-6).

Interestingly, the rat recognition strategy was positively (and, in most case, significantly) correlated with the strategy of a linear ideal observer (see Figure 2-10), while human strategy tended to be negatively correlated with the ideal observer's one (see Table 2-1). In addition, human saliency maps for matching views of the two objects lacked the anti-correlation that was observed for both rat and ideal observer saliency maps (compare Tables 2-1 and 3-2). This leads to the somewhat surprising conclusion that rats made much closer-to-optimal use of the discriminatory information afforded by the two objects than human did (at least when the benchmark is a linear, template-matching-based extraction of such information). A similar conclusion was reached by the above-mentioned study of Gibson et al. (2007), who found that both humans and pigeons discriminated four target objects based on nonaccidental features, although such features did not convey the most diagnostic pictorial information according to a simulated ideal observer (also operating according to a linear, template-matching recognition algorithm). To properly interpret these findings, it must be considered that the ideal observer processes the input images as arrays of pixel intensity values, but does not explicitly take into account the relation between

the pixels—i.e., it does not extract edges, estimate curvatures, orientations, etc. In other words, the ideal observer does not parse the input images into elemental shape features, but relies on the full knowledge of all possible appearances each object can take. On the other hand, the human visual system processes shape information through a cascade of cortical visual areas that operate as banks of local contrast detectors with increasingly complex featural tuning (DiCarlo et al., 2012; Brincat and Connor, 2006; Pasupathy and Connor, 1999; Brincat and Connor, 2004; Kourtzi and Connor, 2011). Therefore, the difference between the recognition strategies of humans and the ideal observer likely reflects the processing of edges, boundaries and other non-accidental properties (e.g., coterminations) that is performed by the human visual system. Such a shape processing is obviously more advanced than the template-matching operation performed by the ideal observer, but, according to our results, less optimal in the use of the discriminatory pictorial information afforded by the object views in our experiment. This should not sound surprising, if we consider that the simulated ideal observer is a recognition system that was built ad-hoc to optimally operate in the context of our stimulus pixel space, with no real capability to generalize its recognition outside the boundary of such a space and, therefore, unable to cope with the variety and variability of real-world visual environments. As a matter of fact, it should be noticed that, in the ideal observer, the view-invariance is trivially built-in by endowing it with the full knowledge of all possible views the two objects can take, which is obviously not a computationally viable solution when dealing with hundreds of thousands of different objects (each projecting a virtually infinite number of images on retina), as the human visual system does. In conclusion, the suboptimal recognition strategy used by human participants in the context of our experiment, likely reflects the price paid by the human visual system to implement a general-purpose shape processing strategy that is able to cope with a tremendous variety of visual objects (and their appearances) in real-world visual scenes. Our data shows that such a strategy is largely based on extraction of contour, boundary and corner information. It remains to be tested whether providing feedback to the human participants, allowing them to familiarize with and memorize each view that the objects can take, would shift their strategy towards a more optimal, template-matching-based strategy (such as the one of the ideal observer).

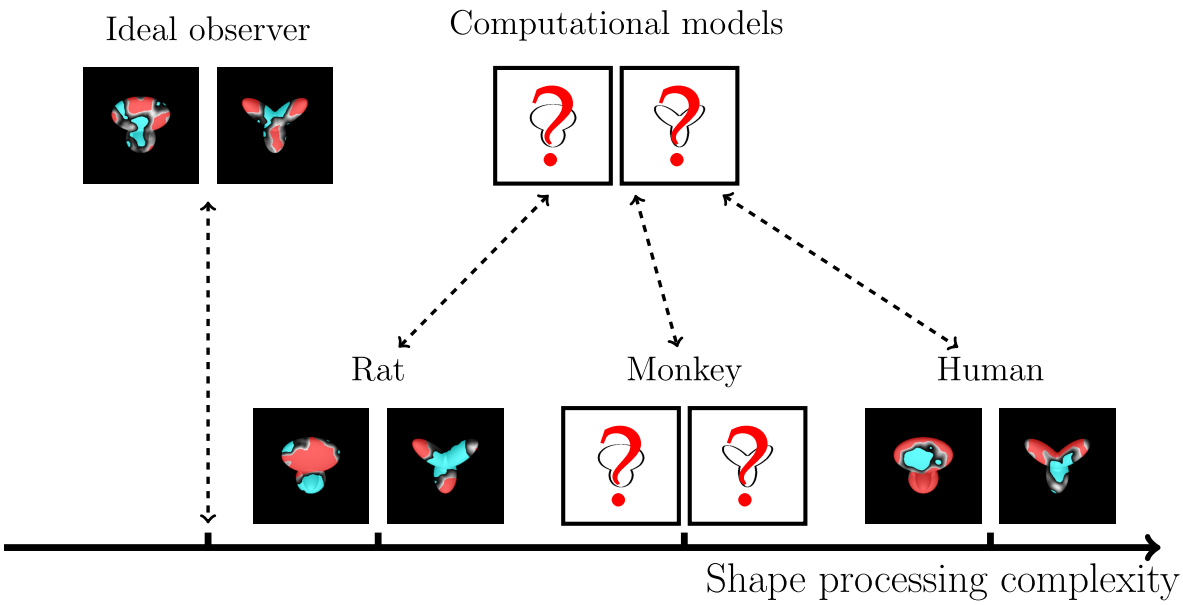

Figure 3-7: Spectrum of shape-processing complexity across species and models. This cartoon qualitatively illustrates where, along a shape-processing complexity axis, the visual object recognition strategies of different species and computational models would sit. The relative ranks of human, rat and the ideal observer recognition strategies are inferred from the data shown in Chapters 2 and 3. Question marks indicate the putative ranks of the recognition strategies of monkeys and state-of-the-art machine vision systems.

As a corollary to what is discussed above, it is important to notice that human recognition strategy also offers a more advanced benchmark (as compared with the ideal observer's strategy), against which rat recognition strategy can be compared. In Chapter 2, based on the results of the rat experiment and the comparison with the ideal observer, we concluded that rats process visual objects through a rather sophisticated shape-based, multi-featural recognition strategy that makes close-to-optimal use of the discriminatory information afforded by the two objects across their various appearances. Although we believe that this conclusion is still largely valid, the comparison with human recognition strategy forces us to reconsider how advanced rat recognition strategy actually is.

On the one hand, the similarity of rat recognition strategy with the ideal observer's strategy, and the dissimilarity from human strategy, suggests that rats may rely on a task-specific, template-matching-based processing of visual shapes, as opposed to the more advanced (and general-purpose) human processing and extraction of edges, boundaries and

coterminations. This would be consistent with the fact that rats, different from human participants, received training with each of the object views and always received feedback about the correctness of their response in both regular and bubbles trials. Therefore, although in Chapter 2 we have argued that rat recognition was likely at least partially based on spontaneous generalization, it could be that the training that the rats received brought them to use a template-matching-based recognition strategy that is very close to the one simulated with the ideal observer. Whether this is the case can only be verified by running another set of rats without providing them any feedback, in both regular and bubbles trials, about the identity of the transformed object views. Although extremely challenging, this could perhaps be achieved by presenting the transformed object views (both unmasked and masked) as prime stimuli in a visual priming paradigm, as recently done by Tafazoli et al. (2012). Similarly, it would be interesting to verify to what extent rat and human recognition strategies are dissimilar because of the dramatically different spatial resolution found in these species. One way to test this would be running additional human participants with low-pass filtered versions of the object views, so to match the resolution of rat retina (as done for the ideal observer analysis in Chapter 2).

On the other hand, as previously discussed in Chapter 2, rat strategy is similar but not identical to the strategy of the ideal observer. Crucially, rats seem to process the most prominent objects' structural parts (i.e., the lobes) as whole features, contrary to the ideal observer, which carved out the negative image of Object 2's top lobes from Object 1's top lobe (see Figure 2-10 and Figure 3-6). This suggests that rats have the tendency to parse the objects into their prominent structural parts, rather than simply rely on the pixels carrying the maximal discriminatory stimulus information (no matter whether these pixels do not match the natural boundaries of the objects' parts), as the ideal observer, by construction, does. Differently from humans, however, such a parsing of object parts appears to involve larger regions on the bulk of the parts, rather than their edges or coterminations. Overall, this places the rat visual system somewhere lower than the human visual system along the spectrum of shape-processing complexity (see cartoon in Figure 3-7), but higher than the linear, template-matching-based ideal observer. One interesting open question is where species with more evolved visual systems (such as monkeys)

would sit along this shape-processing complexity axis (whether closer to humans or to rats). Answering this question would require testing monkeys with the same paradigm and visual objects used in our rat and human studies. However, an indirect comparison between these three species can be attempted by relying on the only study that compared invariant shape recognition in humans and monkeys using the Bubbles method (Nielsen et al., 2008). Although not explicitly addressed by the authors, from the location and extent of the reported saliency regions, it can be inferred that humans preferentially relied on shape boundaries and features' coterminations more than monkeys did. More interestingly, while humans, in agreement with our conclusions, tended to rely on the same features independently of shape orientation, monkeys' tendency was to rely on unique, view-dependent shape regions that were roughly located in the same spatial region of the stimulus display. This is similar to what observed, in our rat experiment, for the horizontal translations, which brought rats to rely more on the part of Object 1's top lobe that was closer to the center of the stimulus display (while in humans the opposite trend was found; compare Figure 2-6 and Figure 3-4). However, the general tendency for both rats and humans, in our study, was to rely, at least partially, on the same features across different object views, as demonstrated by the fact that the aligned overlap values were consistently larger than the raw overlap values. Therefore, the fact that for monkeys, in Nielsen et al, (2008), the raw overlap was larger than (or comparable with) the aligned overlap raises the question of how much closer to humans than to rats monkeys really are along the shape-processing complexity spectrum (see bottom question mark in Figure 3-7). Finally, it would also be interesting to test where state-of-the-art computational models of the visual system and machine vision algorithms [e.g., see (Serre et al., 2007; Pinto et al., 2008)] sit along such a shape-processing complexity axis (see top question mark in Figure 3-7).

3.4.2 Validity and limitations of our findings

The use of the Bubbles methods allowed us to address the question of what critical features underlie invariant visual object recognition in the two different species. However, some

caution should be taken when inferring recognition strategies from the partially occluded visual stimuli produced by the Bubbles method. In fact, the occlusion produced by the bubbles masks could potentially induce observers to rely on a different shape processing strategy, as compared with cases in which they have to discriminate intact objects. As a matter of fact, whether the Bubbles method can induce some substantial strategy changes has been under debat (Murray and Gold, 2004; Gosselin and Schyns, 2004; Murray, 2004). In particular, some authors expressed concerns about whether the repeated presentation of shape fragments (rather than whole shapes) may induce observers to adopt a strategy based on local object features rather than on more global (or holistic) shape processing (Nielsen et al., 2008). No study, so far, has fully clarified this issue. Doing so, would require comparing the saliency maps produced by the Bubbles method with those produced by some other, alternative approaches, such as other classification image (or reverse correlation) methods based on additive pixel-level noise.

Another limitation of our current work is that we only tested two objects. It is unclear to what extent the saliency maps obtained through the Bubbles method are task- and stimulus set-dependent. In the case of the rats and of the ideal observer, the negative correlation between the saliency maps obtained for matching views of the two objects suggests that the pattern of diagnostic features obtained for a given object is strongly influenced by the specific comparison/discrimination the animal/model is performing. In other words, it is possible that, if rats had to recognize many more visual objects, in addition to the two objects used in this study, different saliency maps would have been obtained, perhaps closer to the ones obtained for humans. A possible experiment to test this would be teaching participants to discriminate one of the objects used in this study from many tens (or hundreds) of other possible distractors. Our expectation is that human recognition strategy may not substantially change (already reflecting general-purpose shape processing mechanisms; see Discussion in the previous paragraph), while rat recognition strategy may converge towards the human one.

Along the same line of thinking, it would be interesting to test a larger number and variety of in-depth rotations, including more extreme azimuth rotations, but also elevation rotations. In fact, the Bubbles method looks like an extremely promising approach to settle

the long-standing debate about whether view-invariant recognition is achieved by means of structural, viewpoint-invariant object descriptions or, rather, view-based object representations (Biederman, 1987; Biederman, 2000; Tarr and Blthoff, 1998; Lawson, 1999). Our study provides some hints about this issue, but a much larger battery of in-depth rotated views is necessary to systematically address this question.

Finally, one open question is what is the relationship between multiple diagnostic object features and whether such a relationship may change as a function of the transformations the objects undergo. With the current analysis of the bubbles trials, it is not possible to know whether the presence of a specific feature (among the ones that are diagnostic for a given object view) is enough for an object to be correctly identified, or whether, on the contrary, all diagnostic features must be all simultaneously visible. At a more quantitative level, it is unclear, when multiple, distinct diagnostic features are found for an object, whether the evidences that these features provide about object identity linearly sum up or, rather, interact in some non-linear way. In the next chapter, we propose a new analysis framework to tackle this problem.

Chapter 4

Non-Linear Interactions Analysis of Diagnostic Features

In this chapter, as an original computational contribution in this thesis, we propose an improvement for the analysis of the data collected using the Bubbles method. We show that the proposed analysis is capable of extracting the non-linear relationship between pairs of shape features extracted through the Bubbles method. To prove the soundness and solidity of the proposed method, we do not abstain from rigorous mathematical treatments when needed. It is important to clarify that the goal in this chapter is to demonstrate the fundamental principles upon which our proposed analysis rests, rather than applying the method to re-analyze the data presented in previous chapters.

4.1 Introduction

Understanding what object features the visual system relies upon, when engaged in visual object recognition, is a longstanding challenge in psychophysical studies of visual perception. Although successful approaches have been developed to address this issue [e.g., image classification methods (Murray, 2011)], they have a major limitation: they can estimate the relationship between input visual images and output behavioral responses only under the assumption of a linear observer model, i.e. an observer performing a weighted sum of the information carried by individual pixels within an image. A closely related method to

image classification is the Bubbles method (Gosselin and Schyns, 2001) that was used in the previous two chapters. This method recovers the salient features used by an observer to identify an object, by presenting the target object partially occluded by opaque masks punctured by transparent windows, and then averaging the trials yielding to correct object identification. However, as in the case of the other image classification approaches, the standard analysis of the Bubbles method is linear, i.e., it cannot tell whether multiple salient features in an object interact non-linearly (e.g. whether those features need to be simultaneously visible for the object to be correctly identified).

Estimating the way in which multiple shape features (or object parts) interact is an important step towards gaining a better understanding of shape coding in visual cortex. This is true both at the level of understanding how visual objects are perceived by a human or an animal observer (i.e., analysis of psychophysicalbehavioral data) and at the level of understanding how visual neurons code shape (i.e., analysis neuronal data). For instance, recent advances on this front led to the conclusion that a fraction of monkey inferotemporal (IT) neurons integrate shape features (i.e., boundary/surface elements of synthetic 2D and 3D objects) in a highly non-linear way (i.e., such neurons fire strongly only if multiple critical features are simultaneously present in a given shape (Kourtzi and Connor, 2011; Yamane et al., 2008; Brincat and Connor, 2004)].

To our knowledge, no similar recovering of the non-linear interaction between multiple shape features has been attempted in psychophysical studies. The Bubbles method, with its sparse sampling of the object feature space, provides, in principle, an excellent opportunity to find possible non-linear interactions between object diagnostic features. The aim of this chapter is to introduce a novel analytical approach that evaluates whether pairs of diagnostic features (extracted through the Bubbles method) interact non-linearly. The proposed analysis is general enough to recover any kind of postulated non-linearity, but, to demostrate its effectiveness, we focused on an implementation that recover AND-like and OR-like interactions in binary functions (which are the more relevant in vision sciences).

Our proposed approach is based on information theory. Since the method makes heavily use of information theory and Bayesian network, we provide a brief mathematical background on these subjects. The reader who is familiar with these subjects can simply jump

into the subsequent section.

4.2 Mathematical Preliminaries

This section introduces a set of basic definitions, concepts, and theorems of that are necessary to understand the proposed analysis.

4.2.1 Information Theory

We begin with the most fundamental concept in information theory—*Entropy*. The entropy of a discrete random variable X with probability mass function $p(x)$ is denoted by $H(X)$ and defined as below (Shannon, 1948; Cover and Thomas, 2006):

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{4.1}$$

Note that when X is a continuous random variable, in practice, we need to discretize it using a binning procedure. Unless otherwise needed, from now on, we will not mention whether the random variables are discrete or continuous. Entropy is a measure of uncertainty (or randomness or disorder), which is obtained by defining a functional taking as input a probability distribution and returning a nonzero value.

The definition of entropy can be extended to incorporate two random variables X and Y with a joint probability distribution $p(x, y)$ and defining the *joint entropy*:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y). \tag{4.2}$$

Likewise the *conditional entropy* is defined as:

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y|x). \tag{4.3}$$

The *mutual information* between two random variables X and Y, denoted by $I(X;\ Y)$,

is defined as following:

$$I(X;\ Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}, \tag{4.4}$$

where $p(x,y)$ is the joint probability distribution and $p(x)$ and $p(y)$ are *marginal* probability distributions. As in the case of entropy, one important property of mutual information is that it is always non-negative i.e. $I(X;\ Y) \geq 0$. The mutual information of X and Y is zero if and only if X and Y are independent (denoted by $X \perp Y$).

The definition of mutual information can be written in terms of entropies:

$$I(X;\ Y) = H(X) - H(X|Y) \tag{4.5}$$

$$= H(Y) - H(Y|X). \tag{4.6}$$

We should notice that the relationship between X and Y is symmetric in the definition of mutual information, meaning $I(X,Y) = I(Y,X)$. What Eq. (4.5) tells us is that the mutual information is the amount of reduced uncertainty due to the fact that Y is known.

Likewise the *conditional mutual information* can be written in terms of conditional entropies:

$$I(X;\ Y|Z) = H(X|Z) - H(X|Y,Z). \tag{4.7}$$

Another important concept in probability theory, which is closely related to mutual information, is the concept of *Markov chain*. If the probability distribution of the random variable Z is conditionally independent from X given Y—which is denoted by ($X \perp Z|Y$)— then both $X \to Y \to Z$ and $Z \to Y \to X$ form first-order Markov chains. This independency results in factorization of the following joint probability distributions

$$p(x,y,z) = p(x)p(y|x)p(z|y) \tag{4.8}$$

$$p(x,z|y) = p(x|y)p(z|y), \tag{4.9}$$

which in itself results in

$$I(X;\ Z|Y) = 0. \tag{4.10}$$

Now we are ready to point out an important theorem, the data processing inequality, which will be used in the proof of our main theorem. We skip the proof for brevity; it can be found in (Cover and Thomas, 2006).

Theorem 4.1 (Data processing inequality) *If random variables X, Y, Z form the Markov chain $X \to Y \to Z$, then $I(X;\ Y) \geq I(X;\ Z)$ and also $I(Y;\ Z) \geq I(X;\ Z)$.*

We skip the proof. It can be found in (Cover and Thomas, 2006).

4.2.2 Probabilistic Graphical Model: Bayesian Network

In this section, we turn into a graphical representation of random variables in order to illustrate their relationships. We use the *Bayesian network* representation from *Probabilistic Graphical Models* framework (Koller and Friedman, 2009). A Bayesian network is a directed acyclic graph whose nodes are random variables and whose edges correspond to the influence of random variables on each other. In general, such an influence takes the form of a conditional probability distribution (in the graph, the special case of a deterministic dependency between a random variable and its parents is indicated by double-line node). One of the benefits of the Bayesian network representation is that this data structure provides a skeleton for compactly representing conditional independencies about a distribution over some random variables. As an example, the Bayesian network in Figure 4-1 captures the following independencies: $(X_i \perp X_j)$, $(X_i \perp B|Y)$, $(X_j \perp B|Y)$. This facilitates the representation of a joint probability distribution in terms of conditional probability distributions. As an example, for the Bayesian network in Figure 4-1 the joint probability distribution over the four random variables can be factorized as the following:

$$P(X_i, X_j, Y, B) = P(X_i)P(X_j)P(Y|X_i, X_j)P(B|Y). \tag{4.11}$$

Beside representing the independencies, Bayesian networks are used for making inference and learning the structure and parameters of models. Since we mainly use the representational power of Bayesian networks in this chapter, we do not cover inference and learning in Bayesian network here.

4.3 The Core of Non-linear Interaction Analysis

To capture the essence of non-linear interaction analysis, we will first illustrate a simplified case of a Bubbles experiment (later we will simulate a more realistic Bubbles experiment). Let's assume in this simplified case, that there are only two pixels, which are affecting the behavioral outcome of a simulated observer. The transparency of these two pixels may change in each trial. In this setting, there are only two possible diagnostic features (i.e., the two pixels) that can be represented as two random variables taking continuous values between zero and one. Let's represent these two random variables as X_i and X_j. Each instantiation of this pair of random variables can be regarded as the transparency values of the pixels in a given trial. Let's assume that a deterministic function of these two variables—but none of these two random variables independently—influences an output behavioral variable B (see Figure 4-1). This output random variable is the behavioral response of the observer, which, in the Bubbles experiments described in the previous chapters, is, by definition, binary (i.e., it and can either be 0 or 1, depending on the identity of the object reported by the observer). Note, however, that the validity of our approach is not restricted only to cases in which B is binary (or even discrete).

As a first step in the development of our approach, we demonstrate that whenever there is a specific relationship between two features X_i, X_j and the behavioral response B, two inequalities hold. These inequalities are the criteria for finding non-linear interaction between any two features.

Theorem 4.2 (Main theorem) *If the dependencies of four random variables X_i, X_j, B, and Y—with Y being a deterministic function f of X_i and X_j—are represented by the Bayesian network in Figure 4-1, then the following inequalities hold:*

$$I(f(X_i, X_j);\ B) \geq I(X_i, X_j;\ B), \tag{4.12}$$

and

$$I(f(X_i, X_j);\ B) \geq I(X_i;\ B) + I(X_j;\ B). \tag{4.13}$$

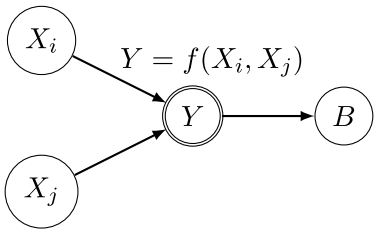

Figure 4-1: Bayesian network of a toy example illustrating a deterministic function of two variables influences an output random variable. This Bayesian network shows the dependencies between random variables X_i, X_j, Y, and B. The random variable Y is a deterministic function of X_i and X_j. This network is used as a toy example in which a non-linear interaction between two features influences an output random variable. The main theorem of this chapter proves that if relationship between these random variables are as shown in the network then the information Y carries about B not only is higher than sum of information carried by each of the two Xs about B but also is higher than the information that the pair (X_i, X_j) carries about B. These inequalities are the basis for extracting non-linear feature interaction method.

Proof: If one applies the data processing inequality to the Bayesian network in Figure 4-1, the following inequality is obtained:

$$I(Y;\ B) \geq I(X_i, X_j;\ B), \tag{4.14}$$

substituting Y by its equivalent in terms of its parents, i.e. $f(X_i, X_j)$, one obtains the first desired inequality:

$$I(f(X_i, X_j);\ B) \geq I(X_i, X_j;\ B). \tag{4.15}$$

In order to derive Eq. (4.13), let's start with the following equality, which holds for every four random variables (Adelman et al., 2003):

$$I(X_i, X_j;\ B) - I(X_i;\ X_j \,|\, B) + I(X_i;\ X_j) = I(X_i;\ B) + I(X_j;\ B). \tag{4.16}$$

Note that the equality can easily be demonstrated using a Venn diagram. In our case $I(X_i;\ X_j) = 0$ which together with the non-negativity property of mutual information gives:

$$I(X_i, X_j;\ B) \geq I(X_i;\ B) + I(X_j;\ B). \tag{4.17}$$

Combining Eq. (4.17) and Eq. (4.15) yields the second desired inequality:

$$I(f(X_i, X_j);\ B) \geq I(X_i;\ B) + I(X_j;\ B). \tag{4.18}$$

■

There are several points that should be noticed. Theorem 4.2 can only be proved under the premises that no direct dependency exists between X_i and B and X_j and B (i.e., under the conditions illustrated in Figure 4-1). The inqualities cannot necessarily hold true, for instance, in a network with some direct link between X_i and B or between X_j and B. In such cases, whether $I(f(X_i, X_j);\ B)$ will be larger than $I(f(X_i, X_j);\ B)$ and $I(X_i, B) + I(X_j, B)$ will depend on the relative strength of influence of the Xs and $f(X_i, X_j)$ over B. Finding cases in which $I(f(X_i, X_j);\ B)$ is larger than larger than $I(X_i, X_j;\ B)$ and $I(X_i, B) + I(X_j, B)$ would imply that the particular functional interaction between the Xs (i.e., the pixels/features) is the main driving force for the output variable B, even though there might be some influence from each of the features, independently.

Another observation is the relation between the three-way mutual information, i.e. $I(X_i, X_j;\ B)$, and the sum of information of the individual Xs about B, i.e. $I(X_i;\ B) + I(X_j;\ B)$. If the latter—which we may called it I_{Lin}—is much less than the former, this means there is a positive *synergy* between the two variables so that the pair conveys more information about B than the sum of the information of individual of them about B (Magri et al., 2009). Whereas if the sum of the individual information is higher than the three-way mutual information [for this to be true we need $I(X_i, X_j) \neq 0$], then there is redundant information or negative synergy between X_i and X_j about B. In general, a large, positive synergy is an indication that the two features X_i and X_j interact in some not trivial way to drive the behavioral outcome. However, to understand what functional form such an interaction takes, it is necessary to explicitly model the interaction itself with some candidate functions f, compute $I(f(X_i, X_j);\ B)$ and then compare it with $I(X_i, X_j;\ B)$ and $I(X_i;\ B) + I(X_j;\ B)$. The function conveying the largest information about B will be the best candidate for the type of underlying interaction between the corresponding pair of features.

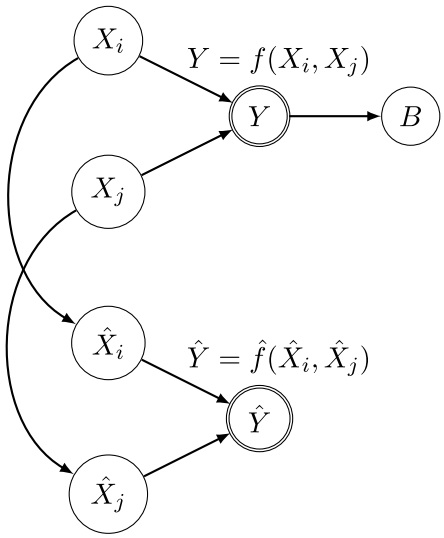

Figure 4-2: Bayesian network of the toy example with surrogate variables. The network illustrates the relationship between the four variables and the approximations of X_i and X_j, i.e. the surrogates $\hat{X}_i$ and $\hat{X}_j$. If one uses the surrogate variables instead of the actual ones, the inequalities in the main theorem may not hold anymore. The better is the approximation of surrogate variables the higher gets the information of $f(\hat{X}_i, \hat{X}_j)$ about B. This toy example illustrates a situation that may be encountered in practice where Eqs. (4.12, 4.13) may not always hold.

An important point for the application of this method is that, practically, we often deal with variables that is an approximation of actual features affecting the behavioral response in a Bubbles experiment. For example, we may analyze the masks instead of bubbled images, or we may binarize the variables or making the masks coarse-grained so as to lower the dimensionality of data. In such conditions, we are dealing with *surrogate* variables $\hat{X}_i$,$\hat{X}_j$, and $\hat{Y}$ instead of X_i, X_j, and Y (Figure 4-2). Depending on the goodness of the approximation, one of the inequalities $I(f(\hat{X}_i, \hat{X}_j);\ B) \geq I(\hat{X}_i, \hat{X}_j;\ B)$ or $I(f(\hat{X}_i, \hat{X}_j);\ B) \geq I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$ or both of them may not hold. Since this may often be the case in practice (at least based on the current implementation of the analysis), it is safer to compute the information of some candidate functions in $I(f(\hat{X}_i, \hat{X}_j);\ B)$ and compare it with $I(\hat{X}_i, \hat{X}_j;\ B)$ and $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$. The function, whose information accounts for large fraction of three way mutual information or sum of the individual information in comparison to the information of other functions, will be the best candidate for the type of interaction between the corresponding pair of random variables.

We have not yet discussed the type of functional interaction between the feature variables X_i and X_j, since inqualities (4.12, 4.13) hold, in general, for any kind of function. In the context of visual neuroscience, the main relevant functions, when it comes to modeling the non-linear interactions between visual features, are AND-like and OR-like functions. In the next sections, we simulate observers operating according to these kinds of non-linear interactions, and we show that computation of mutual information allows correctly recovering the simulated non-linearity.

Throughout this chapter for simplicity, we use the symbol $\wedge$ to refer to either an AND-like operation (such as PRODUCT operation) between continuous random variables or an AND operation between binary random variables. Similarly, the symbol $\vee$ is used to mean either an OR-like operation (such as MAX operation) between continuous random variables or an OR operation between binary random variables.

4.4 Simulation Design and Analysis

To assess the ability of the proposed analysis to recover non-linear interactions, we implemented a simple toy example and also carried out a more realistic simulation of a Bubbles experiment. In this section, we provide details of these simulations by relying on the mathematical core of the analysis explained in the previous section.

4.4.1 Implementattion and analysis of a Toy model

As explained in Section 4.3, the main theorem mandates when the relationship between random variables can be represented by the Bayesian network in Figure 4-1, then the Eqs. (4.12, 4.13) hold true. But if we use an approximation of X_i and X_j in those inequalities, then they may not necessarily hold true anymore. To investigate under what condition we can expect results similar to the ideal situation depicted in Figure 4-1, we implemented the toy model shown in Figure 4-2. Our goal was to show to what extent we can expect $I(f(\hat{X}_i, \hat{X}_j);\ B) \geq I(\hat{X}_i, \hat{X}_j;\ B)$ and $I(f(\hat{X}_i, \hat{X}_j);\ B) \geq I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$ to be true, if we use surrogate variables $\hat{X}_i$ and $\hat{X}_j$ instead of X_i and X_j.

One thousand samples from two real-valued random variables X_i and X_j with uniform

distribution between zero and one were generated. To implement an AND-like function, the product of them, i.e. $Y = X_i.X_j$, was computed which was then thresholded to obtain the binary random variable B as the output (threshold was set to the middle of the range of values that Y could take, which was close to 0.5). We may think of this toy example as an extreme simplified version of an observer that uses an AND-like interaction between the visibilities of two features to compute an output behavior B. To probe Eqs. (4.12, 4.13), we computed four information quantities: $I(X_i, X_j;\ B)$, $I(X_i;\ B) + I(X_j;\ B)$, $I(X_i \wedge X_j;\ B)$, $I(X_i \vee X_j;\ B)$. All of the above procedure was repeated to implement. An OR-like interaction function between X_i and X_j was also implemented, by taking the MAX operation between them ($Y = max(X_i, X_j)$), and the same four mutual information quantities were estimated.

Since the random variables X_i, X_j and Y in the toy example were continuous (because this was more similar to the non-linear observer implemented in the simulated bubbles experiments; see next section), in order to compute mutual information we needed to discretize them. We binned the variables into four bins using an equi-probable binning procedure (Magri et al., 2009). All the information quantities were divided by the entropy of B so as to make the comparison easier. The results of this divisions are unitless, yet, in the following, we still refer to them as information for the sake of simplicity.

Importantly, in real-case scenarios, one typically does not have access to the actual variables underlying the measured behavior of an observer. Moreover, it is often practical to approximate variables that are measurable to make computation of mutual information easier or even possible (e.g., by reducing the dimensionality or computational complexity of a given problem; see examples in the next two section). To simulate such situations and understand to what extent inequalities (4.12, 4.13) hold true under such conditions, we built surrogate variables $\hat{X}_i$,$\hat{X}_j$ and $\hat{Y}$, starting from the actual variables underlying the behavioral output of the toy model. $\hat{X}_i$ and $\hat{X}_j$ were obtained by binarizing X_i and X_j, so that: $\hat{X}_i = 0$ if $X_i < 0.5$, and $\hat{X}_i = 1$ if $X_i > 0.5$. Y was also computed as the logical AND (or logical OR) of the surrogates of X_i and X_j, i.e., $\hat{Y} = \hat{X}_i \wedge \hat{X}_j$ (or $\hat{Y} = \hat{X}_i \vee \hat{X}_j$).

4.4.2 Simulation of a Bubbles Experiment with a non-linear observer

In the Bubbles experiments described in the previous chapters, data are processed under the assumption that an observer relies on some diagnostic shape features of an object view to identify it. By construction, the observer model relies on multiple diagnostic features (refer to previous chapters to see some examples in humans and rats e.g. Figure 2-6 and 3-4) and the evidence of each feature can contribute to the identification of an object view. Evidence of a diagnostic feature corresponds to the sum of the luminance of that feature through a bubbles mask in any given trial. The evidences conveyed by different features can then be combined through a function that can either be linear (e.g., the sum of the evidences) or non-linear (e.g., the sum of the pair-wise products of evidences; see the block diagram in Figure 4-4). In the previous chapters, the diagnostic features underlying rat and human object recognition were extracted by processing the bubbles mask data under the assumption of a linear amplifier model (Green and Swets, 1966). That is, the function underlying the interaction of different features was assumed to be summation. Here, we assume that an observer can use a non-linear interaction function as a way to integrate the evidences about the identity of a given object provided by any given pairs of features. For simplicity, we limit our analysis to treat pairwise interactions, although, in general, the non-linearity could be of higher order—three-way interactions or more instead of pairwise interactions. Such a non-linear observer model was used in simulated bubbles experiments, where the two types of non-linear feature interactions mentioned in the previous sections (i.e., AND-like and OR-like) were implemented.

Each observer had to discriminate bubbles-masked input images of the two objects. In total, for each observer we simulated 3000 bubbles trials, 1500 for Object 1 and 1500 for Object 2. Some examples of simulated trials and responses are shown in Figure 4-5A. Two features of Object 2 were chosen as the diagnostic features of this object. In Figure 4-3, a grid is shown superimposed on Object 2 to show the location of these features and their extent (see the colored circles in the grid squares B2 and B7). In each trial, the evidence of each feature was computed as the sum of the luminance, in the bubbles-masked image, within the area defined by the corresponding circle and divided by the number of pixels

within that circle. Since luminance values ranged between 0 and 1, the evidence of each feature also ranged between these same extremes. To keep the simulations simple, we just set the perceptual threshold for detection of each features to 0.1. This yielded binarized evidence values—a value of 0 meant evidence < 0.1 (i.e., feature not perceived/detected), while a value of 1 meant evidence > 0.1 (i.e., feature perceived/detected). Finally, a binary operation (i.e., either an AND or an OR; see Figure 4-3) was carried out between the binarized evidences to obtain a decision variable "d". If d was equal to one, then the output behavioral decision was Object 2, if d was zero the decision was Object 1 (note that, in this instance of the simulated non-linear observer, d corresponds to both the Y and B variables used in figures 4-1 and 4-2 and in the equations). Finally, we also simulated a thresholded linear observer, i.e., an observer performing a SUM operation over the feature evidences (yielding $Y = X_i + X_j$), which was then discretized to produce the binary output B, so that $B = 1$ if $Y > 0.1$; and $B = 0$ if $Y < 0.1$.

The number of bubbles was set to 30 and kept fixed throughout of the simulation experiment. Bubbles were generated with standard deviation of 10 pixels.

4.4.3 Information theoretical analysis of the simulated bubbles-masked trials

To prepare the dataset for the non-linear feature extraction analysis, we built 1500 trials in which Object 2 was occluded by bubbles masks (e.g., see examples in Figure 4-5A) and we simulated the response of the non-linear observer according to the description provided in the previous paragraph (see also Figure 4-3). The behavioral outcome of these trials (i.e., whether the simulated observer correctly responded Object 2, or, instead, responded Object 1) took the form of a binary vector of 1500 elements (i.e., either 1 or 0, depending on the correctness of the response). This behavioral outcome vector was treated as binary random variable and was denoted by B.

Each bubbles mask was a 193-by-193 element matrix with continuous transparency values. To reduce the dimensionality of the problem and to make the complexity of computation of mutual information easier, each bubbles mask was then coarse-grained and

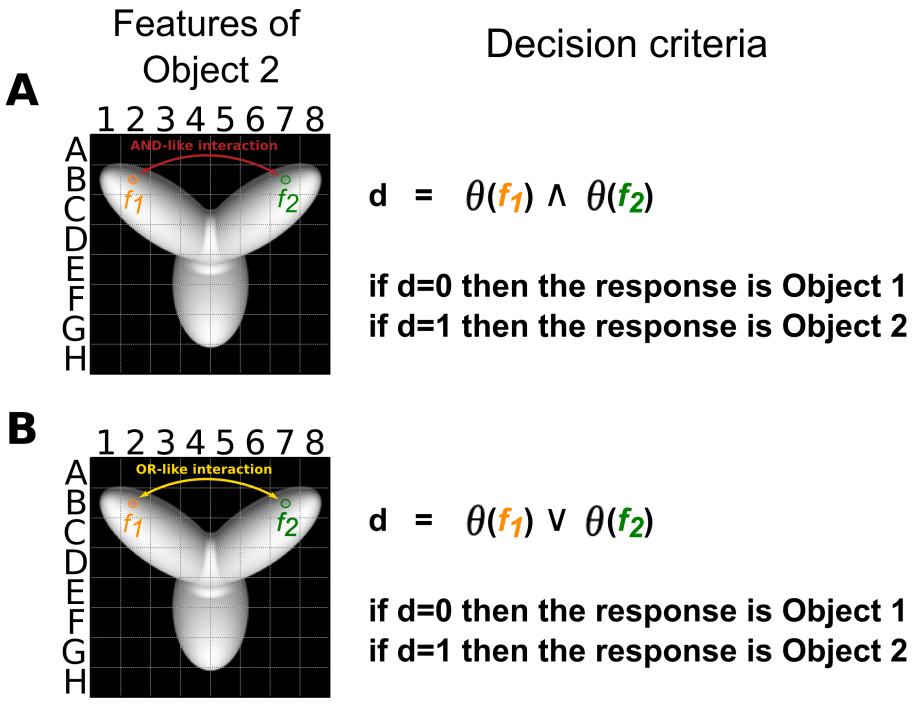

Figure 4-3: Features and their interaction in two simulated observers. We defined two features of Object 2, i.e. f_1 and f_2 that are used in the decision criteria of the simulated observer performing a bubbles task to discriminate the occluded versions of the default views of Object 1 and Object 2. **A.** An AND-like interaction is defined between the two features by first comparing the amount of visibility of that feature against a threshold (implemented by the function θ), and then taking the AND of the thresholded evidence. The outcome is the decision variable "d" the value of which determines the identity of the object (shown in the second column). **B.** Same as **A**, except that an OR-like interaction is defined between the two features. Please note that since the proposed method cannot differentiate between OR and SUM (because it is working with the binary-valued vectors), we have focused our simulations and results only on observers having AND and OR interactions.

binarized into an 8-by-8 binary grid. A grid square was set to one if at least (the center of) one bubble appeared in that square. Each grid was then converted into a binary vector of 64 elements (Figure 4-5B). Concatenating the binary vectors for all the 1500 masks yielded a 64-by-1500 binary matrix (Figure 4-5C). This trial matrix along with the behavioral outcome random variable B was used to compute the mutual information quantities.

Each row of the trial matrix can be considered as a binary random variable with 1500 samples. We selected every pair of rows denoted by $\hat{X}_i$ and $\hat{X}_j$ of this matrix (i.e., the binary values taken by every pair of squares of the 8-by-8 grid) and computed the following four information quantities: $I(\hat{X}_i, \hat{X}_j;\ B)$, $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$, $I(\hat{X}_i \wedge \hat{X}_j;\ B)$,

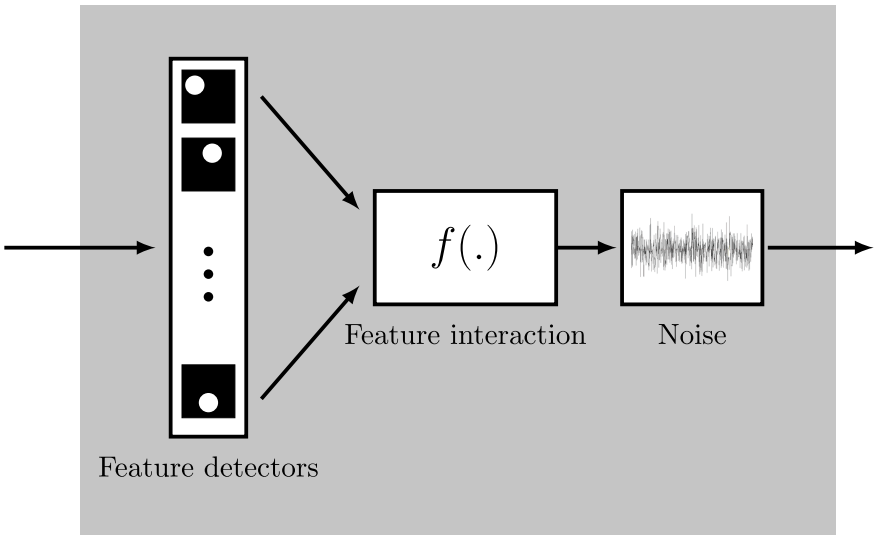

Figure 4-4: Block diagram of the observer model performing a bubbles task. In each trial, the evidences of the features of an object view are extracted from an input image. The evidences of the features are then passed through a function that defines the interactions between features. The output of the function is the overall evidence about the object identity. The function in the simplest case can be just the sum of feature evidences or could be a nonlinear function of them.

$I(\hat{X}_i \vee \hat{X}_j;\ B)$. It should be noted that here, due to the downsampling and binarization process of the bubbles masks, we are dealing with the surrogate variables of those variables that affected the decision of the simulated observer. The candidate functions in the analysis are chosen to be AND and OR because these are the non-linearity that are built into the simulated observer and that we want to recover through computation of the mutual information quantities. Notice that, since a linear SUM is not defined for binary variables (in single-bit digit, a SUM can be taken as equivalent to an OR), we could not compute the mutual information between a SUM operation and the behavioral outcome.

A permutation test was carried out in order to check whether $I(\hat{X}_i \wedge \hat{X}_j;\ B)$ is significantly higher than $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$. This was achieved by randomly shuffling the labels of the behavioral outcome variable B one hundred times and then, for each repetition, we computed the four information quantities. In each repetition, the maximum of each information quantity among the tested pairs was chosen, so as to obtain four null distributions for the four information quantities and their differences. The actual difference between two given information quantities (for example the difference between AND infor-

mation and sum of individual information) was then compared against the null distribution of this difference at 0.05 significance level (Bonferroni corrected; 2016 comparison).

All the mutual information quantities were computed first by a custom-written code in MATLAB, then cross-checked by the IBTB information theory toolbox (Magri et al., 2009). All the simulations and analyses were performed on an Apple Mac Pro computer (with 2 CPUs at 2.93GHz Quad-Core Intel Xeon; Mac OS 10.6.8).

4.5 Results

We have presented an information-theoretic method that can extract non-linear interactions of diagnostic features. The method relies on computing a couple of mutual information quantities in a Bubbles experiment. In order to evaluate the proposed analysis method, two types of simulation were performed. First, we built a toy example and verified the theorem of inequalities (4.12, 4.13) when either the variables underlying the behavior of the simulated toy model are used or, instead, surrogate variables are used (see Section 4.4.1). Second, we simulated a bubbles experiment in which the simulated observer carried out AND/OR-like computation between pairs of features (see Sections 4.4.1,4.4.3). In both cases, we found that the method can recover AND-like and OR-like interactions between diagnostic features.

4.5.1 Toy Example's Result

To assess to what extent and under what conditions the proposed analysis can recover the non-linear interactions, we implemented the toy example illustrated by the Bayesian network of Figure 4-2. The relevance of this toy example in the context of the Bubbles experiment becomes clearer when the two real-valued variables X_i and X_j are considered as evidences provided by two features about the identity of an object. Y can be interpreted as the outcome of the non-linear function that processes those evidences/features to determine behavior B (which is a binary random variable).

If the underlying non-linear function of Y in the toy example is PRODUCT (AND-like) which most influences the output B, then the information that Y carries about B [denoted

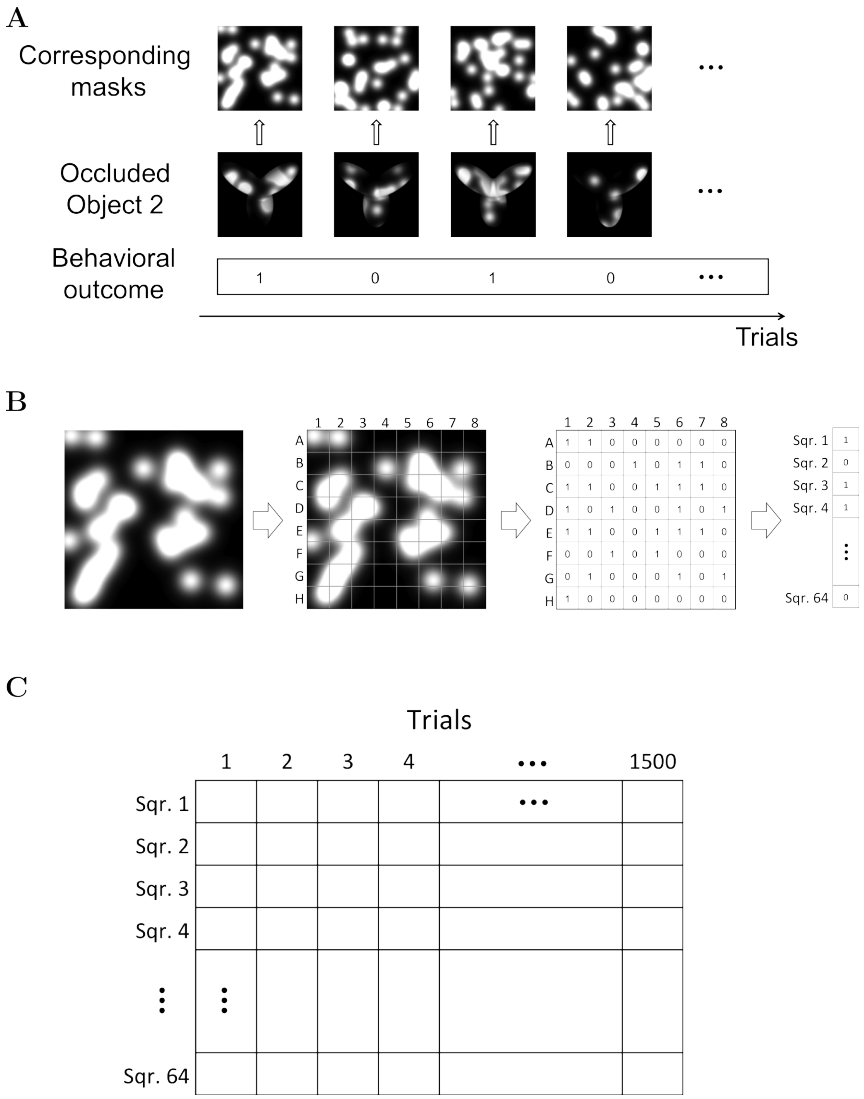

Figure 4-5: Preparing the trials of simulated observers for extracting non-linear interactions. **A.** Examples of trials in the simulation in which Object 2 was presented are shown. The corresponding bubbles mask along with the behavior success is shown for each trial. **B.** For reducing computational complexity, we converted the mask into an 8-by-8 grid with binary values. The rows of the grid are codes by eight letters from A to H and the columns by numbers from 1 to 8. Each square of the grid is set to one if a center of at least one bubble appears in that square; otherwise it is set to zero. The two important squares that contain the two features of the simulator are B2 and B7. The binarized mask was then transformed into a binary vector containing 64 elements. **C.** All the binary vectors related to the mask of trials were piled up in a matrix, the columns of which represent different trials. Since we had 1500 trials in which Object 2 was represented, the matrix turns into a 64-by-1500 binary matrix. This matrix along with the behavior vector was used in the analysis.

by $I(Y;\ B)$ or equivalently shown by $I(X_i \wedge X_j;\ B)$; note that here $\wedge$ means PRODUCT and not logical AND] is higher than both the three-way information $I(X_i, X_j;\ B)$ and the sum of the individual information $I(X_i;\ B) + I(X_j;\ B)$ (Figure 4-6A, the left-most chart; compare the red and dark green bars). Instead, the information that an OR-like operation such as $Y = max(X_i, X_j)$ would convey about B [i.e. $I(X_i \vee X_j;\ B)$; note that here $\vee$ means MAX and not logical OR] is less than the other three quantities by at least two orders of magnitude (compare the yellow bar with other bars in the same chart). This is indeed consistent with the main theorem. The inequalities, however, do not hold if mutual information is computed using the surrogate variables $\hat{X}_i$,$\hat{X}_j$ and $\hat{Y}$, as defined in Section 4.4.1 (i.e., by binarizing X_i and X_j and computing $\hat{Y}$ as the logical AND of $\hat{X}_i$ and $\hat{X}_j$). In fact, $I(\hat{X}_i \wedge \hat{X}_j;\ B)$ becomes lower than $I(\hat{X}_i, \hat{X}_j;\ B)$, although it is still higher than $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$ (Figure 4-6A, the centeral chart). The better the approximation between the surrogate and original variables is, the higher is $I(\hat{X}_i \wedge \hat{X}_j;\ B)$, compared with $I(\hat{X}_i, \hat{X}_j;\ B)$ and $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$. In case the surrogate variables are contaminated by noise, $I(\hat{X}_i \wedge \hat{X}_j;\ B)$ can become smaller than $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$ (Figure 4-6A, the center chart). In spite of such a decrease $I(\hat{X}_i \wedge \hat{X}_j;\ B)$ [or $I(X_i \wedge X_j;\ B)$] remains consistently higher than the $I(\hat{X}_i \vee \hat{X}_j;\ B)$ [or $I(X_i \vee X_j;\ B)$] across all the three conditions by a factor of 3.5, 3.8, and 2.1, respectively. Notice that all the information quantities in Figure 4-6 are normalized by dividing them by the entropy B and are unitless.

Similarly, when the underlying determinist function driving the output is a MAX (OR-like), then the information that $Y = max(X_i, X_j)$ carries about B becomes the highest (Figure 4-6B, the left-most chart; compare yellow bars to all other bars). Using surrogate variables instead of the actual ones (see Section 4.4.1) lowers $I(\hat{X}_i \vee \hat{X}_j;\ B)$, so that it may become equal or lower than the three-way information $I(\hat{X}_i, \hat{X}_j;\ B)$ (compare the yellow bar with the light-green bar in Figure 4-6B, the centeral chart), yet $I(\hat{X}_i \vee \hat{X}_j;\ B)$ is nevertheless higher than sum of the individual information $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$ (compare the yellow bar with the dark-green bar in the same chart). Adding noise reduces all the information and $I(\hat{X}_i \vee \hat{X}_j;\ B)$ may become lower than $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$ or $I(\hat{X}_i, \hat{X}_j;\ B)$ (Figure 4-6B, the right-most chart). Again what remains stable is the fact that $I(\hat{X}_i \vee \hat{X}_j;\ B)$ [or $I(X_i \vee X_j;\ B)$] is consistently higher than the $I(\hat{X}_i \wedge \hat{X}_j;\ B)$ [or

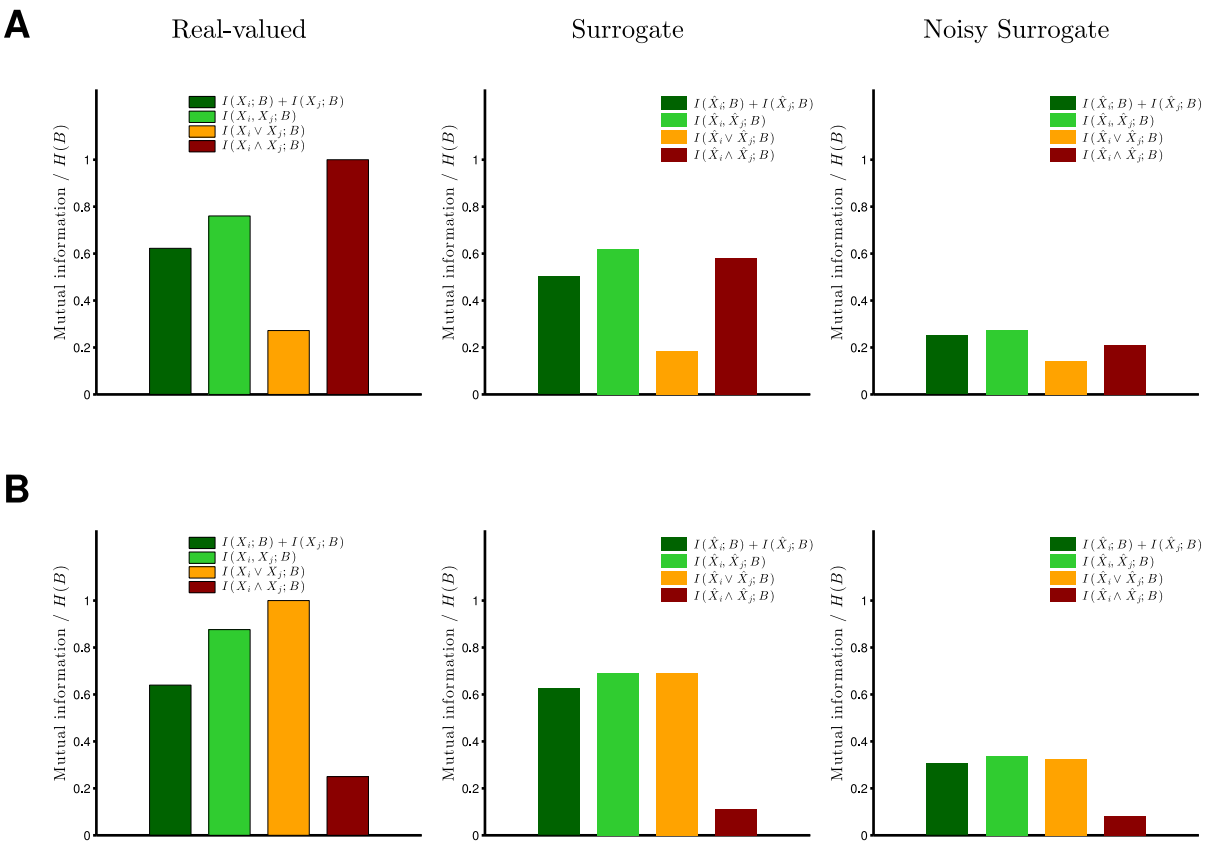

Figure 4-6: Result of evaluating the main theorem under different conditions in the toy example. The charts shows four information quantities between variables of the toy example illustrated by the Bayesian network in Figure 4-2. AND-like information (dark-red), OR-like information (yellow), the three-way mutual information (light-green), the linear sum of individual information (dark-green) on the abscissa. The ordinate shows the amount of mutual information divided by the entropy of variable B. **A.** The function f of the Bayesian network was set to be an AND-like interaction i.e. the product of them.

$I(X_i \wedge X_j;\ B)$] across all the three conditions by a factor of 3.5, 7.1, and 4.5, respectively.

4.5.2 Results of the Simulated Bubbles Experiment

We simulated a Bubbles experiment with the default view of Object 2 used in previous chapters. As explained in the method, we defined two features of Object 2, located in grid squares B2 and B7 (Figure 4-3), to be relevant for the decision of the simulated observers. Two different observers were simulated: one performed an AND-like interaction strategy between these features, and another one performed an OR-like interaction. The observer performing the AND-like interaction computed the binary evidence of each feature (i.e.

decided whether each feature was present or not) in the bubbles-masked image, and then computed the logical AND of the binary evidences (details are provided in Section 4.4.3). This was done for all possible pairs of grid squares (i.e., potential features) in the down-sampled bubbles-masked images (see Figure 4-5). The observer performing the OR-like interaction did the same, with the only difference that the OR operation was used as the interaction between the binary evidences. It should be noted that, since our simulated non-linear observer integrates feature evidences that are binary, it did not make sense to simulate a linear sum interaction (which is not well defined for binary variables).

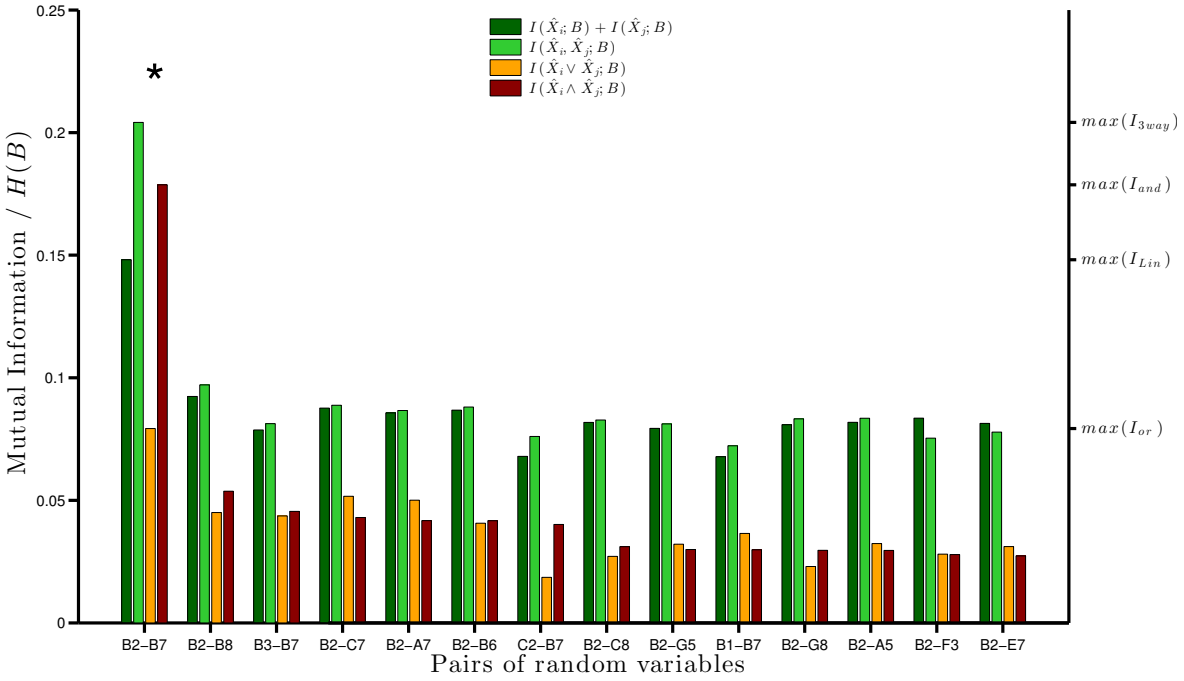

Figure 4-7: Extracting non-linear interaction of diagnostic features in a simulated observer carrying out an AND-like interaction strategy. The information quantities were computed between every pair ($\hat{X}_i$ and $\hat{X}_j$) of rows in the matrix of trials shown in Figure 4-5C and the behavior correctness variable B of the observer . The four information quantities which are the same as quantities in Figure 4-6 are computed and shown with the same color code for different pairs of grid squares coded as shown in Figure 4-5B. The ordinate is the amount of mutual information normalized by the entropy of the behavior random variable B. The simulated observer was implemented with an AND-like interaction between two features located in squares B2 and B7. Here in the result, we see only for the pair B2-B7 the information AND (dark-red) is significantly higher (* $p > 0.5$; permutation test) than the sum of individual information (dark-green) and than the OR information (yellow).

Figure 4-7 shows the result of extracting different kinds of non-linear feature interaction

(i.e., both AND-like and OR-like; see, respectively, the red and the yellow bars) when the behavior of the simulated observer was driven by an AND-like strategy. The abscissa shows different pairs of random variables (more specifically pairs of grid squares in the downsampled stimulus image) and the ordinate shows the amount of mutual information divided by the entropy of the behavioral outcome variable B. For each pair, we computed the four information quantities: (1) the sum of the individual information $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$ shown in dark-green, (2) the three-way mutual information $I(\hat{X}_i, \hat{X}_j;\ B)$ shown in light-green, (3) the OR information $I(\hat{X}_i \vee \hat{X}_j;\ B)$ shown in yellow, and (4) the AND information $I(\hat{X}_i \wedge \hat{X}_j;\ B)$ shown in dark-red color. The pairs are ranked based on the magnitude of the AND information values (dark-red bars). Only for the pair B2-B7 the AND information is close to the three-way information (~86%) and significantly higher than sum of the individual information (* $p > 0.5$, Bonferroni corrected; permutation test). The maximums of all four quantities across 64 possible pairs are shown on the right-hand side axis. For example $max(I_{Lin})$ corresponds to the maximum value of dark-green bars (the linear summation of individual information within pairs) across all 64 pairs. Not only does the pair B2-B7 have the highest AND information across pairs, but it possesses the maximum of other three information quantities across all pairs. The AND information of the pair B2-B7 is 2.3 times higher than the corresponding OR information. Another observation is that only for the pair B2-B7 the three-way information is significantly higher than the sum of the individual information ($p > 0.5$, Bonferroni corrected; permutation test). In fact the three-way mutual information for all other pairs except for B2-B7 are lower than 0.1. As explained in the Section 4.3, this means that there is a synergy between the pair of features located in B2 and B3. The large value of the AND information reveals that the interaction underlying this synergy is consistent with a non-linear AND-like operation. It should be noted that the fact that $I(\hat{X}_i \wedge \hat{X}_j;\ B)$ is lower than $I(\hat{X}_i, \hat{X}_j;\ B)$ has to be attributed to the fact that mutual information quantities are computed using surrogated variables (as explained in the case of the toy model, shown in the previous chapter). Overall, Figure 4-7 shows that our information-theoretic analysis is able to find the AND interaction when an observer is using such a strategy. But this is not enough to show that the analysis is valid. We need to show that if there is no AND interaction at all, the analysis does not erroneously

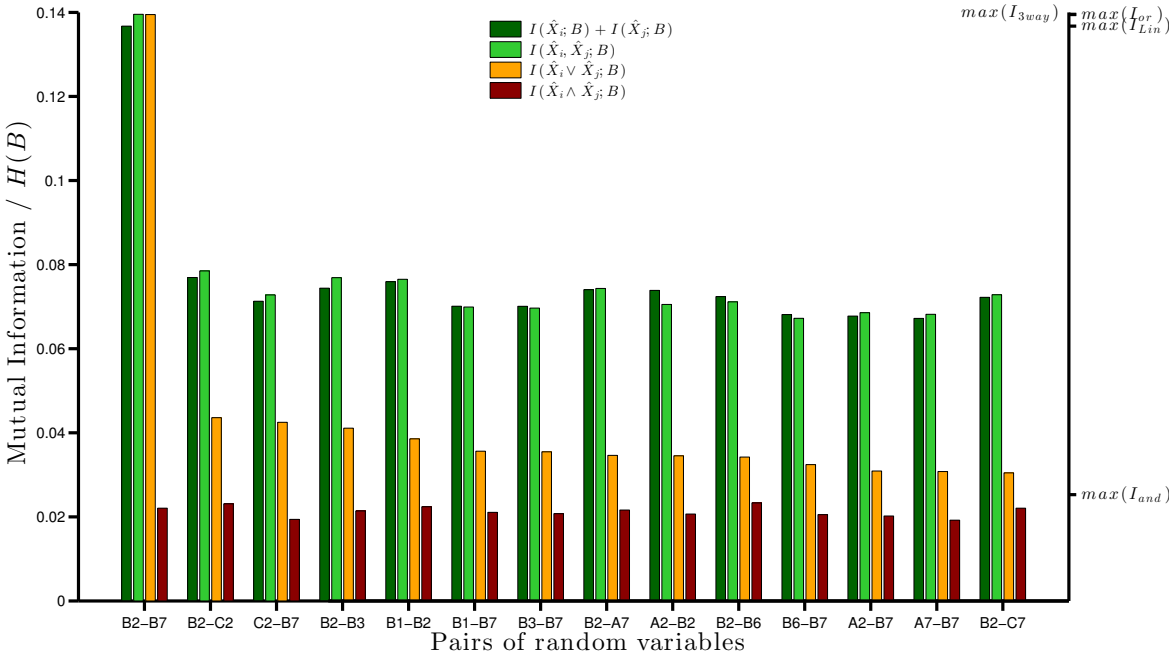

Figure 4-8: Extracting non-linear interaction of diagnostic features in a simulated observer carrying out an OR-like interaction strategy. The same quantities as in Figure 4-7 were computed for every pair of rows ($\hat{X}_i$ and $\hat{X}_j$) in the matrix of trials shown in Figure 4-5C and the behavior variable B of the observer. The four information quantities which are the same as quantities in Figure 4-6 are computed and shown with the same color code for different pairs of grid squares coded as shown in Figure 4-5B. The ordinate is the amount of mutual information normalized by the entropy of the behavior random variable B. The simulated observer was implemented with an OR-like interaction between two features located in squares B2 and B7. Here in the result, we see only for the pair B2-B7 the OR information (yellow) is higher than the sum of individual information (dark-green), and than the three-way information (light-green) and the AND information (dark-red).

extract any AND interactions.

To this aim, we carried out a simulation in which an observer performed an OR-like interaction strategy. The result of applying our feature interaction analysis to such a case is shown in Figure 4-8. The figure is similar to Figure 4-7 except that here the pairs are ranked based on the OR information values (yellow bars). No pair has high AND information value compared with the other information quantities (look at the max of AND information on the right-side axis). In fact, the AND information values of all the pairs are less than half of their three-way mutual information values. Therefore, the analysis recovers no AND interaction. Instead, it shows that the OR information of the pair B2-B7 is almost equal to

the three-way information and is higher than the sum of individual information values. This is consistent with the simulated strategy of the observer. Again, the fact that $I(\hat{X}_i, \hat{X}_j;\ B)$ is higher than $I(\hat{X}_i;\ B) + I(\hat{X}_j;\ B)$ indicates that there is a synergy between features located in these two grid squares in determining the behavioral outcome. Such a synergy is accounted by an OR-like interaction model.

4.6 Discussion

In this chapter, we developed a novel information-theoretic approach that systematically extracts non-linear interactions of diagnostic features in a Bubbles experiments. We presented a mathematical framework which sets two criteria for finding non-linear interactions, explaining under what conditions the framework may not give the satisfactory results (Figure 4-6). By implementing a toy example, we found that when we use the actual real-valued features (upon which the behavior of the toy model is based), then Eqs. (4.12, 4.13) hold, verifying the Main theorem. When we deviate from the ideal situation and we use (noisy) surrogate variables instead of real features, all the information quantities decrease and one or both of the Eqs. (4.12, 4.13) may not hold anymore. Yet we were able to recover the AND-like/OR-like interaction by comparing the AND/OR information with three-way mutual information and the sum of individual information and choose the function (i.e., either AND or OR), whose information is closest to these quantities.

The simulations of non-linear observers performing either an AND-like or an OR-like feature integration confirm that our approach can recover such non-linear interactions in a simulated Bubbles experiment (Figure 4-7, 4-8). We showed that when a simulated observer performed an AND-like interaction between two features, only for the pair of grid squares that have the features, the information of the AND was close to three-way mutual information and significantly higher than other quantities—but not for the other pairs. On the contrary, when the observer performed an OR-like strategy, the OR information was almost equal to the three-way mutual information and was higher than the others (again this was true only for the pair of grid squares containing the simulated features).

Overall, these simulations show that the proposed information theoretic approach re-

liably tells whether the recognition strategy of an observer is more consistent with the requirement that two diagnostic features be simultaneously visible/present (AND-like strategy) or, rather, with the requirement that either feature be visible/present (OR-like strategy). One limitation of our current approach is that we did not carry out extensive simulations of observers performing a linear feature integration (i.e., a SUM between a pair of simulated features). The reason is that implementing a linear observer is not trivial, given that any discretization of the sum of two evidences (i.e., $Y = X_i + X_j$) to produce a binary behavioral output B introduce, by definition, a strong non-linearity. Intuitively, we think that such a discretization will bring the behavior of the simulated observer close to an OR-like or to and AND-like non-linear feature integration strategy, depending on the threshold of the discretization (e.g., depending on threshold δ, where $B = 0$ if $Y = X_i + X_j < \delta$, and $B = 1$ if $Y = X_i + X_j > \delta$). For example, when δ was set to 0.1 (with X_i and X_j ranging from 0 to 1), we observed that the thresholded SUM strategy of the simulated observer was more consistent with an OR strategy than with an AND strategy (see Figure 4-9). We have not exhaustively simulated thresholded SUM observers with varying thresholds, but we think that higher δ values may bring the observers recognition strategy to be more consistent with an AND-like feature interaction. Further simulations are necessary to verify this and, more in general, to verify how a linear observer can be modeled and a linear recognition strategy can be analyzed, so as to provide a better ground of comparison for non-linear feature integration strategies such as the AND and the OR.

Finally, it is important to notice the analysis can be generalized to incorporate three-way feature interactions (or even more) instead of pair-wise interactions. This may require, however, more trials to properly estimate the information quantities.

4.6.1 Comparison with previous studies

An information-theoretic analysis similar to ours (but not in a Bubbles experiment study) has been used to quantify how clusters of neurons in the rat barrel cortex encode the kinetic features of sinusoidal whisker vibrations (Arabzadeh et al., 2004). Based on this approach, the authors reported that barrel neurons mainly encode the product of frequency and am-

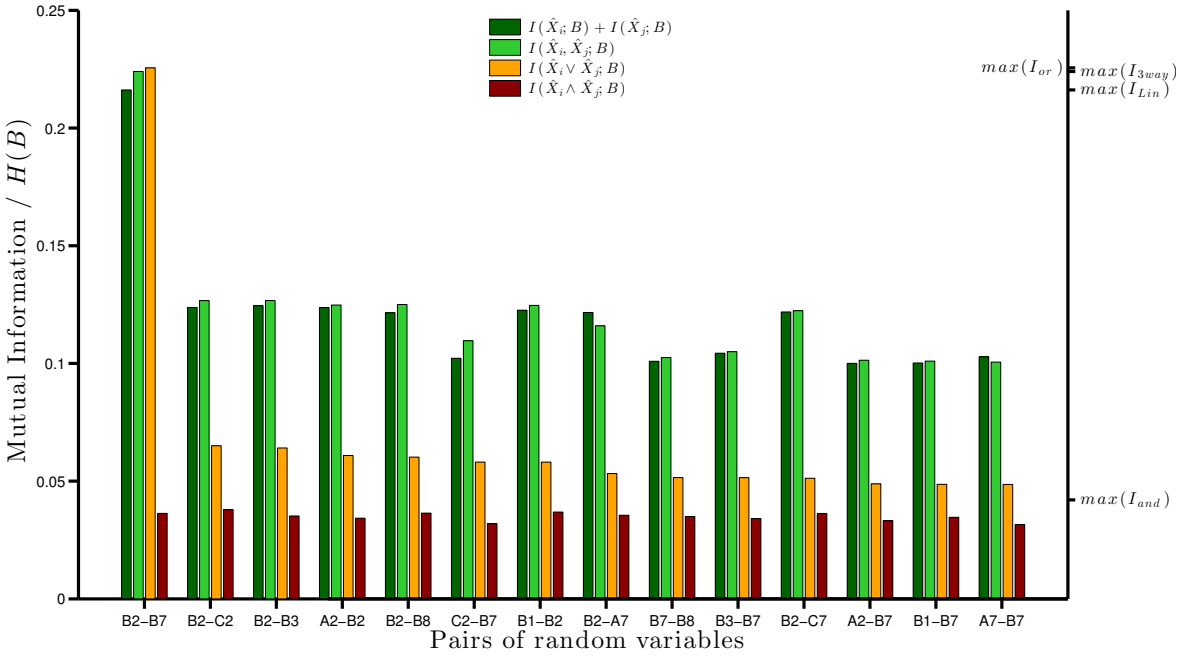

Figure 4-9: Extracting non-linear interaction of diagnostic features in a linear simulated observer carrying out an SUM strategy. The same quantities as in Figure 4-7 were computed for every pair of rows ($\hat{X}_i$ and $\hat{X}_j$) in the matrix of trials shown in Figure 4-5C and the behavior variable B of the observer. This figure shows the limitation of the analysis in the current setting as it shows the OR information to be high, therefore not being able to differentiate between OR-like strategy of the observer and SUM strategy.

plitude of the stimuli. It should be noticed that the inequalities that the authors relied upon are different from ours because they relied on different dependencies between random variables and they had different stimulus set design.

Another study applied information theory for analyzing data of a Bubbles experiment (Schyns et al., 2011). In this study, participants were asked to categorize eight facial expressions of emotion. The authors quantified information carried by three parameters of EEG oscillatory signals (i.e. power, phase, frequency) about the behavior. In particular, they computed two three-way mutual information quantities: the information of the feature pair (power, phase) about behavior, and the information of that pair about pixels. They found that the pair (power, phase) codes specific expressive features across different oscillatory bands. The study is related to our approach but because of the different nature of the features (not so meaningful to multiply power quantity by a phase quantity) only the three-way mutual

information was considered (i.e., no attempt was done at estimating the information carried by function of power and phase). Moreover, in contrast to our study, the authors were not interested in finding the simultaneous presence of specific stimuli features important for the categorization task.

The standard analysis of Bubbles experiment leading to saliency maps (see previous chapters) is, in essence, very similar to the family of reverse correlation and spike-triggered average (STA) techniques, which are used for characterizing receptive of neurons (Dayan and Abbott, 2001). In the analysis of the Bubbles experiments data, the saliency maps essentially show which pixels of the bubbles-masks were more correlated with the correct response of the observer (in other words, a weighted average of response-trigged masks is computed; see 2.3.1). In STA, one looks for the simplest deviation of spike-triggered ensemble stimuli from the raw stimuli, which is the deviation in mean.

This relationship may hint at a relationship between an extension of spike-triggered techniques known as spike-triggered covariance (STC) and our non-linear interaction analysis. STC investigated by some studies for neural characterization (Pillow and Simoncelli, 2006; Schwartz et al., 2006), extends STA this by seeking directions in the stimulus space that most changes the variance of the spike-triggered ensemble in comparison of the raw stimulus ensemble. In both STA and STC, one can find the nonlinearity function underlying Linear-Nonlinear Poisson (LNP) model of neuron by computing the ratio between the probability of a stimulus given a spike and the probability of the raw stimuli. However, these methods rely on the assumption of Gaussian stimulus set that is not valid in Bubbles experiment, and therefore, the subspace that STC finds may lead to artifacts (Pillow and Simoncelli, 2006). Moreover, the non-linearity that STC finds is applied to the outcome of the dot product of each subspace (each subspace is an hyperplane in a high-dimensional stimulu space) and the stimuli in order to compute the firing rate of an LNP neuron model. This nonlinearity is different from the non-linearity that we aim at recovering in the Bubbles experiments where just a sparse image is shown to the observer and the nonlinearity between just a limited set of diagnostic features of the stimulus image is sought—as opposed to non-linearity between the linear combinations of features and the output in STC. We ran the STC analysis on our simulation results and did not achieve the desired result.

Nevertheless, this method could potentially be changed, in future, so as to be applied to Bubbles experiments.

4.7 Concluding Remarks and Future Work

As a future development, to furthur expand the proposed analysis, the strategy of the simulated observer can be extended to more complicated, realistic scenarios. For example, we may ask how the information quantities look like if more than one pair of features interact, or both objects have multiple feature interactions, or an object feature in conjunction with the lack of presence of a feature in the other object help an observer to recognize an object. Simulations under different noise conditions can also be carried out to see how these conditions may affect the feature interaction recovery and where it may fail.

In summary, our simulations show that our information-theoretic approach can recover the type and the strength of the non-linear interactions among the salient features of an object. We are currently applying the analysis to human and rat data obtained from the studies described in previous chapters. It would be interesting to obtain non-linear feature interaction in different views of an object and see how it may change across views. This can shed more light on the strategies used by different species in invariant object recognition.

Bibliography

Adelman TL, Bialek W, Olberg RM (2003) The information content of receptive fields. *Neuron* 40:823–833. 83

Afraz A, Cavanagh P (2009) The gender-specific face aftereffect is based in retinotopic not spatiotopic coordinates across several natural image transformations. *J Vis* 9:10.1–17. 2, 3, 46

Afraz SR, Cavanagh P (2008) Retinotopy of the face aftereffect. *Vision Res* 48:42–54. 2, 3, 46

Andermann ML, Kerlin AM, Roumis DK, Glickfeld LL, Reid RC (2011) Functional specialization of mouse higher visual cortical areas. *Neuron* 72:1025–1039. 43

Arabzadeh E, Panzeri S, Diamond ME (2004) Whisker vibration information carried by rat barrel cortex neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 24:6011–6020. 100

Biederman I (1987) Recognition-by-components: a theory of human image understanding. *Psychol Rev* 94:115–47. 3, 67, 68, 75

Biederman I (2000) Recognizing depth-rotated objects: A review of recent research and theory. *Spatial vision* 13:2–3. 3, 67, 68, 75

Biederman I, Cooper EE (1991) Evidence for complete translational and reflectional invariance in visual object priming. *Perception* 20:585–593 PMID: 1806902. 2, 3, 67

Biederman I, Gerhardstein P (1993) Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance; Journal of Experimental Psychology: Human Perception and Performance* 19:1162. 3, 67, 68

Biederman I, Cooper EE (1992) Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance* 18:121–133. 2, 46

Bonin V, Histed MH, Yurgenson S, Reid RC (2011) Local diversity and fine-scale organization of receptive fields in mouse visual cortex. *The Journal of Neuroscience* 31:18506–18521. 43

Brincat S, Connor C (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat Neurosci* 7:880–6. 70, 78

Brincat S, Connor C (2006) Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron* 49:17–24. 70

Bülthoff H, Edelman S (1992) Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences* 89:60–64. 3, 67, 69

Busse L, Ayaz A, Dhruv NT, Katzner S, Saleem AB, Schlvinck ML, Zaharia AD, Carandini M (2011) The detection of visual contrast in the behaving mouse. *The Journal of Neuroscience* 31:11351–11361. 43

Blthoff HH, Edelman SY, Tarr MJ (1995) How are three-dimensional objects represented in the brain? *Cerebral Cortex (New York, N.Y.: 1991)* 5:247–260 PMID: 7613080. 3, 46, 67, 69

Chelazzi L, Rossi F, Tempia F, Ghirardi M, Strata P (1989) Saccadic eye movements and gaze holding in the head-restrained pigmented rat. *Eur J Neurosci* 1:639–646. 23, 28, 43

Cover TM, Thomas JA (2006) *Elements of Information Theory 2nd Edition* Wiley-Interscience, 2 edition. 79, 81

Dayan P, Abbott L (2001) Theoretical neuroscience: Computational and mathematical modeling of neural systems. 2001. 102

DiCarlo J, Zoccolan D, Rust N (2012) How does the brain solve visual object recognition? *Neuron* 73:415–434. 2, 5, 11, 70

Edelman S (1999) *Representation and Recognition in Vision* The MIT Press. 3, 69

Felleman D, Van Essen D (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* 1:1–47. 2, 4, 5

Fortune B, Bui BV, Morrison JC, Johnson EC, Dong J, Cepurna WO, Jia L, Barber S, Cioffi GA (2004) Selective ganglion cell functional loss in rats with experimental glaucoma. *Investigative Ophthalmology & Visual Science* 45:1854–1862. 9

Foster DH, Gilson SJ (2002) Recognizing novel three-dimensional objects by summing signals from parts and views. *Proceedings of the Royal Society B: Biological Sciences* 269:1939–1947. 3

Freedman D, Riesenhuber M, Poggio T, Miller E (2005) Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cereb Cortex* . 3, 69

Gao E, DeAngelis GC, Burkhalter A (2010) Parallel input channels to mouse primary visual cortex. *The Journal of Neuroscience* 30:5912 –5926. 43

Gibson B, Wasserman E, Gosselin F, Schyns P (2005) Applying bubbles to localize features that control pigeons' visual discrimination behavior. *J Exp Psychol Anim Behav Process* 31:376–82. 6, 16, 18, 19, 40, 42, 68

Gibson BM, Lazareva OF, Gosselin F, Schyns PG, Wasserman EA (2007) Nonaccidental properties underlie shape recognition in mammalian and nonmammalian vision. *Current Biology* 17:336–340. 6, 18, 40

Gosselin F, Schyns P (2004) No troubles with bubbles: a reply to murray and gold. *Vision Res* 44:471–7; discussion 479–82. 74

Gosselin F, Schyns P (2001) Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Res* 41:2261–71. 4, 6, 10, 16, 18, 19, 40, 42, 78

Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics* Wiley, New York. 88

Greenberg DS, Houweling AR, Kerr JND (2008) Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nat Neurosci* 11:749–751. 43

Hayward WG (2003) After the viewpoint debate: where next in object recognition? *Trends in cognitive sciences* 7:425–427. 3

Histed MH, Carvalho LA, Maunsell JHR (2012) Psychophysical measurement of contrast sensitivity in the behaving mouse. *Journal of Neurophysiology* 107:758–765. 43

Hubel D, Wiesel T (1959) Receptive fields of single neurones in the cat's striate cortex. *J Physiol* 148:574–91. 2

Hubel D, Wiesel T (1968) Receptive fields and functional architecture of monkey striate cortex. *J Physiol* 195:215–43. 2

Huberman AD, Niell CM (2011) What can mice tell us about how vision works? *Trends in Neurosciences* 34:464–473. 4, 5, 44

Hung C, Kreiman G, Poggio T, DiCarlo J (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310:863–6. 2

Jack RE, Garrod OGB, Yu H, Caldara R, Schyns PG (2012) Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* . 42

Jacobs G, Fenwick J, Williams G (2001) Cone-based vision of rats for ultraviolet and visible lights. *J Exp Biol* 204:2439–46. 9

Keller J, Strasburger H, Cerutti D, Sabel B (2000) Assessing spatial vision - automated measurement of the contrast-sensitivity function in the hooded rat. *J Neurosci Methods* 97:103–10. 17

Kerr JN, Nimmerjahn A (2012) Functional imaging in freely moving animals. *Current Opinion in Neurobiology* 22:45–53. 43

Koller D, Friedman N (2009) *Probabilistic Graphical Models: Principles and Techniques* The MIT Press, 1 edition. 81

Kourtzi Z, Connor CE (2011) Neural representations for object perception: Structure, category, and adaptive coding. *Annual Review of Neuroscience* 34:45–67. 70, 78

Kravitz DJ, Kriegeskorte N, Baker CI (2010) High-level visual object representations are constrained by position. *Cerebral Cortex* 20:2916 –2925. 46

Lawson R (1999) Achieving visual object constancy across plane rotation and depth rotation. *Acta Psychologica* 102:221–245 PMID: 10504882. 3, 67, 75

Lee AK, Manns ID, Sakmann B, Brecht M (2006) Whole-Cell Recordings in Freely Moving Rats. *Neuron* 51:399–407. 44

Li N, Cox DD, Zoccolan D, DiCarlo JJ (2009) What response properties do individual neurons need to underlie position and clutter Invariant object recognition? *Journal of Neurophysiology* 102:360–376. 11

Logothetis NK, Pauls J, Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. *Current biology : CB* 5:552–563. 2

Logothetis N, Pauls J, Bülthoff H, Poggio T et al. (1994) View-dependent object recognition by monkeys. *Current biology* 4:401–414. 3, 67, 69

Logothetis N, Sheinberg D (1996) Visual object recognition. *Ann. Rev. Neurosci.* 19:577–621. 2, 5

Lowe DG (1987) Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence* 31:355 – 395. 68

Magri C, Whittingstall K, Singh V, Logothetis NK, Panzeri S (2009) A toolbox for the fast information analysis of multiple-site LFP, EEG and spike train recordings. *BMC Neuroscience* 10:81. 84, 87, 92

Marshel J, Garrett M, Nauhaus I, Callaway E (2011) Functional specialization of seven mouse visual cortical areas. *Neuron* 72:1040–1054. 43

Meier P, Flister E, Reinagel P (2011) Collinear features impair visual detection by rats. *Journal of Vision* 11. 43

Melcher D, Colby CL (2008) Trans-saccadic perception. *Trends in Cognitive Sciences* 12:466–473. 23

Minini L, Jeffery K (2006) Do rats use shape to solve "shape discriminations"? *Learn Mem* 13:287–97. 4, 21, 27, 40, 45

Murray R (2004) Reply to gosselin and schyns. *Vision Res* 44:479–482. 74

Murray R, Gold J (2004) Troubles with bubbles. *Vision Res* 44:461–70. 74

Murray RF (2011) Classification images: A review. *Journal of Vision* 11. 42, 46, 77

Naarendorp F, Sato Y, Cajdric A, Hubbard NP (2001) Absolute and relative sensitivity of the scotopic system of rat: Electroretinography and behavior. *Visual Neuroscience* 18:641–656. 9

Nassi JJ, Callaway EM (2009) Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience* 10:360–372. 4, 5

Niell C, Stryker M (2008) Highly selective receptive fields in mouse visual cortex. *J Neurosci* 28:7520–36. 43

Niell CM (2011) Exploring the next frontier of mouse vision. *Neuron* 72:889–892. 5, 44

Niell CM, Stryker MP (2010) Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron* 65:472–479. 43

Nielsen K, Logothetis N, Rainer G (2006) Dissociation between local field potentials and spiking activity in macaque inferior temporal cortex reveals diagnosticity-based encoding of complex objects. *J Neurosci* 26:9639–45. 31, 40

Nielsen K, Logothetis N, Rainer G (2008) Object features used by humans and monkeys to identify rotated shapes. *J Vis* 8:9 1–15. 2, 40, 46, 67, 73, 74

Ohki K, Chung S, Ch'ng Y, Kara P, Reid R (2005) Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* 433:597–603. 44

Op de Beeck H, Wagemans J, Vogels R (2001) Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience* 4:1244–1252. 2

Orban G (2008) Higher order visual processing in macaque extrastriate cortex. *Physiological Reviews* 88:59. 2, 5

Pasupathy A, Connor CE (1999) Responses to contour features in macaque area v4. *Journal of Neurophysiology* 82:2490 –2502. 70

Paxinos G (2004) *The Rat Nervous System* Gulf Professional Publishing. 43

Pillow JW, Simoncelli EP (2006) Dimensionality reduction in neural models: An information-theoretic generalization of spike-triggered average and covariance analysis. *Journal of Vision* 6:9–9. 102

Pinto N, Cox D, DiCarlo J (2008) Why is real-world visual object recognition hard? *PLoS Comput Biol* 4:e27. 3, 5, 28, 73

Poggio T, Edelman S (1990) A network that learns to recognize three-dimensional objects. *Nature* 343:263–266. 3, 67, 69

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* 136:339–348. 17

Riesenhuber M, Poggio T (2000) Models of object recognition. *Nat Neurosci* 3 Suppl:1199–204. 3, 69

Rolls E (2000) Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* 27:205–18. 5

Rust NC, DiCarlo JJ (2010) Selectivity and tolerance ("Invariance") both increase as visual information propagates from cortical area v4 to IT. *J. Neurosci.* 30:12978–12995. 11

Sawinski J, Wallace DJ, Greenberg DS, Grossmann S, Denk W, Kerr JND (2009) Visually evoked activity in cortical cells imaged in freely moving animals. *Proceedings of the National Academy of Sciences* 106:19557 –19562. 43, 44

Schwartz O, Pillow JW, Rust NC, Simoncelli EP (2006) Spike-triggered neural characterization. *Journal of Vision* 6:13–13. 102

Schyns P, Bonnar L, Gosselin F (2002) Show me the features! understanding recognition from the use of visual information. *Psychological science : a journal of the American Psychological Society / APS* 13:402–9. 42

Schyns P, Thut G, Gross J (2011) Cracking the code of oscillatory activity. *PLoS biology* 9:e1001064. 101

Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences* 104:6424 –6429. 52, 73

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27:379–423. 79

Tafazoli S, Di Filippo A, Zoccolan D (2012) Transformation-tolerant object recognition in rats revealed by visual priming. *The Journal of Neuroscience* 32:21–34. 4, 6, 18, 22, 27, 42, 43, 45

Tanaka K (1996) Inferotemporal cortex and object vision. *Annual Review of Neuroscience* 19:109–139. 2, 5

Tarr MJ, Blthoff HH (1998) Image-based object recognition in man, monkey and machine. *Cognition* 67:1–20. 3, 46, 67, 69, 75

Thomas BB, Samant DM, Seiler MJ, Aramant RB, Sheikholeslami S, Zhang K, Chen Z, Sadda SR (2007) Behavioral evaluation of visual function of rats using a visual discrimination apparatus. *Journal of Neuroscience Methods* 162:84–90. 9

Van Hooser SD (2006) Lack of Patchy Horizontal Connectivity in Primary Visual Cortex of a Mammal without Orientation Maps. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 26:7680–7692. 43, 44

Vermaercke B, OpdeBeeck H (2012) A multivariate approach reveals the behavioral templates underlying visual discrimination in rats. *Current Biology* 22:50–55. 2, 4, 16, 27, 40, 42, 43, 45, 46, 67

Wang Q, Gao E, Burkhalter A (2011) Gateways of ventral and dorsal streams in mouse visual cortex. *The Journal of Neuroscience* 31:1905 –1918. 43

Yamane Y, Carlson E, Bowman K, Wang Z, Connor C (2008) A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat Neurosci* . 78

Zoccolan D, Cox D, DiCarlo J (2005) Multiple object response normalization in monkey inferotemporal cortex. *J Neurosci* 25:8150–64. 11

Zoccolan D, Oertelt N, DiCarlo J, Cox D (2009) A rodent model for the study of invariant visual object recognition. *Proc Natl Acad Sci U S A* 106:8748–53. 4, 6, 8, 18, 22, 42, 43, 45

Zoccolan D, Graham BJ, Cox DD (2010) A self-calibrating, camera-based eye tracker for the recording of rodent eye movements. *Frontiers in Neuroscience* 4:193 PMID: 21152259. 23, 28, 43

Zoccolan D, Kouh M, Poggio T, DiCarlo JJ (2007) Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 27:12292–12307. 11