3-2023

Reference Frames in Human Sensory, Motor, and Cognitive Processing

Dongcheng He
University of Denver

Reference Frames in Human Sensory, Motor, and Cognitive Processing

Abstract

Reference-frames, or coordinate systems, are used to express properties and relationships of objects in the environment. While the use of reference-frames is well understood in physical sciences, how the brain uses reference-frames remains a fundamental question. The goal of this dissertation is to reach a better understanding of reference-frames in human perceptual, motor, and cognitive processing. In the first project, we study reference-frames in perception and develop a model to explain the transition from egocentric (based on the observer) to exocentric (based outside the observer) reference-frames to account for the perception of relative motion. In a second project, we focus on motor behavior, more specifically on goal-directed reaching. We develop a model that explains how egocentric perceptual and motor reference-frames can be coordinated through exocentric reference-frames. Finally, in a third project, we study how the cognitive system can store and recognize objects by using sensorimotor schema that allows mental rotation within an exocentric reference-frame.

Document Type

Dissertation

Degree Name

Ph.D.

Department

Computer Engineering

First Advisor

Haluk Ogmen

Second Advisor

Mario Lopez

Third Advisor

Mohammad Mahoor

Keywords

Binocular combination, Canonical theory, Mental rotation, Motion perception, Reference frames, Sensorimotor control

Subject Categories

Artificial Intelligence and Robotics | Cognition and Perception | Cognitive Psychology | Computer Sciences | Other Computer Engineering | Psychology | Robotics

Reference Frames in Human Sensory, Motor, and Cognitive Processing

---

A Dissertation

Presented to

the Faculty of the Daniel Felix Ritchie School of Engineering and Computer Science

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Dongcheng He

March 2023

Advisor: Haluk Ogmen, Ph.D.

Author: Dongcheng He
Title: Reference Frames in Human Sensory, Motor, and Cognitive Processing
Advisor: Haluk Ogmen, Ph.D.
Degree Date: March 2023

Abstract

Reference-frames, or coordinate systems, are used to express properties and relationships of objects in the environment. While the use of reference-frames is well understood in physical sciences, how the brain uses reference-frames remains a fundamental question. The goal of this dissertation is to reach a better understanding of reference-frames in human perceptual, motor, and cognitive processing. In the first project, we study reference-frames in perception and develop a model to explain the transition from egocentric (based on the observer) to exocentric (based outside the observer) reference-frames to account for the perception of relative motion. In a second project, we focus on motor behavior, more specifically on goal-directed reaching. We develop a model that explains how egocentric perceptual and motor reference-frames can be coordinated through exocentric reference-frames. Finally, in a third project, we study how the cognitive system can store and recognize objects by using sensorimotor schema that allows mental rotation within an exocentric reference-frame.

Acknowledgements

I would like to express my deepest gratitude to Dr. Haluk Ogmen for guiding me towards this life milestone with constant support, patience, and encouragement. Studying in such a reassuring environment during the past five years is the most valuable and inspirational experience of my life.

I wish to thank me very much for doing all this hard work.

I wish to thank all these members of my doctoral committee for their help and feedback.

I wish to thank my friends, colleagues, and subjects for their help in my life and research. Specially, I wish to thank Mr. Zijia Liu, my best friend, for giving my faith back to me every time when I lost it. Without them, I couldn't make it.

In the last, I wish to thank my family, especially my parents, for their accompany and belief in me. Also, I wish to extend my thanks to Mr. Ziang Zhou. The day we met was when the new chapter of my life began.

Table of Contents

List of Figures

List of Tables

Chapter One: Introduction

1 Introduction

In physics, reference-frames, or coordinate systems, are used extensively to express properties and relationships of objects in the environment. For example, the position and motion of an object can be described in terms of a variety of reference-frames: The motion and position of a traveler in a train can be expressed with respect to a reference-frame on the platform or a reference-frame on the moving wagon, among many choices. Planetary motions can be described according to a reference-frame based on earth or a reference-frame based on the sun. Although one reference-frame can be converted to the other, it is clear that the trajectories obtained according to the sun-based reference-frame are much simpler than those according to the earth-based reference-frame. Hence, the choice of the reference-frame can influence the complexity of resulting representations and computations.

Similarly, in processing the information about the environment and in producing actions, the brain, not only needs to use reference-frames, but also needs to choose judiciously the type of reference-frame that is most appropriate for a given task, so as to simplify its processing requirements. Whereas research established basic properties of reference-frames used by the brain, how the brain chooses reference-frames, how it converts representations in one reference-frame to those in another one remains largely unknown.

The goal of this dissertation is to investigate the use of reference frames in three broad aspects of neural processing: (i) sensory, (ii) motor, and (iii) cognitive. In the following sections, we will provide a review of our current knowledge about reference-frames used by the brain and identify the gaps that this work aims to fill in. Finally, at the end of this chapter, we will provide the research questions that we will investigate within the context of sensory, motor, and cognitive systems.

1.1 Reference frames in visual and motor systems

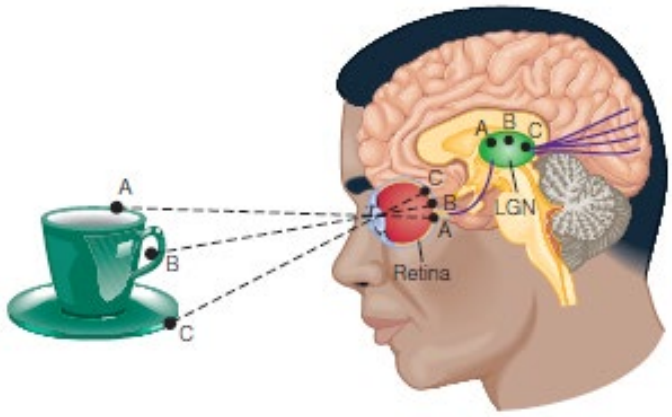

Figure 1 :Retinotopic maps from retina to visual cortex. The three spots on the cup, A, B, and C, are projected on the retina following the optics of the eye. The connections from retina to post-retinal areas are such that their relative spatial positions are preserved across the early visual cortex including lateral geniculate nucleus (LGN) and V1. These spatial representations are called retinotopic maps (Figure from Goldstein & James, 2016).

Our visual system uses a variety of reference frames. Beginning from the retinas, through the optics of the eye, images of neighboring points in the environment are mapped onto neighboring photoreceptors in the retina (Figure 1). The resulting retinal stimulus representations are called *retinotopic maps* because they are based on reference-frames placed on the retina, i.e., they represent positions *relative to the eyes*. Assume that we have an object that is stationary in the environment. The movement of the eyes

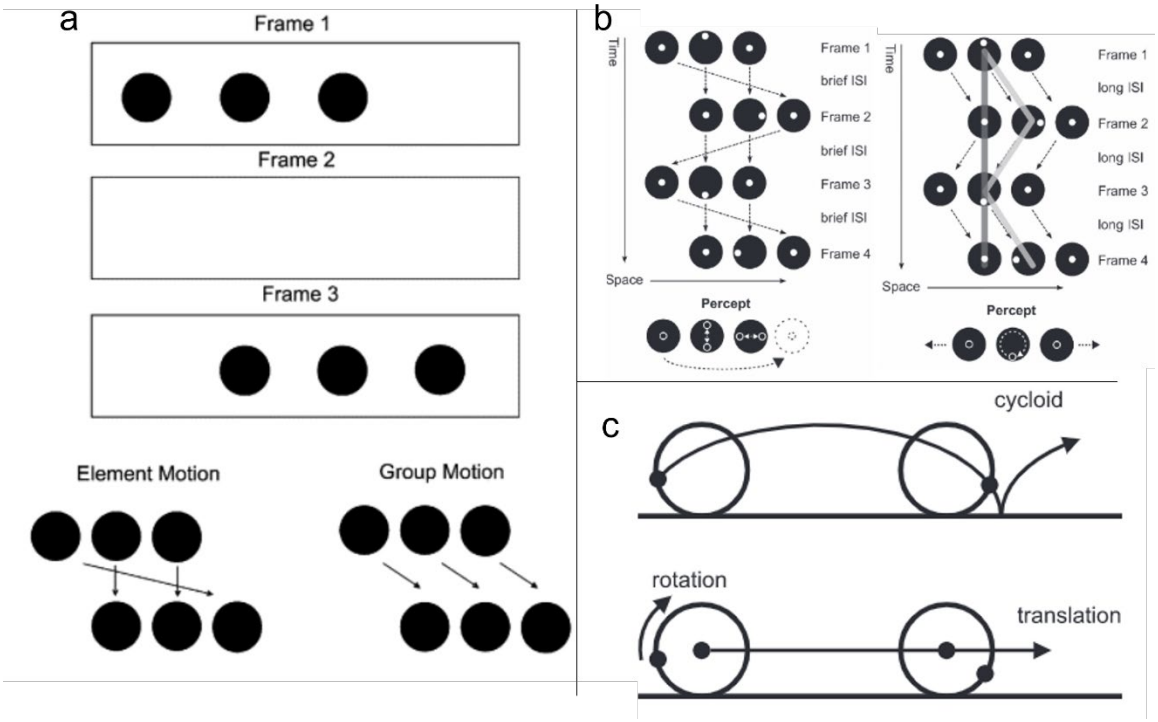

Figure 2: Visual displays used to demonstrate non-retinotopic processing a. Ternus-Pikler display. b. Motion correspondences in Ternus-Pikler display. c. Duncker's rotating wheel paradigm. See text for explanations of how these displays reveal non-retinotopic reference-frames.

will shift the retinotopic position of this object even though the position of the object in space remains constant. Although the early visual areas possess retinotopic organization, non-retinotopic reference-frames must be developed to perceive and understand complex our visual perception is non-retinotopic: For example, when we move our eyes, we do not perceive the stationary environment shifting with our eye movements. Another example can be found in the display illustrated in Figure 2 a. Three discs are shown as the first frame of a sequence of three frames. The second frame is blank and presented for a variable duration, the so-called ISI (inter-stimulus interval). In the final frame, the three (same) discs are shown in shifted positions so that two of the disks occupy the same position in the first and last frames. With a short ISI (e.g., 5 ms), observers perceive

"element motion" (Figure 2 a, bottom left), in which the left disc appears to move to the right while the other two discs appear stationary (Picciano & Picciano, 1976). When the ISI is long (e.g., 200ms), observers instead perceive "group motion" (Figure 2 a, bottom right), i.e., all three discs appear to move rightward in tandem (Picciano & Picciano, 1976). This paradigm is called the Ternus-Pikler Paradigm (Ternus, 1926; Pikler, 1917). The arrows in the bottom panels of Figure 2 a and in Figure 2 b indicate motion-correspondences between the elements in the two frames according to element and group motion conditions. When a small dot is inserted inside the Ternus-Pikler disks (the white dots in the black disks in Figure 2b), the motion of these dots is not perceived according to their retinotopic coordinates, but instead according to motion-correspondences in the Ternus-Pikler display (Boi et al., 2009). In other words, the motion correspondences serve as the reference-frame for the perception of the dot. Based on motion correspondences for the element motion case, the dot in the central disk appears to move up and down. On the other hand, based on motion correspondences for group motion, this same dot appears to rotate inside the central disk. Another example is shown in Figure 2 c. Assume that two reflectors are placed on the wheel of a bicycle, one on the rim and a second one at the center. As shown in the figure, when the wheel rotates, the corresponding retinal trajectory of the reflector on the rim is a cycloid. However, the reflector on the rim is not perceived to move according to its cycloid retinotopic trajectory, but instead is perceived to rotate around the reflector in the center of the wheel as depicted in Figure 2 b (Duncker, 1929). This is because the reflector on the rim and the reflector on the center are perceived as a single perceptual group and the linear motion of the central reflector serves as the reference-frame for the reflector on the rim.

The use of multiple reference-frames and their coordination can also be observed in the motor system. The motor system uses body-centered “vector representations”, where vectors, in the form of neural activities, represent joint angles (Engle et al., 2002; Beurze et al., 2006; Pouget et al., 1976). For example, the activity of neurons can encode the wrist angle, which represents the position of the hand relative to the lower arm-segment. Similarly, the activity of the neuron(s) encoding the elbow joint-angle represents the position of the lower arm-segment with respect to the upper arm-segment. To generate a hand movement towards a visual object, our brain must collect visual sensory information and generate motor commands to drive the body and the limb to reach the target. The visual image of the target starts with retinotopic reference-frames but it is transformed to an object-based reference-frame. This non-retinotopic (object-based)

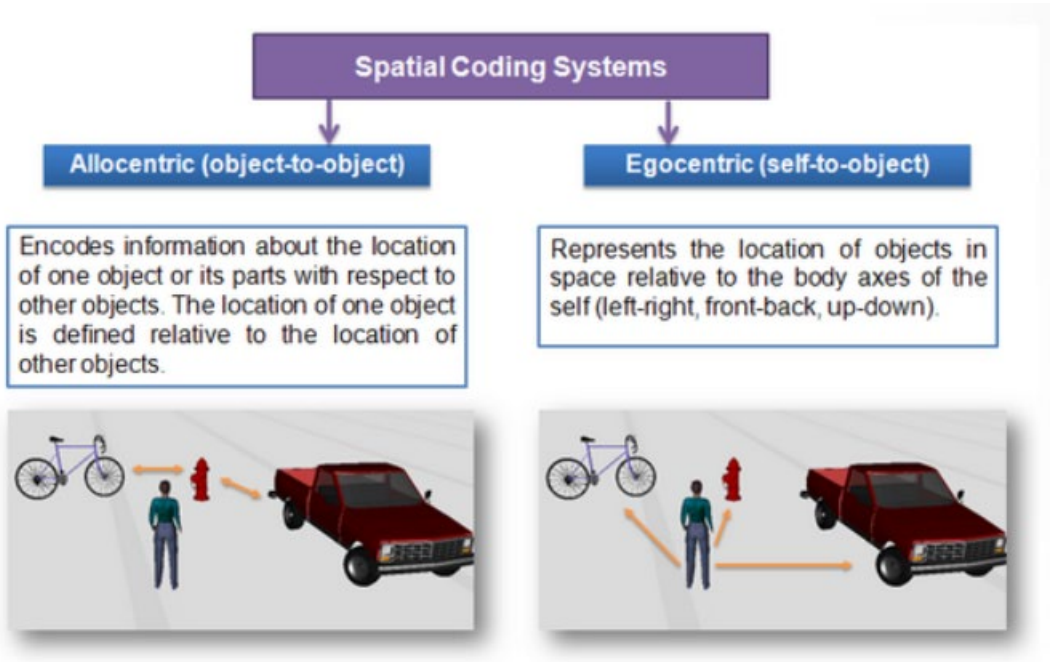

Figure 3: Egocentric and exocentric reference-frames. Figure from: https://www.nmr.mgh.harvard.edu/mkozhevnlab

representation needs to be converted to a visuomotor space to be expressed in terms of body-centered reference-frames used in the motor system. Therefore, both object-centered and body-centered reference-frames and their correspondence are needed in the motor-control process.

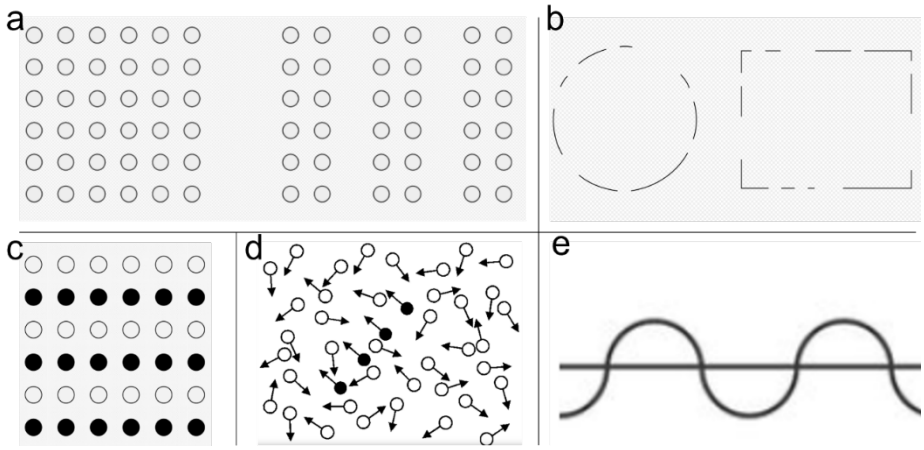

Figure 4: Gestalt grouping principles. a. Proximity. Compared to the left panel where all dots have same inter-column distance, dots on the right panel are perceived as three groups based on the large inter-column distance. b. Closure. Line segments are divided into two groups as the lines on the left formalize a circle and the lines on the right formalize a square if all gaps are filled. c. Similarity. Black dots and white dots are grouped respectively by their similarity in color. d. Common fate. Arrows indicate the velocity of dots where they are attached and all dots are in same color. Four dots are highlighted in black to illustrate that these four dots will be perceived as a group since they have the same motion direction. e. Continuation. The straight line and the curved line are separated due to the continuation within each line.

1.2 Egocentric vs. exocentric reference-frames used in the human brain

As illustrated in Figure 3, an egocentric reference-frame is defined relative to the body, like the head, the eyes, or other body parts. In contrast, an exocentric or allocentric reference-frame is defined with respect to a reference *outside* the subject. The term allocentric, instead of exocentric, is more frequently used in the literature. The prefix "allo" means "other" and is used to refer generically to reference-frames other than egocentric ones. However, the prefix "exo" is more specific and informative since it not

only puts these reference-frames in contrast with egocentric reference-frames, but also highlights the important property that the reference frame is “outside”, i.e., “external” to the organism. Hence, in the rest of this chapter, we will use the term “exocentric” reference-frame.

1.2.1 Egocentric and exocentric reference-frames in visual perception

As stated, our visual system uses both retinotopic and non-retinotopic reference-frames. The former is a type of egocentric reference-frame since its representation is relative to the retina, whereas the latter is exocentric since it is relative to the stimulus. For example, when various elements in a complex display move in an organized manner, interrelations among the elements generate a grouping effect. This grouping effect can shift our reference-frame in perceiving the display from retinotopic to a group-oriented non-retinotopic processing. Multiple grouping principles have been provided by Gestalt psychologists including proximity, similarity, closure, continuation, and common fate, as illustrated in Figure 4 (Koffka, 2013). Based on the perceptual grouping effects, Johansson (1973) proposed three principles applied to retinotopic motion-detection to synthesize reference-frames that lead to non-retinotopic motion perception. Detailed

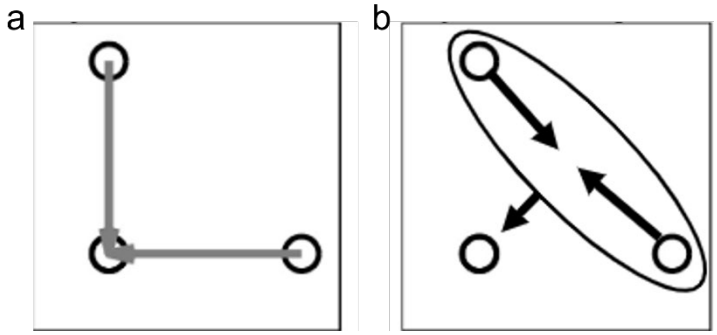

Figure 5: Common-fate direction serves as the reference frame.

information about these principles will be discussed in Section 3.1. Here, we provide a simple example to explain how perceptual grouping induces non-retinotopic motion-perception by changing reference-frames.

In Figure 5 a, two dots move towards the same end-point, one moving downward and the other moving leftward. Each dot's motion direction can be decomposed into a component towards the other dot and a second component pointing from the midpoint of the connecting line of two dots to the meeting point. The second component is shared by both dots and thus called common-fate motion direction. This common-fate direction serves as the reference-frame according to which each dot's motion is decomposed and perceived. Therefore, the two dots are perceived to move together (grouping) along the common-fate direction. At the same time, relative to the reference-frame, each dot behaves by its residual motion component, as shown in Figure 5 b. In this case, we don't perceive these dots' motion relative to our eyes (retinotopically or egocentrically). Instead, we use an exocentric reference-frame, relative to the common-motion of these dots, and perceive them in a non-retinotopic way.

1.2.2 Egocentric and exocentric reference-frames in object recognition

Due to the geometry of optical projections that form the retinotopic image, a given object may have drastically different appearances in its retinotopic representations. For example, when the distance between the object and the observer increases, the retinotopic size of the object becomes smaller. Similarly, due to different perspective views, the object's retinotopic geometry can change drastically. A bicycle wheel can appear as a circle, as an ellipse, and even as a line, depending on its relative angle with respect to the observer. The brain needs to unite these different retinotopic appearances as different

representations of the same object, a phenomenon known as “invariant object recognition”. One approach to invariant object-recognition is to transform retinotopic representations to an object-based representation. Accordingly, shapes are compared according to an exocentric reference-frame that is defined by the intrinsic properties of the object (Rock, 1973; Marr & Nishihara, 1978; Palmer, 1975). Two shapes are considered as same or different depending on whether or not they are same relative to their own reference frames. Therefore, the larger and smaller appearances of an object can be perceived as the same object since, while their reference-frames are in different scales, their properties *relative* to their reference-frames remain the same. In another example, we can clearly read the time from a rotated clock with markers indicating 12 and 9 o’clock positions. This is because we use an object-centered reference frame that is determined by the markers of 12 and 9 o’clock, relative to which we can judge the position of the hands.

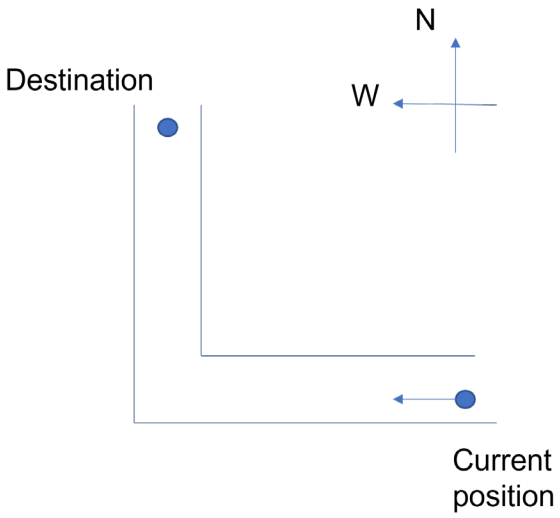

Figure 6: Egocentric and exocentric reference frames in navigation

1.2.3 Egocentric and exocentric reference-frames in motor action

Human action also benefits from reference-frames. As we mentioned, visually guided movement requires both egocentric and exocentric reference frames and their transformations. We can precisely control our arm to not only reach an arbitrary position relative to our body, but also to move around any external spatial axis. Another example of human action under reference-frame coding is navigation. Imagine when we want to navigate to a familiar destination from the current position shown in Figure 6. Based on the egocentric reference-frame, we can navigate from the current position to the destination by going straight first for a given distance and turn right afterwards. However, using an exocentric reference-frame, we can be guided by going towards west (external reference with respect to a coordinate system in the external space) first and then by using the corner as an external reference and turning north (external reference) at the corner.

1.3 Development of reference-frames and the sensorimotor schema

For the question of how humans generate egocentric and exocentric reference-frames, Piaget (1952) proposed a developmental path from egocentric reference-frames to exocentric reference-frames during the sensorimotor stage of development. According to this theory, egocentric reference-frames used in sensory and motor systems become coordinated through sensorimotor schema. The initial sensorimotor explorations are genetically encoded in reflexes, such as reaching, grabbing, and sucking. When a baby holds an object, it creates an egocentric representation in tactile maps (position of the object with respect to fingers) and an egocentric (retinotopic) representation in visual maps. The position of the arm and the hand are encoded by motor vectors that indicate

relative position of the fingers with respect to the hand, the hand with respect to the arm, etc. When the baby moves the object, all these representations change in different ways, while they are all correlated and unified in the external world as they reflect the properties of the *same* object in the environment (e.g., the position and shape of the object). It is through these sensorimotor explorations that exocentric representations that reflect invariant properties of external world emerge. For example, when the baby holds an object at a stationary position, looks at it while she is moving her eyes, according to the egocentric (retinotopic) reference-frame, the object moves, whereas according to the motor vectors (motor encoding), the object remains stationary. This conflict is resolved by transforming egocentric visual representations to exocentric ones so that the

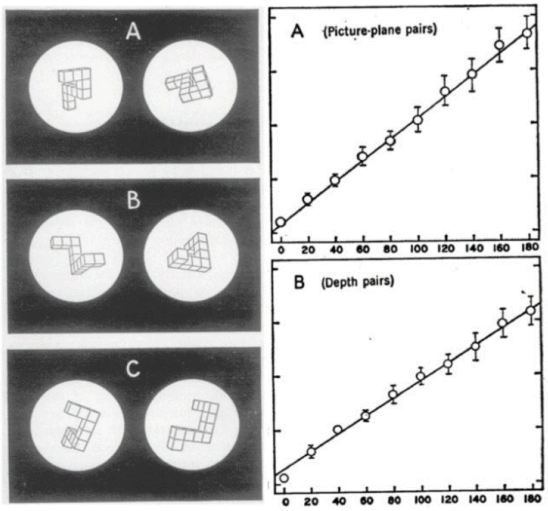

Figure 7: Mental-rotation paradigm. The left side of this figure shows three examples of the stimuli used in Shepard and Metzler's experiment (1971). In each trial, two images showing two either same or mirror-image objects in different viewpoints were presented. Subjects were asked to report if the two objects are the same or mirror-image pairs. As shown in the right side of the figure, mean reaction time across subjects was linearly related to the angular disparity of two objects' viewpoints, supporting the hypothesis of mental rotation: Subjects mentally rotated one of them to align with the other via either picture-plane or in-depth rotation to determine whether these were the same object; hence, their reaction times were a linear function of the required rotation angle.

stationarity of the object in the motor representation becomes consistent with the stationarity of the object in the visual exocentric representation.

Piaget also suggested that these sensorimotor schemas and their underlying ego- and exocentric reference-frames are internalized to constitute the building blocks of higher cognitive functions. The "internalization" refers to mental operations that are equivalent to motor actions. For example, we can rotate an object with our hand; we can also rotate an object in our mental representation without ever touching it. This is called "mental rotation". One cognitive function proposed to be built on internalized sensorimotor schema is object recognition. An object in the environment can project drastically different images on our retinae due to changes in distance and perspective. The brain needs to recognize the object as being the same object despite these drastic changes in its appearance. As mentioned before, a natural approach for this purpose is to use exocentric reference-frames that are free from the changes happening in egocentric reference-frames. For example, if a reference-frame placed on the object is used, the rotation of the object will lead to a rotated representation in retinotopic reference-frames, but no change with respect to the object-based reference frame. Hence, one theory of object recognition proposes that, when an object is stored in memory, it is stored according to an object-based reference-frame. We refer to this storage as "canonical storage". Accordingly, the recognition of a target object requires that an object-based reference-frame is deployed to the target object, and that this reference-frame along with the object is rotated mentally to align it with the canonical storage in memory. Strong empirical evidence for this theory came from the studies of Shepard and Metzler (1971). In their study, they presented subjects with pairs of objects with different orientation angles (Figure 7) and asked to

report whether the two objects are same or not. When they plotted Reaction Times (RTs) as a function of the rotation angle between the two objects, they found a linear relationship (Figure 7). This direct relationship between object-recognition time and rotation angle provides strong evidence for the canonical storage and recognition theory mentioned above. However, the exact mechanisms underlying this process are still largely unknown.

In addition, neuroimaging studies have found the correlates between mental rotation and multiple cognitive processes, including motor control, reference-frames, and memory. Osuagwu and Vuckovic (2014) compared EEG signals collected from subjects in experiments during the mental rotation task and explicit motor imagery task. They hypothesized that, if mental rotation involved implicit motor imagery, similar activities from sensorimotor cortex should be observed in both tasks. In the results of EEG data, multiple sensorimotor areas and motor areas were identified to be involved in both mental rotation and motor imagery, including the precentral gyrus, postcentral gyrus, and inferior parietal lobe. These findings have also been reported in fMRI and PET studies (Parson et al., 1995; Vingerhoets et al., 2002; a review: Zacks, 2008). In another study, subjects performed the mental rotation task with EEG applied on their scalp, and time-frequency analysis was conducted on the EEG data to address how subjects' strategies were correlated with their performance (Gardony et al., 2017). They observed a main effect of angular disparity on both the μ power from sensorimotor cortex and frontal midline θ power, suggesting the involvement of motor simulation and working memory, respectively (Francuz & Zapala, 2011; Llanos et al., 2013; Cavanagh & Frank, 2014; Hsieh & Ranganath, 2014). Moreover, a significant negative effect of angular disparity

was identified on the bilateral parietal α power, which indicated the relation to visuospatial representation and reference-frames (Zacks & Michelon, 2005).

1.4 Goals of the dissertation

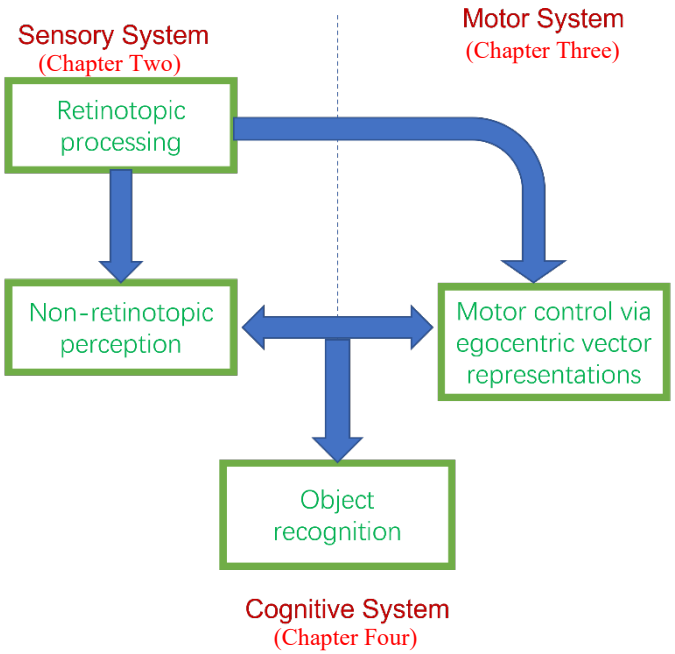

Figure 8: Gaps in the investigated field of study and aims of the present dissertation.

In this dissertation, we will focus on reference-frames associated with perception, action, and object recognition. In the first part (Chapter 2), we studied reference-frames in sensory systems, more specifically in the visual system. We developed and tested a neural model that explains how we perceive motion according to exocentric reference-frames. In the second part (Chapter 3), we developed a model for sensorimotor coordination through egocentric and exocentric reference-frames. More specifically, we proposed and tested a model for visually-guided reaching. Finally, in the third part (Chapter 4), we studied through psychophysical experiments reference-frames in cognition to test the canonical storage and recognition theory.

1.5 Significance

As discussed in the brief review above, reference-frames are crucial in the operation of the brain. Whereas there is significant research in understanding the anatomy and function of early visual retinotopic areas, much less is known about non-retinotopic representations. The retinotopic reference-frames are not sufficient to support sensory perception, object recognition, and visually-guided motor behaviors (Ogmen, 2007). Instead, our visual and motor systems efficiently benefit from non-retinotopic and exocentric processing. Therefore, a fundamental question in neuroscience and psychology has been the understanding of the link between egocentric reference-frames and exocentric processing.

Figure 8 depicts schematically the gaps that our work aims to fill in. Whereas significant knowledge exists on retinotopic processing, non-retinotopic perception, motor control, and object recognition processes (green boxes in Figure 8), how these processes communicate with each other through consistent reference-frames remains a significant gap that our work aims to fill in (blue arrows in Figure 8). This work will provide a better understanding of how non-retinotopic reference-frames are synthesized and deployed, thereby connecting sensory, motor, and cognitive processing under a common theme. Most research focuses only on one of these three essential components of brain function. Our approach aims to integrate these studies and show how the transformation and coordination of reference-frames can provide the bases for synergistic and complementary operations of these three systems in order to solve complex problems. In this dissertation, we will use computational (neural) modeling and psychophysical experiments to address the research questions outlined above.

Chapter Two: A Neural Model for Vector Decomposition and Relative-motion Perception[1]

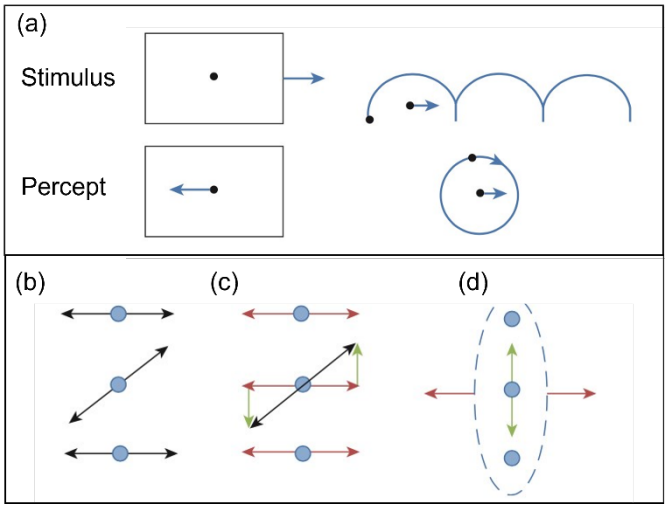

Figure 9: Examples of Stimuli a. Stimuli used by Karl Duncker (top panels) and the percepts that result from these stimuli (bottom panels). On the left, the large rectangle serves as a reference-frame and the static central dot is perceived to move in the opposite direction with respect to the direction of the rectangular frame (induced motion). On the right, two dots move as if they were placed on the rim and the center of a moving wheel. The percept does not follow the retinotopic motion shown on top right but instead the relative motion shown on bottom right. b. A stimulus that illustrates the vector decomposition approach. The top and middle dots move horizontally and the middle dot moves in an oblique direction such that its horizontal motion component is equal to the top and bottom horizontal motion. c. The motion vectors of the three dots are decomposed to produce a common horizontal motion vector (simultaneous and equal horizontal motion vectors that constitute the common motion for the group of the three dots). d. The percept consists of all three dots moving as a single Gestalt, left and right according to the common motion vector. The central dot is also perceived to move up and down relative to the group, due to its residual motion vector shown in panel b.

[1] The contents of this chapter have been published in a peer-reviewed journal: He & Ogmen (2023). A neural model for vector decomposition and relative-motion perception. *Vision Research,* 202 (2023) 108142 *https://doi.org/10.1016/j.visres.2022.108142*

1 Introduction

Neighboring points in the environment are mapped onto neighboring photoreceptors in the retina and these neighborhood relationships are preserved in early visual cortical areas through precise retinocortical projections (Engel, et al., 1997). This eye-based spatial representation is called retinotopic organization. Retinotopically organized areas in the early visual cortex contain motion-tuned neurons that compute locally the direction and speed of moving stimuli. Several computational models have been proposed to explain mechanistically how these neurons compute motion (Grzywacz & Yuille, 1990; Simoncelli & Heeger, 1998; Heeger et al., 1996; Baker & Bair, 2016; Lu & Sperling, 2001).

However, it has been long known that our perception of motion does not correspond to retinotopically computed motion of the stimuli. For example, the Gestalt psychologist Karl Duncker (Duncker, 1929) used the stimuli shown in the Figure 9 a to highlight the importance of non-retinotopic reference-frames used in perceiving motion. The stimulus on the left is a dot surrounded by a rectangular frame. When he kept the dot stationary and moved the frame (Figure 9 a, top left panel), the percept (Figure 9 a, bottom left panel) was that of a stationary frame and a dot moving in the opposite direction of the frame's physical motion. Using this "induced motion" paradigm, he proposed that the frame serves as a reference according to which the motion of the dot is perceived. Thus, the motion of the dot is not perceived according to its retinotopic motion but instead it is perceived relative to the reference frame established by the larger rectangle stimulus. The top right panel of Figure 9 a shows another example studied by Duncker. Here one dot moved horizontally whereas a second dot underwent cycloid motion. This type of motion

would happen when, for example, two bright spots are placed on a moving wheel, one on the rim and a second one on the center of the wheel. According to a retinotopic reference frame, we should perceive the dot on the rim to undergo a cycloid motion; however, we perceive this dot to rotate around the central dot (Figure 9 a, bottom right panel). In other words, the two dots are not analyzed individually but as parts of a single Gestalt. The dot at the center of the wheel serves as a reference according to which the dot on the rim is perceived.

A strong impetus for this line of research came from Johansson's studies (Johansson, 1973; Johansson, 1976). Johansson placed bright spots on an otherwise invisible person and recorded the movements of these spots when the person executed various actions, such as walking. The analysis of the retinotopic trajectories of the dots revealed complex patterns that seemed difficult to interpret. However, instead of these complex retinotopic motion patterns, human observers readily perceived a coherent set of motions reflecting the movement of the walking individual. As in previous examples, the dots were perceived as parts of a Gestalt. The global motion of the walking individual serves as a reference- frame and the various dots are perceived relative to this reference frame. For example, a dot placed on the hand is perceived both moving with the body but also swinging back and forth with respect to the body.

Johansson proposed a theory based on three principles to explain these effects ((Johansson, 1973), p.205): First, he stated that stimulus elements that are in motion are always perceptually related to each other. Second, equal and simultaneous motions of proximal elements perceptually connect these elements into Gestalts composed of rigid perceptual units. Finally, when the motion vectors of proximal elements can be

decomposed to produce equal and simultaneous motion components, these components will be perceptually united into the "one unitary motion". This unitary motion is called the common motion of the grouped elements. The stimulus shown in Figure 9 b provides a simple demonstration of Johansson's vector analysis theory. The stimulus consists of three dots. The upper and the lower dots move horizontally back and forth while the dot in the middle moves in an oblique direction. According to Johansson's theory, the motion of these dots are perceptually related to each other (Principle 1), the three dots are part of a rigid whole (Principle 2), and the equal and simultaneous components (horizontal motion vectors shown in red in Figure 9 c) form one unitary motion, i.e., the "common motion vector" of the group (Principle 3). Accordingly, all three dots are perceived to move as a group left and right according to the common motion vector and, in addition, the middle dot appears to move up and down relative to the group (Figure 9 d). While the vector analysis theory offers a simple explanation for several relative motion percepts, how it may be implemented in the nervous system remains an open question. Motion detectors in early visual areas are organized retinotopically, i.e. they compute motion according to an egocentric (eye-based) reference frame. In this manuscript, we propose and test a neural model in order to explain how egocentric (retinotopic) motion signals are transformed to give rise percepts based on exocentric reference-frames (based outside the observer).

The general structure of this manuscript is as follows: In the introduction, we reviewed classical data showing the necessity of reference-frames in motion perception. In Section 2, we introduce step-by-step a model designed to extract reference-frames based on the Gestalt common-fate principle. This model uses the reference-frame to

decompose motion signals and computes the relative motion of parts of a Gestalt. We then compare the predictions of the model to classical data, viz., the "three-dot", the "rotating wheel", and the more complex the "point walker" paradigms. While simulating the three-dot paradigm, we derive a prediction of the model that differs from the prediction of Johansson's classical vector-analysis theory. We introduce two psychophysical experiments to test this prediction and compare the results with the predictions of our model. We conclude our paper by discussing the explanatory power of the model as well as its shortcomings in comparison to other models and data. For example, we discuss the fact that our model does not include form factors and how it can be extended to address this shortcoming. Finally, we conclude that, whereas form factors can play a role in relative-motion perception, a model without any form factor can go a long way explaining both classical and new data on motion perception.

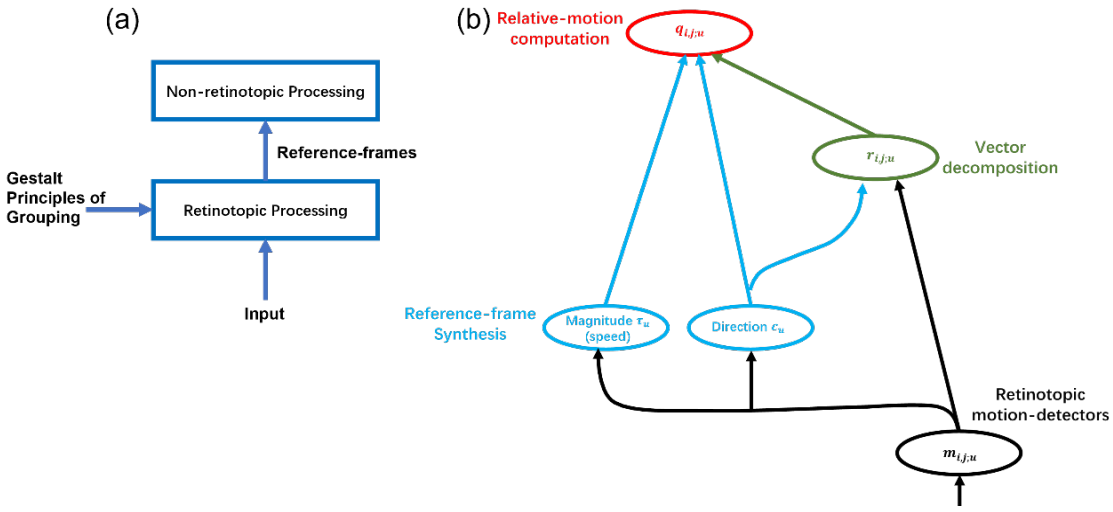

Figure 10: Schematic Description of the Model. a. Block diagram representation of these operations. b. The corresponding neural architecture of the network proposed in this study. Panel a reproduced from (Clarke et al., 2016).

2 Description of the Model

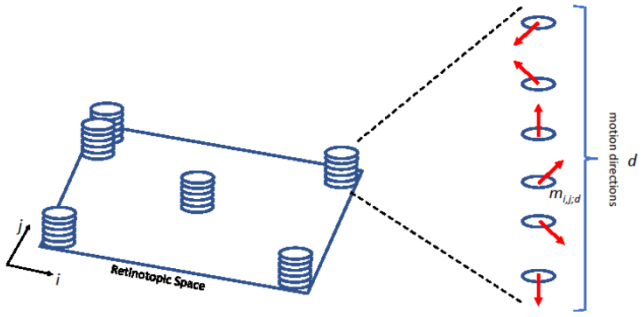

Figure 11: Retinotopic directionally-selective motion detector layer. The first layer of the proposed neural network. It is organized retinotopically and each retinotopic location contains a set of directionally-tuned motion detectors sampling all motion directions at that retinotopic location.

Based on a series of experiments studying reference-frames (Clarke et al., 2016; Agaoglu et al., 2016; Ogmen & Herzog, 2010), we proposed the schematic "two stage" model shown in Figure 10 a, which is a two-stage instantiation of Johansson's approach. The first stage extracts local retinotopic motion information; Gestalt grouping principles (e.g., common fate) are used to establish groups (the first two principles). A common motion is extracted for each group (third principle) and this common motion serves as a reference frame to compute relative motions of elements belonging to the same group. Figure 10 b, in turn shows the neural architecture we used in this study to implement the ideas shown in Figure 10 a into a neural model. The bottom layer consists of retinotopically organized directionally-tuned motion detectors. The Gestalt common-fate principle is applied to the outputs of these neurons to synthesize the direction and

magnitude (speed) of the reference-frame by two layers of networks shown on the left. The direction of the reference-frame is combined with the outputs of the retinotopic motion detectors to decompose these motion vectors along the reference-frame direction and its perpendicular direction. The motion-vectors that are expressed in terms of the coordinates of the reference-frame are then combined with the direction and magnitude of *relative* motion at the top layer in Figure 10 b. We will now explain one by one the architecture, connectivity, and the function of these layers.

2.1 Retinotopic directionally-selective motion detectors.

The first stage of the model consists of retinotopic directionally-selective motion detectors (Figure 10 b, bottom layer). Each retinotopic location contains multiple motion detectors (Figure 11), each tuned to a different direction but as a group covering 360 deg of motion directions (for simplicity, we consider motion in two-dimensional space).

Let $m_{i,j;u}(t)$ represent the activities of these directionally selective neurons. The indices i, j represent the two-dimensional coordinates of the location of this neuron in the retinotopic space whereas the index u denotes the motion direction to which the neuron is tuned to (Figure 11). For simplicity, the receptive fields in the simulation were partially-overlapping squares on the retina (more details will be given in each simulation). In the following context, we use the index u to denote the directionally-selective neuron tuned along the direction of the unit vector $\overrightarrow{d_u}$. These unit vectors are sampled with an angular interval of $\Delta\theta$ with coordinates:

$$\overrightarrow{d_u} = \big(cos(u * \Delta\theta), sin(u * \Delta\theta)\big), u = 0,1,2,3, \dots until\ u * \Delta\theta \geq 2\pi \quad (1)$$

Several models have been developed to describe the activities of retinotopic directionally-selective motion detectors (Grzywacz & Yuille, 1990; Simoncelli & Heeger, 1998; Heeger et al., 1996; Baker & Bair, 2016; Lu & Sperling, 2001). For simplicity, in this work we will not implement detailed motion detector models since our goal is to build a model of how non-retinotopic motion is processed based on the outputs of retinotopic motion-detectors and how non-retinotopic reference-frames and the vector decomposition principle can be applied to explain perceptual data on relative motion perception. A directionally-tuned motion detector responds maximally to motion along its tuned direction. However, it also responds to motion directions that are close to its tuned direction. This is expressed as a tuning curve, typically following a Gaussian distribution. Since the direction vectors are circular, a von Mises distribution is used. Hence, we describe the activities of retinotopic motion detectors as follows:

$$m_{i,j;u}(t) = \frac{|\overrightarrow{v_{i,j}}(t)| \cdot \exp\left(\kappa_1 \cdot cos\left(\theta_{\overrightarrow{v_{i,j}}(t),\overrightarrow{d_u}}\right)\right)}{2\pi I_0(\kappa_1)} \quad (2)$$

where t is time, I_0 is the modified Bessel function of order 0, and the expression

$$\theta_{\overrightarrow{v_{i,j}}(t),\overrightarrow{d_u}} = \arccos\left(\frac{\overrightarrow{v_{i,j}}(t) \cdot \overrightarrow{d_u}}{|\overrightarrow{v_{i,j}}(t)| \cdot |\overrightarrow{d_u}|}\right) \quad (3)$$

provides the angular difference between the stimulus velocity vector at (i, j) at time t , $v_{i,j}$ (t), and the unit vector representing the preferred motion-direction $\overrightarrow{d_u}$ of the neuron. The resulting tuning curve is plotted in Figure 14[2].

[2] Motion detectors can have tuning curves of different widths and also may contain antagonistic surrounds. Here, for simplicity, we used a single width and no antagonistic surround. For more complex stimuli and simulations, this layer can be modified to include multiple populations with different tuning curve widths and antagonistic regions.

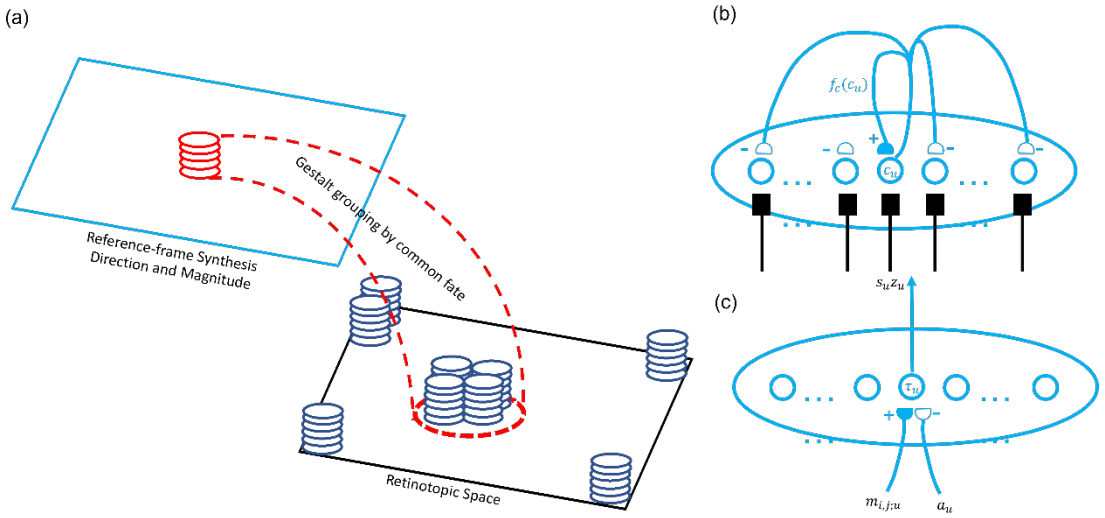

Figure 12: Reference-frame synthesis. a. The reference-frame synthesis layer implements the Gestalt common-fate principle. Neurons in this layer receive inputs from the motion-detector layer according to its receptive field that determines the spatial extent over which activities are pooled to compute common motion activities (common-fate principle). b. Reference-frame motion-direction computation. The network consists of cells tuned to different directions, u. Each neuron receives habituating input (depicted by rectangular connections) that represents the sum of motion energy along its direction. Each neuron excites itself (filled synaptic symbol) and inhibits all other cells in the network (open synaptic symbols) via recurrent (feedback) connections. To avoid clutter, only connections from one cell is shown. All other cells have identical connection patterns. With a faster-than-linear feedback function $f_c(c_u)$ and habituating inputs, it reaches a 0-1 distribution, i.e., the neuron with the highest input reaches the maximum activity whereas all other cells' activities tend to zero. c. Reference-frame motion-magnitude (speed) computation. The network consists of cells tuned to different directions, u. Each neuron receives excitatory inputs $m_{i,j;u}$. And an inhibitory input a_u. To avoid clutter, only one excitatory input is shown. The cell receives excitatory inputs from all $m_{i,j;u}$ s that fall in its receptive field defined on the retinotopic space (indices i, j).

In summary, these equations reflect the activities of direction-tuned neurons according to their tuning-curve expressed as a von Mises distribution centered at the preferred-direction of the neuron. The tuning curve we used is similar to those of retinotopic motion detectors observed in the directional tuning curve of MT neurons recorded from macaque monkeys (Albright, 1984).

Next, we discuss how the reference-frame is synthesized based on the activities of these retinotopic motion-detectors.

2.2 Gestalt grouping, common-fate computation, and the establishment of a non-retinotopic reference-frame.

Gestalt psychologist describes a collection of low-level cues of perceptual grouping (Wertheimer, 1923; Koffka, 1935). For instance, the common-fate principle states that objects with common motion, in terms of speed and/or direction, tend to be grouped together in perception. However, these low-level cues alone are not enough to explain all the grouping and segmentation effects on a daily basis of human perception. Our visual system also relies on high-level grouping cues. For example, we understand all the visual elements of a body are from a human when we see a picture with human portraits. This can be accounted by our familiarity with human body, which has important effect on grouping and segmentation (e.g., Ullman, 1996). Therefore, the neural mechanisms of perceptual grouping have been thought to be complex since low- and high-level cues interact stimulated by the complex interactions among visual elements. One theory that explains the cortical mechanism is "incremental grouping theory" (Roelfsema et al., 2000; Roelfsema, 2006). According to this view, our visual system uses both base grouping and incremental grouping approaches, involving neurons with local receptive-fields that process low-level features, and neurons sensitive to the context and recurrently process information across different receptive-fields, respectively. Recently, some neural models based on similar hierarchical approaches have been proposed to illustrate how low- and high-level processing can be embedded in neural plausible ways (Roelfsema, 2006; Grossberg et al., 1997; Roelfsema et al., 2000; Ross et al., 2000). However, in our

model, we used low-level cues and didn't include the context-based neural interactions. As a matter of fact, herein we applied the common-fate principle of grouping only and explained this principle using neural networks.

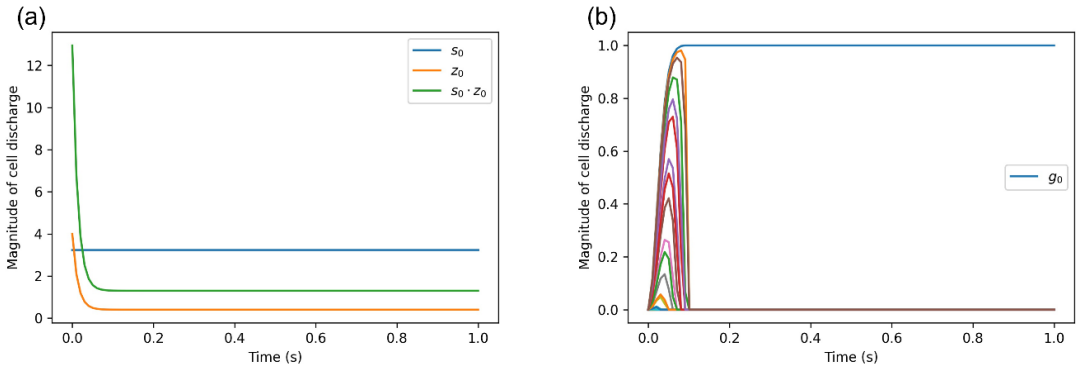

Figure 13: Example of inputs to the reference-frame direction cells. a. This figure shows results from the three-dot stimulus moving with a constant velocity along $\overrightarrow{d_0}$. Each cell s_u sums the retinotopic motion-activities along its preferred direction u. In this example, we plot s_0 which shows a constant horizontal rightward motion activity. In order to make this signal weaker to allow the winner-take-all network to determine the direction of maximum activity (see Section 6), the signal is transmitted by a habituation process z. The net input signal to the winner-take-all network, $s_0 \cdot z_0$, exhibits an initial overshoot to initialize the winner-take-all network, followed by a decay to allow the selection of the winner by feedback connections. b. Outputs $g_u = f_s(c_u)$ of the winner take-all-network, with different colors corresponding to different directions. After a brief transient competition, the horizontal direction g_0 wins the competition (activity = 1) whereas all other directions loose (activities = 0).

As shown schematically in Figure 10 a, Gestalt grouping principles are applied to retinotopic motion-detector activities in order to synthesize reference-frames that transform retinotopic representations into non-retinotopic ones. According to the Gestalt common-fate principle, stimuli with similar motion characteristics are grouped together. We used a simple implementation of the common-fate principle as follows: A group of neurons collects direction and speed information from a region of the retinotopic space in the first layer as shown in Figure 12 a. This region corresponds to the receptive field of

these neurons in the computation of the common-motion of the stimuli that fall in that region. The common-motion vector of the group serves as the non-retinotopic reference-frame for that group, which has two components, direction (motion direction) and magnitude (speed). First, we describe how the direction of the common-motion vector is computed.

2.2.1. Direction of the reference-frame (direction of the common motion)

For simplicity, we used a receptive field that covered the entire retina since our stimuli consisted of a single Gestalt group. The idea in computing the common-fate motion direction is to find the motion direction along which motion activity (or loosely speaking, "motion energy") is highest compared to other directions. To do this, we first compute motion activities along each and every direction $\overrightarrow{d_u}$ and then select the direction with maximum activity. Thus, at the first step, motion summation cells denoted by s_u sum the activities of motion detectors tuned to all directions with weights depending on the angular difference between motion detector's tuned direction and its own. This is analogous to the motion energy along each direction:

$$s_u(t) = \sum_{i,j} \frac{f_s\left(\left|\overrightarrow{v_{i,j}}(t)\right|\right)\cdot \exp\left(\kappa_2 \cdot cos\left(\theta_{\overrightarrow{v_{i,j}}(t),\overrightarrow{d_u}}\right)\right)}{2\pi I_0(\kappa_2)} \quad (4)$$

$$f_s(x) = \frac{2}{1+exp(-\beta x)} - 1 \quad (5)$$

After that step, a group of reference-frame direction neurons, c_u, dynamically compute the direction along which the motion energy is maximal. A recurrent winner-take-all neural network (Grossberg, 1982; Ogmen, 1993) is used to determine the maximum. In the recurrent winner-take-all network (Figure 12 b), each neuron excites

itself and inhibits all other neurons. It can be proven mathematically that if the feedback function is faster-than-linear, the network activities approach asymptotically a 0-1 distribution, i.e., the neuron with the largest input reaches the maximum activity whereas all other neurons' activities are suppressed. The reader is referred to Section 6 for the details of this network. The activity c_u is governed by the shunting differential equation (see Section 6) expressed as:

$$\frac{dc_u}{dt} = -A \cdot c_u + (B - c_u) \cdot [f_c(c_u) + s_u \cdot z_u] - (C + c_u) \cdot \sum_{u' \neq u} f_c(c_{u'}) \quad (6)$$

where parameters A, B, and C are positive constants (see Section 6 for their physiological interpretation and Table 1 for their values). The term $[f_c(c_u) + s_u \cdot z_u]$ is the excitatory input to the neuron and contains the external input $s_u \cdot z_u$ as well as the self-feedback $f_c(c_u)$. For winner-take-all network to reach 0-1 distribution, the feedback function needs to be faster-than-linear. Here we chose

$$f_c(x) = \alpha x^2 \quad (7)$$

, where α is a positive constant. The variable z_u represents a habituation process and follows the differential equation (see Section 6 for details):

$$\frac{dz_u}{dt} = D \cdot (E - z_u) - F \cdot s_u \cdot z_u \quad (8)$$

where D, E, and F are positive constants (Section 6 and Table 1).

The activity c_u undergoes a nonlinearity function f_s, defined in Equation (5) above to produce the output:

$$g_u = f_s(c_u) \quad (9)$$

Figure 13 a illustrates the operation of this stage. It shows one of the inputs, s_0, which is constant in time (blue trace). The habituating variable z_0 transforms this signal into one

with a rapid overshoot and decay to a lower plateau (green trace; see Section 6). In this simulation, s_0 corresponds to the direction with maximum activity. Figure 13 b shows the outputs g_u for all directions, different colors corresponding to different directions. As one can see from this plot, after a transient competition that lasts about 100ms, g_0 wins the competition and the network converges to a 0-1 distribution, with g_0 =1, and g_u=0, with $u \neq 0$.

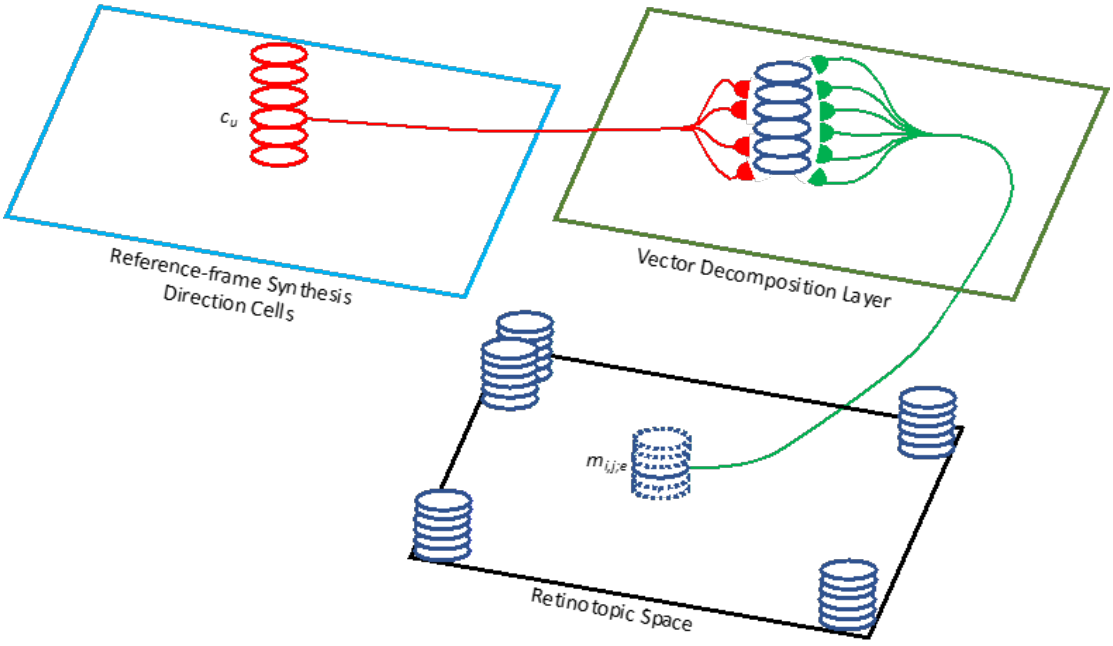

Figure 14: Synaptic projections from reference-frame direction cells to the vector decomposition layer. Each motion detector projects to a vector-decomposition cell located at the same retinotopic position (green connections) with weights that are proportional to the cosine of the angle between its own preferred direction and the preferred direction of the vector-decomposition cell. Each reference-frame direction cell sends inhibitory connections to all vector decomposition cells (red connections) except to those tuned to the same and perpendicular directions relative to its own preferred direction.

2.2.2. Magnitude (speed of the common motion) of the reference frame

The magnitude of the reference-frame vector (i.e., the speed of the Gestalt group) is computed in two steps. First, a running average of motion energies for each direction is computed with the additive equation (see Section 6 for neural correlates of this equation):

$$\frac{da_u}{dt} = -G \cdot a_u + s_u \tag{10}$$

where a_u represents the running average and G is a positive constant. The speed along the direction u, τ_u, (Figure 12 c) obeys the shunting equation:

$$\frac{d\tau_u}{dt} = -H \cdot \tau_u + (I - \tau_u) \cdot \sum_{i,j} m_{i,j;u} - G' \cdot \tau_u \cdot a_u \tag{11}$$

where G, H, I, and G' are positive constants (Table 1).

In our model, the reference-frame synthesis layer contains neurons summing the neural activities across multiple receptive fields. This spatial summation has been found in MT as well as in other visual cortical neurons (Britten & Heuer, 1999; Ghose & Maunsell, 2008; Kay et al., 2013; Oleksiak et al., 2011; Kumano & Uka, 2012). Moreover, the neural computations following the spatial summation in our model can be compared to physiological findings as follows: In our model, information is derived from the spatial summation via two mechanisms, winner-take-all and normalization. Both of these mechanisms are found in the physiological and psychological studies. First, the spatial summation in the parietal area of macaque has been found to follow the winner-take-all rule (Oleksiak et al., 2011). A similar effect was also reported in perceptual studies (Zohary et al., 1996; Salzman & Newsome, 1994). This is consistent with our approach in determining the common motion direction. Second, normalization was found to occur in MT cells of monkeys during the spatial-summation process (Britten & Heuer,

1999). These observations support the plausibility of our approach in calculating the common-motion speed, where we normalized the summed outputs from motion detectors across receptive fields to approximate the average speed.

2.3 Vector decomposition.

The reference-frame established in the previous section is used for transforming retinotopic motion activities into non-retinotopic ones. Hence, motion vectors computed in the first stage according to the retinotopic reference-frame are now expressed according to the Gestalt-based reference frame. This in turn involves two steps: In the first step, the retinotopic motion vector is expressed according to the new reference-frame by computing its projections with respect to this reference-frame (see red and green vectors in Figure 9 c). For this step, we use the direction vectors of the reference frame. This step can be called "vector decomposition", since the retinotopic motion vector is decomposed according to the reference-frame vectors by geometric projection (Figure 9 c). After this vector decomposition stage, *relative* motion is computed by comparing the magnitudes of the decomposed vectors to that of the reference frame. In this section, we discuss the vector decomposition step. As depicted in Figure 9 c, the motion vectors are decomposed by projecting them along the axis of the reference-motion to produce the effective common motion for the elements of the group. The component perpendicular to the axis of the reference-motion in turn represents the relative motion of an element with respect to the other elements that belong to the same group. This is achieved by a combination of two sets of projections to the next layer of the neural network as shown in Figure 14.

This layer also contains directionally-tuned neurons at each retinotopic location. Let $r_{i,j;u}$ represent the activities of these neurons. Each neuron $m_{i,j;u'}$ projects to a neuron $r_{i,j;u}$ by an excitatory synaptic weight $w_{u',u}$, which is proportional to the cosine of their motion-direction angle difference, i.e.,

$$w_{u',u} = cos(\overrightarrow{d_{u'}}, \overrightarrow{d_u}) \tag{12}$$

Since the effective input to the neuron $r_{i,j;u}$ from the neuron $m_{i,j;u'}$ is given by the product of the output (activity) and the synaptic weight, this input represents the projection of the motion vector in the retinotopic space to the axis represented by the preferred direction of the neuron$r_{i,j;u}$. Since $m_{i,j;u'}$projects to all neurons $r_{i,j;u}$, one obtains projections of the retinotopic motion to *all possible* common motion directions via these synaptic connections. However, vector decomposition requires only two projections: The first is along the direction of common motion and the second is perpendicular to the common motion. This constraint is implemented by selective synaptic connectivity from the direction layer of the reference-frame to neurons in the vector decomposition layer as shown in Figure 14.

The goal of these projections is to inhibit all motion directions except two: The same and the perpendicular directions to the common motion direction.

$$\frac{dr_{i,j;u}(t)}{dt} = -J \cdot r_{i,j;u}(t) + \left(K - r_{i,j;u}(t)\right) \sum_{u'} w_{u',u} \cdot m_{i,j;u'}(t) - L \cdot$$

$$r_{i,j;u}(t) \sum_{u'} \delta_{u',u} \cdot c_{u'}(t) \tag{13}$$

where the inhibitory synaptic weight $\delta_{u',u}$ is defined by

$$\delta_{u',u} = \begin{cases} 0, & if\ \overrightarrow{d_{u'}}, \overrightarrow{d_u}\ are\ parallel\ or\ perpendicular\ to\ each\ other \\ 1, & otherwise \end{cases} \tag{14}$$

Given the discrete sampling of the tuning directions, the directions may not be exactly parallel and perpendicular to another. Hence, we consider two directions $\overrightarrow{d_{u'}}, \overrightarrow{d_u}$ parallel or perpendicular if $\theta_{\overrightarrow{d_{u'}},\overrightarrow{d_u}} \pm \gamma = 0, \pi/2, or\ \pi$, where $\pm\gamma$ is a small positive constant (Table 1) representing the tolerance to compensate for the sampling of the directions.

After these computations, the vector decomposition neurons at each retinal location represent the motion direction components that are either parallel or perpendicular to the reference frame (common-fate direction).

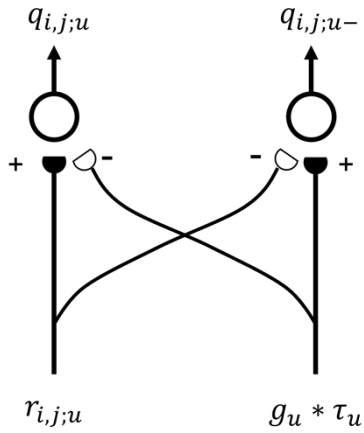

Figure 15: Motion opponency. Opponent connectivity leading to Equations 15 and 16.

Our model uses a population of vector-decomposition neurons that receive excitatory synapses based on cosinusoidal rules and inhibitory synapses from cells tuned to parallel and orthogonal directions. This cosinusodial neural relationship has been also theoretically proposed to explain the "aperture problem", known as the intersection of constraints (IOC) (Bradley, 2001; Chey et al., 1997; Nakayama & Silverman, 1988). The IOC theory states that all local velocity samples have a cosinusoidal relationship in their perceived direction with the object's true direction. On the other hand, it has also been suggested that neurons in monkey's MST have tuning properties along both the preferred

direction and its vertical direction (Duijnhouwer et al., 2013). The existence of these types of neurons suggest that the vector decomposition layer of the present model is neurally plausible.

2.4 Computation of relative motion

The direction and the magnitude of the reference frame indicate the direction and the speed, respectively, of the group. Each element in the group, while moving along with the group, can also exhibit *relative* motion with respect to the group. As an example, the central dot in Figure 9 d appears to move left-and-right with the group (red arrows) as well as up-and-down relative to the group (green arrows). Similarly, the arm of a walking person moves along with the body of the person; however, it can also have a *relative* back-and-forth motion pendulum type motion with respect to the body. Thus, to calculate the relative motion of an element with respect to its group, its residual motion is computed by subtracting the group motion. Moreover, if the residual motion along one direction is negative, this residual motion should be projected to a motion detector with opposite motion tuning, since a hyperpolarization of a rightward motion detector does not signal itself leftward motion; but instead, the depolarization of the motion detector tuned to the leftward motion does. This is typically achieved by an opponent arrangement of motion detectors. Here we implemented a simple opponency network through opposite polarity inputs as shown in Figure 15. The neuron on the left is excited by $r_{i,j;u}(t)$ whereas the neuron on the right is inhibited by the same input. The reference frame signal, represented by the product of the direction and magnitude inhibits the cell on the left while it is exciting the one on the right. As a result, a depolarization in one cell will correspond to a hyperpolarization on the other and vice-versa. The outputs of these cells

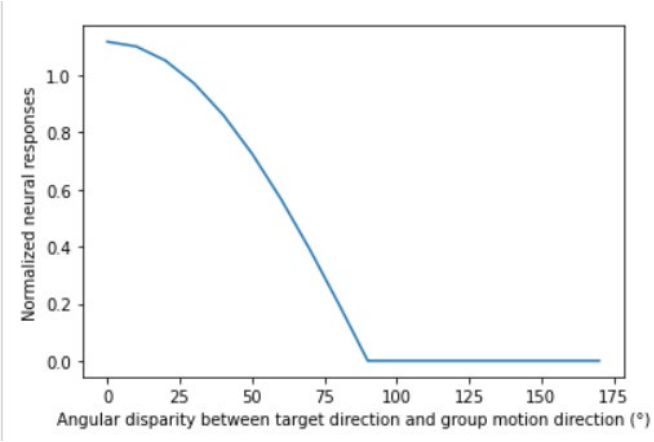

Figure 16: Relative motion selection in vector decomposition layer.

are thresholded. Let $q_{i,j;u}(t)$ represent the activity of the cell computing the relative motion. As can be seen from Figure 7, this cell receives excitatory and inhibitory inputs from $r_{i,j;u}(t)$ and $g_u(t)\tau_u(t)$, respectively. Hence the cell computes the difference between its velocity and the velocity of the reference-frame. This can be expressed by a simple additive equation:

$$\frac{dq_{i,j;u}(t)}{dt} = -Oq_{i,j;u} + r_{i,j;u} - g_u\tau_u \tag{15}$$

where O is a positive constant. Similarly, the opponent direction can be expressed by:

$$\frac{dq_{i,j;u^-}(t)}{dt} = -Oq_{i,j;u^-} + g_u\tau_u - r_{i,j;u} \tag{16}$$

The activities of these cells are passed to the nonlinear function f_q to produce their output. For simplicity, we approximated these additive differential equations at steady-state, i.e.:

$$q_{i,j;u}(t) = f_q\left(\frac{r_{i,j;u}(t) - g_u(t)\tau_u(t)}{O}\right) \tag{17}$$

and

$$q_{i,j;u^-}(t) = f_q\left(\frac{-r_{i,j;u}+g_u(t)\tau_u(t)}{O}\right) \quad (18)$$

with

$$f_q(x) = \frac{x}{1+exp(-\epsilon x)} \quad (19)$$

Note that only one of these can be active at a time, i.e., the instantaneous relative motion can only be in one of the two opponent directions.

The last layer of our model computes the relative motion using a motion-opponent structure. Opponency is a wide-spread mechanism used in the visual system, starting from retinal neurons and continuing throughout the cortex. It was shown that MT neurons responses to their preferred direction were strongly suppressed by local stimulus moving in the opposite direction (Heeger, 1987; Heeger et al., 1999). Similar opponency effects were also found in V1 neurons (Geisler et al., 2001). In addition, our model also exhibits selectivity to relative motion-direction reported in MT neurons by multiple studies (Davidson & Bender, 1991; Allman et al., 1985; Joly and Bender, 1997). Davidson & Bender (1991) reported a tuning curve that reflects selective suppression. In this study, a monkey observed a relative motion paradigm composed of a background stimulus, produced from an array of light dots with random positions, and a target square.

During the display of the stimuli, the background stimulus and the target moved in the same speed along either the same or different directions, while neural activities from the monkey's superior colliculus were recorded. It was found that the normalized magnitude of cells was smaller with decreased angular disparity between the target motion direction and the background motion direction, and the maximum suppression was observed when the target moved along the same direction as the background. A

similar relative-motion selectivity can be found in the activities of the vector decomposition layer in our model, as shown in Figure 16.

Table 1: Model parameters

Parameter	Description	Value
$\Delta\theta$	Angular interval of motion-direction sampling	10°
κ_1	Concentration of tuning curve	3
κ_2	Concentration of tuning curve	7
β	Saturation rate of the sigmoid nonlinearity	2
A	Passive decay rate	4
B	Nernst potential for depolarization	25
C	Nernst potential for hyperpolarization	2
α	2nd order polynomial coefficient	1
D	Replenishment rate	10
E	Maximum level of transducing agent	3
F	Depletion rate	20
G	Decay rate	20
H	Decay rate	30
I	Nernst potential for depolarization	50
G'	Depletion rate	490
J	Passive decay rate	150
K	Nernst potential for depolarization	40
L	Nernst potential for hyperpolarization	800
γ	Tolerance of compensation	2°
O	Passive decay rate	1
	Saturation rate of sigmoid nonlinearity	1.8

3 Simulations

We simulated the model with multiple experimental paradigms using dot motion as the stimuli, including the "Three-Dot" paradigm shown in Figure 9 b, the "Wheel-Rotation" paradigm shown in Figure 9 a, and the "Point-Walker" Display shown in Figure 20. In this section, we introduce the simulation procedures, report the results, and evaluate the performance of our model using the three paradigms mentioned above.

In each case, the stimulus was described by mathematical equations providing the coordinates of each dot with respect to time. Moreover, all the stimuli were defined on a

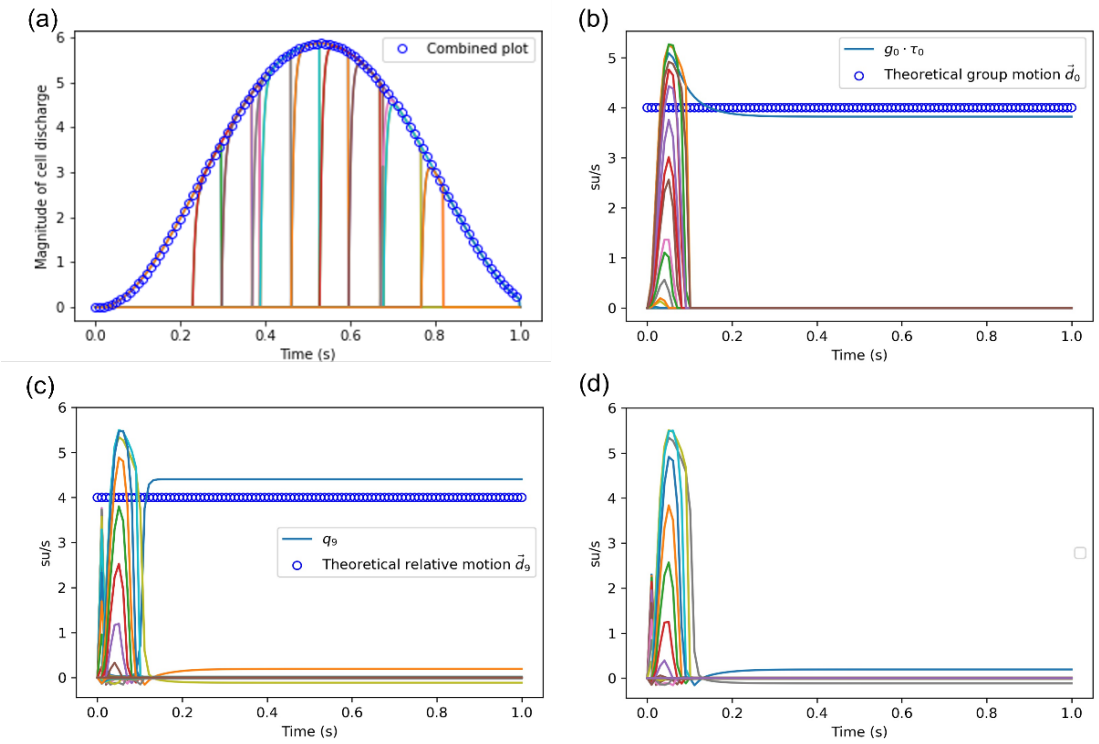

Figure 17: Simulation Results for the Three Paradigm. a. Example of combined plot of cell activities. The results are from the simulation 3.2. Solid lines with different colors are cell activities stimulated by a dot from motion detectors in different positions. Unfilled markers show the combined activities. b. Illustration of the reference-frame layer's output providing both the direction and magnitude of the reference-frame ($g_d \cdot \tau_d$) in the Three-dot paradigm with constant velocity. After an initial transient period of competition, the horizontal direction wins the competition and produces an output proportional to the speed of the reference-frame (speed of the common motion). We calibrated the activities so that they reflect speeds in su/s. The estimated speed of the common motion is close to the theoretical value of 4 su/s. Relative-motion activities in the Three-dot paradigm with constant velocity for the middle (c) and the flanking dot (d). Relative to the reference frame, the middle dot was perceived to move along the upward direction with a steady speed of about 4 su/s. This coincides the theoretical value. The flanking dot had no relative motion with respect to the reference-frame since the flanking dots served as the reference frame.

two-dimensional space, and in this article, we will use i and j axes to represent the horizontal and vertical coordinates, respectively.

The parameters of the model used in the simulation are shown in Table 1. Importantly, the $\triangle\ \theta$ we used was $10°$, so there were total 36 directions used in the simulations. The directions $\overrightarrow{d_0}, \overrightarrow{d_9}, \overrightarrow{d_{18}}$, and $\overrightarrow{d_{27}}$ indicate rightward, upward, leftward, and downward respectively. Also, the system of ODEs used to describe the neural network was numerically solved using the programing language Python (version 3.7.4). We applied the LSODA, a classic solver for stiff or non-stiff systems of first-order ODEs, from SCIPY (https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.odeint.html).

To evaluate the simulation performance, we compared the simulated neural responses to the theoretical neural response of each $r_{i,j;u}$ and g_u. The theoretical common motion for each paradigm is determined by the velocity of one dot in the stimulus, which is the upper dot in the Three-Dot Paradigm, the dot on the center in the Rotating-Wheel Paradigm, and the average motion in the Point-Walker Paradigm.

It is important to mention that we plot the activities of all the cells of the same type together and present these results in one combined plot. This makes it easier to visualize and interpret the results. An example can be found in Figure 17 a. In this paper, all the cell-activity profiles with respect to time will be shown in this manner. Therefore, regardless of i, j, cell activities will be compared according to their direction tuning. This means, in all the graphs, we only show the direction index. For example, when we show r_0, it means the activity of vector-decomposition cell tuned to $\overrightarrow{d_0}$. For each experimental paradigm, we will show the contribution of different layers in the model by plotting the activities of reference-frame direction neurons, g_u, reference-frame speed, $g_u.\ \tau_u$, vector-

decomposition activities, r_u, and finally relative-motion activities q_u. Also, we will use X and Y to denote the spatial position along the horizontal and vertical axes of the receptive field.

Video demos of all the following simulations can be found in https://github.com/hedch/Vector-Decomposition-Model.git.

3.1 The Three-Dot Paradigm

We did two simulations based on this paradigm. First, we ran the traditional stimulus where three dots move by a constant velocity (Figure 9) like the stimulus used in several previous studies (Gershman et al., 2016; Grossberg et al., 2011; Gershman et al., 2013). Then, we used a version where velocities were not constant. We first describe here the results for the traditional constant velocity case. Following that, we will explain the rationale of the modified version.

3.1.1 Constant Velocity

Methods.

The stimulus was similar to the one shown in Figure 9 b. Instead of an image input, we used analytical equations to describe the stimulus. We used an arbitrary spatial unit (su) to indicate distances and su/s to express the speed of the dots, with s representing time in seconds. In the constant velocity condition, we let all dots' horizontal speed and the middle dot's vertical speed to be 4 su/s. The starting horizontal position of these dots was 0, and the starting vertical positions of these dots from top to bottom were 6 su, 1 su, and 0, respectively. They ended at 4 su on the horizontal axis, and the middle dot ended at 5 su along the vertical axis.

The temporal sampling-interval was 0.01 s. The size of the retinotopic space in our simulation was 4 su by 4 su. The receptive fields of retinotopic motion-detectors covered an area of 0.4 su by 0.4 su. There was a spatial overlap of 0.2 su in both horizontal and vertical directions between the receptive fields. So, a total of 361 receptive fields were used.

Results.

Synthesis of the reference-frame: Figure 17 b shows the activities of cells computing the direction and the speed of the reference frame. In the three-dot paradigm, stimuli move from left to right, which means that the theoretical common-motion direction is horizontal, i.e., $\overrightarrow{d_0}$. As can be seen, initially the activities corresponding to all directions rise; however, within ca. 100ms, g_0 starts winning the competition and establishing itself as the direction of the reference-frame. As we will discuss in the Discussion section, the computation of common motion is not instantaneous but takes time, especially for novel stimuli. After repetitive exposures, humans may be able to predict and accelerate their computation-time for common motion; however, these learning effects are not included in our model. Note that, speeds in all directions, τ_u, are computed; however, the product $g_u \cdot \tau_u$ effectively retains only the speed along the direction of the reference-frame (due to 0-1 activities of g_u) zeroing the speed for all other directions. After an initial transient period, $g_0 \cdot \tau_0$ reaches an asymptotic level which is close to the theoretical speed of 4 su/s.

Relative-motion perception: As depicted in Figure 9 d, only the middle dot has relative motion with respect to the group. Figure 17 c and d show the relative-motion

activities for the middle and the flanking dot, respectively. The panel c shows q_9(upward) reaching its theoretical value whereas all other activities being low; the panel d shows no relative motion for the flanking dot.

Overall, these results show that the model can capture well the perception of relative motion in the classical three-dot paradigm.

3.1.2 Variable Velocity

Rationale

As mentioned in the Introduction, the vector analysis approach proposed by Johansson consists of three principles. The second and the third principles rely on *equal and simultaneous* motions: “Principle 2: *Equal and simultaneous* motions in a series of proximal elements automatically connect these elements to *rigid* perceptual units.” and “Principle 3: When, in the motions of a set of proximal elements, *equal simultaneous motion vectors* can be mathematically abstracted (according to some simple rules), these components are perceptually isolated and perceived as one unitary motion”((Johansson, 1976), p. 205, emphases added). Further, he went on to describe in more details the meaning of equal motion directions and velocities: “Equal motion directions and velocities for translatory motion are therefore not only Euclidean parallel motions with the same velocity. (This latter description is valid only for projections from fronto-parallel motion.) The term, "equal," also includes all motions (1) that follow tracks that converge to a common point (a point at infinity) on the picture plane, and (2) whose velocities are mutually proportional relative to this point.” ((Johansson, 1976), p. 205). As an example, Figure 5 of this study ((Johansson, 1976), p. 205) illustrated three cases of motion vectors considered “equal” according to Johansson’s definition. In Figure 5 a,

the four vectors are equal in terms of two-dimensional Euclidean geometry. The cases in Figure 5 b, and c on the other hand show four "equal motion vectors" in terms of their convergence to a common point with proportional velocities. These considerations are motivated by the fact that, in an ecological setting, stimuli are defined in a three-dimensional space and that the projections of three-dimensional stimuli onto two-dimensional proximal (retinotopic) representations can transform three-dimensionally equal motion vectors into two-dimensionally unequal vectors. Whereas these observations take into account how *rigid* motions are perceived, it remains to be determined whether and how non-rigid motion follows also a vector decomposition. A priori, one would expect so because of the existence of non-rigid motion in nature. Moreover, if an object is composed of multiple rigid-components, the proximal stimuli they generate may not follow the above assumptions. For example, consider a point-walker stimulus walking to the right. The body, the segments of the arms and legs may all undergo rigid motion themselves and will have an equal horizontal velocity on the *average.* Instantaneously, however, the swinging arm can have faster or slower horizontal velocity compared to the body. Because our perceptions arise *during* and not *after* stimulus presentation, the theory has to also address these transient effects. As shown in the previous section, our model computes activities in "real-time" and can be analyzed in terms of how it responds to these transient effects. For this purpose, we designed a slightly modified version of the three-dot stimuli where we violated the "equal and simultaneous" velocity constraint.

Methods.

In the variable velocity condition, the stimulus was similar to the classical constant-velocity condition with the following exceptions: Instead of a single flanking dot at the top and bottom positions, we used a pair of flanking dots to strengthen the reference-frame. We used a single central (middle) dot. The starting horizontal position of these

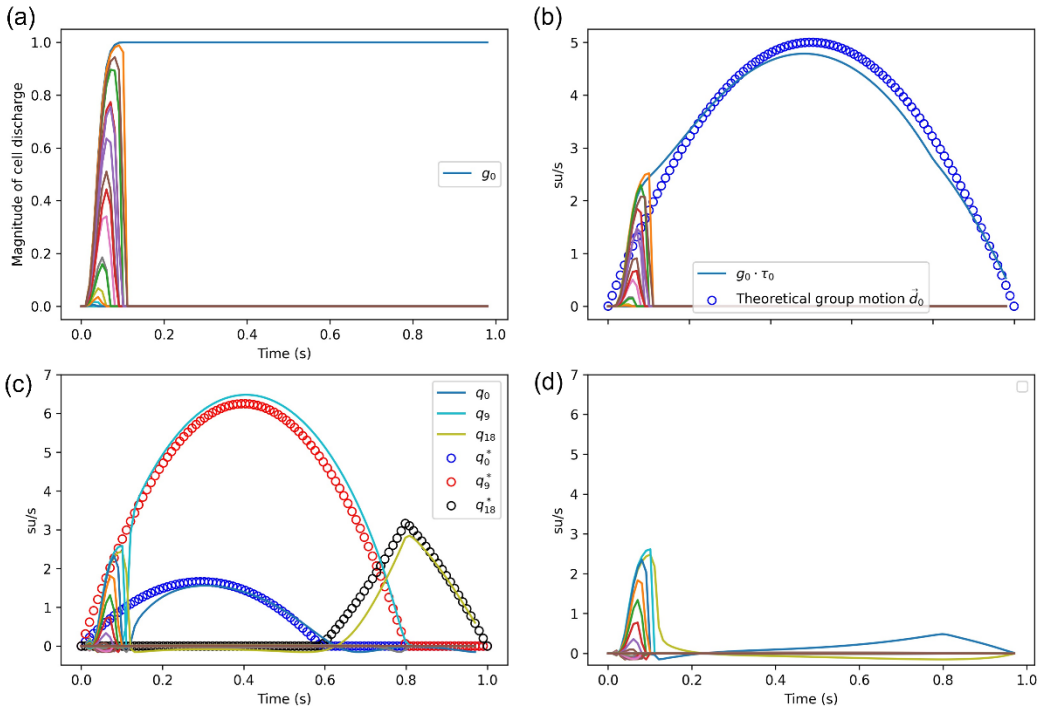

Figure 18: The three-dot paradigm with variable velocity. a. Outputs of the winner take-all-network, with different colors corresponding to different directions. b. Illustration of the reference-frame layer's output providing both the direction and magnitude of the reference-frame $(g_d \cdot \tau_d)$. Relative-motion activities in the Three-dot paradigm with variable velocity for the middle (c) and the flanking dot (d). Relative to the reference frame, the middle dot was perceived initially rightward and upward (blue and red curves, respectively) and towards the end of its motion, it is perceived to move leftward. Together, these directions indicate the perception of a curved trajectory. The flanking dot had no relative motion with respect to the reference-frame, since the flanking dots served as the reference frame.

dots was 0, and the starting vertical positions of these dots from top to bottom were 8 su, 7 su ,2 su ,1 su, and 0, respectively. They ended at 3.33 su on the horizontal axis, and

the middle dot ended at 5.33 su along the vertical axis. We used a single central (middle) dot. We used the following equations to describe the movements of the dot in the middle (denoted by M) and two other dots that are on the top and bottom denoted by S.

$$\begin{cases} \frac{dX_S(t)}{dt} = -20t^2 + 20t \\ \frac{dY_S(t)}{dt} = 0 \end{cases} \quad (20)$$

$$\begin{cases} \frac{dX_M(t)}{dt} = -\frac{20}{t_T^3}t^2 + \frac{20}{t_T^2}t \\ \frac{dY_M(t)}{dt} = -\frac{20}{t_T^3}t^2 + \frac{20}{t_T^2}t \end{cases} \quad (21)$$

where X and Y are the positions in su along horizontal and vertical axes, respectively. Mathematically, these equations describe the speed of the dots with respect to time. In the simulation, the horizontal speeds of all dots and the vertical speed of the middle dot began from 0 at $t = 0s$, changed smoothly, and then went back to 0 as they finished their movements. We used a second or- der polynomial to describe this motion (Equations 20 and 21). The simulation time-interval (t) of flanking dots was from 0 to 1s, and for the middle dot, it was from 0 to t_T . Equation 21 makes the middle dot have the same motion dynamics along X and Y axes. More importantly, it also makes the middle dot end at the same horizontal position as flanking dots although the middle dot may reach the end-point later ($t_T > 1$) or earlier ($t_T < 1$), since $\frac{dX_S(t)}{dt}$'s integral from 0 to 1 is equal to $\frac{dX_M(t)}{dt}$'s integral from 0 to t_T. We used five different values for t_T : 0.8, 0.9, 1, 1.1, 1.2 in the simulations. In this section we show the results for $t_T = 0.8$. With $t_T = 0.8$, the middle dot arrives earlier than the other four flanking dots by 0.2 s. Later

in the manuscript, we compare model predictions for all values of t_T to data collected in the psychophysical experiment that will be described in Section 4.

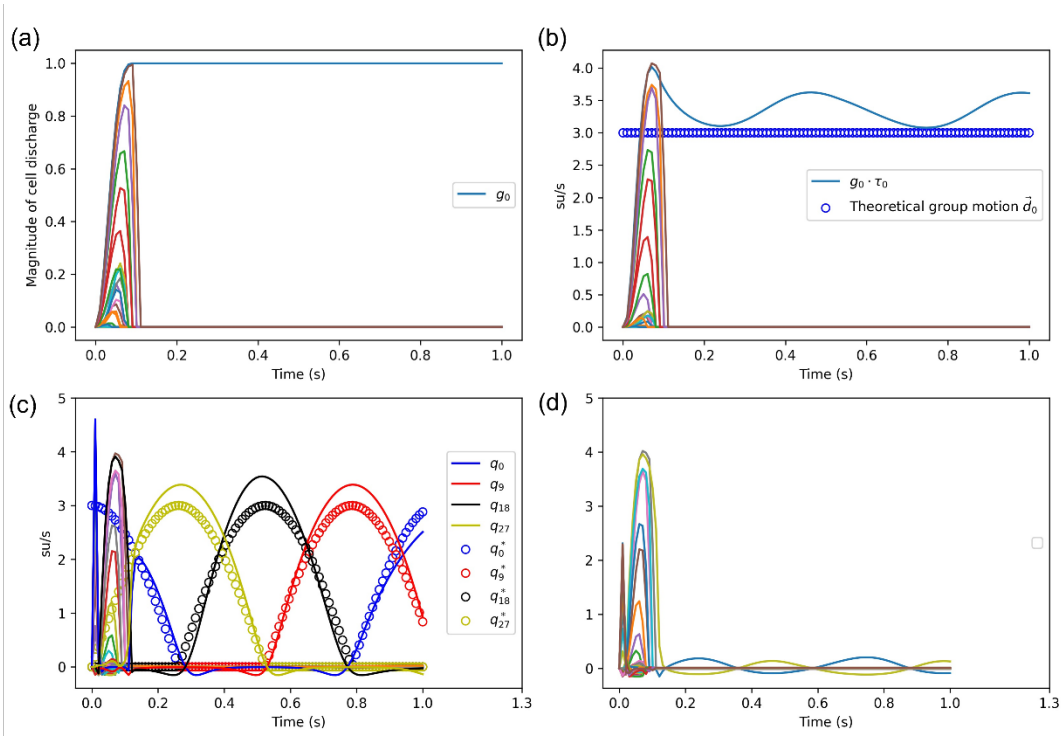

Figure 19: The rotating-wheel paradigm. a. Outputs of the winner take-all-network, with different colors corresponding to different directions. b. Illustration of the reference-frame layer's output providing both the direction and magnitude of the reference-frame ($g_d \cdot \tau_d$). Cell activities from the vector decomposition layer in the Rotating-wheel paradigm, with (c) and (d) panels corresponding to the dots on the rim and the center of the wheel, respectively. The dot in the center of the wheel had only one significant directional component, rightward, corresponding to the horizontal speed of the rotating wheel (3 su/s). Relative-motion activities in the Rotating-wheel paradigm with left and right panels corresponding to the dots on the rim and the center of the wheel, respectively. The dot on the rim activated relative-motion cells tuned along $\overrightarrow{d_0}, \overrightarrow{d_9}, \overrightarrow{d_{18}}$, and $\overrightarrow{d_{27}}$. During a period of rotation, relative-motion cells tuned sequentially to $\overrightarrow{d_9} - \overrightarrow{d_{18}}$, $\overrightarrow{d_9} - \overrightarrow{d_0}$, $\overrightarrow{d_{27}} - \overrightarrow{d_0}$, and $\overrightarrow{d_{27}} - \overrightarrow{d_{018}}$, as the dot rotated around the center towards upper-left, upper-right, lower-right, and lower-left directions. The dot in the center of the wheel had negligeable relative motion with respect to the reference-frame, since it served as the reference frame.

Results.

Synthesis of the reference-frame: Figure 18 plots the direction (a) and speed (b) of the reference-frame cells. As in the classical version of this stimulus, after an initial transient competition, the rightward direction wins the competition. The right panel shows that the network is able to follow closely the time-varying parabolic speed-profile of the group motion.

Relative-motion perception: Figure 18 c and d shows the activities of relative-motion cells for the middle and the flanking dots, respectively. The flanking dots show only negligeable activity, meaning no significant relative-motion perception, whereas the middle dot activities produce neural responses for q_0, q_9, and q_{18}. In other words, it shows that the dot on the middle is perceived to move rightward (along $\overrightarrow{d_0}$) initially and then gradually turning leftward (along $\overrightarrow{d_{18}}$), and at the same time moving upward (along $\overrightarrow{d_9}$) until the lateral dots stop. This is because, when the dot in the middle stops, the other dots are still moving. At the same time, the previous grouping effect doesn't change, so the middle dot is perceived as moving in the *opposite* direction along the horizontal axis with respect to the group. These observations will be tested by a psychophysical experiment described in the Section 4.

3.2 The Rotating-Wheel Paradigm

Methods.

The radius of the wheel we used was 1 su, and this wheel was made to roll rightward with constant speed, 3 su/s. In this simulation, three dots were used, one was in the center of the virtual wheel (denoted by C) and the other two were on the rim positioned

with centro-symmetry (denoted by R). We used two dots on the rim to strengthen the grouping effect, but we will only show the information and activities from one of these two, since they exhibit the same activity except for the phase. To simulate this rotating wheel paradigm, we first defined the movements of the stimulus by:

$$\begin{cases} \frac{dX_C(t)}{dt} = 3 \\ \frac{dY_C(t)}{dt} = 0 \end{cases} \tag{22}$$

$$\begin{cases} \frac{dX_R(t)}{dt} = 3\big(1 + cos(6t)\big) \\ \frac{dY_R(t)}{dt} = -3sin(6t) \end{cases} \tag{23}$$

As in the previous situation, X and Y are the positions in su along the horizontal and vertical axes, respectively. We used a time period from 0 to 1 s with a sampling interval of 0.001 s. We also used a space with a size of 8 su by 2 su, and the size of the receptive field of the retinotopic motion-detectors was 0.4 su by 0.4 su. These receptive fields overlapped with each other by 0.2 su in both X and Y axes. As a result, a total of 351 receptive fields were used.

Results.

Synthesis of the reference-frame: Figure 19 a and b show the activities corresponding to reference-frame direction and magnitude, respectively. After a brief transient period, the reference-frame picks up horizontal rightward direction corresponding to $\overrightarrow{d_0}$ (left panel). As can be seen in the right panel, the horizontal speed (along $\overrightarrow{d_0}$) approaches 3, the theoretical horizontal speed of the wheel. The oscillation came from the shape of the tuning curve. In this paradigm, two dots on the rim rotated around the center. The dot on the top began with the same direction of the center, while the dot on the bottom began

with upward direction. During the motion, their directions were deviating and approaching the common motion direction respectively. As described in Equation 4, these two paths (angular disparity increases and decreases) undergo different tuning properties. Therefore, the resulted activities of two rim dots can't fully offset each other, and the sum of them then behaved trigonometric oscillation smoothly around the supposed value.

Table 2: Point Walker Paradigm Parameter X

Parameter / Marker	x								
	c1	c2	c3	c4	w	ph1	ph2	ph3	k
LWRB	-3.12	-1.85	-0.57	-0.45	5.61	-4.43	-2.08	14.04	14.55
RELB	0.21	-0.94	0.76	-0.74	5.76	-3.2	268.08	-2	13.76
LSHO	-0.34	0.32	-0.5	0.38	5.89	0.09	-56.54	0.17	13.71
LELB	-0.7	-0.76	-0.12	-0.08	5.71	6.79	2.24	-1.4	14.11
C7	0.07	0.02	1.35	1.28	5.73	-1.73	2.65	5.73	13.68
RSHO	2.26	2.06	0.72	-0.56	5.69	-3.18	9.73	9.7	13.62
LKNE	1.28	-0.56	0.39	0.03	5.76	-2.57	2.23	-1.97	13.78
RWRB	-0.68	2.66	0.18	-0.1	6.04	-4.06	3.68	3.22	13.37
RBWT	-0.44	0.33	0.13	0.11	5.63	6.11	0.37	-1.24	13.8
LBWT	-2.32	2.4	0.78	-0.68	5.31	0.04	5.22	-1.17	13.78
LANK	-0.24	-2.8	-0.2	0.2	5.82	-3.35	-1.8	17.66	13.94
LTHI	0.9	-0.55	-0.24	-0.1	5.7	-2	-0.15	-0.37	13.82
RTHI	0.08	-0.8	0.13	-0.22	5.81	6.27	-4.78	-0.3	13.9
RKNE	-1.81	-0.02	-0.24	0.3	5.81	4.71	-2.14	-3.68	13.77
RANK	1.18	3.56	-0.35	0.43	5.68	3.44	0.06	-0.25	14.18

Relative-motion perception: Figure 19 c and d show the results for relative-motion cells corresponding to the dot on the rim and the dot in the center, respectively. As can be seen from the figure, the dot on the rim activated relative-motion cells tuned along $\overrightarrow{d_0}$,

$\overrightarrow{d_9}$, $\overrightarrow{d_{18}}$ and $\overrightarrow{d_{27}}$, while the center dot had very few relative motion components. This coincides with the perception of the stimulus in that the wheel as a whole should be perceived to move rightward with the rim rotating around the center. In other words, the group motion velocity should be equal to the center dot's velocity, and the rim dot should show a circular *relative* motion.

Table 3: Point Walker Paradigm Parameter Y

Parameter	y							
Marker	c1	c2	c3	c4	w	ph1	ph2	k
LWRB	0.16	-0.55	0.28	-0.06	5.78	0.05	-1.87	0.27
RELB	1.1	-1.07	3.66	-3.66	5.73	0.07	-0.44	0.22
LSHO	-0.2	0.16	0.1	0.17	5.72	-0.26	-5.1	0.23
LELB	-0.17	0.1	0.1	0.02	5.82	-0.52	-6.51	0.24
C7	-0.34	0.35	0.02	0.23	5.73	-0.19	-5.19	0.21
RSHO	-0.3	0.35	-0.1	0.36	5.73	-0.25	-4.53	0.23
LKNE	0.11	-0.11	0.66	-0.6	5.63	-0.72	-3.37	0.23
RWRB	0.19	-0.12	0.1	-0.18	5.54	-1.08	-1.3	0.24
RBWT	0.12	-0.06	0.11	0.13	5.73	0.13	-5.07	0.22
LBWT	-0.59	0.58	0.1	0.15	5.72	-0.17	-4.94	0.21
LANK	2.61	-2.69	0.7	-0.62	5.84	-0.14	0.52	0.24
LTHI	0.09	0.01	-0.28	-0.21	5.68	-1.63	-6.98	0.23
RTHI	-0.06	0.01	0.07	0.11	5.84	-1.3	-6.11	0.23
RKNE	-0.1	0.13	0	-0.11	5.68	-0.88	-2.71	0.22
RSHN	0.32	-0.51	-0.05	0.1	5.91	0.59	-4.16	0.24

3.3 The Point-Walker Paradigm

Methods.

For the Point Walker paradigm, to express analytically the motion of each dot, we used the following generic equation proposed in (Troje, 2002).

$$\frac{dC(t)}{dt} = c_1 w cos(wt) + c_2 w cos(wt + \phi_1) + 2c_3 w cos(2wt + \phi_2) + 2c_4 w cos(2wt + \phi_3) + k \quad (24)$$

where C is the movement in su along either X or Y axis and the different choices of parameters allow the adaptation of motion trajectories for different body parts.

Figure 20: An example of Point-Walker Display. From top to bottom, this figure shows successive frames of body postures during the movement from left to right. In the first four frames, the left foot of the walker moved very little while the right foot took a large step from behind to the front. In the next four frames, the right foot moved very little while the left foot took a large step. Note that in the sixth frame, several dots are temporarily superimposed. The body moved rightward all the time.

We used 15 marker positions, which are located at the major joints of the body (shoulders, elbows, wrists, knees, ankles), at the center of the neck, at the centers of two thighs, and at the buttocks. Only the movements of the left view were simulated, which can be visualized by Figure 20. These movements are described using Equation 24 with different parameters, which are fitted using the trajectory data from Carnegie Mellon University's Graphics Lab motion-capture database available at http://mocap.cs.cmu.edu.

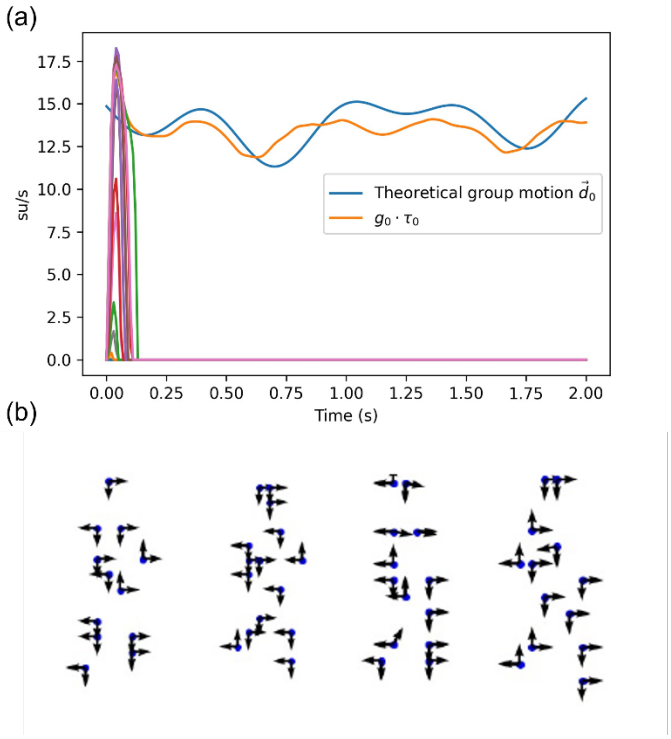

Figure 21: The point-walker paradigm. a. Illustration of the reference-frame layer's output providing both the direction and magnitude of the reference-frame ($g_d \cdot \tau_d$). b. Illustration of perceived relative motion by showing relative-motion vectors for each point on the walker.

We provided the fitted parameters in Table 2, 3. This method of simulating the human walking trajectory has been used in many previous studies (Karg et al., 2010; Davis & Gao, 2004; Zell & Rosenhahn, 2015). The walker is moving from left to right as shown

by successive frames stacked vertically. We used the average motion speed along the rightward direction to represent the theoretical reference-frame speed.

We used a retina with the size of 16 su by 8 su, and the receptive field of the motion detectors was 1 su by 1 su, overlapping by 0.5 su in both horizontal and vertical directions with each other. So, a total of 465 receptive fields were used.

Results.

Synthesis of the reference-frame: As shown in Figure 21 a, the reference-frame direction cells rapidly select horizontal rightward-direction (g_0 corresponding to $\overrightarrow{d_0}$) as the direction of the common group motion. The model's reference-frame speed estimation (orange curve) approximated the dynamics of theoretical speeds (blue curve) well. Hence, notwithstanding the complexity of the stimulus in terms of the number of markers and their individual trajectories, the model was able to determine the direction and the magnitude of the group (common) motion.

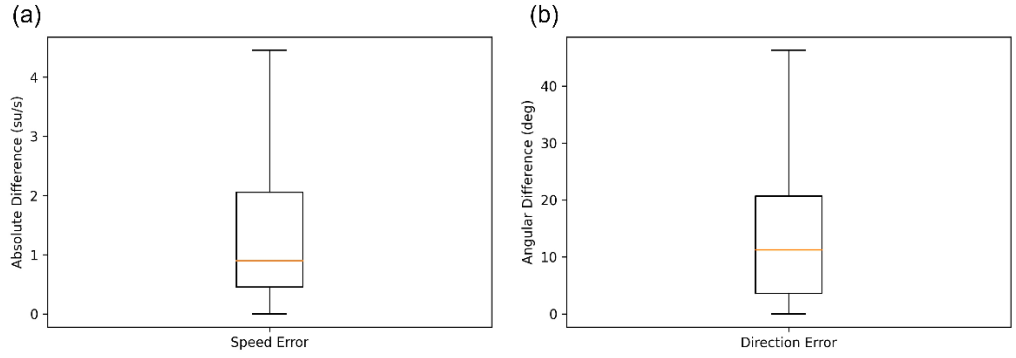

Figure 22: Evaluation of model's performance for the point walker paradigm. Each plot shows a box chart with five bars from the top to the bottom representing the maximum, third quarter value, median, first quarter value, and the minimum values. Outliers are not shown. a. The speed error is calculated by the difference between theoretical and perceived relative motion velocities. b. The directional error is calculated by the difference between theoretical and perceived relative motion directions.

Relative-motion perception: We show some examples of the predicted relative motion perception in Figure 21 b. The figure shows the point walker at four time-instants (t=0.76 s, 0.86 s, 1.16 s, and 1.46 s) corresponding to the third, fifth, seventh, and the last frame in Figure 20. For each marker, the arrows represent the direction and magnitude of the active relative-motion cells. By visual inspection we can see that the arrows attached on each marker reflect its supposed locomotion direction for a walking person. For example, the points on a limb that is moving forward are associated with right arrows. The arms are more marked with up arrows while the legs are more marked with down arrows as during this phase of the movement arms (legs) are going up (down) with respect to the body.

We calculated the error statistics across to 15 markers by comparing each marker's model-predicted behavior to theoretical values. The errors are shown in Figure 22. The median localization error was 0.88 su, compared to the resolution of 1 su. The median speed error was 0.9 su/s. The median direction error was $11.32°$.

In many experiments, the walking-direction of the PLW is used as the dependent variable. However, let us note that detecting the walking direction of a PLW is not a difficult problem, in particular, in the absence of noise. Even with noise, humans can detect the facing direction of a PLW using only the motion of the feet, as was shown by Troje & Westhoff (2006). Our model can also detect the facing direction using only the motion of the feet. However, detecting facing direction using only local cues doesn't necessarily mean that the model can decompose the whole body motion into relative motions. In the simpler three-dot paradigms, at least one dot in the display has a motion direction that is exactly the same as the theoretical common-motion direction. But for

PLW, this is not the case. Therefore, by using PLW, we show that our model generates robust results given a complex display composed of large number of visual elements moving in a complex irregular way. Finally, notwithstanding the fact that we simplified the problem by considering the whole body as a group and ignoring the hierarchical grouping effects, the model is still able to predict correctly the perceived motion of each dot.

4 Psychophysical Experiments

In this section, we report two experiments using a modified version of Johansson's Three-Dot stimulus to test the predictions of our model when the "equal and simultaneous" velocity constraint is violated. In the classical version of this stimulus, the constraint is satisfied in that all three dots have the exact same horizontal velocity at all times and hence appear to move together from start to finish. In our simulations of the model, the central dot had a different horizontal velocity profile than the lateral two dots and the results predicted the perception of *curved* trajectories in the case when the central dot moved slower or faster than the other dots.

4.1 Experiment 1

4.1.1 Subjects

Three adult subjects who were naïve to the hypothesis of the experiment and one of the authors (DH) participated in the experiment. They all had normal or corrected-to-normal visual acuity. This experiment followed a protocol approved by the University of

Denver Institutional Review Board for the Protection of Human Subjects. Each observer gave written informed consent before the experiments.

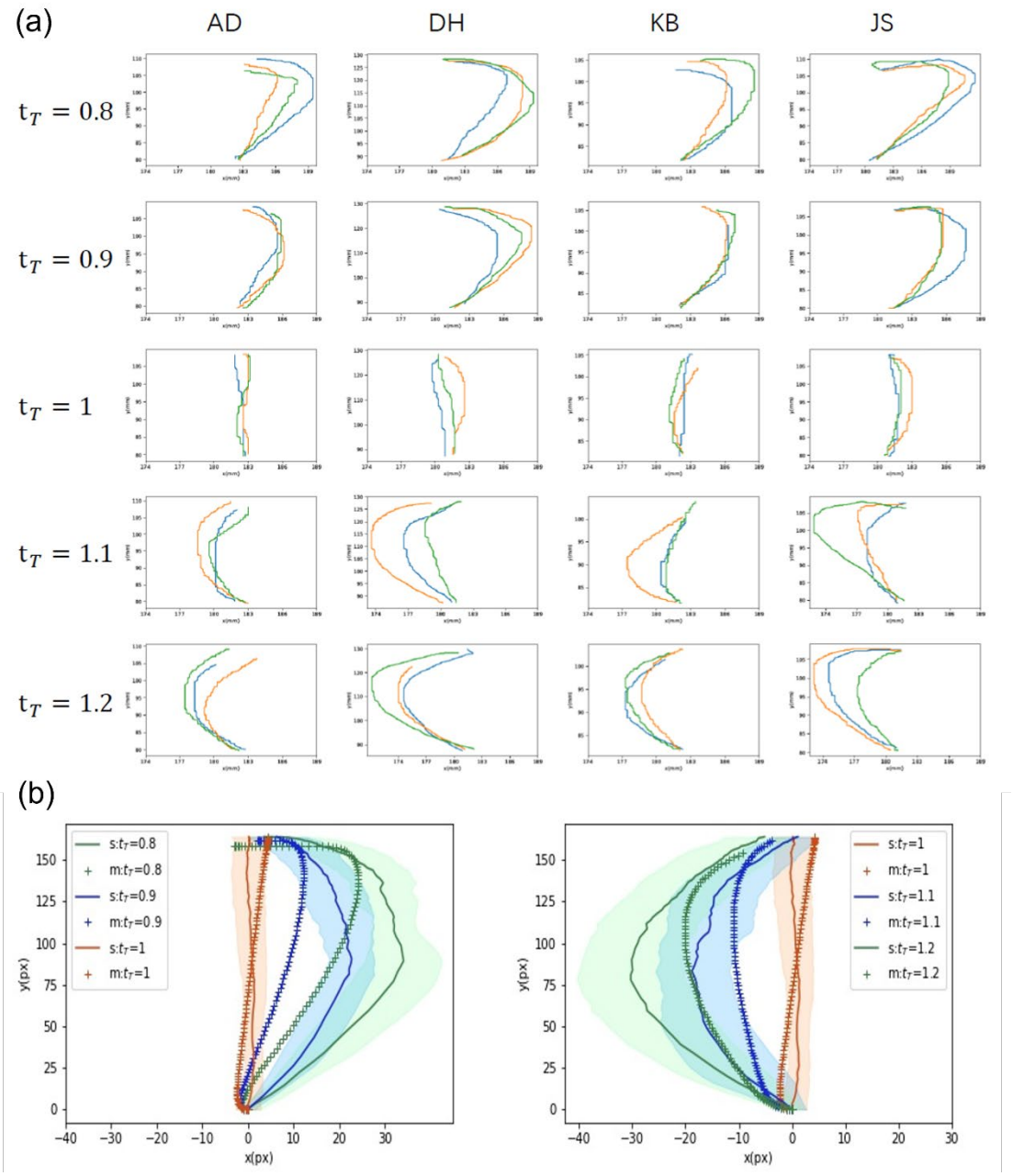

Figure 23: Results of experiment 1 and model predictions. Experimental results. Columns and rows correspond to subjects and different values of t_T, respectively. Three repetitions are shown for each case. b. Trajectories predicted by the model compared to experimental data averaged across the observers. In this figure, "s" indicates subjects' data, and "m" indicates model predictions. The shaded areas around the data indicate ± 1 SEM.

4.1.2 Stimuli and procedure

Participants were seated at approximately 50 cm from the monitor whose resolution was 1920×1080 and its frame rate was 144 Hz, and viewed the five-dot motion stimuli. We increased the number of dots from three to five to strengthen the grouping effect, since the difference of velocity profile between the center dot and the lateral dots might weaken it. The dots were circular and had a size of 0.2°. In each trial, all dots were initialized at a position at 830 pixels (12.14°) along the horizontal axis, and the vertical positions in pixels (visual angles) from top to the bottom were 418 (6.11°), 442 (6.46°), 630 (9.21°), 654 (9.56°), and 678 (9.91°), respectively (choosing the upper-left corner of the monitor as the origin). These dots started moving as soon as the subject pushed a pre-defined button. Depending on the different values of t_T in Equations 20 and 21, the center dot could arrive at the stopping point earlier, later, or at the same time than the flanking dots, as a function of its horizontal velocity with respect to that of the lateral dots. The duration of each flanking dot's movement was 1 s. At the end of their trajectories, the dots stopped and remained visible at their final position. When all stimuli's motion stopped, the subjects were asked to use the mouse to draw the center dot's perceived motion trajectory relative to the other dots. The curves drawn by the subjects were then recorded. After the end of a trial, the subject pushed a button to start the next trial.

This experiment contained three conditions based on five different values of t_T (0.8, 0.9, 1, 1.1, 1.2). Each condition was repeated three times in random order, and thus each subject had to finish 15 trials within one session. Each subject completed three sessions run on different days.

4.1.3 Results

The three curves in each panel of Figure 23 a represents the results of the three sessions by each subject. Overall, one can see both session to session variability in each subject as well as variability across subjects. Notwithstanding these quantitative differences, all subjects reported very similar patterns. They drew curved motion trajectories for stimuli where the center dot moved slower or faster than the others. Figure 23 b shows data averaged across the observers compared to model predictions. Overall, data confirm qualitatively the predictions of the model. We do find however quantitative differences between the model predictions and the data. The extremum points of the theoretical curves occur at higher vertical positions compared to data. In order to obtain better quantitative results, we used a more elaborate experimental approach together with eye-tracking techniques, to measure quantitatively the left- and right-extrema positions of the perceived trajectory of the relative motion.

4.2 Experiment 2

4.2.1 Subjects

Three adult subjects who were naïve to the hypothesis of the experiment and who didn't participate in Experiment 1 and one of the authors (DH) participated in Experiment 2. They all had normal or corrected-to-normal visual acuity. This experiment followed a protocol approved by the University of Denver Institutional Review Board for the Protection of Human Subjects. Each observer gave written informed consent before the experiments.

4.2.2 Apparatus

Visual stimuli were created using Psychopy Toolbox and displayed on a monitor at a resolution of 1920x1080 with a refresh rate of 60 Hz. Gaze-position monitoring for both eyes was performed by an Eyelink- || eye-tracker at 150 Hz sampling rate. The distance between the observer's eyes and the monitor was 70 cm and the dimensions of the display at this distance were 41.11x23.12 deg^2. A head/chin rest was used to help stabilize fixation and to reduce noise due to head movements in eye-movement recordings.

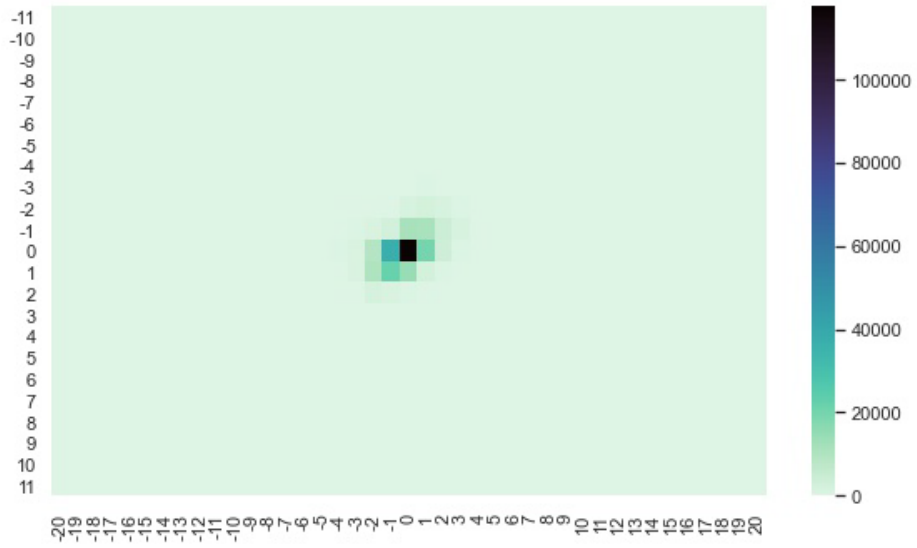

Figure 24: A summary of gaze positions across all the subjects and trials after drift correction during the stimuli's movement phase. X and Y axes indicate the dimension of the monitor screen in deg, and 0 indicates the center. Color bar shows the number of records.

4.2.3 Stimuli and procedures

At the beginning of a trial, a white fixation square (0.5 deg, 154 cd/m^2) was shown at the center of the screen on a dark background (0.27 cd/m^2). Observers were required to

fix their gaze on the square. After 2000 ms, the stimulus was displayed, in which five white disks (0.5 deg, 108 cd/m^2) moved in the same way as the stimuli used in the first experiment, and the movements of the dots are described by Equations 20 and 21. These dots began their motion at a position 2.6 deg to the left of the center of the screen, and their vertical positions were +4.6, +3.6, -2.6, -3.6, -4.6 deg upward (+) or downward (-) with respect to the center. The middle dot moved for 5.2 deg. along both the horizontal and vertical directions, while all flanking dots moved horizontally for 5.2 deg. As in Experiment 1, with the ending time of flanking dots fixed at 1 s, five ending times of the middle dots, t_T, were used (0.8, 0.9, 1, 1.1, 1.2 s), and thus the display lasted from 1000 ms to 1200 ms, depending on the condition. During the display of stimuli, subjects were asked to keep their fixation steady at the center of the screen, and a trial was aborted if the subject's gaze position moved beyond a 2x2 deg^2 rectangular area around the center of the screen. The records of gaze positions during all trials without being aborted across all subjects are presented in a heatmap (Figure 24). After the display of stimuli, a white horizontal or vertical line was shown at a position within 5 deg from the ending position of the middle dot. The trials showing horizontal and vertical lines were separated into two blocks, each was set up to make subjects identify the vertical and horizontal position, respectively, of their perceived extreme point of the middle dot. As in the trajectories drawn by subjects in Experiment 1 (Figure 15), the extreme point is the leftmost or rightmost point of the perceived curved trajectory when $t_T > 1$, and when $t_T < 1$, respectively.

We used interleaved 1-up/1-down staircases to determine the perceived locations of the extrema in each condition: In the block using vertical lines, trials of different t_T

values were generated randomly and for each t_T we had a separate staircase. For each t_T condition, the initial position of the comparison line was selected randomly. Subjects were asked to press the right- or left-arrow key on the keyboard if they perceived the extreme point on the right or left side of the line respectively. Once they reported, the staircase for this t_T condition registered the response and shifted the comparison line for 2 deg along the same direction as reported. This spatial step in moving the comparison

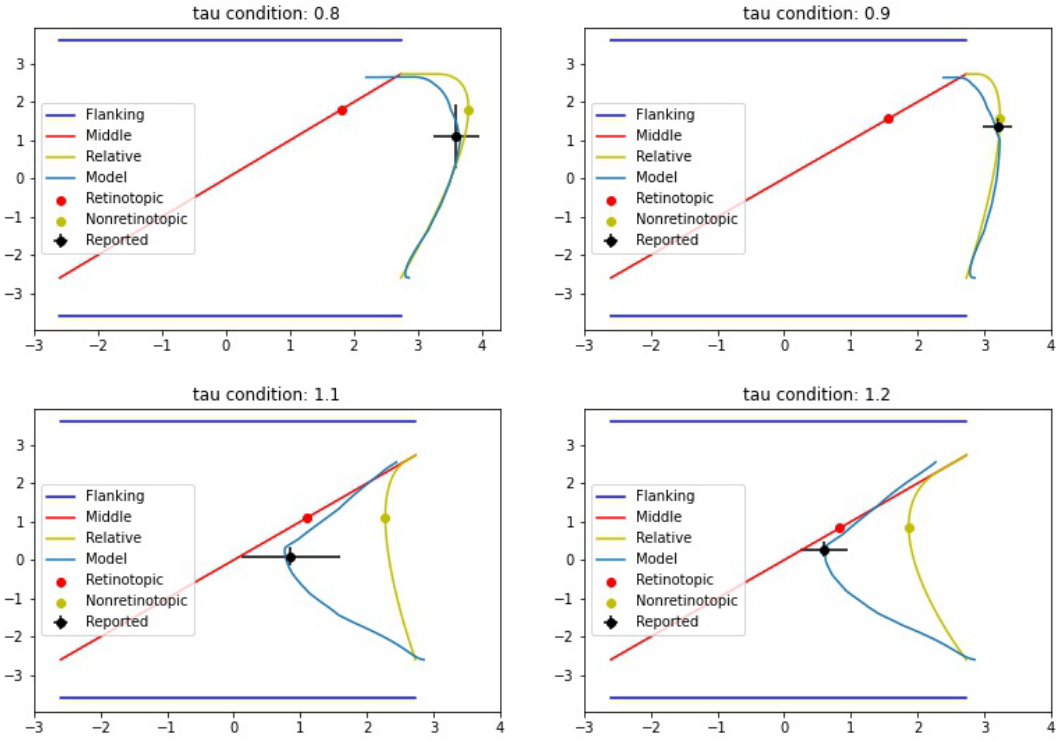

Figure 25: Results of experiment 2 and model predictions. X and Y axes indicate the dimensions of the monitor screen in deg, and 0 is the center. The label "Flanking" indicates the trajectories of the dots above and below the middle dots during the display of the stimulus. "Middle" indicates the trajectory of the middle dot. "Relative" is the physical relative motion of the middle dot relative to the flanking dots defined in coordinates of the dots' ending positions. "Retinotopic " is the projected position on the retina when the middle dot is at its extreme position relative to the flanking dots. "Nonretinotopic" indicates the extreme position of the relative motion. "Reported" marks the subjects' reported position of perceived extreme position with vertical and horizontal error bars. "Model" indicates the trajectories predicted by the model.

line was reduced gradually by halving it after three reversals of the staircase. The staircase, and hence trials of a particular t_T condition finished when the step size was smaller than 0.5 deg. The maximum number of trials under each t_T condition in a block was 30. The block finished when staircases for all t_T conditions converged, or they reached the maximum trial limit. Subjects were asked to run another block with only un-converged t_T conditions included until all $t_T \neq 1$ staircases converged. The final comparison-line position was considered as the horizontal position of subject's perceived extreme point. Similarly, in the horizontal comparison line block, subjects used up- and down-arrow keys to determine the vertical position of perceived relative motion trajectory.

4.2.4 Behavioral Results

For all subjects and all $t_T \neq 1$ conditions, subjects finished all staircases within one block. Including all blocks and t_T conditions, the M (SD) number of trials used across subjects was 225(15.62). The M (SD) number of trials aborted due to unsatisfied gaze movements across subjects was 96.5(27.85).

Figure 25 shows the reported extreme positions under each t_T condition. To visualize the perceived extreme position relative to the stimulus, we also show in the Figure the physical (i.e., according to screen-based reference-frame; with the fixed gaze-position in the experiment, this corresponds to the retinotopic trajectory) trajectories of the flanking and middle dots, together with the physical relative motion (i.e., according to a reference-frame based on flanking dots, i.e., non-retinotopic relative motion) and the theoretical extreme positions in each condition. We also marked the retinotopic and non-retinotopic extreme positions, which correspond to the extreme positions for two aforementioned

conditions. We found that the reported extreme positions were closer to the non-retinotopic extreme position if $t_T < 1$, and the retinotopic extreme position if $t_T > 1$.

4.2.5 Model Predictions

Our model is designed to compute the perceived positions according to a non-retinotopic reference-frame based on Gestalt common-fate principle. It is known that humans do not in general use a single reference-frame, but instead they use an amalgamation of various reference-frames according to the prevailing stimulus conditions (Agaoglu et al., 2015; Huynh et al., 2017; Freeman, 2001; Freeman & Banks, 1998; Baker & Braddick, 1982). Hence, in order to fit data quantitatively, we used the reference-frame *combination* model that we proposed earlier (Agaoglu et al., 2015; Wade & Swanston, 1982; Gogel, 1977). According to this model, the effective reference-frame for motion perception emerges from a weighted summation between non-retinotopic motion-based (P_{mb}), and the retinotopic reference-frames (P_r), as shown in Equations 25 and 26 (Agaoglu et al., 2015). The weights, w, are linearly dependent on the distance (d) between the target stimulus and its neighboring stimuli affected by the Gestalt grouping.

$$Perceived\ motion = w_r(d)P_r + w_{mb}(d)P_{mb} \tag{25}$$

$$w(d) = k \cdot d + c \tag{26}$$

where k is the weight's dependency coefficient on the distance, and c is a constant term. In our application, P_r was determined by the neural outputs from the retinotopic motion detector layer m_u along different tuning directions, and the P_{mb} was determined by the neural outputs from relative-motion computation layer q_u, which can be found in Figure 4. We used the minimum distance between the middle dot and flanking dots above and below it to represent d. Therefore, the model can be written as:

$$PV_u = w_r(d) \cdot m_u + (1 - w_r(d))q_u \quad (27)$$

We fitted the model, and obtained the weights shown in the following equation:

$$w_r(d) = \begin{cases} -0.05d, & t_T < 1 \\ 0.27d - 0.32, & t_T > 1 \end{cases} \quad (28)$$

Relative motion trajectories generated from the model in each t_T condition can be found in Figure 25. The minimum and maximum values of d were 1 and 3.68 respectively. Overall, the model provides a good quantitative fit to the data, as the cartesian errors between the reported positions and the positions of maximum curvature in the curves predicted by model are 0.09 deg ($t_T = 0.8$), 0.33 deg ($t_T = 0.9$), 0.15 deg ($t_T = 1.1$), and 0.07 deg ($t_T = 1.2$).

5 Discussion

5.1 Reference-frames and relative-motion perception

How we perceive motion has a long history that can be traced to antiquity (e.g., Zeno's paradoxes). An important aspect of motion perception was revealed by Gestalt psychologists who, by applying the grouping principles, showed that the perceived motion consists of both the motion of groups and the relative motion of parts within groups. By using biological motion, Johansson demonstrated relative-motion perception with much more complex stimuli. Furthermore, he proposed a theory of vector decomposition that can explain both group and part motion. All these studies revealed the central role the reference-frames (or coordinate systems) play in the computation and perception of motion. In Piagetian theory of cognitive development, reference-frames have a fundamental role in explaining the development of not just sensory or motor competencies but of intelligence in general: He proposed that newborns start with

egocentric reference-frames and synthesize exocentric reference-frames through developmental stages (Piaget, 1952). Indeed, several studies (Swanston et al., 1987; Freeman, 2001; Fink et al., 2003; Gramann et al., 2010) showed that reference-frames used in perception and cognition can be classified into egocentric and exocentric reference-frames[3]. In addition to cognitive and psychological studies, neurophysiological studies have also identified egocentric and exocentric reference frames in the primate nervous system (Olson, 2003). In a recent study, Sasaki et al. (2020) showed that neurons in the ventral intraparietal area respond according to both egocentric and exocentric reference-frames based on task-demands.

5.2. Reference-frame selection and combination

Behavioral research shows that humans use an amalgamation of different reference-frames (Swanston et al., 1987; Agaoglu et al., 2015; Huynh et al., 2017). In fact, we used the reference-frame *combination* model (Agaoglu et al., 2015) for quantitatively fitting our results. However, this does not mean that we can always arbitrarily select a desired reference-frame[4]. Some of the reference-frame selection processes appear to be "automatic" in that they cannot be modified by the observer or by the task. We proposed here a mechanistic model for reference-frame selection, vector decomposition, and relative motion perception. Whether and how reference-frame selection can be modulated

[3] The term allocentric, instead of exocentric, is more frequently used in the literature. The prefix "allo" means "other" and hence is not as informative as "exo", which not only puts these reference-frames in contrast with egocentric reference-frames but also highlight the important property that the reference frame is 'outside", i.e., "external" to the organism.

[4] For example, try perceiving the retinotopic motion of a static object in the scene when you are actively moving your eyes; we perceive the object as static despite its retinotopic motion. On the other hand, if you move your eye passively by pushing gently with your finger (cover the other eye), you will indeed observe the retinotopic motion of the static object.

by task demands in this model depends on the operation of the layer computing the common motion.

5.3. Multiple Gestalt groups

One limitation of our model is that the implementation of this layer is relatively simple in that it computes only the common-fate aspect of the stimuli, and this computation is carried out on the entire extent of the retinotopic motion-detector space. The application of the model for cases where multiple Gestalt groups are present simultaneously will require more sophisticated implementation of Gestalt grouping principles. The projections from the retinotopic motion-detectors to the vector-decomposition layer generate all possible decompositions and the projections from the reference-frame direction layer to the vector-decomposition layer selects the relevant components according to the prevailing reference-frame. Hence, the model can accommodate task-dependent reference-frames by allowing this layer to be modulated by task-dependent signals from higher areas.

5.4. Figural aspects and form factors

Our model was successful in explaining how local motion vectors are decomposed and perceived in the point-walker stimuli. However, one aspect of point-walker stimuli is that the percept is not just relative motions of dots but also a vivid percept of the figural aspects of the body during the execution of these movements. Our model was focused solely on motion signals and did not include form factors; it shows that the basic phenomenology of relative motion perception in these classical examples can be explained by motion signal analysis. In fact, Gilaie-Dotan et al. (2015) showed that patients with damage to ‘form areas” (ventral cortex) performed as well as controls in

perceiving biological motion, supporting the view that motion signals are sufficient in explaining our relative-motion perception. However, this does not completely exclude possible inputs from form processing. For models that use form factors, the reader is referred to (Clarke et al., 2016; Grossberg et al., 2011; Lange & Lappe, 2006). For example, our model is similar to Grossberg et al.'s (2011) model in terms of mathematical formalism. Grossberg et al.'s model uses additional factors such as boundary and depth and explains the data in terms of interactions between form and motion systems. As stated above, our model is based exclusively on the motion system.

5.5. Dynamics of reference-frame formation

Our simulations show that determining the reference-frame can take time, especially if the stimulus is ambiguous in terms of a net common directional motion signal. With the current parameters, we found that it takes around 100 ms to determine and establish the common-motion. Lange and Lappe (2006) varied the duration of the stimulus in a biological motion task (to determine whether the walker moved forward or backward) and found that the performance gradually improved with stimulus duration. The time-constant estimated from their data is about 333 ms, which is the time needed to reach approximately 2/3 of the steady-state value. Neri et al. (1998) showed long temporal-integration times for biological motion, up to 3 seconds. Of course, 3 second integration time includes both the determination of the reference frame and computation of relative motion from their limited-lifetime motion stimuli. It is clear that our model can benefit from a better calibration of temporal parameters. However, determining the exact real-time dynamics of trajectory perception is difficult, in particular when the stimulus is complex and ambiguous. Prior exposure and/or perceptual learning can affect

significantly the temporal aspects of performance. A naïve subject who sees a point-walker for the first time may require several trials before the percept emerges.

5.6. Hierarchy of reference-frames

One of the challenging future directions for this work consists of developing further the reference-frame detection layer by introducing additional Gestalt principles, by allowing the determination of multiple groups simultaneously, and by allowing task-specific modulations where appropriate. For complex stimuli, not only separate reference-frames are needed for separate groups, but also a hierarchy of reference-frames can be established. For example, for a walker a simple interpretation is to have the lateral movement of the walker as the reference-frame and interpret all other motions relative to that reference-frame. A more detailed analysis, however, may consider a hierarchy of reference-frames: The arm moves with respect to the torso, the hand moves with respect to the arm, and the finger moves with respect to the hand. There have been models focusing on this hierarchical decomposition problem. Restle (1979) used the coding theory (Buffart et al., 1981) and considered a parametric description of moving elements in terms of physical variables such as amplitude and phase. Each parameter is assumed to contribute to the information load in processing the stimulus. Different hierarchical combinations of stimuli can result in different information loads and the goal of the coding theory is to select the hierarchical configuration that minimizes the information load. In a sense, this theory implements the "Good Gestalt" principle. Shum & Wolford (1983) adopted the same approach as Restle but used a different description of the stimulus (Fourier components). Gershman et al. (2016) analyzed the hierarchical decomposition problem by adopting a tree-structure of hierarchy and applying Bayesian

inference to this tree structure. These models do not inform on whether and how their theoretical operations are carried out by the nervous system. They may be able to explain the curved trajectories we reported here if their formulations keep the central dot as part of the Gestalt despite the differences in the velocities (cf. Johansson's the "equal simultaneous" constraint). More generally, as with other Gestalt principles, there is no all-encompassing theory of factors that determine a good Gestalt when one moves from simple to more complex stimuli. Our approach was not motivated by high-level organization processes but instead by low-level stimulus processing constraints. We argued that moving stimuli require moving reference-frames (Ogmen, 2007) and based on empirical evidence (Ogmen & Herzog, 2010), we proposed the model shown in Fig. 10 and integrated this model to the memory systems. The relative-motion computations considered in our model are low-level and are based on the activities of early motion detectors. Our model does not include form factors, nor does it take into account the hierarchy of the reference-frames. We believe that those factors will require the introduction of short-term (STM) and long-term (LTM) memories. The processing of the point-walker, for example, can be transferred to STM where it can be compared to templates from LTM feeding back to STM. This comparison process may match the stimulus pattern with learned templates and hence bring in both form-factors and hierarchy in the organization of the stimulus into Gestalts. In fact, template matching is used in Lange & Lappe (2006)'s model which uses form factors. In such a model, one can expect differences in the perception of upright versus upside-down walker, as our experience is heavily biased in terms of upright observations.

5.7. Neural correlates

Currently, there is not sufficient neurophysiological data to map our model directly with a cortical network with identified neurons. However, as we discussed in the sections that introduce the model, both the mechanisms and neural response properties associated with various stages of the model have neurophysiological support.

Whereas relative motion is computed in general, the case of biological motion, as in the point-walker stimulus, occupies a privileged position (like face processing) due to its ecological importance. Indeed, there is evidence for special brain areas devoted to biological-motion processing (Saygin et al., 2004; Pelphrey et al., 2005; Bonda et al., 1996). While our model is not specialized for biological motion, it can explain the perception of relative-motion in those displays but lacks form-related processing components. Hence, our model can be part of a larger network devoted to biological motion processing.

Based on their psychophysical study, Shioiri et al. (2002) suggested the existence of two pathways, one specialized for relative motion and the other for uniform global motion. Bex et al. (1998) provided evidence for a functional hierarchy that starts with the computation of local-motion direction and speed, followed by a global mechanism that integrates these signals according to the configuration of the local motions. Our model has this hierarchy. Our first layer, the retinotopic directionally-tuned motion detectors compute local motion direction and speed. The reference-frame synthesis layer integrates the outputs of local motion-detectors globally. The outputs of these layers feed to the next levels of the hierarchy where vector decomposition and relative-motion computations take place. The retinotopic motion-detectors are likely to correspond to directionally-

selective neurons with relatively small receptive-fields in V1 or MT. An anatomical segregation of neurons with antagonistic receptive fields and those that summate motion over large areas had been reported (Born & Tootell, 1992). Vector-decomposition/relative motion cells and reference-frame synthesis cells may fall into this

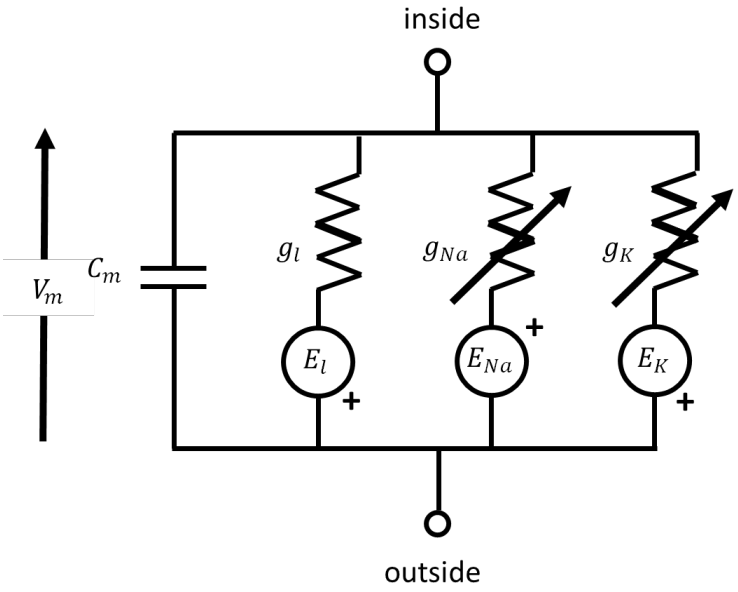

Figure 26: Equivalent electrical-circuit for the Hodgkin-Huxley model

segregation. It is possible that vector-decomposition neurons are in MT or MST. The characteristic of these neurons is that their surround inhibition comes from reference-frame direction neurons. These reference-frame direction neurons summate retinotopic motion in a directionally selective manner and thus may be directionally selective neurons with large receptive-fields. They inhibit vector-decomposition cells in a directionally-selective way thereby forming an antagonistic surround whose directional tuning is different than the directional tuning of the center. These model neurons are likely to correspond to neurons with center-surround organization with different

directional tuning. These considerations remain speculative at this point as more research is needed to directly map the architecture of our model to the functional anatomy of the visual system.

6. Mathematical Background

6.1. Multiplicative and Additive Equations of Neural Dynamics

The basic equations used in our modeling are derived from the Hodgkin-Huxley electrical-circuit model (Figure 26) for a membrane patch. The variable V_m represents the membrane potential, C_m the capacitance of the membrane, g_{Na} , g_K, and g_l correspond to sodium, potassium, and leak conductances, respectively. E_{Na} , E_K , E_l are the Nernst, or reversal, potentials for each of these channels, respectively. The differential equation corresponding to this circuit can be derived as follows:

$$C_m \frac{dV_m}{dt} = -(E_l + V_m)g_l + (E_{Na} - V_m)g_{Na} - (E_K + V_m)g_K \quad (25)$$

where *t* is time. This model was originally developed to describe how the membrane potential is controlled across a small membrane patch. Later, it had been generalized to represent the entire membrane and thus the entire neuron (with parametric variations to take into account different types of channels, etc.). For this purpose, the potential difference across the membrane patch V_m is replaced by a variable, x_i, that represents the membrane potential of the *i*th neuron, rather than just the voltage difference across a small membrane patch. Let $x_i = E_l + V_m$ so that x_i denotes the membrane potential for the neuron *i,* shifted by a constant E_l for mathematical convenience. Instead of specific ionic labels for the conductances, they are grouped into three categories: (i) those for which an increase in conductance leads to depolarization from the resting potential (cf.

g_{Na} in Equation 25), (ii) those for which an increase in conductance leads to hyperpolarization from the resting potential (cf. g_K in Equation 25), and (iii) passive, i.e., fixed conductances (cf. g_l in Equation 25). Also, the generalized model includes ligand-gated (ionotropic) channels converting synaptic inputs into post-synaptic depolarization (excitatory post-synaptic potentials, EPSPs) or hyperpolarization (inhibitory post-synaptic potentials, IPSPs). Depolarizing and hyperpolarizing conductance are then represented by excitatory and inhibitory inputs to the neuron, I_{exc} and I_{inh} , respectively. By making these substitutions into Equation 25 above, we obtain:

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)I_{exc} - (D + x_i)I_{inh} \tag{26}$$

where $A = \frac{1}{C_m} g_l$, $B = E_{Na} + E_l$, $I_{exc} = \frac{1}{C_m} g_{Na}$, $D = E_K + E_l$, and $I_{inh} = \frac{1}{C_m} g_K$.

Note that *A, B, D,* I_{exc} and I_{inh} are all non-negative. This equation has been called shunting model, and multiplicative model (Grossberg, 1988). The equations for c_u, τ_u, and $r_{i,j;u}$ are all of this type.

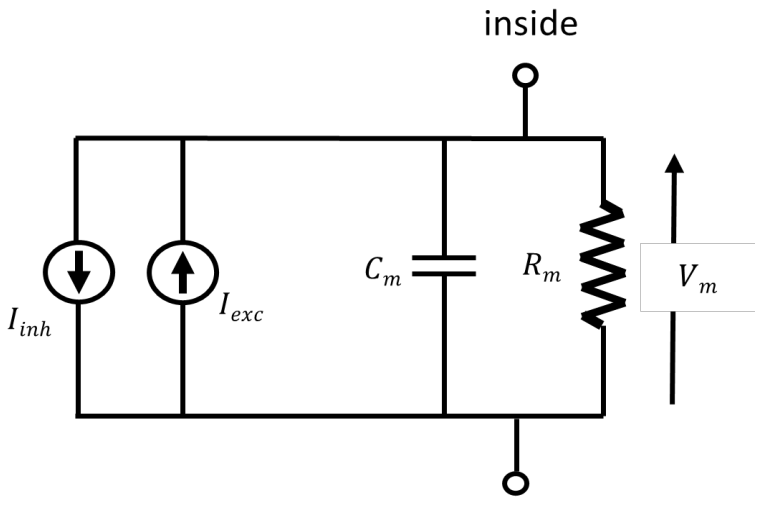

Figure 27: Equivalent electrical circuit of the additive model of the membrane potential

A simpler version of this model is called the additive or leaky-integrator model. As can be seen from its equivalent electric circuit diagram (Figure 27), voltage-gated channels are omitted and only passive channels are used, leading to a passive RC-circuit described by the simpler first-order constant-coefficient linear differential equation:

$$C_m \frac{dV_m}{dt} = -\frac{1}{R_m} V_m + I_{exc} - I_{inh} \tag{27}$$

Generalizing this membrane voltage to a neuron with activity x_i

$$\frac{dx_i}{dt} = -Ax_i + I^*_{exc} - I^*_{inh} \tag{28}$$

where $A = \frac{1}{C_m R_m}$, $I^*_{exc} = \frac{1}{C_m} I_{exc}$ and $I^*_{inh} = \frac{1}{C_m} I_{inh}$.

At steady-state, the additive equation provides a linear relationship between its inputs and output:

$$x_i = \frac{1}{A}(I^*_{exc} - I^*_{inh}) \tag{29}$$

We used this simpler equation for equations for a_u and $q_{i,j;u}$.

6.2. Winner-take-all Networks

Grossberg (1973) studied the following version of the multiplicative equation:

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i) f(x_i) - \sum_{k \neq i} f(x_k) \tag{30}$$

which can be derived from Equation 26 by setting $D = 0$, with the net excitatory input $I_{exc} = f(x_i)$, i.e., self positive-feedback. The net inhibitory input consists of the surround of feedback inhibition from other cells. Grossberg proved that (Grossberg, 1973, Theorem 2) if $f(x)$ is a "faster-than-linear" function, i.e., if $\frac{f(x)}{x}$ is a continuous, non-negative, strictly increasing function, than the asymptotic activities $x_i(\infty)$ approach a 0-1 distribution. In other words, the cell with the largest initial value will approach B in

Eqn. (30) whereas all other cells will reach 0. The system described by Eqn. (30) does not have external inputs. In our application, there are persistent inputs coming from the external world. When external inputs I_i are added

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)[f(x_i) + I_i] - \sum_{k \neq i}[f(x_k) + I_k] \quad (31)$$

The network has two tendencies: The feed-forward connections generate activities that are proportional to the inputs whereas feedback connections generate 0-1 activities. In other words, the inputs prevent the network reaching the desired 0-1 distribution. Ogmen (1993) proposed the use of habituating inputs in order to reduce the strength of input signals over time to allow feedback connections to dominate and reach the 0-1 distribution. Let $z(t)$ denote the available amount of transmitter to send the pre-synaptic signal $I(t)$ to postsynaptic receptors. As the intensity of $I(t)$ increases (corresponding to higher spike-frequency at the synapse) more transmitters will be used and depleted. A simple differential equation can express the depletion process as follows:

$$\frac{dz}{dt} = -FzI \quad (32)$$

where *F* is a positive constant (depletion rate constant). Equation 32 states that *z* is depleted proportional to its available amount and the input. A parallel process uptakes and replenishes the transmitter. This can be accounted by adding the replenishment term to Equation 32:

$$\frac{dz}{dt} = D(E - z) - FzI \quad (33)$$

where *D* and *E* are positive constants representing the replenishment rate constant and the maximum amount of *z,* respectively. The postsynaptic signal is proportional to the intensity of the input and the amount of available transmitter, i.e., $I(t) \cdot z(t)$.

As shown in Figure 13, left panel, a constant input (s_0 in the example) starts depleting its transmitter (z_0), and the resulting signal ($s_0 z_0$) consists of an initial overshoot followed by a lower plateau activity. The initial overshoot rapidly initializes the cells in the winner-take-all network and as the signal decays to a lower plateau, the feedback connections of the winner-take-all network can dominate and produce a 0-1 behavior (Figure 13, right panel).

Chapter Three: Sensorimotor Self-organization via Circular Reactions[5]

1. Introduction

Reaching for a desired target in space is a fundamental human sensorimotor ability. Many processes are involved in this ability, including stereovision to recover the position of the target in the three-dimensional space, motor control to move the arm, and visuospatial sensorimotor learning to coordinate sensory and motor representations (Mackrous and Proteau, 2016). Goal-directed reaching involves the detection and recognition of the object of interest among surrounding objects in space, determining its spatial position, and finally guiding the arm toward that position.

Goal-directed reaching has many applications in robotics and has been approached both from the perspective of physical modeling (forward and inverse arm kinematics) (Goldenberg et al., 1985; Manocha and Canny, 1994; Parikh and Lam, 2005; Mohammed and Sunar, 2015; Srisuk et al., 2017; Reiter et al., 2018) as well as from the perspective of biological-system modeling. Physical modeling approaches rely heavily on the accurate and explicit model of the arm (length of limbs, etc) and require re-calibration when physical parameters undergo unforeseen changes. On the other hand, biological systems exhibit remarkable adaptability; for example, the size of a growing child's arm changes but the brain can adapt and "automatically re-calibrate" its sensorimotor control

[5] The contents of this chapter have been published in a peer-reviewed journal: He & Ogmen (2021). Sensorimotor Self-organization via Circular Reactions, Frontiers in Neurorobotics, 10.3389.

processes. This is one reason why several researchers focused on biologically-based approaches to sensorimotor control (Saxon and Mukerjee, 1990; Asuni et al., 2003, 2006; Laschi et al., 2008; Hoffmann et al., 2017).

Approaches to biologically-based sensorimotor control are influenced by psychological theories of intelligence. According to behaviorism, the motor system produces observable behaviors which constitute the fundamental level of analysis (Graham, 2000; Sherwood and Lee, 2003). Strict behaviorism proposes that the analysis of intelligence should be based solely on observable variables, viz., stimuli and responses (behavior), without any reference to the system itself. In other words, the biological system is treated as a "black box," and learning is defined as changes in behavior as a result of two associative processes: In classical, or Pavlovian, conditioning, changes in behavior result from associating one stimulus (conditioned stimulus, CS) with another one (unconditioned stimulus, US), which is contingent on CS. In instrumental or operant conditioning, changes in behavior occur as a result of a reinforcing stimulus which is contingent on the behavior produced by the organism. This approach is exemplified with the currently popular deep-learning models that use a dataset containing inputs (stimuli) and desired outputs (reinforcement signals) and train a multi-layer network whose architecture is defined mostly in an ad-hoc manner. In contrast, constructivist theories put a central role on the internal processes of the organism that actively structures its inputs (Piaget, 1952). Hence, constructivist approaches place a central role for structural and functional properties of the organism. Our approach follows this latter theoretical perspective by incorporating modules that are inspired from the structure, i.e., functional

neuro-anatomy of the primate brain, and the functional principle of "circular reactions" (Piaget, 1952).

The primate cortex consists of two general pathways: the dorsal and ventral pathways (Goodale and Milner, 1992). These two pathways carry out complementary information processing: The ventral pathway is specialized for processing "what" information, i.e., the detection and recognition of objects. The dorsal pathway is specialized for the "where" information, i.e., the localization of objects in space. From its definition, it is clear that goal-directed reaching necessitates both the ventral (detection and recognition of the desired target) and the dorsal pathways (localization of the desired target in order to guide arm movements). The joint operation of these two pathways suggests interactions between them. Indeed, neurophysiological findings suggest that the "what" and "where" specializations are not binary exclusive properties of these pathways but are shared to some extent (Mishkin et al., 1983; Wang et al., 1999; Sereno et al., 2014). Neurophysiological studies also indicate heavy connectivity between these areas, possibly underlying their joint synergetic operations (Rosa et al., 2009; Wang et al., 2012; Van Polanen and Davare, 2015). In our model, we start with egocentric visual representations that reflect the coding of visual information in early visual areas of the cortex. Through the optics of the eye, neighboring points in the environment are projected to neighboring points on our retinas. These neighborhood relationships are preserved by retino-cortical projections. This organization is called retinotopic organization (Engel et al., 1997). The retinotopic cortical areas constitute a map representation in the sense that the location of active neurons indicates the location of the

stimulus with respect to the positions of the eyes. The eye-based representation is an egocentric map because the location is encoded with respect to the eyes of the observer. An egocentric reference frame is one that is relative to the subject, e.g., eye-, head-, body-, limb-based reference-frames. In the next stage of the model, disparity information is used to combine the two egocentric retinotopic maps into an exocentric "cyclopean map." Exocentric reference-frames are those that are relative to a reference outside the subject1. For example, in the cyclopean map, the position of an object is with respect to its position in the external world and hence its coded position does not change when the eyes move. This exocentric representation is then coordinated with motor representations by using the functional principle of circular reactions. Newborns start with genetically encoded reflexes, which consist of actions like sucking. These reflexive motor behaviors form circular reactions in that their end point becomes the beginning of a new cycle and this closed cycle repeating itself for autonomous learning and self-organization. These circular reactions allow the coordination of different senses and motor actions to guide the movements of our body (Piaget, 1952). Beginning with reflexes or random body explorations and repeating these procedures circularly, sensory and motor representations are gradually coordinated.

To model and simulate this sensorimotor self-organization, we propose and test an integrative model that combines several neural-network modules that are based on neuro-anatomical and functional properties of the primate visual system.

2 Related Work

As discussed in the previous section, several studies use the physical modeling to characterize the known structure of the arm and joints and the application of forward and inverse kinematics can be used to determine and move the arm to a desired location (Goldenberg et al., 1985; Manocha and Canny, 1994; Parikh and Lam, 2005; Mohammed and Sunar, 2015; Srisuk et al., 2017; Reiter et al., 2018). Biologically motivated studies that follow the behavioristic approach do not use a model of the arm but "learn" its structure through stimulus-response training. Our approach follows the constructivist tradition and incorporates structural and functional properties of the system, in this case the primate brain. Hence the key elements of our approach are egocentric and exocentric maps, motor vector representations, local associative coordination of maps and vectors through circular reactions.

Many studies indicated that in human, the developmental functions of brain is modulated by the sensory-motor experience (Barsalou, 2008; Schillaci et al., 2016). This skill is thought to be acquired through the active interactions with the external environment (Piaget, 1952). A typical scenario of this is the reaching behavior under the guidance of visual information, which has been widely simulated by various of modeling structures. For instance, Saxon and Mukerjee (1990) studied sensory-motor coordination by self-organizing neural networks, and created associations between an egocentric visual map and a motor map via circular reactions. The problem was simplified into a two-dimensional working space and was simulated with a robotic arm consisting of three degrees of freedom. Another study, described in Asuni et al. (2003, 2006), offered a more

developed system working in three-dimensional space and simulated with a DEXTER robotic arm. But the visual space was effectively two-dimensional since the target objects were located on a planar table. Similarly, the model learned through circular reactions and the motor system was represented by vectors. However, the objects visual locations were represented by gaze positions (vectors).

Some studies focused more on the learning mechanisms. For instance, Santucci et al. (2014) proposed a model incorporating a novel reward mechanism that used the dopaminergic neurons to strengthen the learning effect of reaching behaviors. This model was simulated on a robotic arm, which moved in a three-dimensional space, even though it was tested with target objects located on a table. The neural networks were trained via reinforcement learning where both the visual inputs and the motor system were represented by vectors. Tanneberg et al. (2019) implemented a stochastic recurrent network to refine the end-effector's motion trajectory to avoid the obstacles on the way during reaching. In their study, visual inputs were not used. In some studies more complex hand movement scenarios, like grasping were implemented (Sarantopoulos and Doulgeri, 2018; Della Santina et al., 2019).

In recent years, the scope of this research area has been expanded to a larger variety of tasks beyond vision-based reaching. For instance, Hoffmann et al. (2017) implemented reaching with tactile stimuli and incorporated a transformation between the tactile map and motor coordinates. The robot learned through self-generated random babbling and a self-touch. Laschi et al. (2008) incorporated a visual processing module that is able to predict the object's tactile properties. The model learned the reaching direction and object

orientation and mapped the arm and hand coordinations to the objects geometrical features so that it was able to predict a suitable movement to grasp the object. Chao et al. (2010) developed an ocular-motor coordination that gradually mapped the gaze space and the motor space of the ocular muscles. This study included the differential resolution found on the retina (fovea vs. periphery) and used eye movements (saccades) to bring the stimulus from the periphery to the fovea. Schmerling et al. (2015) used a robot with head motion and suggested that head-arm coordinations would improve learning. The neural network drove the goal-directed reaching of an arm of a robot and were trained through circular reactions. The objects positions were represented by head's rotation vectors and thus the study did not address how exocentric reference frames are produced. Pugach et al. (2019) proposed a "gain field" neural model where tactile information is included to establish a mapping between visual and motor spaces. Our model also uses gain field neurons and processes the motor commands through neural population encoding. This computational principle has been found to play an important role in goal-directed sensory-motor transformation (Andersen and Mountcastle, 1983; Salinas E, 2001; Pouget et al., 2002; Blohm and Crawford, 2009).

The novelty of the present model is that we provide a neurally plausible solution to the coordination between motor configurations and an exocentric map, which is generated and associated with joint vectors simultaneously. Currently, some approaches to build and expand the visual map are reported. For instance, Jamone et al. (2014) presented a strategy by which the robot learns to expand and associate the visual maps in different body positions by goal-directed reaching movements. However, the visual map

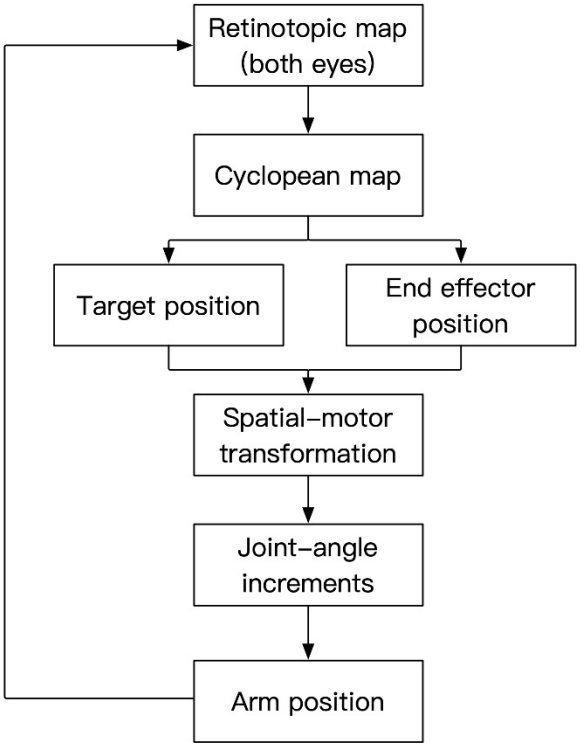

Figure 28: Processing stages of the model. Two-dimensional retinotopic maps from the left and right eyes are combined into a cyclopean map to reveal depth information. This information is used to represent target and end-effector positions in an exocentric reference-frame. The spatial position information is then converted to equivalent motor representations, which in turn drive the movements of the joints to have the end effector reach the target position. The visual input corresponding to the moving arm established a feedback loop to control the action.

in their study is not a map of neurons representing the spatial relationships among all the objects in the external environment. Instead, it is a map representing the reachability of each fixation point. The model was simulated on a robot in a three-dimensional space. Chao et al. (2016) proposed a robotic system utilizing a visual processing approach inspired from human retina. The method uses a head motor system to transform the spatial locations to the head joint vectors. In this method, spatial locations are reflected by the motor vectors instead of the inter-spatial relationships in the retinotopic map.

Other studies reproducing this method also include (Shaw et al., 2012; Law et al., 2013). In contrast to those approaches, the exocentric reference frame in our model is built by fusing two retinotopic maps and by compensating for eye movements.

3 Description of the Model

The cortical organization reflects interactive functioning of many specialized modules, anatomically corresponding to various "areas" of the cortex. In a similar way, as shown in Figure 28, our model consists of interacting modules. In this study, for simplicity, we limited senses to vision and motor control to one arm. Our model receives the visual input through its two "eyes" and encodes this information retinotopically as in human early visual areas (Engel et al., 1997). In human vision, the environment is projected on the retina through the optics of the eyes following perspective geometry. Hence, neighboring points in the environment are imaged on neighboring retinotopic positions. These neighborhood relations are preserved through the precise connections from retina to early visual cortex giving rise to the "retinotopic organization" of early visual areas. Retinotopic areas provide an egocentric map representation for the stimulus. This is called a "map" representation (Bullock et al., 1993) because the relative position of each neuron with respect to its neighbors carries information about the position of the stimulus, much like the representation of cities on a map carries information about their relative locations. It is an egocentric map because the location information is relative not only to the position of the stimulus in space but also relative to the position of the eyes. The next step in the model is to recover the position of the stimulus in space in a way that is invariant with respect to the positions of the eyes, i.e., an exocentric representation.

Hence, at the next stage, the two retinotopic maps are fused by using the binocular disparity information to build a "cyclopean map" (Julesz, 1971). The arm position is represented by neurons that encode joint angles in a vector format. In this "vector representation" (Bullock et al., 1993), each group of neurons is associated with a joint and the activities of these neurons encodes in an analog way the joint angle, e.g., the higher the activity the larger the joint angle. Synaptic connections between sensory map-representations and motor vector-representations allow the coordination of these activities through circular reactions. To initiate the circular reaction, we send a random command to joint angles which then moves the arm accordingly. As the arm moves, its image is represented in the retinotopic maps, creating a visuo-motor feedback loop. Through this self-generated action, the system activates motor and sensory representations and these simultaneous activities provide the input for associating sensory and motor activities that are congruent with the physics of the external world. This way, we do not need to incorporate physical models of the arm, eyes, etc., the system learns the relationships of the joints, limbs, eyes, etc. by perceiving the consequences of self-generated actions. An important implication of this type of learning is that the system does not need explicit models, parameters but constantly adapts and recalibrates through action-perception-learning loops. Hence, self-organization, adaptation, and re-calibration are emergent properties of this approach. The synaptic plasticity of the connections between these sensory and motor representations coordinates them through associative learning. Once these coordinations are learned, a target position can predict the corresponding associated joint angles for the arm to reach the target and vice-versa.

3.1 Visual Processing: Retinotopic and Cyclopean Maps

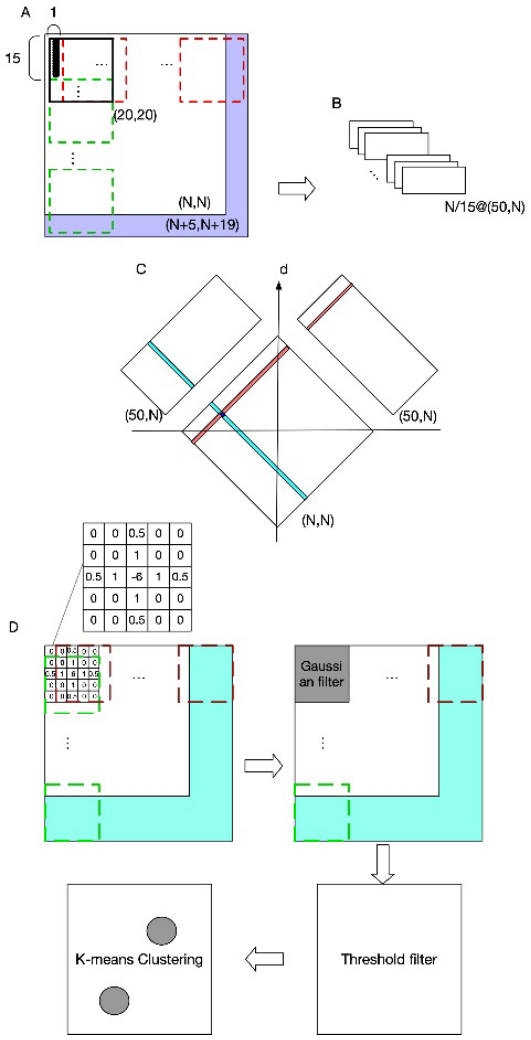

Figure 29: Visual processing. (A) The purple area shows zero pads. The red dashed frame indicates transformer's spatial extent along columns and the green dashed frame along rows. (B) Spatial frequency response maps (SfRM) with size of (50, N). Each row of UA is projected to a SfRM, within which each column is the frequency response of an UA. Therefore, a LumM possessing N/15 rows and N columns of UA creates N/15 SfRM in size of (50, N). A SfRM has 50 rows because 50 frequency responses are obtained, and N columns come from the corresponding N columns of UA. (C) Binocular correspondence map (BCM). Frequency-based comparisons determine the correspondence relations between the left and right UAs, and also their depth. (D) Spatial localization of the target object and the end-effector.

A central feature of our model is that we process the egocentric (retinotopic) maps from the stereopair image to produce an internal visuomotor spatial representation-the cyclopean map (CPM) (Julesz, 1971), serving as an exocentric map to guide reaching activities. Our model implements this process in three steps: (1) coordinate the stereopair mapping; (2) compensate the binocular disparity and expand the depth dimension; (3) synthesize the luminance profile by a weighted sum of the luminance profiles of the stereopair after disparity-compensation.

3.1.1 Coordination of Retinotopic Mappings

A stereopair consists of two retinotopic images that have mostly horizontally-shifted luminance profiles. This is because, given the horizontally displaced positioning of the two eyes, a point in the three-dimensional environment is projected onto horizontally-shifted locations in the left and right retina. This relative spatial displacement is called binocular disparity. Our model determines the binocular disparity information according to the criterion of spatial-frequency similarity. This process can be intuitively described as follows: The retinotopic maps are firstly demarcated into non-overlapping unit areas (UA) with uniform shapes. Thereafter, a spatial frequency filter that is slightly larger in size is applied to all the unit areas to obtain the response signals of each of them. After this step, a UA in one image can then be matched to another UA in the other image if they have the most related spatial frequency response across all areas. Importantly, the filter is larger than those unit areas, which means that the correspondence among unit areas are determined with considerations of not only the unit area itself but also its

neighbors. In the following illustrations, we use$(a, b)$to indicate the size of neural maps, where “a” is the number of cells in rows and “b” in columns.

Figure 29 shows the visual processing beginning with a retinotopic map, with a size of (N, N), and it is firstly converted to a luminance map (LumM) by transforming the colored image to a gray scale image. This LumM was demarcated into $N * N/15$ UA with a size of $(15,1)$. This UA can be thought as the resolution by which binocular disparity and depth are determined. Then spatial-frequency filters corresponding to multiple frequency-channels are applied to LumM. Here we used a two-dimensional Fast Fourier transformer with 50 frequency responses (from 1 to 50 $cycles/deg$). LumM was zero-padded to$(N + 5, N + 19)$ and then scanned by a transformer in size of$(20,20)$, which covered a UA and its surround. As a result, $N/15$ spatial-frequency response-maps (SfRM) of size $(50, N)$ were obtained. $N/15$ Binocular Correspondence Maps (BCMs) in (N, N) were then generated using Equations (34) and (35), where $SfRM_{l,i}(j, k)$ represents $k\ cycles/deg$ spatial-frequency response of an area registered by a UA in ith row and jth column of left retinal LumM (replace l by r to represent right retina). In Equation 34, the squared difference between the frequency-contents of the corresponding left- and right-eye patches are computed and its minimum provides the best matching binocular pair in terms of frequency contents. $BCM_i(x, y)$ was used to indicate the correspondent pairs in ith row of left and right retinal LumM. For a row i and each x from 1 to N,

$$y = \min_j \sum_{k=1}^{50} \left(SfRM_{l,i}(x, k) - SfRM_{r,i}(j, k)\right)^2 \quad (34)$$

$$\begin{cases} BCM_i(x,y) = 1 \\ BCM_i(x,\hat{y}) = 0,\ \forall \hat{y} \notin y \end{cases} \tag{35}$$

According to Equations 34 and 35, each activate cell in a BCM encodes a binocular-correspondence relationship between a UA on left retinal LumM and another one on the right.

3.1.2 Disparity Compensation and Depth Recovery

The depth of a UA can be determined by its representative neuron's relative position on the BCM map with respect to the d axis as shown in Figure 29C. In other words, on BCM, a UA's representative neuron's position along d will be equal to the depth of activated neurons when projected onto CPM. Since there are N placeholders along the depth dimension, the size of CPM is(N, N, N). This depth recovery approach follows Hirai and Fukushima's neural network model for extracting binocular parallax (Hirai and Fukushima, 1978). This process can alternatively be explained by the existence of a group of “binocular depth neurons” that are selectively sensitive to a binocular stimulation with a specific amount of parallax. Take the activated neuron in Figure 29C for example, this neuron in BCM becomes active only when it receives simultaneous stimulation of a UA whose position in the column is indicated by teal color and the UA marked by red. In fact, “binocular-depth neurons” have been found in many species including monkey and mouse (Poggio et al., 1988; La Chioma et al., 2020).

3.1.3 Binocular Combination

Stereo image pair's luminance profiles are combined by summing them with a pair of physiologically plausible weights defined by a simplified version of Ding-Sperling model (Ding and Levi, 2017). This model was built based on the principle that each eye uses a

gain-control on the other eye's signal in proportion to the contrast energy of its own input. After this procedure, the fused luminance profiles are then filtered and clustered based on the contrast change of the luminance. According to Ding-Sperling model, each spot of the luminance map is allocated with respect to the contrasts of that spot from the two eyes. This means that contrasts will be rebalanced toward the eye carrying the larger contrast energy. This contrast rebalancing is used, because simply taking the luminance distribution from single eye, without the contrast rebalance, weakens the effects of contrast-based clustering. It deteriorates the precision of cluster center localization when two objects are spatially closed. In this model, for each row, the total luminance I0 was determined by Equation 36, in which x is the number of the column, and I indicates the luminance. The contrast at x, $C(x)$, is determined by Equation 37. Equation 38 provides the definition of the spatial-frequency filter$LoG_i(x)$. The contrast at ith spatial frequency $C_i(x)$ can be calculated by the convolution between ith spatial frequency filter and luminance profile of this row. The contrast energy in the ith channel, ϕ_i, can then be calculated by Equation 39. The total contrast energy Φ is defined as the sum of contrast energies for all spatial frequency channels. And the total luminance energy $\mathcal{L}$ is determined by Equation 40. Φ_L and $\mathcal{L}_L$ are the total contrast and luminance energies for a row in a left retinal UA, and energies for right retinal UA are given byΦ_R and $\mathcal{L}_R$. The weights w_L and w_R, which are used to combine a left retinal UA and its paired right UA, are calculated using Equation 41. The luminance profile of CPM is then determined by summing all UAs on the left $LumM$ with their corresponding UAs on the right $LumM$ according to these weights.

The above processes can be abstracted and summarized as follows: A UA on the left retina is firstly matched with a corresponding UA on the right retina based on the similarity of their spatial-frequency contents. These two UAs are assumed to be projected from the same environmental stimulus. These two UAs of size $(15,1)$ are then projected to 15 neurons on the CPM, where their relative locations indicate their spatial position in the environment. The output of these 15 neurons on CPM are determined by weighted sum of two UAs' luminance profiles.

$$I_0 = \sum_x I(x) \tag{36}$$

$$C(x) = \frac{I(x+2)+I(x)-2I(x+1)}{I_0} \tag{37}$$

$$LoG_i(x) = -\frac{1}{\pi\sigma_i^2}\left(1-\frac{x^2}{\sigma_i^2}\right), \sigma_i = 5*i, i = 1,2,3,\dots,10 \tag{38}$$

$$\begin{cases} C_i(x) = \frac{1}{I_0} LoG_i * I(x) \\ \phi_i = \sum_x \left(w_c(x)\cdot C_i(x)\right), w_c = \frac{1}{1+x^2} \\ \Phi = \sum \phi_i \end{cases} \tag{39}$$

$$\begin{cases} \mathcal{L} = \sum_x \left(w_{lum}(x)\cdot I(x)\right) \\ w_{lum} = \frac{1}{1+x^2} \end{cases} \tag{40}$$

$$\begin{cases} w_L = \frac{\Phi_L \mathcal{L}_\mathcal{L}}{\Phi_L \mathcal{L}_\mathcal{L} + \Phi_R \mathcal{L}_R} \\ w_R = \frac{\Phi_R \mathcal{L}_R}{\Phi_L \mathcal{L}_\mathcal{L} + \Phi_R \mathcal{L}_R} \end{cases} \tag{41}$$

3.2 Object Recognition and Localization on the CPM

The discharge of neurons in CPM reflects the luminance information from stereovision. The task requires the model to recognize an object's or end-effector's

luminance patterns and return their locations within the CPM. The heuristic used here assumes that an object is usually observed by a closed contour, and the center of this contour might be used to represent the location of this object. Since the focus of this work is on sensorimotor coordination rather than complex object recognition, the more complex cases of occlusion and boundary ownership are not taken into account (Heydt et al., 2003; Layton and Yazdanbakhsh, 2015; Dresp-Langley and Grossberg, 2016). To achieve our goal, we first applied a center-surround filter and convolved CPM's luminance profile regardless of depth, as shown in the grid in Figure 29D, where the filled pattern is an example and corresponds to what we actually used in the experiments. This filter is able to reduce the noise and enhance the edges. After that, a Gaussian filter with size of (5,5) was used to suppress the noise in the map. The Gaussian filter is defined by

$$G_i = \frac{\exp\left(-(d_i-2)^2\right)/2.42}{\sum_{i=1}^{25} G_i} \tag{42}$$

where i is the index of values on the filter kernel and can be from 1 to 25, and di is ith value's position on the kernel relative to the center and can be 1, 2, or 3. After this, a threshold filter eliminated the discharges of all neurons on the map whose discharges represented luminance values below $130\ cd/m^2$. The resulting map was tested to be clean enough for object localization based on clustering approaches. Luminant spots emitted from an identified object are clustered with the same label through K-means algorithm, and the returned objects' cluster centers are used as their position in CPM.

3.3 Motor Planning: Neural Networks

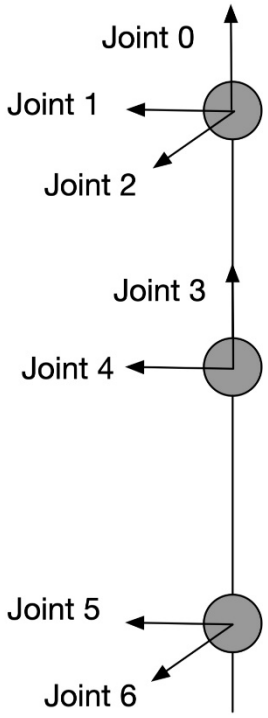

Figure 30: Kinematic information of the arm model. This is a 7-DOF arm with each joint and its rotation axe shown in the figure. All the angular rotations in this paper are in terms of counter-clockwise except for joint 2 which works clockwise.

Visual processing and recognition stages compute the spatial information of the target-object and the end-effector and encode this information as the position of discharging neurons in the cyclopean map. Our model connects the neurons in the cyclopean map via adaptive synaptic connections to motor neurons to guide the movement of a humanoid arm with 7 degrees of freedom (DOF) and let the end-effector (wrist) reach arbitrary positions that are reachable and visible. The arm model is shown in Figure 30. Interactions of two functionally complementary subsystems are needed to process this: one subsystem controls the upper limb (position controller) and the other

subsystem adjusts postures in response to the environmental conditions (posture controller). This mechanistic property is in line with physiological findings showing that two main systems in human parieto-frontal networks play a major role in visually guided hand-object interaction (Lega et al., 2020). These two systems are associated with controlling upper-limb positions and with coding hand postures, respectively.

3.3.1 Position Controller

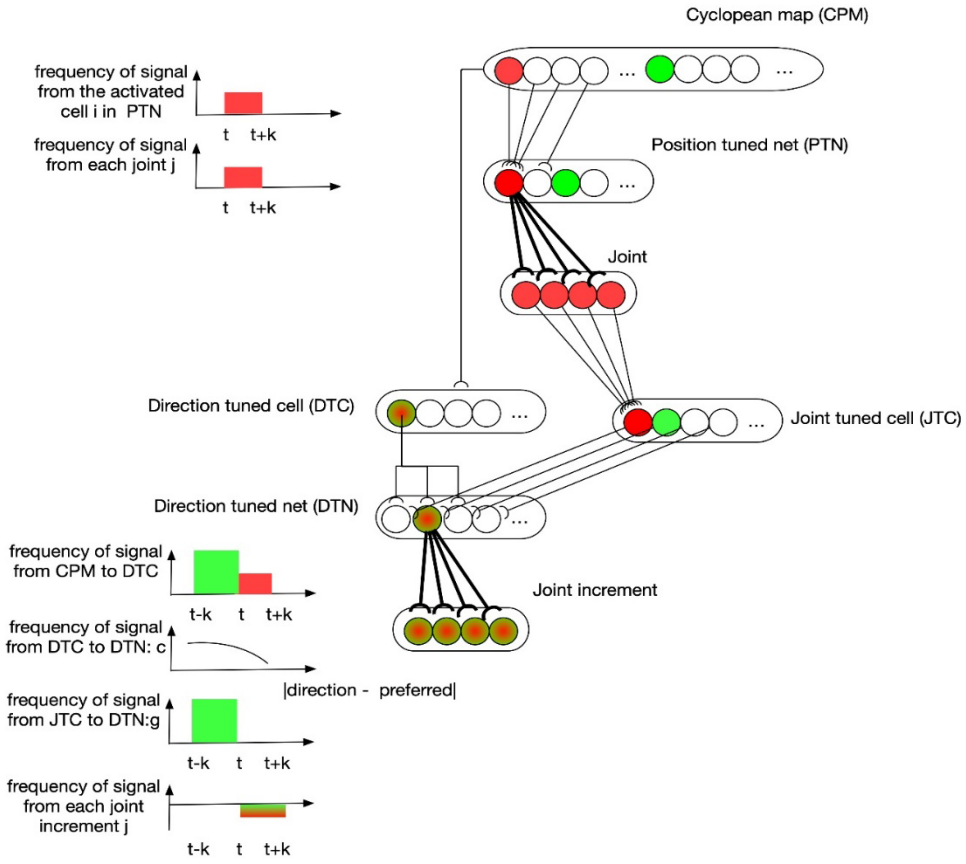

Figure 31: Position controller: learning of neural networks. At the time t-k, the visual information of the end-effector activates the green cell in the CPM. Then the end-effector is moved to another position represented by the red cell in CPM at time t. A cell in PTN down-samples the CPM by selectively receiving signals from a group of spatially-neighboring neurons in the CPM. At the same time, the state of arm joints stays in response to the red PTN cell, and this simultaneous activation of both PTN cell and joint cell strengthens the synapses between them through associative-learning rules (bolded connections between PTN and joint neurons in the figure). In the subsequent motion of the end-effector, the direction of motion is captured by some DTC cells, as shown by a green-red cell for example, because this direction lies in their receptive fields. Their discharges are dependent on the angular difference between their preferred direction and the perceived direction. A cell in DTN receives two signals, one from a DTC and the other from a JTC, leading to a selective tuning for a specific joint configuration. The number of cells in DTN equals the product of the number of cells in DTC and JTC so that DTN captures all possible situations. A DTN cell's discharge is equal to the multiplication of two signals it receives. At time t, DTN cells receive JTC's signals in green state and the DTC signals capturing the end-effector's motion direction. The synapses between activated DTN cells and the joint increments resulting in this motion are then learned accordingly.

The position controller works as a modified version of Grossberg-Bullock reaching model (Bullock et al., 1993), where direction-tuned neural networks code a 3-DOF arm to successfully reach spatial targets with a satisfactory error on a 2D working surface. Here we expand it with an additional dimension and make the modified model capable of taking spatial information from CPM and to send movement control signals to arm's joints 0, 1, 2, and 4 accordingly. Two neural networks are embedded: a position-tuned net (PTN) possessing neurons sensitive to specific spatial zones in CPM; and a supplementary direction-tuned net (DTN) possessing neurons sensitive to specific ranges of spatial direction vectors. In practice, PTN generates the first command every time when a new target appears, followed by DTN which corrects the arm's positions based on the spatial vectors from the end-effector to the target.

An example describing how learning takes place in these two neural nets is depicted in Figure 31. At the time t-k, the visual information of the end-effector activates the green cell in the CPM. Then the end-effector is moved to another position represented by the red cell in CPM at time t. A cell in PTN down-samples the CPM by selectively receiving signals from a group of spatially-neighboring neurons in the CPM. At the same time, the state of arm joints stays in response to the red PTN cell, and this simultaneous activation of both PTN cell and joint cell strengthens the synapses between them through associative-learning rules (bolded connections between PTN and joint neurons in the figure). In the subsequent motion of the end-effector, the direction of motion is captured by some DTC cells, as shown by a green-red cell for example, because this direction lies in their receptive fields. Their discharges are dependent on the angular difference

between their preferred direction and the perceived direction. A cell in DTN receives two signals, one from a DTC and the other from a JTC, leading to a selective tuning for a specific joint configuration. The number of cells in DTN equals the product of the number of cells in DTC and JTC so that DTN captures all possible situations. A DTN cell's discharge is equal to the multiplication of two signals it receives. At time t, DTN cells receive JTC's signals in green state and the DTC signals capturing the end-effector's motion direction. The synapses between activated DTN cells and the joint increments resulting in this motion are then learned accordingly.

Mathematically, using PTN_{ij} to represent the synaptic weight from a PTN cell i to a joint $j(j = 0,1,2,4)$, the learning rule for PTN is described by Equations 43 and 44. With a desired position pos^*, PTN generates a motor command by Equation 45, where $\hat{p}_i$ is cell i 's sensitive position zone in BCM, epos is the end-effector's position in BCM, θ_j is joint j's position in degrees. η was 0.5 in our simulation.

$$p_i(t) = \begin{cases} 1, if\ \ epos(t)\ \ in\ \ \hat{p}_i \\ 0, otherwise \end{cases} \tag{43}$$

$$PTN_{ij}(t+1) = PTN_{ij}(t) + p_i(t) \cdot \Big((1-\eta) \cdot \theta_j(t) + (\eta - 1) \cdot PTN_{ij}(t)\Big) \tag{44}$$

$$\theta_j(0) = PTN_{ij}, i: pos^*\ in\ \hat{p}_i \tag{45}$$

Using $DTN_{i,j}$ to represent the synaptic weight between a DTN cell i to a joint j $(j = 0,1,2,4)$, the learning rule for DTN is given by Equations 46–51, where $adif(v_1,\ v_2)$ is the angular difference between two vectors v_1 and v_2, $\hat{v}_i$ represents the spatial direction range for which cell i is tuned, $\hat{\theta}_{i,j}$ is cell i's sensitive angular range of joint j, t is time or step in the learning and testing dynamics, $c(v, v^*)$ is a direction-tuned neuron's tuning

curve given a stimulated vector v and its selective vector v^*, and $a \rightarrow b$ represents a spatial vector defined by position a and b with a direction from former to latter.

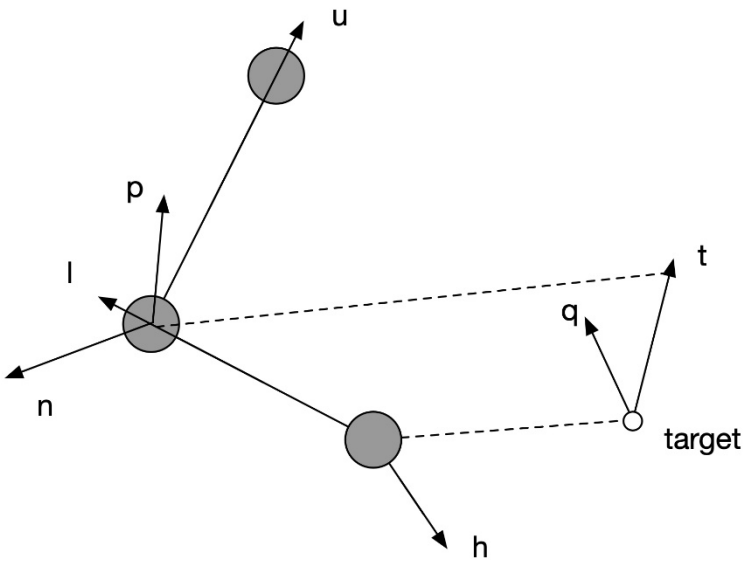

Figure 32: The explanation of spatial relations by which the posture controller works. Vectors $\vec{u}, \vec{l}, \vec{h}$ are along the spatial directions of upper arm, lower arm and hand, respectively. Vectors $\vec{n}$ and $\vec{p}$ are the normal vectors of planes defined by $(\vec{u}, \vec{l})$ and $(\vec{l}, \vec{n})$, respectively. Vector $\vec{t}$ represents the desired direction when the end-effector reaches the target. $\vec{q}$ is the normal vector of the plane defined by $(\vec{l}, \vec{t})$.

After enough learning, DTN is expected to have the ability to drive the arm to move the end-effector to a desired position pos^* by Equation 51. In the above functions, both γ and ρ are step-sizes and δ is a parameter of regularization. β represents the tuning activity of DTN neurons. In our simulation, we used $\gamma = 1, \delta = 0.1, \rho = 0.05$, and $\beta = 0.001$.

$$dv(t) = epos(t) \rightarrow epos(t+1) \tag{46}$$

$$\Delta\theta_j(t) = \theta_j(t+1) - \theta_j(t) \tag{47}$$

$$c(v, v^*) = \begin{cases} \exp\left(-\beta \cdot adif(v, v^*)^2\right), if\ adif(v, v^*) < 90^\circ \\ 0, otherwise \end{cases} \tag{48}$$

$$g_i(t) = \begin{cases} 1, if \quad \theta_j(t) \quad in \quad \hat{\theta}_{i,j} \quad for \quad j = 0,1,2,4 \\ 0, otherwise \end{cases} \tag{49}$$

$$DTN_{ij}(t+1) = DTN_{ij}(t) + \gamma \cdot c\left(\hat{v_i}, dv(t)\right) \cdot g_i(t) \cdot \left(\Delta\theta_j(t) - \delta \cdot DTN_{ij}(t)\right) \tag{50}$$

$$\theta_j(t+1) = \theta_j(t) + \rho \cdot DTN_{ij}, i: \begin{cases} \theta_j(t)\ in\hat{\theta}_{i,j}\ for\ j = 0,1,2,4 \\ \min 1_i\ adif(\hat{v_i}, epos(t) \rightarrow pos^*) \end{cases} \tag{51}$$

3.3.2 Posture Controller

Figure 32 explains the functioning of the posture controller, which adjusts the orientation of the palm to parsimoniously reconcile the environmental requirement defined by specified geometrical relations. Vectors $\vec{u}, \vec{l}, \vec{h}$ are along the spatial directions of upper arm, lower arm and hand, respectively. Vectors n⟶ and p⟶ are the normal vectors of planes defined by $(\vec{u}, \vec{l})$and$(\vec{l}, \vec{n})$, respectively. Vector t⟶ represents the desired direction when the end-effector reaches the target. q⟶ is the normal vector of the plane defined by $(\vec{l}, \vec{t})$. The rotation of each joint is defined by the angular difference of a pair of vectors as shown in Equation 53. The rotation of joint 3 in degrees was defined to be equal to the angular difference between $\vec{p}$ and $\vec{q}$, so that when end-effector reached the target, the plane of the palm could be perpendicular to the plane formalized by lower arm and target direction. After this, joint 6 can be rotated to align the palm to the target direction by a degree equal to the angular difference between $\vec{l}\ and\ \vec{t}$. However, due to limited rotation capacity of all the joints, the hand might not perfectly align with the target's direction; so joint 5 can further adjust the hand's direction by rotating by a degree equal to the angular difference between $\vec{h}$ and $\vec{t}$ at last to make the alignment as close as possible. However, each joint is limited in its range of rotations. If the desired rotation

angle exceeds their limit, they rotate up to the maximum or minimum of the range, as shown by Equations (52) and (53).

$$\begin{cases} \vec{n} = \vec{u} \times \vec{l} \\ \vec{p} = \vec{l} \times \vec{n} \\ \vec{q} = \vec{t} \times \vec{l} \end{cases} \tag{52}$$

$$\begin{cases} joint\,3 = adif(\vec{p}, \vec{q}),\ joint\,3 \in [0{,}105°] \\ joint\,5 = adif(\vec{h}, \vec{t}),\ joint\,5 \in [-15°, 15°] \\ joint\,6 = adif(\vec{l}, \vec{t}),\ joint\,6 \in [-50°, 50°] \end{cases} \tag{53}$$

4 Biological Evidence

The learning in this model is autonomous, unsupervised, and local. The model autonomously generates movements, by which the activities in sensory and motor representations can be associated to predict each other. This learning procedure is inspired by multiple studies suggesting infants acquire spatial and motor knowledge and their associations by self-exploration and object manipulation (Needham, 2000; Soska et al., 2010; Schwarzer et al., 2013; Soska and Adolph, 2014). Like infants, the model learns autonomously in an unsupervised manner. The learning equations are not based on error-correction following teacher-provided learning targets but rather on associative learning, which simply associates correlated sensory and motor activities. Learning feedback is provided directly by the environment through action-perception loops. This type of sensorimotor organization is believed to underly the more abstract concepts of space. For instance, some investigators found that, after object exploration, infants' performance in mental spatial imagination is improved, which suggests an importance contribution from exploration experience to spatial development (Slone et al., 2018). The sensory representation herein is projected in CPM created by two retinal luminance maps.

This stereoscopic sensing recovers the depth information through a specified geometrical definition based on binocular disparity and enables the motion detection in three-dimensional aspects. This three-dimensional motion sensing has been investigated using various of paradigms including direction selectivity, temporal resolution and changing size, etc. (Beverley and Regan, 1974; Gray and Regan, 1996; Portfors and Regan, 1997). Among those studies, Beverley recorded electrical brain responses to stimuli in motion along the depth-axis and found these responses to be different with respect to different binocular disparities. The explicit map representations in the model allow local learning. Maps provide a representation for space and sensitivity-zones (e.g., Equation 44) determine local regions in this space. For example, learning for PTN cells occur only when the end-effector is within their local sensitivity zone (Equations 44 and 45). That way highly nonlinear relationships across the entire space can be simplified by local approximations, resulting in a much simpler learning approach.

One important technique used in this model is to coordinate spots on two retinas by spatial-frequency similarity. The spatial-frequency channels in human vision and their psychometric functions are well known (Sachs et al., 1971). On the motor-control side, our model has neurons selectively tuned to different spatial directions, abstracting neurons identified in the primary motor cortex (M1). In one study that reported recordings from monkey's motor cortex, researchers found cells that code the direction of movement in a way dependent on the position of the arm in space (Caminiti et al., 1990). Similarly, DTN in our model combine information from both arm configuration and

motion direction. More recent studies reported similar neurons found in human cortex M1 (Tanaka et al., 2018; Feldman, 2019).

5 Experiments

5.1 Platform and Simulation Procedures

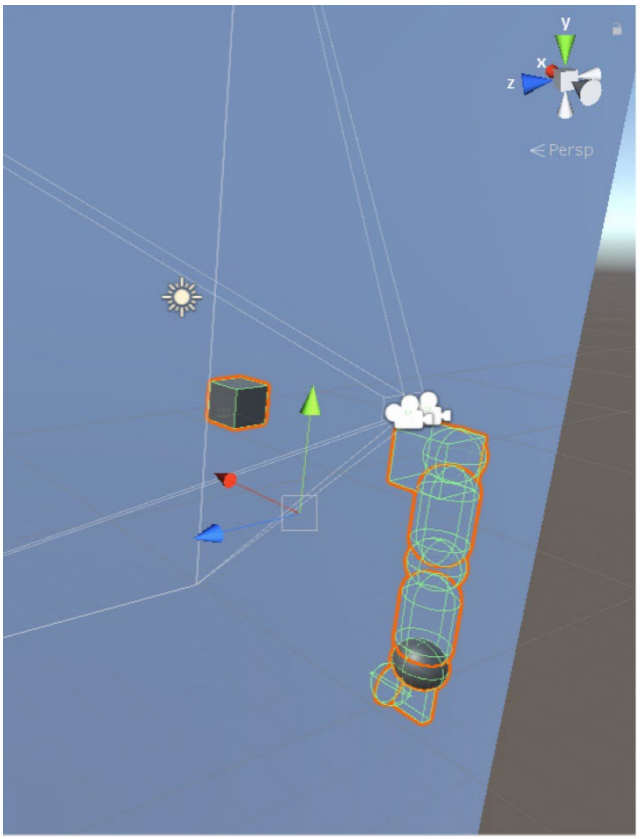

Figure 33: Simulation environment in Unity3D.The axes indicated by the green and the red arrows are aligning with the axes marked by joint 2 and joint 0 in Figure 30, respectively. The blue arrow is pointing inward the shoulder while joint 1 in Figure 30 is pointing outward the shoulder.

Our model was simulated on Unity3D, where we programmed the objects in the environment to make the arm work following the kinematic rules described in this paper. We calibrated the measurements by assuming that a unit scale in Unity3D scene is equal to 10 cm. As shown in Figure 33, a pair of cameras were placed both in 10 cm upward

than the center of shoulder joint and 9.5 cm and 14.5 cm leftward than the center of shoulder joint, respectively. These cameras took pictures at a resolution of 300*300 pixels. The gaze position of two eyes was 120 cm forward, 15.7 cm rightward, and 13.7 cm below the left eye. We used capsules with a diameter of 10 cm for upper and lower limbs and spheres with equal diameter as the joints connecting two equally long arm limbs. The end-effector herein was the wrist that is the only visible body part. We kept other limbs transparent due to the requirement of precise prediction of end-effector's

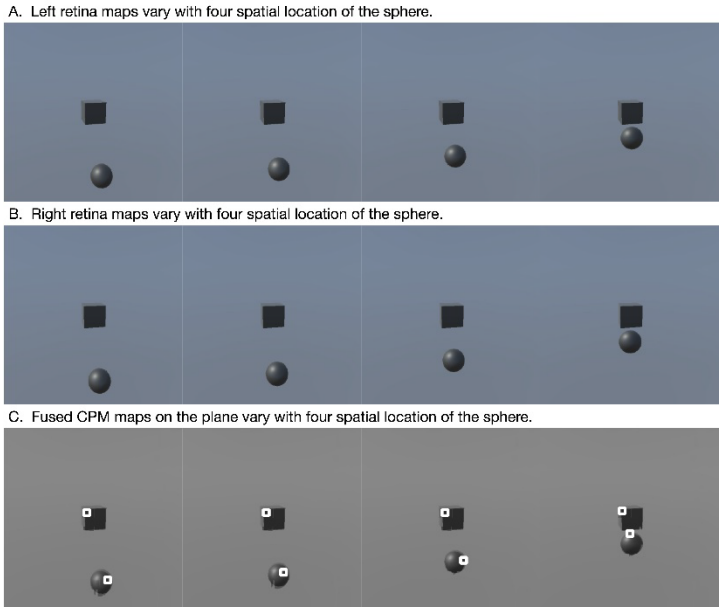

Figure 34: Examples of visual localization. This figure contains three rows and four columns. Each column indicates an example, and three rows are left retina map (A), right retina map and fused CPM's projection on the plane (B, C). In the third row, two white circles are marked on two objects, respectively. These marks indicate the spatial location of two objects determined by the model.

spatial position during visual processing. We also placed a cube with 10 cm long, 5 cm wide and 2 cm thick to serve as the palm. The lengths of limbs were 28 and 10 cm, respectively.

We trained the model using a uniformly distributed random-variable that generated self-exploratory movements. The neural networks embedded in the model learned spontaneously in the way described. The model was tested by using a black cube as target with sides 10 cm long placed randomly somewhere within the view range of “eyeballs” at the beginning of each trial. The system was then expected to deliver the end-effector to contact the cube, and when it reached the cube, the palm was expected to point in the forward direction as the red arrow axis shown in Figure 33.

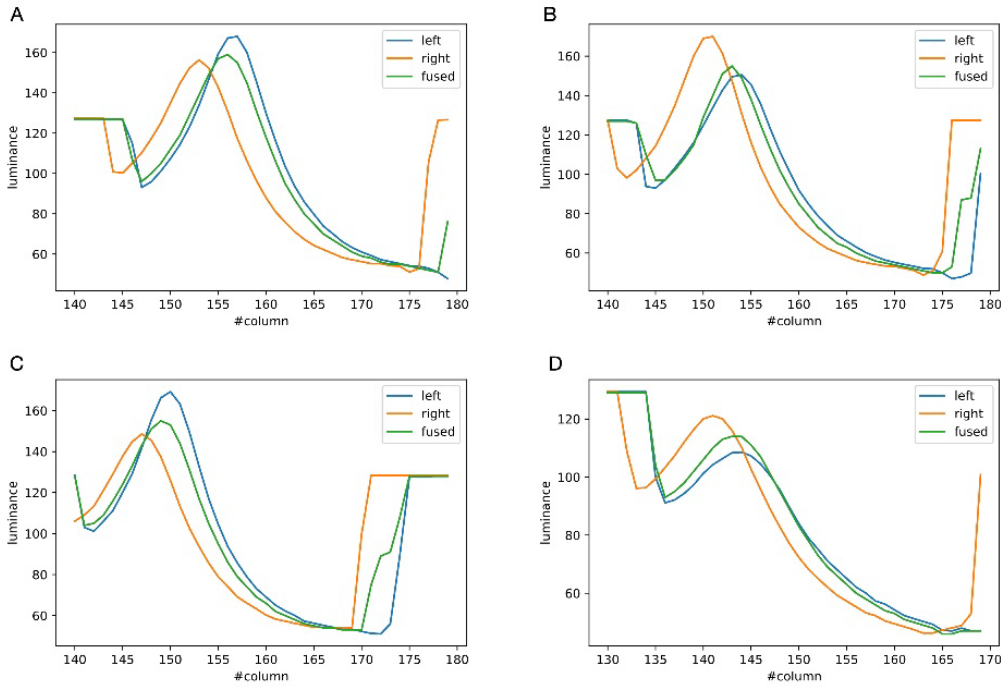

Figure 35: Examples of binocular combination. The four (A–D) show the luminance profiles of a single row where the four black spheres in Figure 34 are located, respectively. Each figure shows the luminance profiles from the left retina, the right retina, and the fused map.

5.2 Validation of Visual Processing

In this part, we show the feasibility of the visual processing methods by an example containing three movements. As shown in Figure 34, two cameras representing two

eyeballs were fixed to the center of the black cube. The cameras were kept stable, while the black sphere was moving for three steps. Eight pictures from four states and two retinas were captured. First, we picked a row of the pixels and plotted its luminance with respect to the index of columns for each of the pictures in Figure 34. As shown in Figure 35, binocular disparities co-varied with the four motion states of the black sphere, as indicated by variations in the misalignments between luminance curves from left and right retinas. As indicated by Equations 34 and 35, this model is dominated by the left

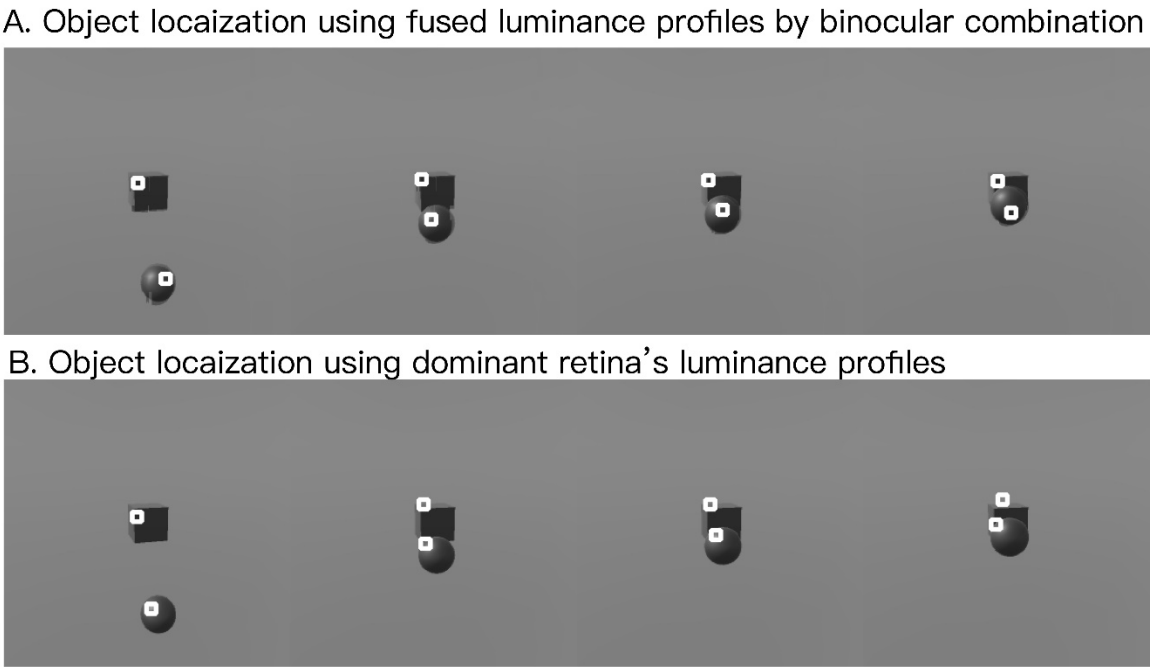

Figure 36: Effects of clustering in binocular combination. The upper row shows the clustering after binocular combination (A), and the lower row shows the clustering using the dominant eye's luminance profile (B). White circles mark the cluster centers.

eye as the luminance profiles are compensated from the right eye to the left eye. The fused profiles' shape is in line with the dominant eye in terms of spatial locations, while the contrast change and absolute luminance value are both rebalanced taking both eyes

into account. Second, the object localization presented with the white circles in the bottom row of Figure 34 reflects their real spatial information. It can also be found that, with illumination unchanged, objects' recognized centers are locally stable. Even when two objects are close to each other, the model is able to differentiate them. Third, Figure 35 shows examples about the effect of contrast-related weighted summation of binocular luminance profiles. In Figure 36, we compared binocular combination with dominant vision in the results of localization. As discussed, Ding-Sperling's model of contrast rebalance optimizes contrast-based localization. Specifically, this method localizes the object on the edges and corners, or on the zones of chiaroscuro of smooth surface. This also stabilizes the localization of objects in motion, as well as when two objects are spatially close. In contrast, monocular vision is more sensitive to the absolute luminance value as objects tend to be localized on the light spot. When two objects are close to each other, the cluster center of the black cube is marginalized, as shown in Figure 36.
Last, Figure 37 shows the three-dimensional spatial vectors indicating these three motion directions in Figure 34. Putting the maps in Figure 34 into the 3D coordinates shown in Figure 37, the depth value of the cube is smaller than that of the sphere. The directions along the depth axis are reconstructed, which coincide with the real motion as the black sphere is moving toward the black cube.

5.3 Experiment 1: Reaching with Fixed Gaze Position

In this experiment, we fixed the gaze position as where it was initialized. We tested the functions of each module as well as the neural networks. In this case, the objects and the arm were all lying on the peripheral retina during the test session. We reported the

performance in terms of the cartesian error between the wrist and the target, and the angular difference of the hand posture and the target orientation.

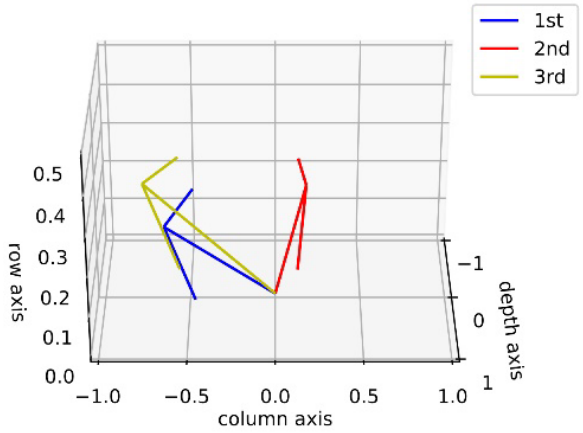

Figure 38: Examples of three-dimensional motion direction. This figure shows three spatial vectors along the column axis, row axis and depth axis, correspondent to the three movements represented in Figure 34.

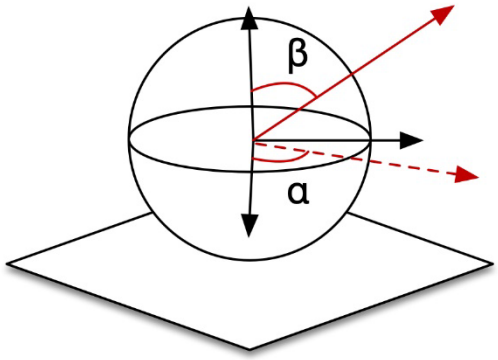

Figure 37: Definition of spatial vectors.

5.3.1 Experimental Parameters and Learning

This system possesses two trainable neural-networks, TPN and DTN, where neurons are sensitive to positional zones in CPM, angular zones in arm joints or spatial directions in CPM. To implement in simulation, we demarcated the CPM of size$(300,300,300)$ into 2.7×10^4 cube zones of size $(10,10,10)$ as the $\hat{p}$ in Equations 51–53. The ranges of

joints driving the end-effector were $[0°, 90°]$, $[60°, 120°]$, $[0°, 25°]$, and$[0°, 90°]$ for joints 0,1, 2, and 4, respectively. We used 6° as the interval to demarcate this hyperspace into $9{,}000(15 \times 10 \times 4 \times 15)$ angular zones, $\hat{\theta}_i$. Spatial vectors are represented using polar coordinates (α_i, β_i) here as shown in Figure 38. We defined 180 selective vectors, $\hat{v}_i$ in Equation 51, by $\alpha_i = 20°,\ 40°,\ 60°, \ldots, 340°,\ 360°$ and $\beta_i = 20°,\ 40°,\ 60°, \ldots, 160°,\ 180°$. Therefore, DTN contains $1.62 \times 106(9{,}000 \times 180)$ cells because of its sensitivity to both spatial direction and arm position. PTN contains 2.7×104 cells each selectively becoming active for a unique zone in CPM. In the learning session, a total amount of 2.916×107 steps of self-exploration were implemented to drive the learning and self-organization processes of the two neural networks. We additionally tested the performance when the system was trained with 1/3 and 2/3 of the total amount. We also trained the system with an additional noisy condition, where a random noise in the range of 0 to 5 degrees was added to each of the joints.

5.3.2 Tests and Results

In the testing session, the system took six groups of tests, within which each group contained 50 trials. Groups of tests are distinguished by three different amounts of learning and two conditions (with or without noise) to test robustness. In each trial, the system operated in the way described above and predicted the increments of arm joints to move the end-effector to a target-object placed in a pseudo-randomly assigned position that is reachable, visible, and also novel, i.e., not experienced during the learning history.

The system was also expected to adjust the direction of the palm toward the forward directions. Importantly, in each trial the system was only allowed to move for four steps.

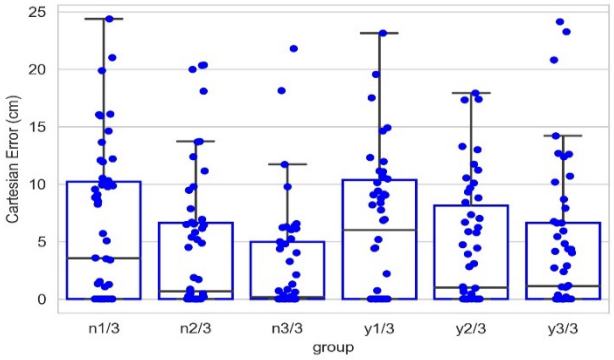

Figure 39: Cartesian errors in all six conditions. Along the vertical axis, "n" means results are obtained under the condition without noise applied to joints, "y" means noisy, and the fractions after n or y show the proportion of the entire learning session experienced after which test results are obtained. Within the chart, each box contains three black bars, and a box body marks two levels by upper and lower edges. From up to down, these five levels indicate the maximum, third quarter (Q3), median, first quarter (Q1) and minimum values of the value set represented, which is called "five-number summary." Points that past Q3+1.5*IQR (interquartile range) or Q1-1.5*IQR are not included in the box.

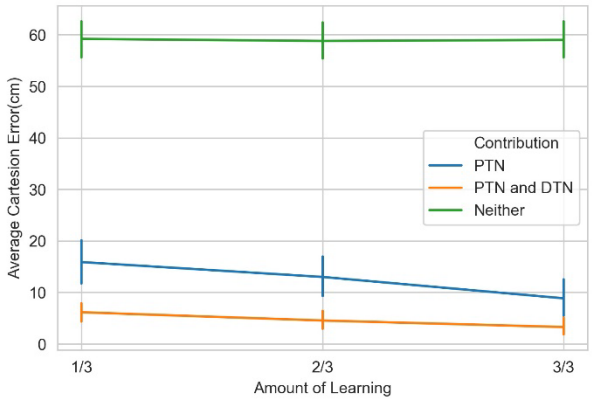

Figure 40: Contributions of position-tuned and direction-tuned neural networks. The three values indicated by the green line are 59.20, 58.80, and 59.00 cm, respectively from 1 to 3 of the learning session to the end of the full session. With PTN involved, as shown by the blue line, these three values are 15.89, 13.00, and 8.86 cm. When both PTN and DTN are functioned, these three values are reduced to 6.74, 4.55, and 3.28 cm.

We evaluated the performance in terms of Cartesian error measured by the Cartesian-distance between the end-effector and the target object, as well as by the angular difference between the palm and the desired direction.

Figure 39 shows the contribution of the position controller by experimental conditions. In this figure, along the vertical axis, “n” means results are obtained under the condition without noise applied to joints, “y” means noisy, and the fractions after n or y show the proportion of the entire learning session experienced after which test results are obtained. Within the chart, each box contains three black bars, and a box body marks two levels by upper and lower edges. From up to down, these five levels indicate the maximum, third quarter, median, first quarter and minimum values of the value set represented, which is called “five-number summary.” Regardless the presence of noise, Cartesian error decreases with larger amounts of learning. The one-way ANOVA also shows the significant effect of the amount of learning ($F(2, 297) = 4.98$, $p < 0.01$). Comparing the conditions with and without noise, as expected, the median levels of Cartesian errors are higher when noise is present. The median error is 0.96 cm without noise applied and 2.54 cm with noise present, although we found no significant effect of the noise condition on the Cartesian error ($F(1,298) = 1.49$, $p = 0.22$). Moreover, the first quarter error values are 0 for all six conditions, and median errors are smaller than 5 cm for all conditions except for the noisy one after 1/3 of the learning session (6.01 cm). In particular, after going through the entire learning session, under the condition without noise, the median Cartesian error is 0.16 cm, and the third quarter error value is 4.96 cm. These results indicate that the position controller is able to deliver the end-effector to

contact or reach the target with satisfactory error levels, compared to other recent studies (Mahoor et al., 2016; Nguyen et al., 2019; Rayyes et al., 2020). As an example, Mahoor reported a median Euclidean distance error of approximately 4 cm achieved by neural-networks learned through motor babbling. Note that the accuracy and precision of the system can be improved significantly by increasing the resolution of internal representations and the learning trials (circular reactions). Here we demonstrated that even with low resolution and fast learning, the network is capable of reasonable performance levels. Figure 40 further shows the contributions of position-tuned net and direction-tuned net, respectively, considering all conditions. After full learning experience, PTN alone reduces the average distance error from 59 to 8.86 cm, and followed by DTN who finally reduces this average error to 3.28 cm. Comparing PTN only and PTN&DTN, there is a significant effect of DTN ($F(1,298) = 37.15$, $p < 0.01$). This suggests that the PTN successfully drives the end-effector to somewhere close to the target, and DTN also behaves effectively in correcting end-effector's position.

Figure 41 A shows the contribution of posture controller under the condition of full learning and without noise. When this component doesn't function, which means palm's direction keeps aligning with lower limb's direction, the median value of angular errors is 58.36°. This median error is reduced to 18.63° with the involvement of posture controller. The effect of posture controller is also significant ($F(1,98) = 55.76$, $p < 0.01$). This suggests that the posture controller is effective in adjusting palm's posture. In some

positions, the desired directions might be awkward for the posture controller to adjust when those can possibly exceed joints' rotation limits.

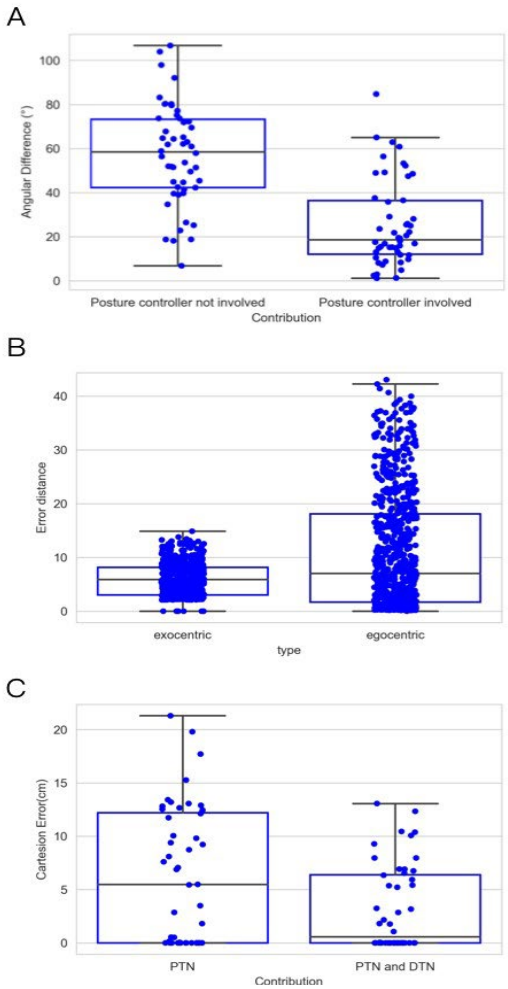

Figure 41:Contributions of PTN and DTN. (A) Average angular errors between palm and desired direction when trials are completed with and without the application of posture controller. Same as Figure 39, each box in this chart summarizes five numbers. For the left box, these numbers are 106.69, 73.30, 58.36, 42.30, and 6.80 in degree from maximum to minimum, respectively, and 84.73, 36.31, 18.63, 12.06, and 1.16 for the right box. (B) Distance error in terms of the number of cells in the CPM. For the left box, these numbers are 43.04, 18.11, 7.04, 1.69, and 0 from maximum to minimum, respectively, and 14.86, 8.14, 5.87, 3.04, and 0 for the right box. (C) Cartesian errors when PTN worked only and when both PTN and DTN worked. Same as Figure 39, each box in this chart summarizes five numbers. For the left box, these numbers are 21.29, 12.2, 5.46, 0, and 0 in cm from maximum to minimum, respectively, and 13.06, 6.39, 0.55, 0, and 0 for the right box.

These results in visually guided reaching demonstrate the effectiveness of sensorimotor coordinations and the three-dimensional exocentric external frame of references.

5.4 Experiment 2: Reaching with Active Fixation

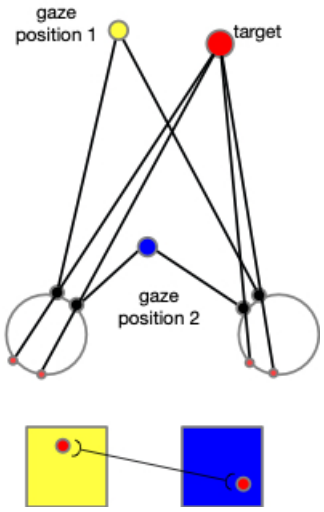

Figure 42: Compensation of gaze position. When the eyeballs rotate to fix on the gaze position 2 from the previous fixation (gaze position 1), a spot's projection on the two retinas moves accordingly. Two cells activated sequentially from the corresponding two CPM are then associated. In this figure, the yellow square represents the CPM of gaze position 1, and the blue square represents the CPM of the gaze position 2. The red disks represent the target and its projections on retinas and CPMs.

In addition to the first experiment, in which only the arm consists of degrees of freedom, we conducted a supplementary experiment adding the motion of eyeballs. By this experiment, we show how the neural networks are expanded when other body joints are included, such as eye movement, head rotation, and other body motions. As explained in the following context, adding any other degree of freedom would necessitate expansions in the same way as in the introduction of eye movement to the model. The cells in each neural layer would not only tune to the arm configurations and the

directions, but they would also tune to the gaze positions. This means there would be different layers of cells, and each layer will correspond to a specific gaze position. When the reaching begins, the gaze position firstly triggers the selected layer, and then the cells within this layer then function according to the input arm configurations and directions. In addition, the CPM formalized by each gaze position will compensate each other in a way that cells from different CPMs that are projected by the identical spot in the environment are connected. As shown in Figure 42, this associated learning makes the spatial localization independent of the gaze position, and thus exocentric.

Regarding the eyeballs' rotation, there are two strategies to cope with the reaching behavior: (1) reaching a target without fixing on it; and (2) fixing on the target and then reach it. This experiment will be tested using the latter, since the former are shown in the first experiment.

5.4.1 Illustration of the Eyeball Rotation Compensation

To illustrate the operation of exocentric representations, we conducted a separated test, in which we compared the errors of spatial localization under both exocentric and egocentric conditions. We tested internal representations of spatial position with two gaze positions and 740 spatial positions. The eyes rotated either 6° or 12° horizontally, with left eye rotating rightward and right eye rotating leftward. We used 3° as the resolution of the CPM cells considering the gaze positions. We compared the distance between target spots' internal representation before and after the rotation with and without the compensation. The results are shown in Figure 41B. The mean error in the egocentric condition is 11.06, which is reduced to 6.01 by the compensation. Due to the limited

resolution regarding the gaze position and the size of the CPM, eyeball rotation have very little impact on some spots' projections, which can be found in the minimum error in the egocentric condition. However, the maximum and third quarter values of the error are greatly reduced. This means that, for the scenarios where eye movements cause drastic spatial localization errors, the compensation effectively reduces the errors in the localization. The ANOVA also shows a significant difference between these two groups of the errors ($F(1,1478) = 146.83$, $p < 0.01$).

5.4.2 Experimental Parameters and Learning

This experiment used the same parameters relevant to the CPM and the cell's tuning properties as what we used in the first experiment. However, since the eye movement was included, both PTN and DTN were expanded by adding cells with gaze position tuned properties, which selectively respond to a specific eyeball configuration represented by the vertical rotation, left eyeball horizontal rotation, and right eyeball horizontal rotation. We considered 0° for an eyeball's both horizontal and vertical rotations when the eyeball points exactly forward. The range of these rotations that the cells were tuned to were [−40°, 0°], [0°, 30°], and [−20°, 20°], respectively, with an interval of 10°. (Here we use positive values for upward and rightward rotations, and negative for downward and leftward rotations.) The cells thus were tuned to 100 (5 × 4 × 5) gaze positions, making the cells arranged in 100 different layers. Within each layer, the cells selectively discharge to 4,096 (8^4) arm configurations. These are given by 8 angles for each of the four joints: 0°, 25°, 50°, …, 150°, 175°. Therefore, within each of the 100 layers, DTN contains 730, 800 (4, 060 × 180) cells and PTN contains 2.7 × 106 cells. Importantly,

from the first experiment to this one, we changed the resolution of arm configuration in each joint from 6° to 25°. In the learning session, a total amount of 1.6 × 108 steps of self-exploration were implemented to drive the learning and self-organization processes of the two neural networks. Other procedures were all identical to the first experiment.

5.4.3 Tests and Results

In the test session, we used the noisy targets group that we used in the first experiment. Figure 41C shows the contributions of the posture and the direction controllers. Compared to the original wrist position, as shown by the green line in Figures 40, 41C, both groups show strong learning effect as the median final reaching errors after five motion steps are strongly reduced. The mean distance from the wrist to the targets during the entire test session was 59 cm, and the mean error of the reaching after first step driven by PTN was 6.21 cm. This is even smaller than the error from the same condition that was tested in the last experiment, which is 8.86 cm. The DTN's contribution afterward reduces the mean error to 3.19 cm. This is also better than the mean error in the last experiment, which was 3.28 cm. The difference between the two groups of errors is also significant ($F(1,98) = 7.77$, $p < 0.01$).

The performance of the model in this test is generally better than the performance found in the first experiment even though we reduced the resolution in the joints from 6 to 25°. On the other hand, we expanded the range of each joint to 180°, which may explain the improvement in the performance. Moreover, the eye motion improves the contribution of the PTN, resulting in a decrease in the error.

6 Conclusions

In this paper, we investigated a model reproducing sensorimotor activities observed in human cognitive development. The model learns to coordinate sensory map representations with motor vector representations thereby generating accurate goal-directed reaching movements. We show that the implementation of the cyclopean map successfully provides the visual information in the guidance of reaching behaviors.

The experimental results with our proposed system show that its contrast-sensitive visual processing is able to locate an object's spatial center. In particular, a contrast balance method, the Ding-Sperling model, improves the object localization in the situation where two objects are spatially close to each other. The experimental results also show a good reaching performance measured by the Cartesian error between the end-effector and the target. With proper amount of learning, the model successfully contacts the target in almost half trials that have been tested, and the errors are within 5 cm in three quarters of the trials (condition “n3/3”). The two neural-networks, PTN and DTN, show their distinct and significant contributions during the test session. We also found our model robust in the noisy condition. Even though the median Cartesian error increases when noise is applied, there is no significant difference in the Cartesian error between “noisy” and “clear” conditions.

At present, our model is able to locate and reach the target in a relative low resolution. The model is not optimized according to the serial computing architectures and principles used in today's computing technology. In contrast, the model is built according to the massively-parallel computing-principle used in the nervous system.

Thus, current computing technology limits the implementation of the model and we had to restrict the resolution of internal representations (i.e., number of neurons and layers) to be able to run our simulations in a reasonable time. Massively parallel analog computers can provide a much better platform for implementing our model. In terms of comparing serial computing technology with massively parallel neurocomputing, it suffices to highlight that even though neurons operate orders of magnitude slower than integrated circuits (time-scale of milliseconds vs. nanoseconds), for many sensory and motor tasks the brain outperforms computers both in accuracy and speed.

As mentioned in the introduction, an alternative approach to sensorimotor coordination could be based on the popular deep-learning methodology (Vos and Scheepstra, 1993; Takemura et al., 2018). Typically, a multi-layer feed-forward neural-network is set and initialized by random weight values. A training set is used, where the inputs represent the visual image or coordinates of the target whereas the outputs consist of joint angles. By using supervised learning, the error between the actual joint-angles and the desired joint-angles can be back-propagated to adjust the synaptic weights. Our approach is different in that it is (i) autonomous, (ii) unsupervised, and (iii) local. Unlike the aforementioned deep-learning approach, we do not need an external teacher who will generate training data and feed the training data to the network. Through circular reactions, our model autonomously generates its own training trials and data. The sensorimotor closed-loop in action automatically provides error signals in real-time. Hence, no supervision is needed. Finally, we embed explicitly map and vector representations in our model based on the neurophysiology of the primate brain, as

opposed to starting “tabula rasa,” i.e., a network with randomly selected weights. These map representations and their coordinates are inspired by the organization and development of sensory systems in biology. As mentioned in the Biological Evidence Section, the explicit map representations in the model allow local learning. Learning is restricted to “sensitivity zones,” which represent local subsets of space. According to this approach, highly nonlinear relationships across the entire visual space can be simplified by local approximations, resulting in a much simpler learning approach.

Chapter Four: Canonical Forms and their Mental Processing in Object Recognition

1 Introduction

1.1. Theories of invariant object recognition

Recognizing a previously learned object requires that we match the current appearance of the object (stimulus) with the memory representations of candidate objects and determine the best match. This task is complicated because objects do not have unique appearances: As the relative position of the object with respect to the observer changes (perspective views), the appearance of the object can undergo drastic changes. Several theories have been formulated to explain how the brain accomplishes this "invariant object recognition" task (reviews: Logothetis & Sheinberg, 1996; Riesenhuber & Poggio, 2000; DiCarlo et al., 2012; Gauthier & Tarr, 2016).

According to "invariant feature/relation" approaches, this problem can be by-passed all together by using memory representations that are invariant to perspective views (Palmer, 1999; Pinker, 1984). In these theories, an object is described by a collection of low-level features such as angles, curves (Sutherland, 1968; Barlow et al., 1972), and/or higher-level structural characteristics such as component parts and their relations (e.g., Palmer 1977; Biederman, 1987), which are independent of viewpoints. The stimulus-memory comparison takes place by matching these invariant feature/relation-based structural descriptions. For example, an edge length or a vertex angle can be compared

directly regardless of the orientation of the stimulus and the memory item, as long as matching edges and vertices between the stimulus and memory representation are found. A horse can be recognized independent of its orientation based on the relations of the parts or components (head, neck, torso, legs, tail, etc.) (Marr & Nishihara, 1978; Biederman, 1987).

In contrast to invariant feature/relation approaches, several theories proposed mechanisms whereby view-variant representations are combined or actively transformed to accomplish object recognition. The "perspective-storage theory" proposes that, as the subject experiences different views/perspectives of the same item, each view/perspective is stored as is under the same label (e.g., my friend Jane). This theory follows the behavioristic approach where stimuli and responses are associated with each other as they appear in the environment. For example, if we see five different perspectives of a given face and each is presented together with a name, those perspective views will be associated with that particular name. As in associative learning, when the subject experiences in the future one of the stored views, it will generate the associated response, such as the name of the person. In this theory, the internal representations consist of a set of stimuli with an associated label. In classical conditioning, the object and the label may be occurring in close spatiotemporal proximity (e.g., while an object is shown and its label/name is verbalized) and in reinforcement learning, the observer's response is either positively or negatively reinforced until the correct response is found. In addition to behaviorism, this approach is also used extensively in artificial neural networks from early versions (Rosenblatt, 1958) to later incarnations (e.g., Fukushima (1975);

Riesenhuber and Poggio (1999); Krizhevsky et al. (2017)). These are hierarchical feed-forward models where each layer filters its input and sends the filtered information to the next layer via simple nonlinearities. In these networks, memory is implicit in that it consists of distributed values of synaptic connections across the network. The "generalization", i.e., the association of the label to novel perspectives occur via some interpolation process using similar perspectives that are already stored (Edelman & Bülthoff, 1992). In other words, memory representations and storage in these models are "passive processes" in the sense that there is no active internal structuring or manipulation of the stimuli during storage and recall.

Whereas invariant feature/relation approaches deal with variances by using invariant features, an alternative approach is to cancel environmental variations by applying inverse transforms. Hence, theories that use this approach posit an active internal structuring during storage and/or recall to compensate for the effects of environmental variances. For example, according to the alignment theory (Ullman, 1989), anchor points are used to align the stimulus with the memory (internal model) by using transformations such as scaling and rotation, a process called "normalization". The "canonical-representation theories" (Palmer et al., 1981) use similar alignment approaches but posit that, the memory storage is not arbitrary, but follows a canonical scheme: When a stimulus appears, it is not stored as is. Instead, a canonical representation is chosen and this canonical form is stored in memory. For example, the canonical form for objects can be according to the symmetry axis for symmetric objects. To carry out the comparison

between a stimulus and the memory representation, the input shape is converted to the canonical orientation through mental rotation.

Several studies provided evidence against invariant features/relations theories(Tarr et al., 1998; Fang and, He, 2005; Kourtzi and Shiffrar, 1999), however, tests of the other theories have been equivocal (Willems and Wagemans, 2001; Ratan Murty and Arun, 2015; Vanrie et al., 2001; Tarr and Hayward, 2017). Note that these theories are not mutually exclusive and there are also hybrid versions: For example, Tarr & Pinker (1989) proposed that we store a small set of orientation-specific representations, as in the perspective storage theory, and the input shape is transformed to match the closest one, as in the canonical storage theory.

1.2. Reference-frames

An underlying concept in these theories is reference-frames. For example, structural description theories use an object-based reference-frame according to which structural descriptions are formulated. In canonical-representation theories, the canonical form can be expressed in terms of a specific reference-frame, for example one aligned with the elongation axis of an object. Previous studies suggested that geometrical properties such as the symmetry and the aspect ratio play an important role in the selection of intrinsic reference-frames (Palmer, 1983; Palmer, 1985; Rock, 1973; Marr & Nishihara, 1978). For example, Mou et al. (2007) investigated how layout geometry affects the selection of intrinsic reference-frame in judging the relative direction of objects within the layout. They found that subjects behaved quicker when the heading direction was parallel to the symmetrical axis of the layout. It was also demonstrated that the axis of symmetry and

the axis of elongation were selected as the intrinsic orientation of the shape (Sekuler & Swimmer, 2000).

1.3. A sensorimotor approach

As mentioned above, mental rotation has been proposed as an active transformation mechanism to compensate for rotational variances in stimulus appearance. This is in line with sensorimotor theories of intelligence. For example, the first stage in the Piagetian theory of cognitive development is the sensorimotor stage. This is the stage underlying the emergence of object concept and constancy. Starting with innate reflexes, such as sucking, infants gradually build a repertoire of sensorimotor schema which are then "internalized" in the sense that the sensorimotor schema do not have to be executed physically but can be "simulated" mentally. Through this internalization, the infant does not need to grope or experiment physically by motor action, but can solve problems through "mental combination" (Piaget, 1952). Mental combination involves internal simulation of sensorimotor schema. In other words, sensorimotor schema is executed (or simulated) mentally without motor action. Whereas Piaget built his theory mainly through behavioral observations, more recent neurophysiological studies provide support for the internalized sensorimotor schema. Previous studies have found evidence showing sensorimotor strategies in object recognition tasks. For example, one study reported correlations between sensorimotor networks and facial expression recognition (Wood et al., 2016). In another study, it was argued that subjects with multimodal agnosia, a visual recognition deficit, preserved some extent of ability for recognition via sensorimotor pathways (Sirigu et al., 1991). However, arguably, the most direct evidence for the

internalized sensorimotor proposal comes from Shepard and colleagues' studies (e.g., Shepard & Metzler, 1971; Cooper & Shepard, 1973). By using two-dimensional alphanumeric characters or two-dimensional projections of three-dimensional objects build from cubes, they assessed Reaction Times (RTs) required to determine whether two samples were identical or mirror-image version of each other. They found that RTs depended linearly on the angular disparity between the two samples to be compared. These results have been interpreted to involve a mental rotation operation whose duration depends linearly with the required rotation angle. The effect was found both with familiar shapes such as letters and digits, and with unfamiliar shapes (Cooper, 1975). For example, Shinar and Owen (1973) taught subjects multiple novel polygonal shapes at their upright orientation and let subjects recognize these shapes when presented with unfamiliar orientations. They found the time to perform this judgement was dependent on the shape's orientation relative to the upright. In another study, Jolicoeur (1988) had subjects repeatedly name images of natural objects in different orientations. It was found that the time required to name objects was dependent on the orientation at the beginning, and the effect disappeared as subjects finished more and more repetitions of objects. The failure to observe mental-rotation effects in some experiments and the vanishing of these effects with practice can be explained by the discriminability of the stimulus (Förster et al., 1996). Förster and colleagues showed that mental-rotation effects can be found not only with mirror-image discrimination tasks but also using complex (polygons) as well as simple (line segments) stimuli, provided that the discrimination task is difficult enough. Mirror-image stimuli are used to eliminate all shape differences with the exception of

mirror-image symmetry to make the task difficult and independent of direct strategies by comparing some specific features of the stimuli. For example, if the sample and the comparison differ from each other by the number of sharp edges they have, the observer can accomplish the task without any detailed shape comparison (hence no need for rotation) based on the number of sharp edges. Similarly, with practice, observers may discover simple local feature differences in the stimuli and base their judgments on that criterion rather than a detailed comparison via rotation. In addition to the aforementioned behavioral evidence, electrophysiological correlates of mental rotation have also been identified. Gardony et al. (2017) found multiple EEG signals in their data collected from subjects performing mental rotation tasks including sensorimotor μ desynchronization, parietal α desynchronization, and frontal θ synchronization. These signatures indicated the employment of motor processing, visuospatial processing, and working memory maintenance. In another study using the images of hands as the stimuli, EEG data from sensorimotor area showed similar pattern between mental rotation task and motor imagery task (Osuagwu and Vuckovic, 2014). Another evidence supporting the involvement of motor processing during mental rotation is the strong correlation between α band suppression and the reaction time of mental rotation, as α suppression was suggested to correlate with motor system activation (Umilta' et al., 2012; Perry et al., 2010; Michel et al., 1994).

1.4. The goals of the study

One goal of this study was to examine systematically the roles of (i) elongation and symmetry (two ubiquitous aspects of natural stimuli) and (ii) boundaries and surface-

texture (two fundamental aspects of natural objects) in the choice of canonical forms. Our stimuli were designed to systematically control elongation and symmetry properties and whether these properties were expressed by boundary and/or texture information. The second goal was to study the relationship between canonical forms and mental rotation. Within this context, we also aimed to test different theories of invariant object recognition: The structural-description theory predicts that observers' performance should be independent of the viewing rotation-angle because its coding is inherently independent of rotational changes. The perspective-storage theory predicts that observers' performance will vary according to the statistics of perspective views; in other words, those views experienced more often should lead to better recognition. Whereas it is difficult to determine the viewing statistics of familiar objects for each and every rotation angle, one can control these statistics by using novel objects in a laboratory environment. If novel objects' presentation follows a uniform distribution in terms of orientation angle, then the perspective-storage theory predicts the same performance independent of rotation angle. The canonical sensorimotor theory, however, predicts that performance will be best for the canonical orientation and will degrade monotonically with the difference between canonical orientation and the viewing orientation. Hence unlike studies that presented novel objects with a pre-selected orientation during training (e.g., Cooper, 1975; Tarr and Pinker, 1989; Gomez et al., 2008; Edelman and Bülthoff, 1992), we used a uniform presentation of all orientations. We provided feedback and analyzed the results when observers reached high levels of accuracy in the task to make sure our analyses reflect the operation of memory and recall at its steady-state with a well-formed

and stable memory representation (cf. Edelman and Bülthoff, 1992). Finally, the choice of canonical orientation or the relevant reference-frame may depend on cues other than the stimulus itself. For example, for stimuli presented on monitors, the edges of the monitor or other references in the laboratory may influence the choice of the reference-frame. To minimize these factors, we used a virtual reality (VR) headset to present our stimuli.

2 Experiment 1: Symmetry and Canonical Orientation

The purpose of this experiment was to study the role of symmetry in determining the canonical orientation for storage and recognition. Two of the fundamental attributes defining an object are boundary and texture. Hence, we used stimuli whose symmetry was defined either according to boundary or texture information.

2.1 Methods

2.1.1 Participants

Five students from the University of Denver and one of the authors (DH) participated in this experiment (one female and five males; age: M[SD]=26.67[2.81] years) and all participants had normal or corrected-to-normal vision. This experiment followed a protocol approved by the University of Denver Institutional Review Board for the Protection of Human Subjects. Each observer gave written informed consent before the experiment.

2.1.2 Equipment & Calibration

In an experiment conducted on a traditional monitor placed in an experimental room, subjects are exposed not only to the experimental stimuli but also to various other

geometric cues that can serve as a reference-frame. For example, the rectangular shape of the display monitor can serve as a reference-frame. To avoid such cues, we presented our stimulus using an HTC VIVE VR headset released in 2017.

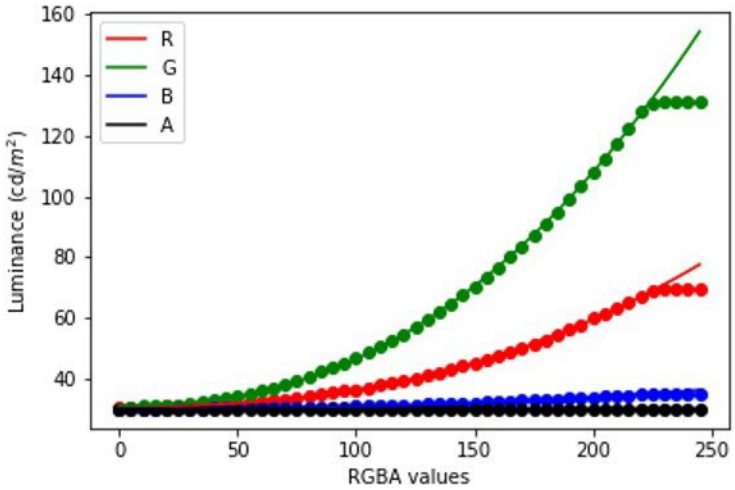

Figure 43: Gamma fitting results of VR headset screen's luminance with respect to the RGBA value inputs in the Unity3D under the experimental condition.

The resolution of the VR headset was 1200×1080 with a 90 Hz refresh rate. The approximate pupil-to-lens distance was 18 mm. The color display of the device was controlled by the development environment Unity3D using RGBA values. Since the left and right displays of the VR headset have the same parameters, we calibrated the color display by measuring the luminance of the left display with a Minolta LS-110 luminance meter while adjusting the RGBA input from Unity3D under our experimental settings. These four channels represent red, green, blue, and alpha respectively, in which alpha indicates the degree of transparency. First, we tested each color-channel separately, i.e., by varying its inputs from 0 to 255, with a sampling interval of 5, while keeping all other channels at 0. The relations are shown in Figure 43. We found the minimum luminance of all four channels to be approximately $30cd/m^2$. The maximum luminance of

individual RGBA channels were 75, 154, 36, and $30 cd/m^2$, respectively. We fitted Gamma functions to the data and obtained the following:

$$\begin{cases} LumR = 8.47 \times 10^{-5} \times R^{2.4} + 30.87 \\ LumG = 3.43 \times 10^{-4} \times G^{2.3} + 31.13 \\ LumB = 1.59 \times 10^{-5} \times B^{2.3} + 30.14 \end{cases} \quad (54)$$

The RMSE of these fits were 1.54, 4.55, and 0.27 cd/m^2. As for the A channel, the luminance doesn't change with respect to the A value (t=3.07×10^{-8}, p = 0.999). We then tested the overall luminance in relation to the RGB values and obtained the following calibration function:

$$Lum = 0.97LumR + 0.99LumG + 0.73LumB - 51 \quad (55)$$

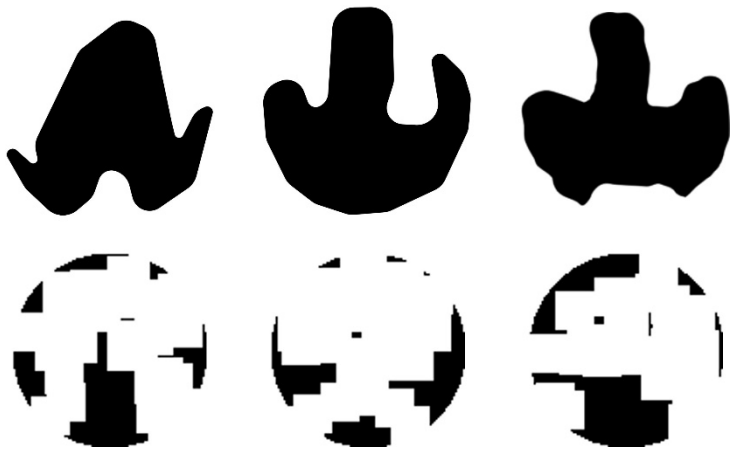

Figure 44: The stimuli used in the Experiment 1. The first row contains three boundaries: B1, B2, and B3, from the left to right. The second row contains three textures: T1, T2, and T3, from the left to right.

2.1.3 Stimuli & Procedure

Subjects observed the stimuli through an HTC VIVE VR headset. They were seated in front of the headset, which was fixed on the table, with their eyes approximately 5 cm from the screens. As shown in Figure 44, the stimuli consisted of three asymmetrical black boundaries (B1, B2, B3), and three asymmetrical black textures (T1, T2, T3). The

boundaries were drawn within an invisible circle (24.81 deg) on whose edge significant proportion of their boundary was lying. Textures were iteratively generated by randomly extracting square patches from a disk until the difference in degree of symmetry between the most symmetrical and the second most symmetrical orientations reached 0.2. Labeling the orientation of the images shown in Figure 44 to be 0°, stimuli were shown by different orientations reported in degrees measured either clockwise or counterclockwise. Although the six images are not perfectly symmetrical, the degree of symmetry along some axes is still higher than for other axes. We calculated the degree of symmetry for each stimulus along each testing axis by the proportion of overlapping areas between the two halves after folding one side to the other along the axis. The results are shown in Figure 45. As can be seen from this figure, the highest degree of symmetry is obtained for the orientation labeled 0° (i.e., the orientation at which they appear in Figure 44).

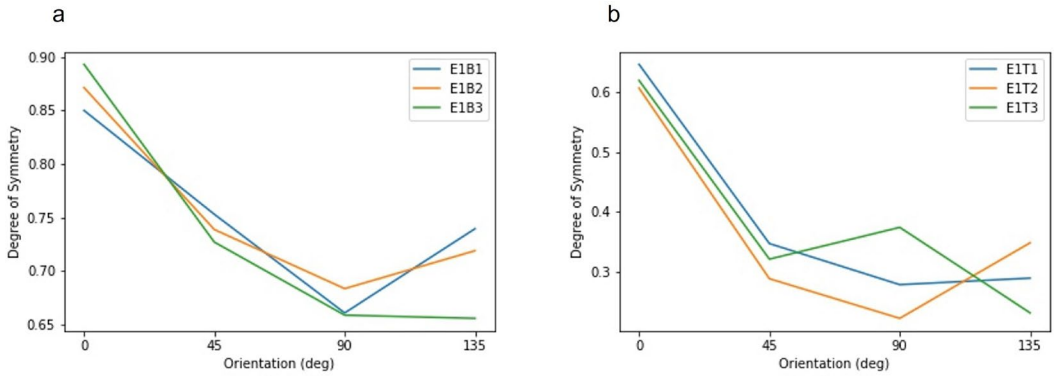

Figure 45: The relationship between the degree of symmetry and orientation of the stimuli in Experiment 1. (a) Boundaries. (b) Textures. In each panel, three lines correspond to the three objects, as shown in the labels, which refer the same object labels shown in Figure 44.

The experiments were separated into two conditions: the boundary condition, and the texture condition, using the boundary stimuli and the texture stimuli respectively. In each condition, each trial contained only one frame, showing a stimulus (RGBA values: 0, 0, 0, 255) at the center of the visual field (RGBA values: 183, 179, 179, 255). The stimulus was presented randomly either in its "original" form as shown in Figure 44 or in its mirror-image form (i.e., "flipped"). The task of the observer was to report whether the stimulus was shown in its original or flipped form by pressing the left (original) or the right (flipped) key of a computer mouse. Subjects were not instructed about the handedness of the stimulus at the beginning. A feedback beep followed wrong responses after each trial, which allowed subjects learn autonomously the original versus the flipped versions of the stimuli. Once the subject clicked the mouse key, the next trial followed automatically. The stimulus was presented in an orientation that was selected randomly from one of $0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}, 180^{\circ}$, by either clockwise or counterclockwise. The distribution of orientations was uniform over these angles. In a given trial, the stimulus shown can be any one of the three shapes, five orientations, two rotation directions, and two handedness. In each block, each case was shown only once, and the types of trials were selected according to a random sequence. Hence, there were sixty trials in each block. Each subject was asked to finish three blocks and was allowed to rest for a brief period between the blocks, which lasted typically less than one minute.

Blocks with the subjects' performance under 85% correct were excluded from the analysis. As in previous studies that used the mental rotation paradigm, in each condition of this experiment, data from all shapes and subjects were pooled together to analyze but

data representing incorrect answers were excluded from the analysis (Shepard & Metzler, 1971). For each subject's reaction time (RT) on each shape, data that were out of the range set by three times standard deviation plus/minus median or longer than ten seconds were not included in the analysis. Moreover, since this experiment examined the effect of symmetry on the selection of a reference-frame, considering the equal level of symmetry between 0° and 180° orientations, we fitted the RT results as a function of orientation angles for each subject and each shape, by combining the orientations with equal symmetries, i.e., the angles $[0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}, 180^{\circ}]$ with the equivalent angles $[-180^{\circ}, -135^{\circ}, -90^{\circ}, -45^{\circ}, 0^{\circ}]$). Therefore, 0° represented the preferred most symmetrical orientation in the plots shown in the results, in which within-subjects mean and SEM were plotted.

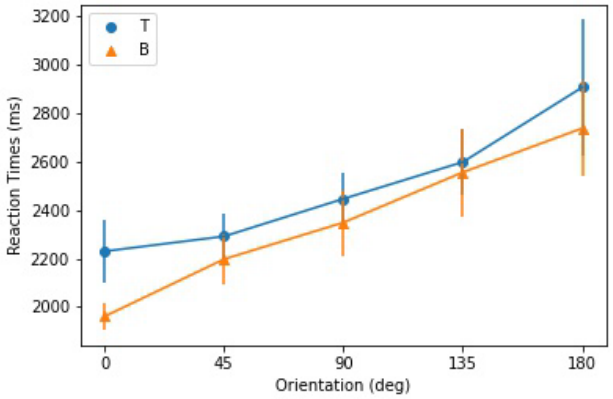

Figure 46: Reaction times with respect to orientation for the stimuli used in Experiment 1. T and B represent to the Textures and Boundaries, respectively.

2.2 Results

A power analysis was performed to determine whether the sample that was obtained in the current study (N=6) would be sufficient to detect a meaningful effect size. This

analysis revealed that the power of the study to detect the significant effect of orientation on the RT was 0.99 ($\eta^2 = 0.44$, $\alpha = 0.05$).

For the boundary condition, the mean [standard deviation] values of subjects' performance in the three blocks were 79.72% [0.11], 96.67% [0.03], and 95.83% [0.01] respectively, indicating, as expected, a learning effect. The first block of four subjects showed performance that was under 85% and thus excluded from further analysis. As for the texture condition, these values for the three blocks were 86.11% [0.13], 94.72% [0.02], and 95.28% [0.03] respectively, and the first block of three subjects was excluded due to unsatisfactory performance (reflecting the learning phase, rather than the "steady state" learned phase). As shown in Figure 46, the shortest mean RT corresponds to the orientation of 0° in both boundary and texture conditions. The mean RT with respect to orientation profiles for both conditions show linearity as reported in previous studies using mental-rotation paradigm (Shepard and Metzler, 1971; Cooper and Shepard, 1973; Cooper, 1975), reflected by the R-squared values: 0.99 (boundary) and 0.94 (texture). A repeated-measures ANOVA with orientation and condition as main factors showed a significant effect of orientation on RT ($F(4, 20)=14.91$, $p<0.01$). The slope (ms/deg) and the intercept (s) for the boundary condition were 4.24 and 1.98 respectively. For the texture condition, the slope and the intercept were 3.68 and 2.16 respectively. Moreover, t-test showed the slopes for both of conditions were significantly different than 0 (texture: $p<0.01$; boundary: $p<0.01$). Generally, data show that the RT curve for texture is slightly above the boundary. However, neither the intercepts nor the slopes under two conditions were significantly different, as repeated-measures ANOVA showed no significant effect

of condition, and no significant interaction between condition and orientation on the RT (condition: $F(1, 5)=0.9$, $p=0.39$; orientation × condition: $F(4, 20)=0.67$, $p=0.62$).

2.3 Discussion

The findings of this experiment provide support for the canonical sensorimotor theory and against the other two theories. Furthermore, it highlights the role of symmetry in determining the canonical orientation: (i) RTs for recognition are at a minimum for the angle representing maximum symmetry, (ii) RTs follow a linear trend, supporting a mental rotation process whose duration is linearly related to the angle of orientation needed to align the stimulus with the memory prototype stored according to its canonical orientation.

The results also show that both boundary and texture characteristics of the stimulus can inform about the symmetry of the stimulus and thus can be used in memory storage and pattern recognition.

3 Experiment 2: Aspect Ratio and Canonical Orientation

Previous research showed that aspect ratio can also play an important role on how shapes are perceived and recognized (Davis et al., 2003; Sekuler, 1996; Sekuler and Swimmer, 2000). Furthermore, elongation and orientation are ubiquitous in nature and in fact the visual system is equipped to analyze orientation information through orientation columns in early cortex. Here, we tested the role of aspect ratio in determining the canonical orientation.

3.1 Methods

3.1.1 Participants

The same group of subjects who attended Experiment 1 participated in this experiment.

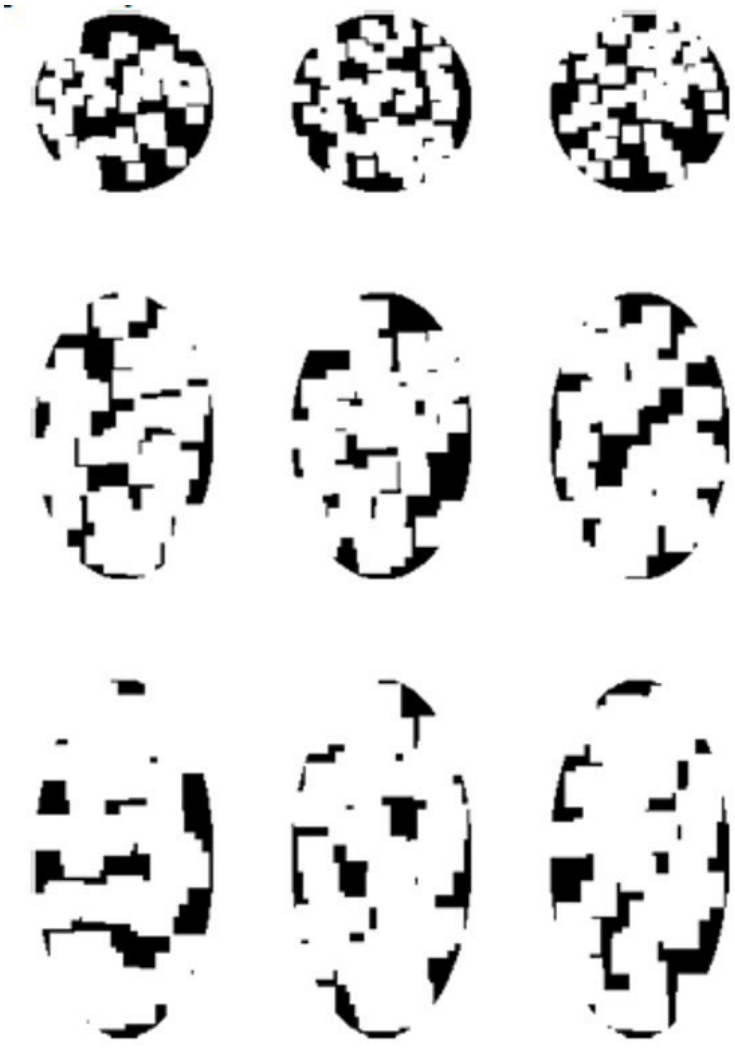

Figure 47: The stimuli used in the Experiment 2. The three rows from top to bottom correspond to three different aspect ratios: 1, 1.6, and 2. For each texture, the degrees of symmetry across all tested orientations are similar.

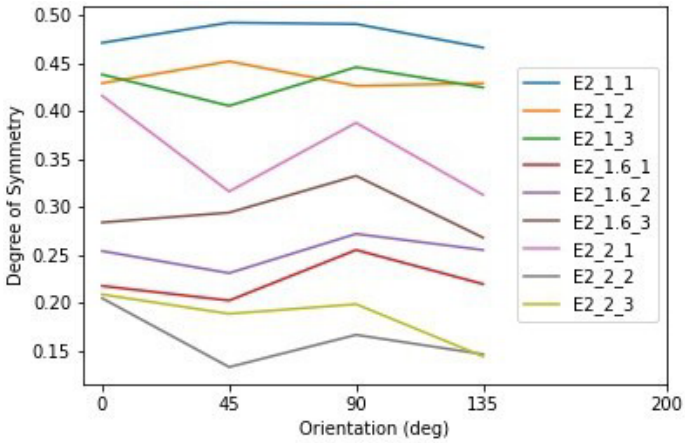

Figure 48 The relationship between the degree of symmetry and orientation of the stimuli in Experiment 2. Each color of line corresponds to an object in Figure 46, as indicated by the label. For example, E2 1.6 3 indicates the object with the aspect ratio of 1.6 and in the third column of the array in Figure 46.

3.1.2 Stimuli & Procedure

As shown in Figure 47, there were nine textures used as the stimuli in this experiment. These textures were iteratively generated by randomly extracting square patches from a disk or an ellipse with a specific aspect-ratio defined by the length of major axis (AR=1: 24.81 deg; AR=1.6: 31.26 deg; AR=2: 35.08 deg) divided by the length of minor axis (AR=1: 24.81 deg; AR=1.6: 19.6 deg; AR=2: 17.54 deg). The iteration stopped once the difference between the max and min symmetry was smaller than 0.05. As we show in Figure 47, elongated along the 0° orientation, the three rows of textures are in three different aspect ratios (AR): 1:1, 1.6:1, and 2:1, and all textures have approximately equal degree of symmetry across all tested axes. The symmetry properties of these textures are plotted in Figure 48. This experiment consisted of three sessions, in which each used the three textures with three different AR from a column in Figure 47, and each session contained three blocks. The procedures and parameters of a block were

exactly the same as in Experiment 1. For each subject's reaction time (RT) on each shape, data that were out of the range set by three times standard deviation plus/minus median were not included in the analysis. Other data pre-processing and statistical analysis procedures used in Experiment 1 were also used here.

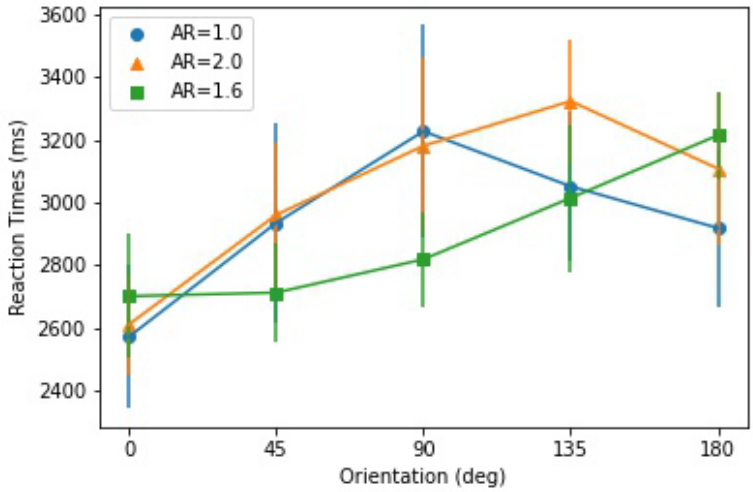

Figure 49: Reaction times with respect to orientation for the stimuli used in Experiment 2. Three colors correspond to three different aspect ratios (ARs).

Table 4: The mean [standard deviation] performance across subjects in each session and block of the Experiment 2.

	Block 1	Block 2	Block 3
Session 1	77% [0.11]	93% [0.09]	94% [0.06]
Session 2	84%[0.06]	95% [0.02]	95% [0.04]
Session 3	94% [0.05]	95% [0.02]	96% [0.03]

3.2 Results

A power analysis was performed to determine whether the sample that was obtained in the current study (N=6) would be sufficient to detect a meaningful effect size. This analysis revealed that the power of the study to detect the significant effect of orientation on the RT was 0.82 ($\eta^2 = 0.21$, $\alpha = 0.05$).

The mean [standard deviation] values of subjects' performance in the three sessions grouped by three blocks are shown in Table 4. In all three sessions and six subjects, with 54 blocks in total, ten blocks were excluded from further analysis for unsatisfactory performance (<85%). A repeated-measures ANOVA with orientation and AR condition as main factors showed a significant effect of orientation on RT ($F(4, 20)=6.64$, $p<0.01$), and a significant interaction between AR condition and orientation on RT ($F(8, 40)=2.51$, $p<0.05$).

As shown in Figure 49, RTs for AR=1.6 and AR=2 show an increasing trend whereas RTs for AR=1 do not, as suggested by the linear regression. The slope (ms/deg) and the intercept (s) for the AR=1.6 condition were 2.95 and 2.63 respectively. As for AR=2, the slope and the intercept were 3.02 and 2.77 respectively. A t-test showed that the slopes for these elongated AR conditions were significantly different than 0 (AR=1.6: $p=0.01$; AR=2: $p=0.01$). However, for AR=1, the slope (deg/ms) and the intercept (s) were 1.84 and 2.79, respectively, and t-test showed that the slope was not significant ($p=0.23$). The mean RT with respect to orientation profiles for both two elongated AR conditions show higher linearity than the AR=1 condition, reflected by the R-squared values: 0.28 (AR=1), 0.91 (AR=1.6), and 0.62 (AR=2).

The shortest mean RT corresponds to the orientation of 45° and 0° in AR=1.6 and AR=2 conditions respectively. For the AR=1.6 condition, RTs for 0° and 45° were not statistically different (t-test: $t=0.01$, $p=0.91$).

3.3 Discussion

To manipulate the aspect ratio, the figure is elongated along a given axis, e.g., the circle becoming an ellipse. This geometrical transformation also creates a symmetry axis based on boundary information along the axis of elongation, thereby creating correlated aspect ratio and symmetry properties. In order to isolate aspect ratio from symmetry, we introduced texture to the figure in a way texture did not have any preferred symmetry axis. Although this does not completely override boundary-based symmetry, it reduces its effect making the aspect ratio more prominent than symmetry. The AR=1 condition presents no preferred canonical orientation. The prediction for this case is that there would be no preferred orientation for storage. In fact, data on the slope and the degree of linearity of the RT results for AR=1 support this prediction. On the other hand, if observers were using aspect ratio in selecting a canonical orientation, we would expect linear RTs with a minimum at 0° orientation. The R^2 values for the linear regression for AR=1.6 and 2.0 support linearity. The minimum occurred at orientation of 0° for the strongest value of AR (AR=2). For AR=1.6, the minimum RT was at 45°; however, RT for 45° was not statistically different than the RT for 0°.

4 Experiment 3: Joint Contributions of Symmetry and Aspect Ratio

Previous experiments showed that both boundary and texture information can guide the selection of the canonical orientation for memory storage and pattern recognition. In this experiment, we studied how the canonical orientation is selected when boundary and texture information provide different solutions. We considered three hypotheses:

(1) Winner-take-all: In this case, one of the two factors, symmetry or aspect ratio, dominates the selection of canonical orientation. This leads to a reaction time profile with respect to the angular disparity between the input orientation and the orientation of maximum symmetry or aspect ratio.

(2) Dual canonical orientation: In this view, both the most symmetrical and the most elongated orientation can serve as the canonical orientation and remain in human memory when they are not parallel. With an input shape is to be recognized, it should be rotated to the nearest canonical orientation. The reaction time is then dependent on the smaller angular disparity between input orientation and two canonical orientations.

(3) Weighted-combination of two canonical-orientation candidates: Following this view, the canonical orientation should be on an axis between the orientation of the maximum symmetry and another one of maximum aspect ratio.

These hypotheses can be described by the following equation:

$$RT = W_{AR} \cdot k \cdot |\theta - \theta^*_{AR}| + W_{Sym} \cdot k \cdot |\theta - \theta^*_{Sym}| + T_e \quad (54)$$

where RT is the reaction time, W_{AR} and W_{Sym} are the steady-state weights of aspect-ratio and symmetry based canonical orientations, respectively, θ is the input orientation of the shape, θ^*_{AR} and θ^*_{Sym} are the aspect-ratio and symmetry based canonical orientation respectively, k is the rotation speed, and T_e is the baseline time determined by multiple factors including the encoding of shape, memory transfer, time spent on determining W_{AR} and W_{Sym}, preparation and execution of the motor response, etc.

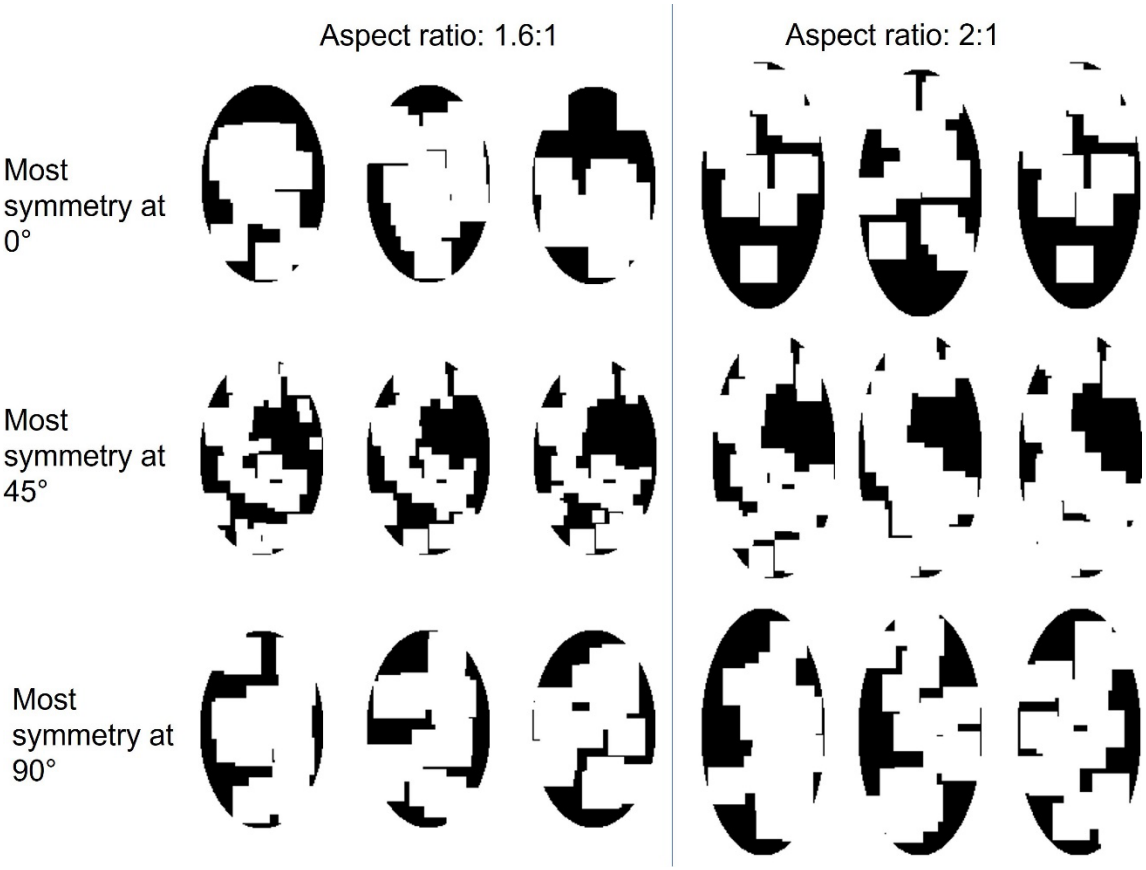

Figure 50: The stimuli used in Experiment 3. There are three symmetrical conditions, two aspect ratios, and three object for each symmetry-elongation combination.

The three hypotheses can then be expressed by:

$$Winner-take-all: \begin{cases} W_K = 1, if\ K\ dominates \\ W_K = 0, if\ K'\ dominates \end{cases} \tag{55}$$

$$Dual\ canonical\ orientation: \begin{cases} W_K = 1, if|\theta - \theta_K^*| \leq |\theta - \theta_{K'}^*| \\ W_K = 0, if|\theta - \theta_K^*| \geq |\theta - \theta_{K'}^*| \end{cases} \tag{56}$$

$$Weighted-combination: W_K = f(K, K') \tag{57}$$

where K can be either aspect ratio or symmetry and K' is the other one, $f(K, K')$ is a normalized function that is dependent on the magnitude of aspect ratio and symmetry.

4.1 Methods

4.1.1 Participants

Eight students from the University of Denver, including one of the authors (DH), participated in this experiment (two females and six males; age: M[SD]=23.17[3.67] years) and all participants had a normal or corrected-to-normal vision. This experiment followed a protocol approved by the University of Denver Institutional Review Board for the Protection of Human Subjects. Each observer gave written informed consent before the experiment.

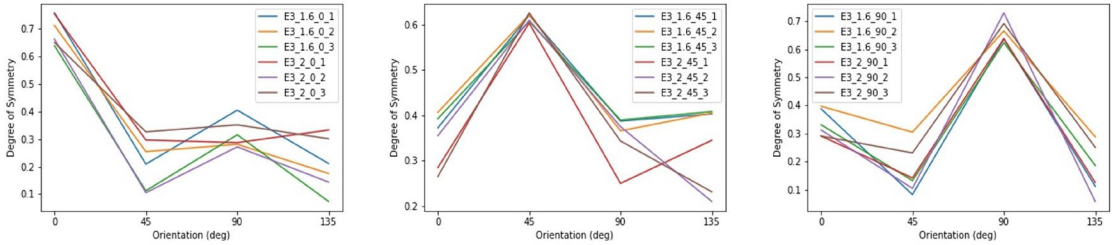

Figure 51: The relationship between the degree of symmetry and orientation of the stimuli in Experiment 3. These three panels correspond to the three symmetrical conditions: most symmetrical at 0°, 45°, and 90°. In each panel, each line represents to an object in Figure 50, as indicated by it label. For example, E3_2_45_1 indicates the object that is most symmetrical at 45°, has an aspect ratio of 2, and is the first item from left to right in Figure 50 under the same.

4.1.2 Stimuli & Procedure

As shown in Figure 50, there were eighteen textures used as the stimuli in this experiment generated in the same way as in previous experiments. The iteration stopped once the most symmetrical axis had a symmetry measure of at least 0.2 unit larger than the second most symmetrical axis. Figure 50 shows these textures in two panels, where the textures with AR of 1.6 are on the left side and those with AR of 2 are on the right side. All were elongated along an 0°. The three rows indicate three dominant symmetrical

axes: $0^{\circ}, 45^{\circ},$ and 90°. Therefore, the canonical orientation according to symmetry and according to aspect ratio could be parallel (most symmetrical at 0°), diagonal (most symmetrical at 45°), or perpendicular (most symmetrical at 90°). The symmetry properties of these textures are plotted in Figure 51.

This experiment consisted of six sessions, in which each session used the three textures with three different symmetrical properties and the same AR from a column in Figure 50. Each session contained three blocks. The procedures and parameters of a block were exactly the same as in Experiment 1 and the same data pre-processing and statistical analysis procedures as in Experiment 1 were used.

Table 5: The mean [standard deviation] performance across subjects in each session and block of the Experiment 4.

	Block 1	Block 2	Block 3
Session 1	75.56% [0.15]	87.5% [0.13]	96.39% [0.03]
Session 2	85% [0.13]	96.67% [0.04]	98.33% [0.02]
Session 3	89.44% [0.11]	97.22% [0.01]	97.22% [0.02]
Session 4	86.11% [0.11]	97.5% [0.03]	98.06% [0.03]
Session 5	93.33%[0.05]	97.5% [0.03]	97.78% [0.03]
Session 6	95.83%[0.06]	97.22% [0.03]	96.67% [0.03]

Table 6: The slopes, intercepts, and p-values of t-test (H0: slope=0) for the RT Orientation linear regression in each condition.

Slope, intercept	parallel	diagonal	perpendicular
AR=1.6	3.55, 2.03, $p<0.01$	4.2, 2.07, $p<0.01$	4.31, 2.17, $p<0.01$
AR=2	3.57, 2.18, $p<0.01$	2.15, 2.03, $p<0.01$	3.85, 2.28, $p<0.01$

4.2 Results

Two subjects gave an overall performance that was under 80% and thus were excluded from the analysis. A power analysis was performed to determine whether the sample that was obtained in the current study (N=6) would be sufficient to detect a meaningful effect size. This analysis revealed that the power of the study to detect the significant effect of orientation on the RT was 0.84 ($\eta^2 = 0.21$, $\alpha = 0.05$).

The mean [std] of other six subjects' overall performance was 93.52% [0.09]. The mean [standard deviation] values of subjects' performance in the six sessions grouped by three blocks are shown in Table 5. In all six sessions and six subjects, 12 blocks were excluded from further analysis for unsatisfactory performance.

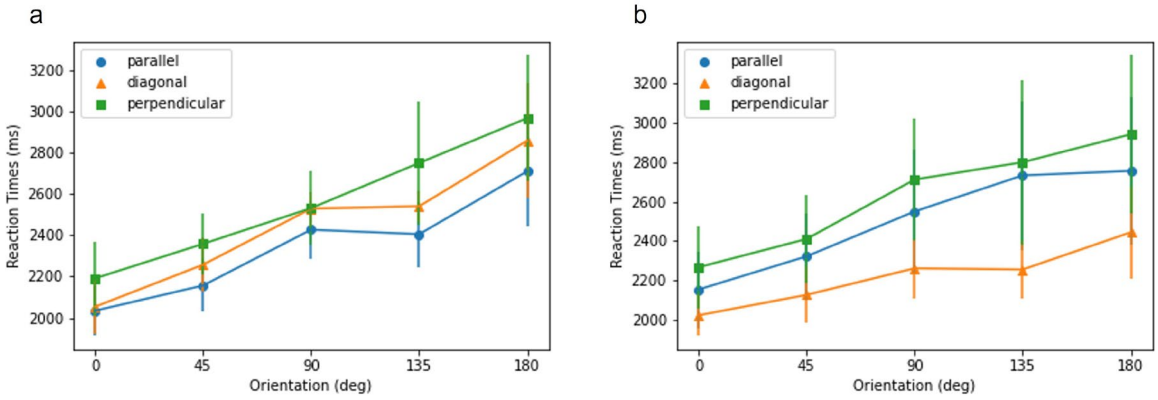

Figure 52: Reaction times with respect to orientation for the stimuli used in Experiment 3. (a) AR=1.6; (b) AR=2.

As shown in Figure 52, for all symmetry and aspect ratio conditions, RTs were lowest at the 0° orientation, in spite of three joint axes conditions. As shown in Figure 52 a, in the AR=1.6 condition, linearity was found in the mean RT~Orientation profiles from all three symmetry-elongation conditions (parallel: R^2=0.93; diagonal: R^2=0.94;

perpendicular: R^2=0.98). In addition, the slopes (ms/deg) and the intercepts (s) are shown in Table 6.

As shown in Figure 52 b, in the AR=2 condition, the linearity of mean RT~Orientation relationship reflected by R-squared values are as follows: parallel: R^2=0.94; diagonal: R^2=0.94; perpendicular: $R^2 = 0.96$. The slopes (ms/deg) and the intercepts (s) can also be found in Table 6. A repeated-measures ANOVA with orientations, AR, and dominant symmetry orientations as main factors showed a significant effect of orientation on the RT [orientation: F=7.16, p<0.01], but didn't find any other significant effect nor interactions. Detailed statistics can be found in Table 7.

Table 7: The results of repeated measure ANOVA with orientations, AR, and dominant symmetry orientations (Condition).

Anova

	F Value	Num DF	Den DF	Pr > F
Orientation	7.1639	4.0000	16.0000	0.0017
AR	0.0000	1.0000	4.0000	0.9982
Condition	1.7544	2.0000	8.0000	0.2335
Orientation:AR	1.0848	4.0000	16.0000	0.3969
Orientation:Condition	0.4078	8.0000	32.0000	0.9077
AR:Condition	3.3310	2.0000	8.0000	0.0886
Orientation:AR:Condition	1.1032	8.0000	32.0000	0.3868

4.3 Discussion

The 0 deg orientation in Figure 52 corresponds to the optimal orientation according to aspect ratio. The optimal orientations according to symmetry are 0, 45, and 90 deg for parallel, diagonal, and perpendicular conditions, respectively. The existence of a minimum at 0 deg in the data support the Winner-take-all hypothesis, and the aspect ratio

as the dominant feature compared to the symmetry. Consistent with the findings of Experiment 2, increased AR didn't cause significant effect on RTs. The data also support the sensorimotor theory of the memory storage as the mental rotation strategy was used in the object recognition tasks, indicated by the linear relation between the RTs and the orientations.

In the case of winner-take-all hypothesis, we were expecting RTs to be lowest in the congruent parallel condition, followed by the incongruent diagonal and perpendicular conditions. This is because, the smaller the difference between the two optimal oorientations, the faster we expected the competition between the two factors to settle. We observe this tendency in the AR=1.6 condition but not in the AR=2.0 condition. One possible reason could be the feature of shapes. When AR was 2, to meet the symmetrical properties along each axis, the texture of shapes was unbalanced comparing the left and right sides, as shown in Figure 50. With one side more filled than the other, the task could become easier by using this cue. However, let us note that the effect of condition (parallel, diagonal, perpendicular) was not significant in our data and these observations are not conclusive.

5 General Discussion

The results of our experiments show the use of canonical orientation during an object recognition task. In all the conditions, in average, subjects spent shortest time on recognizing objects in those orientations that were salient according to the factors we studied, i.e., symmetry and elongation. In Experiments 1 and 2, our data indicate that subjects selected either the most symmetrical or the most elongated orientation as the

canonical template when these factors were present in isolation. Moreover, our data suggest a Winner-Take-All process when the two factors were simultaneously present. In Experiment 3, the shortest reaction times correspond to the elongated orientation regardless of different symmetrical axes. Finally, our results support the sensorimotor theory of memory storage as our data reflect a mental rotation strategy based on the linear reaction time profile with respect to the orientation.

These results are inconsistent with invariant feature/relation approaches (Palmer, 1999; Sutherland, 1968; Barlow et al., 1972; Biederman, 1987), which predict equal time spent in recognizing a shape in different orientations. However, our results show a shortest reaction time associated with a specific (canonical) orientation and a linear relation as a function of the difference between the stimulus orientation and this canonical orientation. Some previous studies claimed that mental rotation is used only when the task is to determine the handedness and irrelevant to the cognitive processing the object recognition (Hinton and Parsons, 1981; Corballis et al., 1978). Our results, along with other studies (Tarr and Pinker, 1989; Jolicoeur, 1985; Charles Leek and Johnston, 2006), provide evidence against this claim: no linearity in reaction time when the objects didn't consist of any reference-frame affecting factors in Experiment 2 . With same handedness judgement task as other conditions, the shapes in the top row of Figure 47 didn't stimulate linear trend in the reaction times. Importantly, in this condition, the mean reaction time profile with respect to the orientation was not a flat curve. Therefore, the non-linearity was considered to be caused by the diversity of canonical orientations across subjects and shapes since there was no salient canonical indicators. This points out

that the canonical orientation was still used with textures without any geometrically solid axis.

Previous studies provided evidence for the role of symmetry and elongation in selecting a reference frame for different objects (Palmer, 1983; Rock, 1973; Marr and Nishihara, 1978; Mou et al., 2007). For example, Sekuler and Swimmer (2000) conducted an experiment and let subject to determine the primary axis of shapes with different symmetrical and elongation axes. They found that both the axis of symmetry and the axis of elongation were sufficient for deriving the primary axis and these two factors affected each other. However, they didn't address the question of how these axes are utilized in object recognition. Our experiment tested these two geometrical factors in object recognition and memory storage by a reinforcement learning process, which help us understand the correlation and interaction between the stored shape template and these two geometrical factors, and how the primary or canonical orientation was utilized through the sensorimotor process, mental rotation, in recognizing other orientations.

The empirical literature on the canonical orientation in object recognition shows that shapes are often most efficiently recognized in their upright orientations (Friedman and Hall, 1996; Corballis and McMaster, 1996). For example, it was shown that tilting objects, which were familiar to subjects according to their upright orientations (e.g., text and face stimuli), produced deleterious effects on the recognition performance (Freire et al., 2000; Jolicoeur, 1985; Rock, 1973; Rossion and Gauthier, 2002). This finding was also reported for unfamiliar objects. For example, Tarr and Pinker (1989) found that subjects stored the upright orientation of novel and meaningless objects in the training phase and

adopted a mental rotation strategy in the following recognition task phases. This is not surprising since the human sensory system is greatly influenced by gravity and balance relative to the vertical axis. Horizontal and vertical are two important axes in our behaviors in natural environments. Even though the vestibular system still signals vertical, our use of a VR headset aimed at reducing the involvement of these two primary environmental axes.

To conclude, our results provide evidence for canonical orientations determined by symmetry and elongation. However, as we mentioned in the Introduction section, this does not rule out strategies that can be driven by environmental cues. Our goal in this study was to focus on figural cues (symmetry and elongation) in isolation. Future studies can examine the combination of figural and environmental cues and assess whether the Winner-Take-All rule also holds for those combinations.

Chapter Five: Summary and Conclusions

Our perception relies on reference-frames in perceptual, motor, and cognitive processing. Although being analogous to coordinate systems in physics, the perceptual reference-frames are not as simple as a geometrical description of spatial relationships. Instead, the reference-frames our brain uses constitute complex mechanisms to compensate for the limitations of information sources, such as our vision that begins from a pair of two-dimensional retinal projections, and encodes information both precisely and continuously, so that we can adapt to and survive in a dynamically changing three-dimensional external environment. This dissertation's focus was on the use of different reference-frames that support sensory perception, motor behavior, and object recognition. In Chapter Two, we investigated the reference-frames involved in relative motion perception. We conducted psychophysical experiments that showed the use of both retinotopic and non-retinotopic reference-frames in the perception of motion. Moreover, we developed a neural model to explain the transition from retinotopic to non-retinotopic reference-frames. In addition to explaining results from several classical paradigms in motion perception, novel predictions of this model were tested and confirmed by our psychophysical experiments.

To perform motor behaviors accurately and precisely, our brain deploys motor reference-frames, according to which limb positions and movements are planned and executed. One of the most common behaviors is visually-guided and goal-directed

reaching. To achieve this, perceptual and motor reference-frames must be coordinated, which has been suggested to emerge via sensorimotor learning that starts in infancy (Piaget, 1952). In Chapter Three, we developed a neural model that establishes coordination between egocentric perceptual reference-frames (visual maps) and motor reference-frames (arm coordination) by a sequence of random arm explorations, inspired by infant sensorimotor organization (circular reactions). The model consists of visual map-representations and motor vector-representations. Synapses that follow outstar learning underwent active unsupervised self-organization through circular reactions. After learning (or self-organization), the model performed goal-directed reaching tasks successfully. We also found that the performance could be improved by using a more exocentric reference-frame. These results and findings validated the plausibility of Piagetian theory of cognitive development in the synthesis of exocentric reference-frames beginning from egocentric reference-frames through developmental stages.

Object recognition involves the matching of the current appearance of an object (stimulus) with the memory representations of candidate objects. This task is complicated because objects do not have unique appearances: As the relative position of the object with respect to the observer changes (perspective views), the appearance of the object can undergo drastic changes. Several theories have been formulated to explain how the brain accomplishes this "invariant object recognition". The invariant feature approach bypasses reference-frames by selecting representations that are independent of specific reference-frames. View-variant representation theories implicitly involve reference-frames since

the interpolations required for generalization necessitate the establishment of relations between various view-variant memory traces.

The theory that uses most explicitly reference-frames is the "canonical representation theory". This theory makes use of an object-centered reference-frame which specifies a "canonical form" for memory storage and recognition. To shed light on the reference-frames involved in this approach, in Chapter Four, we used a reinforcement learning paradigm with a mental rotation task to test if and how subjects use canonical forms during object recognition. We hypothesized that, if the canonical form was used, sensorimotor transformation must be adopted. Therefore, the reaction times should be linearly dependent on the angular disparity between the visual stimulus' orientation and the canonical orientation. Our results suggest that subjects align the observed objects and the memory traces according to canonical reference-frames. Furthermore, we showed that both boundary and texture information can be used to determine canonical forms. Symmetry and aspect ratio were found to influence the selection of canonical forms. When both aspect ratio and symmetry were present on a single shape, a winner-take-all strategy was adopted in deciding the canonical form.

Taken together, in this dissertation we used neural modeling along with psychophysical experiments to provide a better understanding of how different reference-frames link and communicate to support sensory, motor, and cognitive processing. Our neural models instantiated processes of how non-retinotopic reference-frames are synthesized from retinotopic reference-frames, and how sensory and motor reference-frames are coordinated.

Chapter Six: Future Directions

The links between retinotopic and non-retinotopic reference-frames, egocentric and exocentric reference-frames, and the correlations between these reference-frames within visual, motor, and cognitive processing are still not completely understood. Based on the experiments and studies discussed in the previous chapters, there are still fundamental questions deserving further investigations to provide a more comprehensive understanding of neural processing. In this section, we will outline several experiments for future work.

1. A quantitative measurement to how geometrical factors determine the canonical orientation

As we have already shown, both symmetry and aspect ratio can impact the canonical orientation and a winner-take-all strategy is used when both factors are present. It is important to know how these two factors interact quantitatively in determining the canonical orientation. To investigate this question, one should test more aspect ratios beginning from 1. In Experiment 3 of Chapter Four, we used two aspect ratios, 1.6 and 2, and found in both cases that elongation determined the canonical axis regardless of symmetrical axes. However, the relations among reaction times associated with congruent geometry-elongation axes and incongruent axes were not consistent within two aspect-ratio conditions. Moreover, reaction times were not significantly different between two aspect-ratio conditions. These all indicated that elongation may not always be the winning factor compared to symmetry in determining the canonical orientation.

Therefore, one can find out how the dominant factor changes with respect to the aspect ratio as well as the relationship between reaction time and inter-factor congruency by testing more aspect ratios lying in the range from 1 to 1.6.

2. Binocular combination and depth perception

As mentioned in Chapter Three, our visual system synthesizes exocentric reference-frames through egocentric reference-frames. In this synthesis, one fundamental stage is the formalization of three-dimensional reference-frames based on a pair of two-dimensional retinotopic representations. It has been suggested that our binocular vision is not as simple as averaging the information from two retinas (Basgoze, Mackey, Cooper, 2018 Review; Wang et al., 2022). In Chapter Three, we used the Ding-Sperling model (Ding and Levi, 2017; Ding and Sperling, 2006, 2007) to replicate the binocular combination by contrast-weighted summation. However, many properties of the visual system have not been considered. For example, it has been shown that the two eyes contribute differently to the combination of dichoptic images from two eyes, and both winner-dominant and loser-dominant strategies exist (Legge and Rubin, 1981). Therefore, computational models explaining these mechanisms are still needed for deeper understanding of the relevant experimental data.

Also, the literature shows that spatial frequency has impact on binocular interactions (Alberti and Bex, 2018). In our model described in Chapter Three, we used spatial frequency responses on different horizontal spots of each retinal image to analyze the binocular correspondence. Our data validated this strategy as our algorithm efficiently

converged the two retinal images with very slight mis-localization. However, direct neural evidence supporting this strategy is still lacking.

Another important element of binocular combination is the interocular velocity differences during the object motion. When an object is moving through the depth dimension, its monocular velocities on two retinas are different. Previous studies suggested that visual system can perform the computation of interocular velocity differences, which indicate that this difference might also serve as a cue for depth perception (Review: Cormack et al., 2017). In Chapter Three, our model used a cyclopean level cue, binocular disparity, to extract the depth information. This operation is sufficient for our model as the objects were static in our simulation. However, one should consider the interocular velocity cues when simulating tasks with objects in motion.

3. Hierarchical structures of reference-frames

In Chapter Two, we provided a neural model for relative motion perception by synthesizing non-retinotopic reference-frames from retinotopic ones. As already mentioned in Chapter Two, one of the challenging future directions for this work consists of developing further the reference-frame detection layer by introducing additional Gestalt principles, by allowing the determination of multiple groups simultaneously, and by allowing task-specific modulations where appropriate. For complex stimuli, not only separate reference-frames are needed for separate groups, but also a hierarchy of reference-frames can be established. For example, for a walker a simple interpretation is to have the lateral movement of the walker as the reference-frame and interpret all other

motions relative to that reference-frame. A more detailed analysis, however, may consider a hierarchy of reference-frames: The arm moves with respect to the torso, the hand moves with respect to the arm, and the finger moves with respect to the hand.

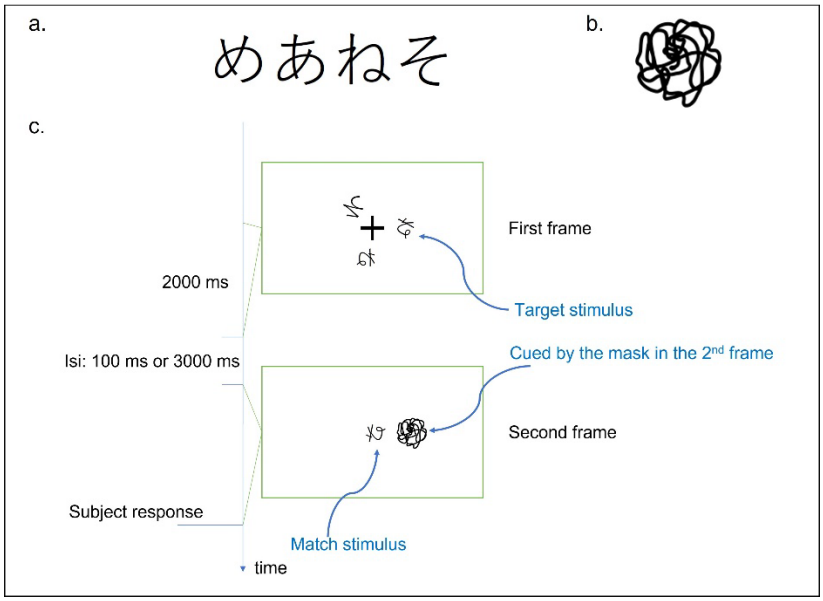

Figure 53: Experimental Paradigm. a. The stimulus set consisted of four Japanese Hiragana letters. In each trial, three letters were randomly selected and presented, in either the original or the flipped form (randomly selected). b. A mask stimulus was used as a cue. c. Schematic of a trial. d. Two values of ISI were used to investigate the role of timing between the stimulus offset and the cue.

4. Correlation between Sensorimotor Transformation and Memory

As we showed in Chapter Four, sensorimotor transformation, more specifically mental rotation, is needed in object recognition when canonical forms serve as the reference-frames in memory storage. Therefore, in order to understand the link between sensory reference-frame and motor processing, one should investigate where and how mental rotation takes place.

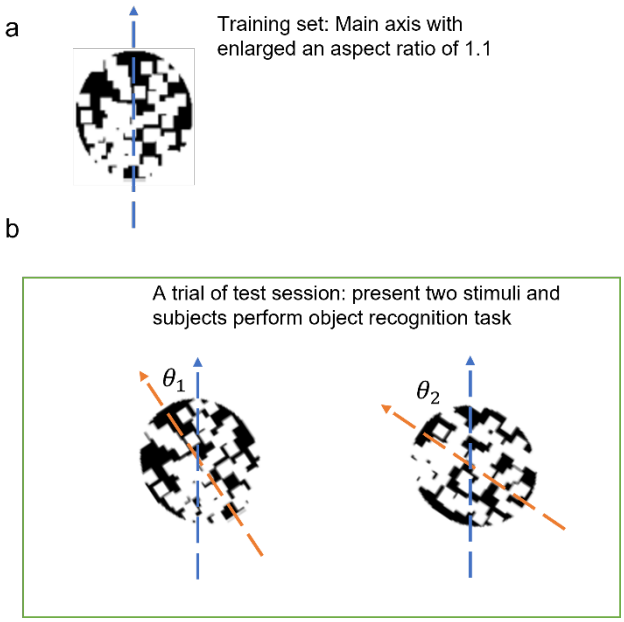

Figure 54: Experimental Paradigm. a. The training set consists of a series of textures with no salient canonical cues but an enlarged orientation as the main axis (marked by the blue arrow that is invisible in the experiment) with an aspect ratio of 1.1. b. In a trial, subjects are supposed to perform object recognition task on two objects presented on two consecutively rotated orientations (main axes are marked by the orange arrows that are invisible during the experiment).

Our preliminary data suggested that mental rotation is carried out in working memory. When the relevant stimulus is not already in working memory, this mental rotation is preceded by a transfer from sensory memory. In a preliminary experiment, as shown in Figure 53, we used a cue-delay paradigm with short and long inter-stimulus-interval (ISI) conditions incorporating the mental rotation task. The main difference between short and long ISI conditions is the possibility of selective-transfer from sensory memory to working memory in the short but not the long ISI condition. In addition to the psychophysical experiments, we used EEG to record and analyze event-related potentials (ERPs) and event-related time-frequency dynamics of electrical activity arising from the

brain. In the data, rotation-related negativity (Provost et al., 2013) was observed in both cue-delay conditions and RTs were similar. Frontal beta band ERD- ERS was found in both conditions. These observations indicate a common sensorimotor strategy involving mental rotation regardless of cue delay (Tatti et al., 2021). Moreover, stronger Frontal P1 in the short ISI condition was found, which was likely to reflect selective transfer from sensory to working memory (Baruth et al., 2010). Additional evidence came from the stronger Frontal N4 and early Theta band ERS in the long cue-delay condition, which can be interpreted as signatures of additional working memory maintenance operation in the absence of selective transfer (Gunter et al., 1995; Missonnier et al., 2006). These results support the hypothesis that mental rotation take place in working memory. However, further evidence from more comprehensive approaches is still needed and the neural mechanisms of selective- and unselective-transfers are still not well understood.

Another question that arises when it comes to the correlation between mental rotation and memory during object recognition is whether people rotate the visual stimulus or the memory trace. One could provide insight to answer this question by the following experimental design as shown in Figure 54. In this experiment, subjects will learn a set of textures without any salient canonical cue but the elongation along one axis with the aspect ratio of 1.1, as shown in Figure 54 a. These textures will be labeled, e.g., texture 1, texture 2, etc. After the learning session, subjects will enter the test session of experiment, during which in each trial they will be presented with two textures with same elongation as those in the learning session but with different orientations, as shown in the Figure 54 b. Note that the orientation of texture on the left will always be less rotated

than the right one deviating from the vertical orientation. The task will be, from left to right, to report whether the presented texture is the same as a specific labeled texture by pressing one of two buttons. We predict that reaction times will be linearly dependent on the maximum of θ_1 and θ_2 if subjects rotate their memory traces of the training textures to recognize the presented textures in the testing session. However, if they rotate each presented texture and compare them with the memory traces, the reaction times will be linearly dependent on the summation of θ_1 and θ_2.

References

Aben, B., Stapert, S., & Blokland, A. (2012). About the distinction between working memory and short-term memory. Frontiers in psychology, 3, 301.

Agaoglu, M. N., Clarke, A. M., Herzog, M. H., & Öğmen, H. (2016). Motion-based nearest vector metric for reference frame selection in the perception of motion. Journal of Vision, 16(7), 14-14.

Agaoglu, M. N., Herzog, M. H., & Öğmen, H. (2015). The effective reference frame in perceptual judgments of motion direction. Vision Research, 107, 101-112.

Alberti, C. F., & Bex, P. J. (2018). Binocular contrast summation and inhibition depends on spatial frequency, eccentricity and binocular disparity. Ophthalmic & Physiological Optics, 38(5), 525–537, https://doi.org/10.1111/opo.12581.

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. Journal of neurophysiology, 52(6), 1106-1130.

Allman, J., Miezin, F., & McGuinness, E. (1985). Direction-and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). Perception, 14(2), 105-126.

Andersen, R. A., and Mountcastle, V. B. (1983). The influence of the angle of gaze upon the excitability of the light-sensitive neurons of the posterior parietal cortex. J. Neurosci. 3, 532–548. doi: 10.1523/JNEUROSCI.03-03-00532.1983

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(5), 1063.

Anguera, J. A., Reuter-Lorenz, P. A., Willingham, D. T., & Seidler, R. D. (2010). Contributions of spatial working memory to visuomotor learning. Journal of cognitive neuroscience, 22(9), 1917-1930.

Asuni, G., Leoni, F., Guglielmelli, E., Starita, A., and Dario, P. (2003). "A neuro-controller for robotic manipulators based on biologically-inspired visuo-motor co-ordination neural models," in First International IEEE EMBS Conference on Neural Engineering, 2003. Conference Proceedings (Capri: IEEE), 450–453.

Asuni, G., Teti, G., Laschi, C., Guglielmelli, E., and Dario, P. (2006). "Extension to end-effector position and orientation control of a learning-based neurocontroller for a humanoid arm," in 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (Beijing: IEEE), 4151–4156.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Psychology of learning and motivation (Vol. 2, pp. 89-195). Academic Press.

Baker, P. M., & Bair, W. (2016). A model of binocular motion integration in MT neurons. Journal of Neuroscience, 36(24), 6563-6582.

Barlow, H., Narasimhan, R., and Rosenfeld, A. (1972). Visual pattern analysis in machines and animals: The same principles may underlie the operation of sensory neurons, computer programs, and perception. Science 177, 567–575

Barsalou, L. W. (2008). Grounded cognition. Annu. Rev. Psychol. 59, 617–645. doi: 10.1146/annurev.psych.59.103006.093639

Baruth, J., Casanova, M., Sears, L., & Sokhadze, E. (2010). Early-stage visual processing abnormalities in high-functioning autism spectrum disorder (ASD). *Translational Neuroscience*, *1*(2), 177-187.

Başgöze, Z., Mackey, A. P., & Cooper, E. A. (2018). Plasticity and adaptation in adult binocular vision. *Current biology*, *28*(24), R1406-R1413.

Berman, N. E., Wilkes, M. E., & Payne, B. R. (1987). Organization of orientation and direction selectivity in areas 17 and 18 of cat cerebral cortex. Journal of Neurophysiology, 58(4), 676-699.

Beurze SM, Van Pelt S, Medendorp WP. Behavioral reference frames for planning human reaching movements. J Neurophysiol. 2006; 96:352–362.

Beverley, K., and Regan, D. (1974). Visual sensitivity to disparity pulses: evidence for directional selectivity. Vision Res. 14, 357–361. doi: 10.1016/0042-6989(74)90095-9

Bex, P. J., Metha, A. B., & Makous, W. (1998). Psychophysical evidence for a functional hierarchy of motion processing mechanisms. JOSA A, 15(4), 769-776.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. Psychological review 94, 115

Blohm, G., and Crawford, J. D. (2009). Fields of gain in the brain. Neuron 64, 598–600. doi: 10.1016/j.neuron.2009.11.022

Boi, M., Ogmen, H., Krummenacher, J., Otto, T. U., & Herzog, M. H. (2009). A (fascinating) litmus test for human retino- vs.non-retinotopic processing. Journal of Vision, 9(13), 5–5. https://doi.org/10.1167/9.13.5

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. Journal of Neuroscience, 16(11), 3737-3744.

Born, R. T., & Tootell, R. B. (1992). Segregation of global and local motion processing in primate middle temporal visual area. Nature, 357(6378), 497-499.

Bradley, D. (2001). MT signals: better with time. Nature Neuroscience, 4(4), 346-348.

Britten, K. H., & Heuer, H. W. (1999). Spatial summation in the receptive fields of MT neurons. Journal of Neuroscience, 19(12), 5074-5084.

Bu¨lthoff, H. H. and 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

Buffart, H., Leeuwenberg, E., & Restle, F. (1981). Coding theory of visual pattern completion. Journal of Experimental Psychology: Human Perception and Performance, 7(2), 241.

Bullock, D., Grossberg, S., and Guenther, F. H. (1993). A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm. J. Cogn. Neurosci. 5, 408–435. doi: 10.1162/jocn.1993.5.4.408

Caminiti, R., Johnson, P. B., and Urbano, A. (1990). Making arm movements within different parts of space: dynamic aspects in the primate motor cortex. J. Neurosci. 10, 2039–2058. doi: 10.1523/JNEUROSCI.10-07-02039.1990

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. Nature Reviews Neuroscience, 13(1), 51-62.

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. Trends in cognitive sciences, 18(8), 414-421.

Chao, F., Lee, M. H., and Lee, J. J. (2010). A developmental algorithm for ocular-motor coordination. Rob. Auton. Syst. 58, 239–248. doi: 10.1016/j.robot.2009.08.002

Chao, F., Zhu, Z., Lin, C.-M., Hu, H., Yang, L., Shang, C., et al. (2016). Enhanced robotic hand-eye coordination inspired from human-like behavioral patterns. IEEE Trans. Cogn. Dev. Syst. 10, 384–396. doi: 10.1109/TCDS.2016.2620156

Charles Leek, E. and Johnston, S. J. (2006). A polarity effect in misoriented object recognition: The role of polar features in the computation of orientation-invariant shape representations. Visual Cognition 13, 573–600

Chey, J., Grossberg, S., & Mingolla, E. (1997). Neural dynamics of motion grouping: From aperture ambiguity to object speed and direction. JOSA A, 14(10), 2570-2594.

Clarke, A. M., Öğmen, H., & Herzog, M. H. (2016). A computational model for reference-frame synthesis with applications to motion perception. Vision Research, 126, 242-253.

Cooper, L. A. (1975). Mental rotation of random two-dimensional shapes. Cognitive psychology 7, 20–43

Cooper, L. A. and Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In Visual information processing (Elsevier). 75–176

Cooper, L. A., & Shepard, R. N. (1973). The time required to prepare for a rotated stimulus. Memory & cognition, 1(3), 246-250.

Corballis, M. C. and McMaster, H. (1996). The roles of stimulus-response compatibility and mental rotation in mirror-image and left-right decisions. Canadian Journal of Experimental Psychology 50, 397

Corballis, M. C., Zbrodoff, N. J., Shetzer, L. I., and Butler, P. B. (1978). Decisions about identity and orientation of rotated letters and digits. Memory & Cognition 6, 98–107

Cormack, L. K., Czuba, T. B., Knöll, J., & Huk, A. C. (2017). Binocular mechanisms of 3D motion processing. *Annual Review of Vision Science*, *3*, 297.

Davidson, R. M., & Bender, D. B. (1991). Selectivity for relative motion in the monkey superior colliculus. Journal of Neurophysiology, 65(5), 1115-1133.

Davis, J. W., & Gao, H. (2004). An expressive three-mode principal components model for gender recognition. Journal of Vision, 4(5), 2-2.

Davis, W., Frederick, C. M., and Valois, K. K. (2003). Building a representation of aspect ratio. Journal of Vision 3, 13–13

Della Santina, C., Arapi, V., Averta, G., Damiani, F., Fiore, G., Settimi, A., et al. (2019). Learning from humans how to grasp: a data-driven architecture for autonomous grasping with anthropomorphic soft hands. IEEE Rob. Autom. Lett. 4, 1533–1540. doi: 10.1109/LRA.2019.2896485

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods, 134(1), 9-21.

D'Esposito, M., Ballard, D., Zarahn, E., & Aguirre, G. K. (2000). The role of prefrontal cortex in sensory memory and motor preparation: an event-related fMRI study. Neuroimage, 11(5), 400-408.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? Neuron 73, 415–434

Ding, J., & Sperling, G. (2006). A gain-control theory of binocular combination. *Proceedings of the National Academy of Science USA*, 103 (4), 1141–1146.

Ding, J., & Sperling, G. (2007). Binocular combination: Measurements and a model. In Harris L. & Jenkin M. (Eds.), *Computational vision in neural and machine systems* (pp. 257–305). Cambridge, UK: Cambridge Unversity Press.

Ding, J., and Levi, D. M. (2017). Binocular combination of luminance profiles. J. Vis. 17, 4–4. doi: 10.1167/17.13.4

Dodd, Michael D., Tara McAuley, and Jay Pratt. "An illusion of 3-D motion with the Ternus display." Vision research 45.8 (2005): 969-973

Dresp-Langley, B., and Grossberg, S. (2016). Neural computation of surface border ownership and relative surface depth from ambiguous contrast inputs. Front. Psychol. 7:1102. doi: 10.3389/fpsyg.2016.01102

Duijnhouwer, J., Noest, A., Lankheet, M., Van Den Berg, A., & Van Wezel, R. J. (2013). Speed and direction response profiles of neurons in macaque MT and MST show modest constraint line tuning. Frontiers in behavioral neuroscience, 7, 22.

Duncker, K. Uber induzierte bewegung, Psychologische Forschung 12 (1) (1929) 180{259}. (1)

Edelman, S. and Bulthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vision research 32, 2385–2400

Engel KC, Flanders M, Soechting JF. Oculocentric frames of reference for limb movement. Archives italiennes de biologie. 2002; 140:211–219.

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cerebral cortex (New York, NY: 1991), 7(2), 181-192.

Fang, F. and He, S. (2005). Viewer-centered object representation in the human visual system revealed by viewpoint aftereffects. Neuron 45, 793–800

Feldman, A. G. (2019). Indirect, referent control of motor actions underlies directional tuning of neurons. J. Neurophysiol. 121, 823–841. doi: 10.1152/jn.00575.2018

Fink, G. R., Marshall, J. C., Weiss, P. H., Stephan, T., Grefkes, C., Shah, N. J., ... & Dieterich, M. (2003). Performing allocentric visuospatial judgments with induced distortion of the egocentric reference frame: an fMRI study with clinical implications. Neuroimage, 20(3), 1505-1517.

Fo¨rster, B., Gebhardt, R.-P., Lindlar, K., Siemann, M., and Delius, J. D. (1996). Mental-rotation effect: A function of elementary stimulus discriminability? Perception 25, 1301–1316

Francuz, P., & Zapała, D. (2011). The suppression of the μ rhythm during the creation of imagery representation of movement. Neuroscience letters, 495(1), 39-43.

Freeman, T. C. (2001). Transducer models of head-centred motion perception. Vision research, 41(21), 2741-2755.

Freire, A., Lee, K., and Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. Perception 29, 159–170

Friedman, A. and Hall, D. L. (1996). The importance of being upright: Use of environmental and viewer-centered reference frames in shape discriminations of novel three-dimensional objects. Memory & Cognition 24, 285–295

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. Biological cybernetics 20, 121–136

Gardony, A. L., Eddy, M. D., Brunyé, T. T., & Taylor, H. A. (2017). Cognitive strategies in the mental rotation task revealed by EEG spectral power. Brain and Cognition, 118, 1-18.

Gardony, A. L., Eddy, M. D., Brunyé, T. T., & Taylor, H. A. (2017). Cognitive strategies in the mental rotation task revealed by EEG spectral power. Brain and Cognition, 118, 1-18.

Gardony, A. L., Eddy, M. D., Brunye´, T. T., and Taylor, H. A. (2017). Cognitive strategies in the mental rotation task revealed by eeg spectral power. Brain and Cognition 118, 1–18

Gauthier, I. and Tarr, M. J. (2016). Visual object recognition: Do we (finally) know more now than we did? Annual review of vision science 2, 377–396

Geisler, W. S., Albrecht, D. G., Crane, A. M., & Stern, L. (2001). Motion direction signals in the primary visual cortex of cat and monkey. Visual neuroscience, 18(4), 501-516.

Gershman, Samuel J., Joshua B. Tenenbaum, and Frank Jäkel. "Discovering hierarchical motion structure." Vision research 126 (2016): 232-241.

Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. Cerebral cortex (New York, NY: 1991), 7(4), 374-385.

Ghose, G. M., & Maunsell, J. H. (2008). Spatial summation can explain the attentional modulation of neuronal responses to multiple stimuli in area V4. Journal of neuroscience, 28(19), 5115-5126.

Gilaie-Dotan, S., Saygin, A. P., Lorenzi, L. J., Rees, G., & Behrmann, M. (2015). Ventral aspect of the visual form pathway is not critical for the perception of biological motion. Proceedings of the National Academy of Sciences, 112(4), E361-E370.

Gogel, W. C. (1977). The metric of visual space. Stability and constancy in visual perception: Mechanisms and processes, 129-181.

Goldenberg, A., Benhabib, B., and Fenton, R. (1985). A complete generalized solution to the inverse kinematics of robots. IEEE Journal on Robotics and Automation 1, 14–20. doi: 10.1109/JRA.1985.1086995

Goldstein, E. Bruce, and James Brockmole. Sensation and perception. Cengage Learning, 2016.

Gomez, P., Shutter, J., and Rouder, J. N. (2008). Memory for objects in canonical and noncanonical viewpoints. Psychonomic bulletin & review 15, 940–944

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. Trends Neurosci. 15, 20–25. doi: 10.1016/0166-2236(92)90344-8

Graham, G. (2000). "Behaviorism," in Stanford Encyclopedia of Philosophy, eds E. N. Zalta, U. Nodelman, C. Allen, R.L. Anderson (New York, NY: Stansford University), 520–630.

Gramann, K., Onton, J., Riccobon, D., Mueller, H. J., Bardins, S., & Makeig, S. (2010). Human brain dynamics accompanying use of egocentric and allocentric reference frames during navigation. Journal of cognitive neuroscience, 22(12), 2836-2849.

Gray, R., and Regan, D. (1996). Cyclopean motion perception produced by oscillations of size, disparity and location. Vision Res. 36, 655–665. doi: 10.1016/0042-6989(95)00145-X

Grossberg, S. (1982). Contour enhancement, short term memory, and constancies in reverberating neural networks. In Studies of mind and brain (pp. 332-378). Springer, Dordrecht.

Grossberg, S., Léveillé, J., & Versace, M. (2011). How do object reference frames and motion vector decomposition emerge in laminar cortical circuits?. Attention, Perception, & Psychophysics, 73(4), 1147-1170.

Grossberg, S., Mingolla, E., & Ross, W. D. (1997). Visual brain and visual perception: how does the cortex do perceptual grouping? Trends in neurosciences, 20(3), 106-111.

Grzywacz, N. M., & Yuille, A. L. (1990). A model for the estimate of local image velocity by cells in the visual cortex. Proceedings of the Royal Society of London. B. Biological Sciences, 239(1295), 129-161.

Gunter, T. C., Jackson, J. L., & Mulder, G. (1995). Language, memory, and aging: an electrophysiological exploration of the N400 during reading of memory‐demanding sentences. *Psychophysiology*, *32*(3), 215-229.

Heeger, D. J. (1987). Model for the extraction of image flow. JOSA A, 4(8), 1455-1471.

Heeger, D. J., Boynton, G. M., Demb, J. B., Seidemann, E., & Newsome, W. T. (1999). Motion opponency in visual cortex. Journal of Neuroscience, 19(16), 7162-7174.

Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (1996). Computational models of cortical visual processing. Proceedings of the National Academy of Sciences, 93(2), 623-627.

Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (1996). Computational models of cortical visual processing. Proceedings of the National Academy of Sciences, 93(2), 623-627.

Heil, M. (2002). The functional significance of ERP effects during mental rotation. Psychophysiology, 39(5), 535-545.

Heil, M., Rauch, M., & Hennighausen, E. (1998). Response preparation begins before mental rotation is finished: Evidence from event-related brain potentials. Acta Psychologica, 99(2), 217-232.

Heydt, R., Krieger, F. T. Q., and He, Z. J. (2003). Neural mechanisms in border ownership assignment: motion parallax and gestalt cues. J. Vis. 3, 666–666. doi: 10.1167/3.9.666

Hinton, G. E. and Parsons, L. M. (1981). Frames of reference and mental imagery. Attention and performance IX, 261–277

Hirai, Y., and Fukushima, K. (1978). An inference upon the neural network finding binocular correspondence. Biol. Cybern. 31, 209–217. doi: 10.1007/BF00337092

Hoffmann, M., Chinn, L. K., Somogyi, E., Heed, T., Fagard, J., Lockman, J. J., et al. (2017). "Development of reaching to the body in early infancy: from experiments to robotic models," in 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob) (Lisbon: IEEE), 112–119.

Hsieh, L. T., & Ranganath, C. (2014). Frontal midline theta oscillations during working memory maintenance and episodic encoding and retrieval. Neuroimage, 85, 721-729.

Huynh, D., Tripathy, S. P., Bedell, H. E., & Öğmen, H. (2017). The reference frame for encoding and retention of motion depends on stimulus set size. Attention, Perception, & Psychophysics, 79(3), 888-910.

Hyun, J. S., & Luck, S. J. (2007). Visual working memory as the substrate for mental rotation. Psychonomic bulletin & review, 14(1), 154-158.

Jamone, L., Brandao, M., Natale, L., Hashimoto, K., Sandini, G., and Takanishi, A. (2014). Autonomous online generation of a motor representation of the workspace for intelligentfor intelligent whole-body reaching. Rob. Auton. Syst. 62, 556–567. doi: 10.1016/j.robot.2013.12.011

Johansson, G. (1976) Spatiotemporal differentiation and integration in visual motion perception. Psychological research 38 (4) 379–393.

Johansson, Gunnar. "Visual perception of biological motion and a model for its analysis." Perception & psychophysics 14.2 (1973): 201-211.

Jolicoeur, P. (1985). The time to name disoriented natural objects. Memory & cognition 13, 289–303

Jolicoeur, P. (1988). Mental rotation and the identification of disoriented objects. Canadian Journal of Psychology/Revue canadienne de psychologie, 42(4), 461.

Jolicoeur, P. (1988). Mental rotation and the identification of disoriented objects. Canadian Journal of Psychology/Revue canadienne de psychologie 42, 461

Joly, T. J., & Bender, D. B. (1997). Loss of relative-motion sensitivity in the monkey superior colliculus after lesions of cortical area MT. Experimental Brain Research, 117(1), 43-58.

Jordan, K., Heinze, H. J., Lutz, K., Kanowski, M., & Jäncke, L. (2001). Cortical activations during the mental rotation of different visual objects. Neuroimage, 13(1), 143-152.

Julesz, B. (1971). Foundations of Cyclopean Perception. Chicago, IL: Chicago Press.

Kail, R., & Hall, L. K. (2001). Distinguishing short-term memory from working memory. Memory & cognition, 29(1), 1-9.

Karg, M., Kühnlenz, K., & Buss, M. (2010). Recognition of affect based on gait patterns. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 40(4), 1050-1061.

Kay, K. N., Winawer, J., Mezer, A., & Wandell, B. A. (2013). Compressive spatial summation in human visual cortex. Journal of neurophysiology, 110(2), 481-494.

Koffka, K. (1935) Principles of Gestalt Psychology. New York: Harcourt Brace

Koffka, K. (2013). Principles of Gestalt psychology. Routledge.

Kourtzi, Z. and Shiffrar, M. (1999). The visual representation of three-dimensional, rotating objects. Acta Psychologica 102, 265–292

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM 60, 84–90

Kumano, H., & Uka, T. (2012). Reduction in receptive field size of macaque MT neurons in the presence of visual noise. Journal of Neurophysiology, 108(1), 215-226.

La Chioma, A., Bonhoeffer, T., and Hübener, M. (2020). Disparity sensitivity and binocular integration in mouse visual cortex areas. bioRxiv. doi: 10.1523/JNEUROSCI.1060-20.2020

Lagae, L., Raiguel, S., & Orban, G. A. (1993). Speed and direction selectivity of macaque middle temporal neurons. Journal of neurophysiology, 69(1), 19-39.

Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. Journal of Neuroscience, 26(11), 2894-2906.

Laschi, C., Asuni, G., Guglielmelli, E., Teti, G., Johansson, R., Konosu, H., et al. (2008). A bio-inspired predictive sensory-motor coordination scheme for robot reaching and preshaping. Auton. Rob. 25, 85–101. doi: 10.1007/s10514-007-9065-4

Law, J., Shaw, P., and Lee, M. (2013). A biologically constrained architecture for developmental learning of eye-head gaze control on a humanoid robot. Auton. Rob. 35, 77–92. doi: 10.1007/s10514-013-9335-2

Layton, O. W., and Yazdanbakhsh, A. (2015). A neural model of border-ownership from kinetic occlusion. Vision Res. 106, 64–80. doi: 10.1016/j.visres.2014.11.002

Lega, C., Pirruccio, M., Bicego, M., Parmigiani, L., Chelazzi, L., and Cattaneo, L. (2020). The topography of visually guided grasping in the premotor cortex: a dense-transcranial magnetic stimulation (tms) mapping study. J. Neurosci. 40, 6790–6800. doi: 10.1523/JNEUROSCI.0560-20.2020

Legge, G. E., & Rubin, G. S. (1981). Binocular interactions in suprathreshold contrast perception. Perception & Psychophysics, 30(1), 49–61, https://doi.org/10.3758/BF03206136.

Llanos, C., Rodriguez, M., Rodriguez-Sabate, C., Morales, I., & Sabate, M. (2013). Mu-rhythm changes during the planning of motor and motor imagery actions. Neuropsychologia, 51(6), 1019-1026.

Logothetis, N. K. and Sheinberg, D. L. (1996). Visual object recognition. Annual review of neuroscience 19, 577–621

Lu, Z. L., & Sperling, G. (2001). Three-systems theory of human visual motion perception: review and update. JOSA A, 18(9), 2331-2370.

Mackrous, I., and Proteau, L. (2016). Visual online control of goal-directed aiming movements in children. Front. Psychol. 7:989. doi: 10.3389/fpsyg.2016.00989

Mahoor, Z., MacLennan, B. J., and McBride, A. C. (2016). "Neurally plausible motor babbling in robot reaching," in 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob) (Cergy-Pontoise: IEEE), 9–14.

Manocha, D., and Canny, J. F. (1994). Efficient inverse kinematics for general 6r manipulators. IEEE Trans. Rob. Autom. 10, 648–657. doi: 10.1109/70.326569

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London. Series B. Biological Sciences 200, 269–294

Marvel, C. L., Morgan, O. P., & Kronemer, S. I. (2019). How the motor system integrates with working memory. Neuroscience & Biobehavioral Reviews, 102, 184-194.

Maunsell, J. H., & Van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. Journal of neurophysiology, 49(5), 1127-1147.

Michel, C. M., Kaufman, L., and Williamson, S. J. (1994). Duration of eeg and meg α suppression increases with angle in a mental rotation task. Journal of cognitive neuroscience 6, 139–150

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. Trends Neurosci. 6, 414–417. doi: 10.1016/0166-2236(83)90190-X

Missonnier, P., Deiber, M. P., Gold, G., Millet, P., Gex-Fabry Pun, M., Fazio-Costa, L., ... & Ibáñez, V. (2006). Frontal theta event-related synchronization: comparison of directed attention and working memory load effects. *Journal of neural transmission*, *113*(10), 1477-1486.

Mohammed, A. A., and Sunar, M. (2015). "Kinematics modeling of a 4-dof robotic arm," in 2015 International Conference on Control, Automation and Robotics (Singapore: IEEE), 87–91.

Mou, W., Zhao, M., & McNamara, T. P. (2007). Layout geometry in the selection of intrinsic frames of reference from multiple viewpoints. Journal of experimental psychology: Learning, Memory, and Cognition, 33(1), 145.

Nakayama, K., & Silverman, G. H. (1988). The aperture problem—I. Perception of nonrigidity and motion direction in translating sinusoidal lines. Vision research, 28(6), 739-746.

Needham, A. (2000). Improvements in object exploration skills may facilitate the development of object segregation in early infancy. J. Cogn. Dev. 1, 131–156. doi: 10.1207/S15327647JCD010201

Neri, P., Morrone, M. C., & Burr, D. C. (1998). Seeing biological motion. Nature, 395(6705), 894-896.

Nguyen, P. D., Hoffmann, M., Pattacini, U., and Metta, G. (2019). “Reaching development through visuo-proprioceptive-tactile integration on a humanoid robot-a deep learning approach,” in 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob) (Oslo: IEEE), 163–170.

Öğmen, H. (1993). A neural theory of retino-cortical dynamics. Neural networks, 6(2), 245-273.

Öğmen, H. (2007). A theory of moving form perception: Synergy between masking, perceptual grouping, and motion computation in retinotopic and non-retinotopic representations. Advances in Cognitive Psychology, 3(1-2), 67.

Ogmen, H., & Herzog, M. H. (2010). The geometry of visual perception: Retinotopic and nonretinotopic representations in the human visual system. Proceedings of the IEEE, 98(3), 479-492.

Okada, Y. C., & Salenius, S. (1998). Roles of attention, memory, and motor preparation in modulating human brain activity in a spatial working memory task. Cerebral cortex (New York, NY: 1991), 8(1), 80-96.

Oleksiak, A., Klink, P. C., Postma, A., van der Ham, I. J., Lankheet, M. J., & van Wezel, R. J. (2011). Spatial summation in macaque parietal area 7a follows a winner-take-all rule. Journal of neurophysiology, 105(3), 1150-1158.

Olson, C. R. (2003). Brain representation of object-centered space in monkeys and humans. Annual review of neuroscience, 26(1), 331-354.

Onton, J., Delorme, A., & Makeig, S. (2005). Frontal midline EEG dynamics during working memory. Neuroimage, 27(2), 341-356.

Osuagwu, B. A. and Vuckovic, A. (2014). Similarities between explicit and implicit motor imagery in mental rotation of hands: An eeg study. Neuropsychologia 65, 197–210

Osuagwu, B. A., & Vuckovic, A. (2014). Similarities between explicit and implicit motor imagery in mental rotation of hands: An EEG study. Neuropsychologia, 65, 197-210.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. Cognitive psychology 9, 441–474

Palmer, S. E. (1983). The psychology of perceptual organization: A transformational approach. In Human and machine vision (pp. 269-339). Academic Press.

Palmer, S. E. (1983). The psychology of perceptual organization: A transformational approach. In Human and machine vision (Elsevier). 269–339

Palmer, S. E. (1985). The role of symmetry in shape perception. Acta Psychologica, 59(1), 67-90.

Palmer, S. E. (1999). Vision science: Photons to phenomenology. MIT press.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long and A. Baddeley (Eds.), Attention and performance IX. Hillsdale NJ: Erlbaum.

Palmer, S., Rosch, E., and Chase, P. (1981). Canonical perspective and the perception of objects. Attention and performance, 135–151

Palmer, T. E. (1975). The effects of contextual scenes on the identification of objects. Memory & cognition, 3(5), 519-526.

Pantle, A., & Picciano, L. (1976). A multistable movement display: Evidence for two

Paolini, M., & Sereno, M. I. (1998). Direction selectivity in the middle lateral and lateral (ML and L) visual areas in the California ground squirrel. Cerebral cortex (New York, NY: 1991), 8(4), 362-371.

Parikh, P. J., and Lam, S. S. (2005). A hybrid strategy to solve the forward kinematics problem in parallel manipulators. IEEE Trans. Rob. 21, 18–25. doi: 10.1109/TRO.2004.833801

Parsons, L. M., Fox, P. T., Downs, J. H., Glass, T., Hirsch, T. B., Martin, C. C., ... & Lancaster, J. L. (1995). Use of implicit motor imagery for visual shape discrimination as revealed by PET. Nature, 375(6526), 54-58.

Pavlova, M., & Sokolov, A. (2000). Orientation specificity in biological motion perception. Perception & Psychophysics, 62(5), 889-899.

Pelphrey, K. A., Morris, J. P., Michelich, C. R., Allison, T., & McCarthy, G. (2005). Functional anatomy of biological motion perception in posterior temporal cortex: an fMRI study of eye, mouth and hand movements. Cerebral cortex, 15(12), 1866-1876.

Perry, A., Troje, N. F., and Bentin, S. (2010). Exploring motor system contributions to the perception of social information: Evidence from eeg activity in the mu/alpha frequency range. Social neuroscience 5, 272–284

Piaget, J. (1952). The Origins of Intelligence in Children. New York, NY: W.W. Norton & Co. doi: 10.1037/11494-000

Pikler, J. (1917). Sinnesphysiologische Untersuchungen. Leipzig, Germany: Barth.

Pinker, S. (1984). Visual cognition: An introduction. Cognition 18, 1–63

Pitts, W., & McCulloch, W. S. (1947). How we know universals the perception of auditory and visual forms. The Bulletin of mathematical biophysics, 9(3), 127-147.

Poggio, G. F., Gonzalez, F., and Krause, F. (1988). Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. J. Neurosci. 8, 4531–4550. doi: 10.1523/JNEUROSCI.08-12-04531.1988

Portfors, C. V., and Regan, D. (1997). Just-noticeable difference in the speed of cyclopean motion in depth and the speed of cyclopean motion within a frontoparallel plane. J. Exp. Psychol. Hum. Percept. Perform. 23, 1074. doi: 10.1037/0096-1523.23.4.1074

Pouget A, Ducom JC, Torri J, Bavelier D. Multisensory spatial representations in eye-centered coordinates for reaching. Cognition. 2002; 83:B1–11.

Pouget, A., Deneve, S., and Duhamel, J.-R. (2002). A computational perspective on the neural basis of multisensory spatial representations. Nat. Rev. Neurosci. 3, 741–747. doi: 10.1038/nrn914

Provost, A., Johnson, B., Karayanidis, F., Brown, S. D., & Heathcote, A. (2013). Two routes to expertise in mental rotation. *Cognitive science*, *37*(7), 1321-1342.

Pugach, G., Pitti, A., Tolochko, O., and Gaussier, P. (2019). Brain-inspired coding of robot body schema through visuo-motor integration of touched events. Front. Neurorobot. 13:5. doi: 10.3389/fnbot.2019.00005

Ratan Murty, N. A. and Arun, S. P. (2015). Dynamics of 3d view invariance in monkey inferotemporal cortex. Journal of Neurophysiology 113, 2180–2194

Rayyes, R., Donat, H., and Steil, J. (2020). “Hierarchical interest-driven goal babbling for efficient bootstrapping of sensorimotor skills,” in 2020 IEEE International Conference on Robotics and Automation (ICRA) (Paris: IEEE), 1336–1342.

Reiter, A., Müller, A., and Gattringer, H. (2018). On higher order inverse kinematics methods in time-optimal trajectory planning for kinematically redundant manipulators. IEEE Trans. Ind. Inform. 14, 1681–1690. doi: 10.1109/TII.2018.2792002

Restle, F. (1979). Coding theory of the perception of motion configurations. Psychological Review, 86(1), 1.

Riečanský, I., & Katina, S. (2010). Induced EEG alpha oscillations are related to mental rotation ability: the evidence for neural efficiency and serial processing. Neuroscience letters, 482(2), 133-136.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature neuroscience 2, 1019–1025

Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. Nature neuroscience 3, 1199–1204

Rock, I. (1973). Orientation and Form. New York: Academic Press

Rodman, H. R., & Albright, T. D. (1987). Coding of visual stimulus velocity in area MT of the macaque. Vision research, 27(12), 2035-2048.

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. Annu. Rev. Neurosci., 29, 203-227.

Roelfsema, P. R., Lamme, V. A., & Spekreijse, H. (2000). The implementation of visual routines. Vision research, 40(10-12), 1385-1411.

Rokem, A., & Silver, M. A. (2009). A model of encoding and decoding in V1 and MT accounts for motion perception anisotropies in the human visual system. Brain research, 1299, 3-16.

Rosa, M. G., Palmer, S. M., Gamberini, M., Burman, K. J., Yu, H.-H., Reser, D. H., et al. (2009). Connections of the dorsomedial visual area: pathways for early integration of dorsal and ventral streams in extrastriate cortex. J. Neurosci. 29, 4548–4563. doi: 10.1523/JNEUROSCI.0529-09.2009

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 65, 386

Ross, W. D., Grossberg, S., & Mingolla, E. (2000). Visual cortical mechanisms of perceptual grouping: Interacting layers, networks, columns, and maps. Neural Networks, 13(6), 571-588.

Rossion, B. and Gauthier, I. (2002). How does the brain process upright and inverted faces? Behavioral and cognitive neuroscience reviews 1, 63–75

Rust, N. C., Mante, V., Simoncelli, E. P., & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. Nature neuroscience, 9(11), 1421-1431.

Sachs, M. B., Nachmias, J., and Robson, J. G. (1971). Spatial-frequency channels in human vision. J. Opt. Soc. Am. 61, 1176–1186. doi: 10.1364/JOSA.61.001176

Salinas, E., and A. L. (2001). Coordinate transformations in the visual system: how to generate gain fields and what to compute with them. Prog. Brain Res. 90, 130–175. doi: 10.1016/S0079-6123(01)30012-2

Salzman, C. D., & Newsome, W. T. (1994). Neural mechanisms for forming a perceptual decision. Science, 264(5156), 231-237.

Santucci, V. G., Baldassarre, G., and Mirolli, M. (2014). "Cumulative learning through intrinsic reinforcements," in Evolution, Complexity and Artificial Life, eds S. Cagnoni, M. Miorlli, and Villani M. (New York, NY: Springer), 107–122.

Sarantopoulos, I., and Doulgeri, Z. (2018). Human-inspired robotic grasping of flat objects. Rob. Auton. Syst. 108, 179–191. doi: 10.1016/j.robot.2018.07.005

Sauseng, P., Griesmayr, B., Freunberger, R., & Klimesch, W. (2010). Control mechanisms in working memory: a possible function of EEG theta oscillations. Neuroscience & Biobehavioral Reviews, 34(7), 1015-1022.

Saxon, J. B., and Mukerjee, A. (1990). "Learning the motion map of a robot arm with neural networks," in 1990 IJCNN International Joint Conference on Neural Networks (San Diego, CA: IEEE), 777–782.

Saygin, A. P., Wilson, S. M., Hagler, D. J., Bates, E., & Sereno, M. I. (2004). Point-light biological motion perception activates human premotor cortex. Journal of Neuroscience, 24(27), 6181-6188.

Schillaci, G., Hafner, V. V., and Lara, B. (2016). Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents. Front. Rob. AI 3:39. doi: 10.3389/frobt.2016.00039

Schmerling, M., Schillaci, G., and Hafner, V. V. (2015). "Goal-directed learning of hand-eye coordination in a humanoid robot," in 2015 Joint IEEE International

Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob) (Providence, RI: IEEE), 168–175.

Schöning, S., Engelien, A., Kugel, H., Schäfer, S., Schiffbauer, H., Zwitserlood, P., ... & Konrad, C. (2007). Functional anatomy of visuo-spatial working memory during mental rotation is influenced by sex, menstrual cycle, and sex steroid hormones. Neuropsychologia, 45(14), 3203-3214.

Schwarzer, G., Freitag, C., and Schum, N. (2013). How crawling and manual object exploration are related to the mental rotation abilities of 9-month-old infants. Front. Psychol. 4:97. doi: 10.3389/fpsyg.2013.00097

Sekuler, A. B. (1996). Axis of elongation can determine reference frames for object perception. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expe´rimentale 50, 270

Sekuler, A. B. and Swimmer, M. B. (2000). Interactions between symmetry and elongation in determining reference frames for object perception. Canadian Journal of Experimental Psychology/Revue Canadienne de psychologie expe´rimentale 54, 42

Sereno, A. B., Sereno, M. E., and Lehky, S. R. (2014). Recovering stimulus locations using populations of eye-position modulated neurons in dorsal and ventral visual streams of non-human primates. Front. Integr. Neurosci. 8:28. doi: 10.3389/fnint.2014.00028

Shaw, P., Law, J., and Lee, M. (2012). An evaluation of environmental constraints for biologically constrained development of gaze control on an icub robot. Paladyn 3, 147–155. doi: 10.2478/s13230-013-0103-y

Shepard, R. N. and Metzler, J. (1971). Mental rotation of three-dimensional objects. Science 171, 701–703

Sherwood, D. E., and Lee, T. D. (2003). Schema theory: critical review and implications for the role of cognition in a new theory of motor learning. Res. Q. Exerc. Sport 74, 376–382. doi: 10.1080/02701367.2003.10609107

Shinar, D. and Owen, B. H. (1973). Effects of form rotation on the speed of classification: The development of shape constancy. Perception & Psychophysics 14, 149–154

Shioiri, S., Ito, S., Sakurai, K., & Yaguchi, H. (2002). Detection of relative and uniform motion. JOSA A, 19(11), 2169-2179.

Shum, K. H., Wolford, g. l. (1983) A quantitative study of perceptual vector analysis, Perception & Psychophysics 34 (1) 17–24.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. Vision research, 38(5), 743-761.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. Vision research, 38(5), 743-761.

Sirigu, A., Duhamel, J.-R., and Poncet, M. (1991). The role of sensorimotor experience in object recognition: A case of multimodal agnosia. Brain 114, 2555–2573

Slone, L. K., Moore, D. S., and Johnson, S. P. (2018). Object exploration facilitates 4-month-olds mental rotation performance. PLoS ONE 13:e0200468. doi: 10.1371/journal.pone.0200468

Soska, K. C., Adolph, K. E., and Johnson, S. P. (2010). Systems in development: motor skill acquisition facilitates three-dimensional object completion. Dev. Psychol. 46, 129. doi: 10.1037/a0014618

Soska, K. C., and Adolph, K. E. (2014). Postural position constrains multimodal object exploration in infants. Infancy 19, 138–161. doi: 10.1111/infa.12039

Srisuk, P., Sento, A., and Kitjaidure, Y. (2017). "Inverse kinematics solution using neural networks from forward kinematics equations," in 2017 9th international conference on Knowledge and Smart Technology (KST) (Chonburi: IEEE), 61–65.

Sutherland, N. (1968). Outlines of a theory of visual pattern recognition in animals and man. Proceedings of the Royal society of London. Series B. Biological sciences 171, 297–317

Swanston, M. T., Wade, N. J., & Day, R. H. (1987). The representation of uniform motion in vision. Perception, 16(2), 143-159.

Takemura, N., Inui, T., and Fukui, T. (2018). A neural network model for development of reaching and pointing based on the interaction of forward and inverse transformations. Dev. Sci. 21:e12565. doi: 10.1111/desc.12565

Tanaka, H., Miyakoshi, M., and Makeig, S. (2018). Dynamics of directional tuning and reference frames in humans: a high-density eeg study. Sci. Rep. 8, 1–18. doi: 10.1038/s41598-018-26609-9

Tanneberg, D., Peters, J., and Rueckert, E. (2019). Intrinsic motivation and mental replay enable efficient online adaptation in stochastic recurrent networks. Neural Netw. 109, 67–80. doi: 10.1016/j.neunet.2018.10.005

Tarr, M. J. and Hayward, W. G. (2017). The concurrent encoding of viewpoint-invariant and viewpoint-dependent information in visual object recognition. Visual Cognition 25, 100–121

Tarr, M. J. and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. Cognitive psychology 21, 233–282

Tarr, M. J., Williams, P., Hayward, W. G., and Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. Nature neuroscience 1, 275–277

Tatti, E., Ferraioli, F., Peter, J., Alalade, T., Nelson, A. B., Ricci, S., ... & Ghilardi, M. F. (2021). Frontal increase of beta modulation during the practice of a motor task is enhanced by visuomotor learning. *Scientific reports*, *11*(1), 1-14.

Ternus, J. (1926). Experimentelle Untersuchungen über phänomenale Identität.

Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of vision, 2(5), 2-2.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. Cognition 32, 193–254

Ullman, S. (1996). High-level vision. Cambridge, MA: MIT Press.

Umilta', M. A., Berchio, C., Sestito, M., Freedberg, D., and Gallese, V. (2012). Abstract art and cortical motor activation: an eeg study. Frontiers in human neuroscience 6, 311

Van Polanen, V., and Davare, M. (2015). Interactions between dorsal and ventral streams for controlling skilled grasp. Neuropsychologia 79, 186–191. doi: 10.1016/j.neuropsychologia.2015.07.010

Vanrie, J., Willems, B., and Wagemans, J. (2001). Multiple routes to object matching from different viewpoints: Mental rotation versus invariant features. Perception 30, 1047–1056

Vingerhoets, G., De Lange, F. P., Vandemaele, P., Deblaere, K., & Achten, E. (2002). Motor imagery in mental rotation: an fMRI study. Neuroimage, 17(3), 1623-1633.

Vos, J., and Scheepstra, K. (1993). Computer-simulated neural networks: an appropriate model for motor development? Early Hum. Dev. 34, 101–112. doi: 10.1016/0378-3782(93)90045-V

Wang M, Ding J, Levi DM, Cooper EA. The effect of spatial structure on binocular contrast perception. J Vis. 2022 Nov 1;22(12):7. doi: 10.1167/jov.22.12.7. PMID: 36326743; PMCID: PMC9645364.

Wang, J., Zhou, T., Qiu, M., Du, A., Cai, K., Wang, Z., et al. (1999). Relationship between ventral stream for object vision and dorsal stream for spatial vision: an fmri+ erp study. Hum. Brain Mapp. 8, 170–181. doi: 10.1002/(SICI)1097-0193(1999)8:4andlt;170::AID-HBM2andgt;3.0.CO;2-W

Wang, Q., Sporns, O., and Burkhalter, A. (2012). Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex. J. Neurosci. 32, 4386–4399. doi: 10.1523/JNEUROSCI.6063-11.2012

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. II. Psychologische forschung, 4(1), 301-350.

Willems, B. and Wagemans, J. (2001). Matching multicomponent objects from different viewpoints: Mental rotation or normalization? Journal of Experimental Psychology: Human Perception and Performance 27, 1090

Williams, D. W., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. ACM SIGGRAPH Computer Graphics, 18(1), 24-24.

Wilson, H. R., Wilkinson, F., Lin, L. M., & Castillo, M. (2000). Perception of head orientation. Vision research, 40(5), 459-472.

Wood, A., Rychlowska, M., Korb, S., and Niedenthal, P. (2016). Fashioning the face: sensorimotor simulation contributes to facial expression recognition. Trends in cognitive sciences 20, 227–240

Zacks, J. M. (2008). Neuroimaging studies of mental rotation: a meta-analysis and review. Journal of cognitive neuroscience, 20(1), 1-19.

Zacks, J. M., & Michelon, P. (2005). Transformations of visuospatial images. Behavioral and cognitive neuroscience reviews, 4(2), 96-118.

Zell, P., & Rosenhahn, B. (2015, October). A physics-based statistical model for human gait analysis. In German Conference on Pattern Recognition (pp. 169-180). Springer, Cham.

Zohary, E., Scase, M. O., & Braddick, O. J. (1996). Integration across directions in dynamic random dot displays: Vector summation or winner take all? Vision research, 36(15), 2321-2331.