A Computational Model For Animacy Perception

Thesis
submitted in partial fulfillment of the requirements for the degree

Master of Science

Graduate School of Neural Information Processing

Faculty of Science
Faculty of Medicine

Eberhard-Karls-Universität Tübingen

Presented by

Nitin Saini

from New Delhi, India

Tübingen, November 30, 2017

Thesis Advisor: Prof. Dr. Martin Giese
Section for Computational Sensorimotorics,
Department of Cognitive Neurology,
Hertie Institute for Clinical Brain Research,
Werner Reichardt Center for Integrative Neuroscience,
University Clinic Tübingen
Tübingen, Germany

Second Reader: Dr. Tjeerd Dijkstra
Section for Computational Sensorimotorics,
Department of Cognitive Neurology,
Hertie Institute for Clinical Brain Research,
Werner Reichardt Center for Integrative Neuroscience,
University Clinic Tübingen
Tübingen, Germany

- I affirm that I have written the dissertation myself and have not used any sources and aids other than those indicated.
- I affirm that I have not included data generated in one of my laboratory rotations and already presented in the respective laboratory report.

Date / Signature: November 30, 2017 Nitin Saini

Acknowledgments

I am thankful to my advisor, Prof. Dr. Martin Giese, for providing me a place in his lab, his full support and guidance throughout the duration of this work. I also appreciate the guidance and support of my second advisor Dr. Tjeerd Dijkstra.
I am thankful to Mohammad Hovaidi Ardestani for his guidance and contribution towards the project. I am grateful to the Computational Sensomotorics group members Leonid Fedorov, Andrea Christensen, Enrico Chiovetto, Albert Mukovskiy, Björn Müller, Jindrich Kodl, Mirjana Angelovska, Nick Taubert, Nima Ghorbani, and Nicolas Ludolph for inspiring discussions and their friendly support.

Abstract

Humans have the ability to tell if an object in their visual field is alive (animate) or not. Even if we are looking in a display with few 2D simple objects (e.g. a circle) making interactive movements, we can tell if any of the object is a living object or not just by looking at the motion of the objects. This perception mechanism is known as animacy perception. Motion parameters of an object heavily contributes to the perceived animacy of the object. For example, If an object suddenly changes its speed or the trajectory without any external interference, it is immediately perceived as animate. Although motion parameters are heavily involved in the animacy perception, the other factors like visual features of the object and its interaction with other objects in the scene plays a vital role in perceiving an object to be alive. Visual features like the anteroposterior axis should be aligned with the motion for the object to be perceived as animate. The interactive motion of two objects instead of the individual motion of each object sometimes leads to the combine perception of animacy or inanimacy.

In this thesis, we introduce a computational model for animacy perception. It belongs to a class of heirarchical models which are inspired from the human visual systems. Such models are used in the perception of biological moements, facial recognintion, recognition of hand actions etc. However, no such model exists for the animacy perception. Our model is a convolutional neural network with partly learned and partly hand-crafted features. It is inspired from the human visual system and uses the biological plausible features in its layers. The intermediate layers extracts the motion and visual features of an object from the scene and these features are combined by a hypothetical animacy neuron to give the animacy of the object. The model can extract the visual features of multiple objects but the motion features for only a single object, that is why it gives the animacy rating for the scenario with the single objects only. However, it can be easily extended to extract the motion parameters of multiple objects to give the final animacy of the scene.

Contents

1 Introduction **1**

1.1 What is animacy perception? . . . 1

1.2 Experiment by Heider and Simmel . . . 1

1.3 Cues for animacy perception . . . 3

1.3.1 Behavioral cues . . . 3

1.3.2 Kinematics . . . 3

1.4 Cues for intention perception . . . 4

1.4.1 Motion cues . . . 4

1.4.2 Viewing angle . . . 4

1.5 Energy violation hypothesis . . . 4

1.6 Separate encoding of animacy and intention in the brain . . . 5

1.7 Animacy perception: a visual processing . . . 6

1.8 Animacy perception: a phylogenetically old capability . . . 6

1.9 Two processes behind the visual perception . . . 6

1.10 The wolfpack effect . . . 7

1.11 The perception of chasing . . . 8

1.12 The intention categorization . . . 11

1.13 Animacy perception in case of a single object . . . 12

1.14 Computational models for visual perception . . . 13

2 The Model **15**

2.1 Form Pathway . . . 17

2.1.1 Orientation Detectors . . . 17

2.1.2 Pooling Neurons . . . 18

2.1.3 Object Detectors . . . 19

2.2 Motion Pathway . . . 21

2.2.1 Position Detectors . . . 22

2.2.2 Motion Detectors . . . 22

2.2.3 Pooling Neurons . . . 23

2.2.4 Acceleration Detectors . . . 23

2.3 Animacy Detector . . . 24

3 Results **27**
3.1 Stimuli Videos 27
3.2 Results 29
3.2.1 Form Pathway 29
3.2.2 Motion Pathway 32
3.2.3 Animacy Neuron 36

4 Discussion **39**
4.1 Challenges and future work 40

Chapter 1

Introduction

1.1 What is animacy perception?

We all have the ability to tell if an object is alive or not. We can tell just by looking at the object if it is a living object or a non-living object. This ability of recognizing a living object is called as "Animacy Perception".

How does the animacy perception in human beings works? How the animacy perception is different in case of other living beings? Is animacy perception a low level mechanism of the brain? What features of the object are responsible for it being perceived as an animate object? These are some of the early questions that have been raised and researched upon by various researchers.

How do we know by looking at an object if it is a living object or not? Does our brain checks for the visual features of the objects and we perceive it to be alive if those matches with that of a previously seen living object? Or our brain focuses on the motion of the object and if the motion looks familiar to that of a previously seen object, we perceive it as an animate object. The preliminary answer to these questions came from the experiment by Heider and Simmel.

1.2 Experiment by Heider and Simmel

In 1944, Heider and Simmel did an experiment where the human participants were shown a video with simple 2D figures in motion. The figures where simply a circle and 2 triangles moving on the screen in some interactive manner (the snapshots of the video are shown in the figure 1.1). The participants not only perceived the objects as alive, they were also telling the intention of the motion of the objects and labeling the personality traits to the objects. The objects used in the videos are simple 2D shapes and have no features of any living object. This suggests that the object's motion features are sufficient to induce a strong animacy perception.

However, visual features do contribute to the animacy perception. The visual features similar to that of an alive object adds confidence to the perceived animacy of an object.

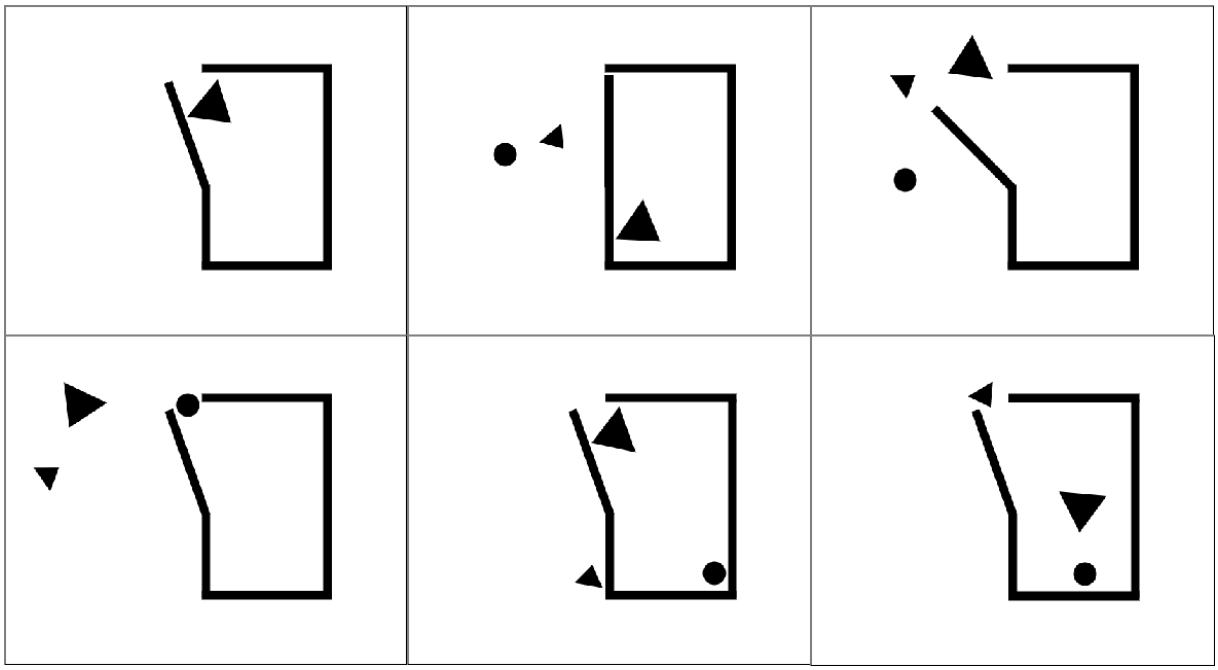


Figure 1.1: Some of the snapshots of the video used by Heider and Simmel for their experiment. The scene contains two triangles and one circle moving and interacting with each other. The participants not only perceived these objects as alive with intentions but also inferred and assigned personality traits to them.

A crucial feature to identify a moving object is the anteroposterior axis. Almost all of the living objects move along their anteroposterior axis (one exception is a crab) and the human visual system is highly tuned to this observation. A moving object whose motion is not aligned with its perceived anteroposterior axis is easily perceived as a non-animate object Tremoulet and Feldman [2000]. The anterposterior axis of an object is identified using both the visual features and the motion features. If an object is elongated in a direction, then its anteroposterior axis seems to be along the elongation. If some eyes like features are present then the anteroposterior axis is identified by assuming the eyes to be in the front facing direction. In case such visual features are absent, the direction of motion of the object is identified as the anteroposterior axis Tremoulet and Feldman [2000] Hernik et al. [2014]. However, if the identified anteroposterior axis using visual features is not aligned with the direction of motion then this leads to the conflict in perception and the object is perceived as non-animate.

You see an object moving behind another object and it is perceived to be an act of chase. If these two objects are continuously colliding, we perceive them to be fighting each other. These are some of the example scenarios and their perceived results from the Heider and Simmel video. This ability of perceiving the intention from any interactive motion is called as 'Intention Perception'.

In the scenarios where there are multiple objects and they have interactive motion, the animacy perception gets more complicated. In a video with more than one object, the individual object might seem like animate/inanimate if looked separately, but might look completely opposite if looked as a whole Tremoulet and Feldman [2006].

1.3 Cues for animacy perception

1.3.1 Behavioral cues

Self propulsion Di Giorgio et al. [2016]

Self propulsion is a strong cue for animacy . If an object starts moving itself without any external help, then it has high chances of being perceived as alive. Humans does not acquire this perceiving ability while growing up, but it is innate to our visual system. Newly born babies are found to have a preference for self propelled motion (self propelled onset or self propelled change in direction) compared to the motion or change in the motion caused by influence from some external source.

Goal directedness van Buren et al. [2015]

Another cue for animacy perception is the goal directedness. If an object's motion seems to be driven by a goal it is perceived to be a living object. In a scenario, where motion of an object is in correlation with another object, the first object is likely to be perceived as an animate object. The object does not even have to pursue or move away from the second object, if it's orientation is always towards the second object, then also it is perceived as animate.

Reactive to social contingencies Dittrich and Lea [1994]

We learn many kinds of interactions and motions throughout the life, for example, the motion behavior of the living objects, the non-living objects and the interaction between any combination of these. The motion interaction between alive objects is linked to the social behavior. If a motion or interaction is similar to that of a previous experience of the observer, then the objects are perceived to be of similar nature as experienced before.

1.3.2 Kinematics

DirectionTremoulet and Feldman [2000]

The alive objects does not have a monotonous non-changing motion. If an object is alive, it is capable of changing its motion direction. On the other hand, a non-animate object can't change its direction of motion without any external influence. A change in the motion direction is a cue for the animacy perception. However, if there is a change in the motion direction but it seems like it is done by an external source, the animacy is severely diminished.

Speed Szego and Rutherford [2007]

A change in the speed is a cue for animacy perception. If an object speeds up it is perceived as animate, as it gives an impression that the object has an energy source and

the speed up is done intentionally. However, if an object slows down, it is perceived as less animate than the object with no change in the speed Tremoulet and Feldman [2000]. The animacy also depends on the absolute speed of the object. Even if there is no change in the speed, the object with higher speed is perceived as more animate than the object with lower speed. Interestingly, the animacy perception depends on the perceived speed of the object rather than the actual speed. If the two objects have the same speed but one is perceived to be faster because of illusion then the faster object is perceived as animate more often than the other object.

1.4 Cues for intention perception

1.4.1 Motion cues

Barrett et al studied the cues relevant for the classification of different types of interactive motion. They generated displays with 7 categories of interactive motions (chasing, courting, following, guarding, attacking, fighting, playing) between two 2D objects. The displays were generated with the two person playing a video game with each one having a control of one of the objects. They found out that these motions can be categorized by any categorization strategy (neural network etc) using the following cues 1) Absolute velocity 2) Relative angle 3) Relative velocity 4) Relative heading 5) Relative vorticity 6) Absolute vorticity 7) Relative distance

1.4.2 Viewing angle

The viewing angle of the observer is not a cue related to the object or other things in the display. This is a feature related to the display or the 2D video seen by the observer. The image on the retina is a 2D image. The depth related raw information is received by the brain and processed to get the depth in the scene. However, in the animacy displays or the scenes where depth related information is not available, the motion cues (discussed above) depends on the viewing angle. If an interactive motion is looked upon from the top then the motion cues are correctly calculated, but if it is looked upon from the side, the motion cues are not calculated correctly as there is no depth information available.

1.5 Energy violation hypothesis

The animacy perception seems to be a complicated mechanism and difficult to explain with some generalized statement. Scholl and tremoulet Scholl and Tremoulet [2000] made an attempt to narrow down the animacy perception to a single sentence. They stated that, if a motion or a change in motion looks like as if it is done by using the energy that comes from the object, or, it seems like object has its own source of energy inside, then the object is perceived as animate. The self-propelled motion is an example of such

motion. In the case of the self-propelled motion there is no external force that results in the motion onset, that is why it is perceived as if the motion onset is due to the energy inside and hence it is perceived as an animate object. This hypothesis of animacy due to the internal energy source is known as 'Energy violation hypothesis'.

The animacy perception highly depends on the motion cues, but the motion cues are not always sufficient to decide if an object is animate or not. The motion information can be ambiguous and additional information is used by humans to correctly identify if an object is animate or not Gelman et al. [1996]. This is similar to the findings of Tremoulet and Feldman Tremoulet and Feldman [2000], Tremoulet and Feldman [2006] where they mentioned that the information such as visual features of the object and the contextual information can change the animacy perception even if the motion remains the same. These studies and the findings are not in conflict with the energy violation hypothesis but suggests that not only the motion information but contextual information also contributes to the animacy perception.

The study from Barrett et al. Barrett et al. [2005] suggested that the motion information is sufficient for the animacy as well as intention perception. These findings are in conflict with the other studies Gelman et al. [1996], Tremoulet and Feldman [2000] where the animacy perception can be highly affected by visual features and contextual information. However, all the stimuli used by Barrett et al. had more than two objects. It might happen that the perceived animacy is due to the mapping of interactive motion to some previously learned social interaction. If one of the object is removed while keeping the same motion for another object, it might be perceived as non-animate.

1.6 Separate encoding of animacy and intention in the brain

Animacy and intention perception are dependent processes. One cannot perceive intention without perceiving the animacy. It is not possible that a human can tell the intention of an object without perceiving it as animate. However, Scholl and Gao Gao et al. [2012] performed experiments to check if the perceived animacy and intention are encoded separately in the brain. They made the animacy displays with the same motion cues for animacy perception but varied the intentions of the object. They found out that the change in intention resulted in higher activations in the right posterior superior temporal sulcus(pSTS). The pSTS was known to be involved in the social perception, however this study clearly showed that it selecetively encodes the intentions rather than the animacy cues.

1.7 Animacy perception: a visual processing

Animacy perception is not a high-level process, it a low-level processing similar to the depth perception Gao and Scholl [2013]. Animacy perception is a visual processing and motion cues are processed in the lower layers of the visual system. Based on this, Gao and Scholl raised questions over the psychophysics of some animacy perception studies (e.g. Tremoulet and Feldman) Tremoulet and Feldman [2000], Tremoulet and Feldman [2006]. In these studies, the animacy response is collected by asking the participant to give an animacy rating to the display. The participant's high level reasoning might affect the rating and hence might not be the actual animacy induced by the motion cues.

Visual perception is severely affected by the visual attention Chun et al. [2010], Treisman [2006]. The object under the attention of the observer is better processed by the visual system than the rest of the visual field. Sometimes, the objects which are not under the observer's attention are completely ignored by the conscious perception Gao and Scholl [2013]. The visual system is designed to shift the attention to the animate objects and the attention shift is driven by the low level visual processing. The animate agents are given priority for visual attention and it is related to the survival related instincts. For example, the visual system has special detection mechanism for snakes and spiders Rakison and Derringer [2008], Öhman and Mineka [2003]. The visual system prioritizes for the visual and motion features of such potential harmful objects.

1.8 Animacy perception: a phylogenetically old capability

Salva et al. studied the animacy perception in naïve chicks compared to humans. In the display, they showed a moving object with changing velocity and the direction. The results showed that the humans and the chicks both shows higher animacy percept for such displays. They carried out further experiments on the chicks only and found out that the moment of the speed change should be visible to the observer in order to perceive the object as animate. If the object is occluded while making the speed change then it is not perceived as animate. Based on their study, they stated that the animacy perception is a phylogenetically old capability shared among the vertebrates.

1.9 Two processes behind the visual perception

Dittrich and Lea Dittrich and Lea [1994] hypothesized that animacy and intention perception are composed of two types of processes, the top-down and the bottom-up process. The bottom-up process is a low-level process where the motion features are captured and encoded. The top-down process combines these features to get the animacy and intention response. The top-down process is a high-level process which can be affected by the prior

information available to the observer. For example, there are multiple objects on the screen and the participant is told about the nature of the objects (predator, prey etc).

They performed experiments with changing motion features (manipulating the bottom-up process) and different prior information given to the participants (manipulating the top-down process) and looked for the change in the perception. They looked for the change in animacy, intention and the degree of interaction between the objects.

The display used in these experiments had three objects: a chaser, a chasee and some distractors. The chaser follows the chasee while the distractors move randomly. It was found that the chaser's motion with higher speed and more directedness towards the chasee looks more intentional. They did another experiment with scenarios where the chasee is present on the screen but is invisible to the participant. In such scenarios, the intention perception is diminished but not entirely gone. In another scenario, the chasee's motion is made random, similar to that of the distractors and it was found that the chaser's motion wasn't perceived as intentional.

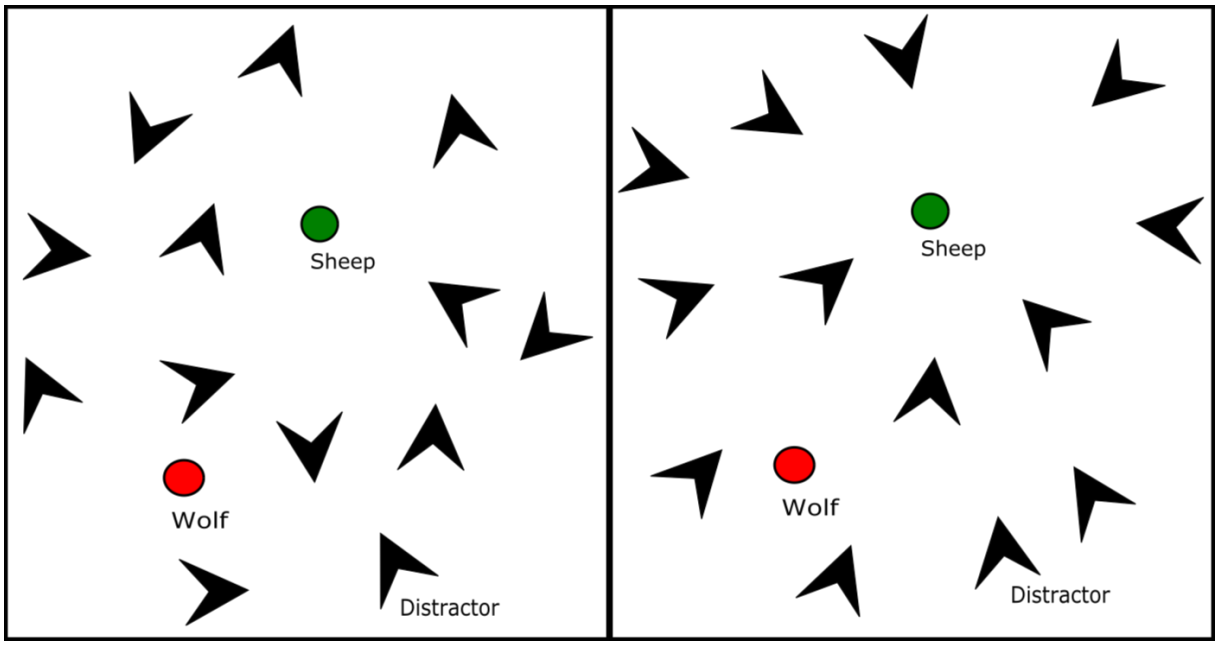
It was concluded that the perceived intention of the object is not affected by the prior information about the scene. The perceived animacy is moderately affected and the degree of interaction is highly affected by it. The perceived intention depends on the absolute motion parameters (velocity, direction etc), the degree of interaction depends on the relative motion parameters and the animacy perception depends on both of these.

1.10 The wolfpack effect

Gao, McCarthy and Scholl Gao et al. [2010] did a study on the effect of the facing direction of an object on the perceived animacy. If an object is making random motion but always facing a second object which is making movements independent of the first object, then the first object seems to be animate with the intention of following the second object.

They used a display similar to that of Dittrich and Lea containing a chaser, a chasee and some distractors. The distractors are pointed objects where the pointing direction is perceived as the facing direction. The chaser chases the chasee and the distractors moves randomly on the screen. In another scenario which is called as the wolfpack scenario, the chaser's, the chasee's and the distractors have the same motion but the distractors are always facing towards the chasee. It was found that the participants' ability to perceive the chase was diminished in the case of wolfpack scenario. Not only this, when the participants were given the control of the chasee and were told to escape from the chaser, their ability to escape from the chaser was also impaired.

In the wolfpack scenario, the distractors have the same random motion as that of the normal scenario but the distractors are always facing towards the chasee. This scenario is perceived as if the distractors are actively pursuing the chasee and this effect is called as "wolfpack effec". A pack of wolf uses the same strategy for hunting their prey by not moving directly towards it but continuously facing it irrespective of their motion. This is



(a) (b)

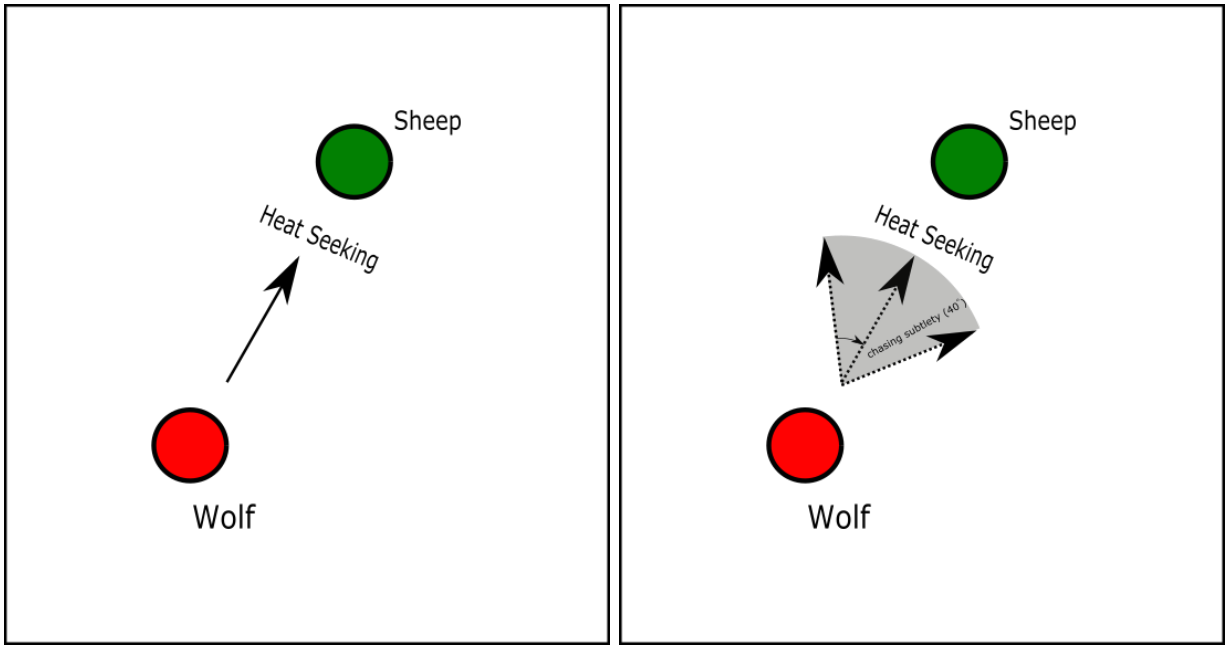
Figure 1.2: a) In the normal scenario, the wolf chases the sheep while the distractors moves randomly. b) In the wolfpack scenario, the wolf, the sheep and the distractors have the same motion as the normal scenario but the distractors are always facing towards the sheep irrespective of their motion. **Adapted from Gao et al. [2010]**

a low-level predatory instinct used to panic and distract the prey such that it would be difficult for the prey to find an escape route.

1.11 The perception of chasing

The animacy perception research has several challenges. Most of the animacy displays are hand-made by the researchers. It is difficult to quantify the parameter changes in the display. For example, the Heider and Simmel's video can't be used to study the effect on animacy due to the motion parameters, as the video is handcrafted and the motion parameters are not available. Another challenge is the quantification of the participants' response. Generally, the participants are asked for the animacy response and there is no standard way of quantizing the rating given by them. As the ratings are explicitly asked, the high-level mechanisms can also interfere with the rating of the perceived animacy. It might happen that the participant's response is affected by some other information apart from the cues under study and their rating is different than that of the actual perception.

Gao et al Gao et al. [2009] designed a display similar to that of Dittrich and Lea with a wolf (chaser), a sheep (chasee) and some distractors. They studied the perception of chasing by varying the chasing subtlety and the directionality. The chasing subtlety is the angle deviation with which the wolf approaches the sheep (see figure 1.3) and the directionality is the deviation of the sheep's facing direction with its direction of motion. The objects are circular in shape except when the directionality is studied. For the



(a) (b)

Figure 1.3: a) Heat seeking situation is when the chasing subtlety angle is 0°. In this situation, the wolf's motion is always directed towards the sheep. b) In case the chasing subtlety angle is non zero (say θ), the wolf's motion is not directed exactly towards the sheep but deviates from the heat seeking direction which is the reference direction and labeled as 0°. At each time, the wolf's motion can be in any direction between $\pm\theta$ and the value is obtained by uniformly sampling from the values from $-\theta$ to θ. **Adapted from Gao et al. [2009]**

directionality experiments, the distractors and the wolf are in the shape of arrowhead with pointing direction perceived as the facing direction. In the directionality study, they looked for the chasing perception with the varying directionality i.e. varying angle of the facing direction with the motion direction. It was found that the chasing perception is reduced as the facing direction deviates from the motion direction. The study suggested the same results as that of the Tremoulet and Feldman [2000], if the facing direction is not aligned with the direction of motion, the motion does not look natural and hence in this case diminishes the chasing perception. In the chasing subtlety study, all the wolf and the distractors are circular and looks alike and the participant has to detect the wolf. To study the effect of chasing subtlety they designed two kinds of experiments:

Find-the-chase: In these experiments, the participants had to respond if they perceive any chasing happening in the scene. The experiment is divided into two types of equal number of trials. In half of them, the wolf was chasing the sheep and in another half the motion of the wolf was same as that of the chasing trials, but the sheep was invisible. The accuracy with which the participant can tell if the chasing is happening in the scene, was used as the measure of perceived animacy. They varied the chasing subtlety and found out that as the angle of chasing subtlety increases the search accuracy decreases. Although the wolf is seeking for the sheep in all the trials, the increase in chasing subtlety angle leads to the lesser perception that the wolf is actively seeking for the sheep.

Don't-get-caught: In these experiments, the participant had the control over the sheep. The participant could control the motion of the sheep on the screen. The task was to escape from the wolf and the percentage of trials in which the participant was able to escape was used as the performance measure of the chase perception. They changed the chasing subtlety and the results showed that the ability to escape first decrease and then increase as the angle of chasing subtlety increases. The result can be explained as following: for a low angle of chasing subtlety, the chasing is easily perceived. The scenario is close to heat seeking and it is easy to detect a wolf among the distractors and that is why it is easy to escape from such a wolf. This type of wolf is called as the "obvious wolf". If the chasing subtlety angle is increased (90-120 degrees), the wolf becomes difficult to detect and hence difficult to escape. Such wolf is called as the "dangerous wolf". As the chasing subtlety angle is further increased (180 degrees), the performance of the participants increases and they are able to escape easily. However, this improvement in the performance is not due to the better detection of the wolf and the chasing, but due to the fact that the wolf's motion deviates too much from the sheep and it is unable to capture the sheep. This type of wolf is called as "incompetent wolf".

All these experiments show that the chasing subtlety is an important cue for animacy perception in the case of chasing. As the chaser is more directed towards the chasee, it's motion is perceived as more intentional and hence it is perceived as more animate.

1.12 The intention categorization

To generate the displays used in the animacy perception studies is a challenging task. As discussed earlier the handmade displays like that of the Heider and Simmel's has rich motion features but they are not quantizable. Also, the experiments generally need many displays with varying cues, and it is difficult to mass produce the handmade displays. Another way of generating displays is to artificially generate them using the computer algorithms. These displays have the advantage that the motion cues are quantizable and can be controlled precisely with the computer programs. Many of such displays with varying cues can be generated easily by feeding the parameters to the computer programs.

McAleer and Love McAleer and Love [2013] disagree with the usage of artificial displays for the animacy perception studies. They state that the artificially generated displays lack the rich motion features and hence does not capture the natural looking motions and interactions. They devised their own way of creating the animacy displays. The people were asked to perform the desired type of interactive motion (e.g. chasing) in a room and the scene was recorded using the camera from the top and the sides. These videos were then processed to replace the people with simple shapes (e.g. circle). Hence, the interactive motion remains intact while replacing the visual features of the people with the simple shapes. They recorded people with different types of interactive movements such as dancing. However, the participants were unable to identify the dancing movements and recognized them as random motions. McAleer and Love concluded that the participants failed to recognize the movements because of not having the experience of seeing that type of movement from that particular viewpoint. For example, seeing the dancing movements from the top view. We perceive any motion to be animate if we have learned from a previous experience of similar motion pattern knowing it to be animate.

For further experiments, they generated animacy displays for the common interactive motion patterns. They recorded and created displays for 6 types of interactive motions mostly known to all: fighting, following, chasing, playing, guarding, flirting.

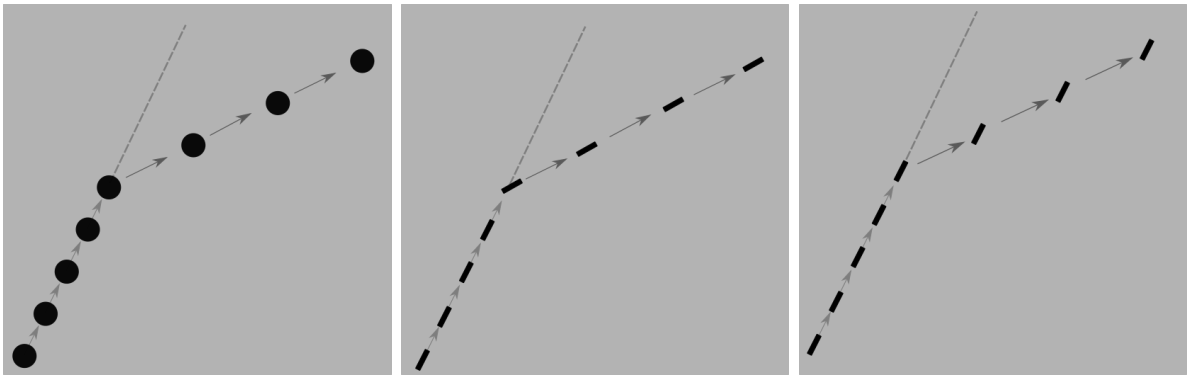
They first did the experiments to see if the animacy perception depends on the viewing angle of the observer. For this, they showed the participants the displays recorded with top view and the side view with the same motion. It was found that the participants were able to identify the motion correctly from the top view but not from the side view. As we look at a scene, we look at the 2D projection of a 3D world. When viewing from the side we lose the relative parameters such as the distance between the object, their relative depth etc, which are crucial in identifying the interactive motion between the objects. This information remains intact when seen from the top.

Next, they wanted to know how quickly humans can identify a particular intention. Participants were shown the top view display of different intentional motions for different time durations (1sec, 5 secs, 10 secs). The results showed that all the intentions can be perceived within 1 sec, except guarding. Even guarding can be perceived within 5 secs and if we increase the time to 10 secs, there is no improvement in the perception. This

shows that the intention judgement in humans is a fast processing. This can be related to the low-level survival related instincts where the quick judgement of intentions based on the motion was necessary to take quick actions against potentially harmful object.

1.13 Animacy perception in case of a single object

In the displays with multiple objects, the animacy might not be entirely because of the motion cues but the interactive motion. As discussed earlier, the extra information other than the motion features such as context, interaction with another object etc might lead to the change in the perceived animacy. To study just the motion cues without interference of any other factors, Tremoulet and Feldman Tremoulet and Feldman [2000] used the displays with just one object and with changing motion cues (speed, direction etc). They used two kinds of objects, a circle and a rectangle. In each video, the object entered into the screen with some velocity and after reaching the middle of the screen it changed its trajectory, speed and orientation.



(a) (b) (c)

Figure 1.4: The display used by Tremoulet and Feldman to study the effect of directionality on the perceived animacy. a) The circular object enters the screen and changes its trajectory by deviating from the original trajectory with some angle but the velocity remains the same. b) The rectangular object follows the exact same motion as in (a) and its orientation is always aligned to its direction of motion. c)The motion of the rectangluar object remains same as (a) and (b) but after the deviaiton from the original trajectory, the object does not change its orientation and keeps it same as before, misaligning its orientation from the direction of motion. The scenario (b) is perceived as the most animate followed by scenario (a) while the scenario (c) is not perceived as animate. **Adapted from Tremoulet and Feldman [2000]**

In the first experiment, they compared the perceived animacy of three conditions: the circle, the misaligned rectangle and the aligned rectangle. Three conditions are shown in figure 1.4. In the misaligned condition, the rectangle enters the screen and after the trajectory deviation, it does not change its orientation keeping it same as before. In the aligned condition, the rectangle changes its orientation such that it gets aligned to the new direction of motion. In the circle condition, the trajectory and the velocity remain exactly

same as that of the first two conditions. The results showed that the misaligned condition is perceived to be almost inanimate, the aligned condition to be the most animate and the circle condition to be moderately animate.

In the next set of experiments, there was no misaligned condition and the rectangle was always aligned to its direction of motion. The videos were prepared with increasing deviation in trajectory and speed. If the initial velocity was 1V, then the four types of final velocity configuration were 0.5V, 1V, 2V and 4V. There were four direction deviation configurations 10 degrees, 20 degrees, 40 degrees and 80 degrees. Two videos for each final velocity and the direction deviation configuration were prepared, one with the rectangular object and the other with the circle. It was found that the animacy increases with the increasing direction deviation and the final velocity. If the final velocity is less than the initial velocity, the object is perceived less animate than the condition when there is no change in the velocity.

1.14 Computational models for visual perception

We've discussed about the existing research on the animacy and the intention perception. These two are highly connected to each other and that is why used interchangeably in some of the existing literature. However, the paper discussed in the previous section Tremoulet and Feldman [2000] studies the simple scenarios and the simple motion cues for animacy perception. In this thesis, a computational model is proposed for the animacy perception which can work for the simple scenarios used by Tremoulet and Feldman. The model is inspired from the heirarchical model for the recognition of biological movements presented by Giese and Poggio Giese and Poggio [2003]. The Giese and Poggio's model does not model the animacy perception mechanism but is used for the recognition of biological movements in case of the point light walker displays. It belongs to the family of heirarchical models used in visual perception Poggio and Edelman [1990] Serre et al. [2005] Riesenhuber and Poggio [1999]. These models are inspired from the human visual systems and are robust in feature extraction from the visual input. Next, we discuss the Giese and Poggio's model, which is used for designing the architecture of the model presented in this thesis.

The Giese and Poggio's model consists of two pathways (the form pathway and the motion pathway) inspired from the two pathways in the visual system (the "What" and the "Where" pathway). The form pathway processes the visual information related to "how the objects in the scene looks?" and the motion pathway processes the motion information related to "how the objects in the scene moves?".

The figure 1.5 shows the Giese and Poggio model. The first layer of the form pathway consists of the local orientation detectors and models the V1 neurons in the brain. These are modeled by the Gabor filters tuned to different orientations. The next layer is the layer of contour and bar detectors. These kind of neurons are found in the V1 and V4

layer of the visual cortex. Next layer is the layer of snapshot neurons each of which are tuned to a particular visual configuration of the object (for example a pose of a human being). These neurons models the neurons present in the IT (inferotemporal cortex), STS (superior temporal sulcus) and FA (Brodmann area 4) areas of the brain.

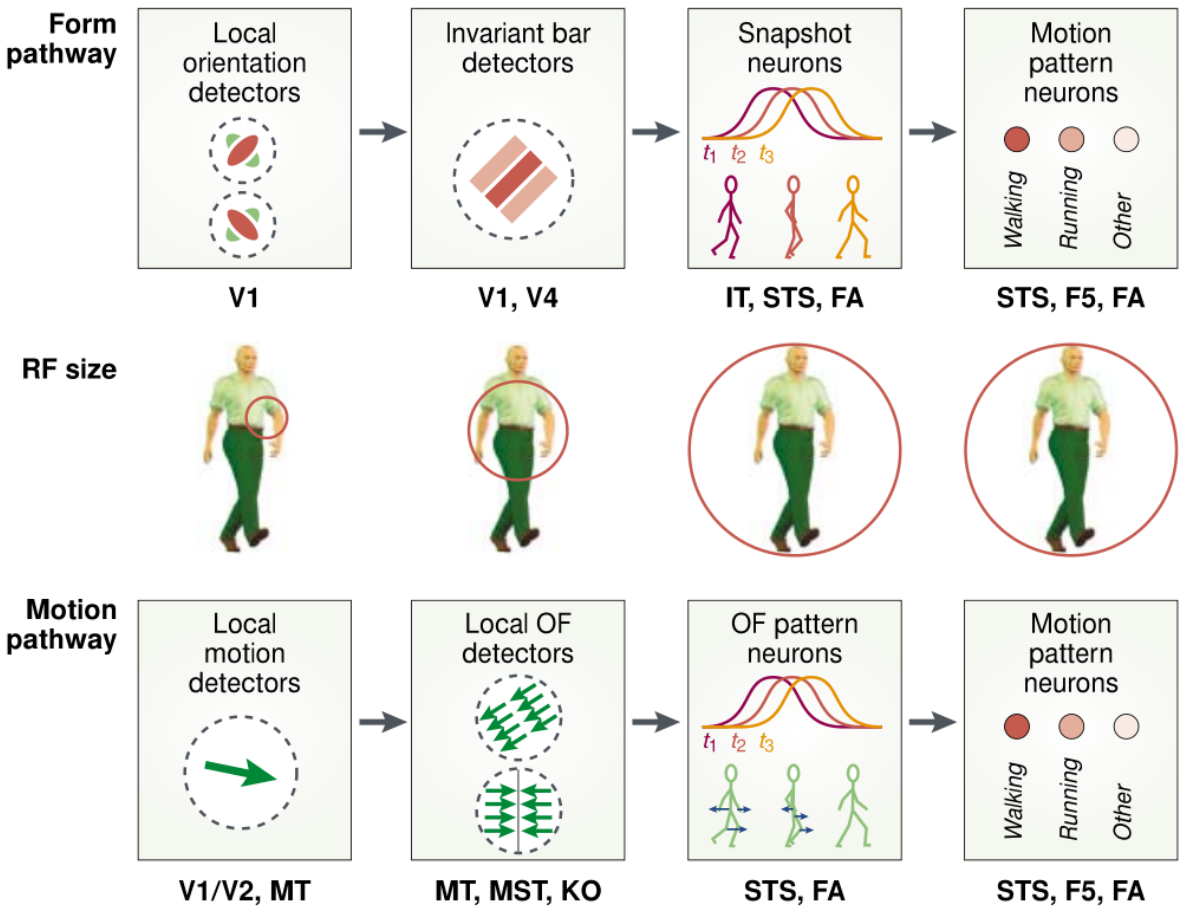


Figure 1.5: The computational model introduced by Giese and Poggio for the recognition of biological movements. **(from Giese and Poggio [2003])**

The first layer of the motion pathway is the local motion detectors which models the neurons present in the V1, V2 and the MT area of the brain. Next layer is the layer of the local optical flow detectors which takes the input from the first layer and pools to encode the optical flow in a local region. These kind of neurons are found in the MT, MST and the kinetic occipital cortex region of the brain. The next layer is the layer of the optical flow pattern neurons. These neurons are tuned to the complete optical flow configuration of the whole scene and models the neurons present in the STS and the FA area of the brain.

The final output is given by the motion pattern neurons which models the neurons present in the STS, F5 and the FA areas of the brain. These neurons takes their input from the final layers of each pathway and are tuned to the different types of biological motions (for example walking, running etc.).

Chapter 2

The Model

This chapter discusses about the computational model for the animacy Perception. The model is a convolutional neural network with multiple layers, some of the layers are with the learned features while others are with the hand crafted features. I will go through the model layer-by-layer and explain the features and the processing done by each layer.

Figure 2.1 shows the complete model with functional blocks. These functional blocks are composed of one or more layers. The model takes video frames as the input and passes it on to the two parallel pathways, i.e. the form pathway and the motion pathway. These two pathways are inspired by the two stream hypothesis in the human visual system Goodale and Milner. The two stream hypothesis says that the visual system gets divided into the two pathways after the V1 layer. One is the ventral pathway ("What" pathway) which is responsible for the processing related to the visual features of the scene and the other one is the dorsal pathway ("Where" pathway) which processes the motion and position related information in the scene. After the processing from these two pathways there is an animacy block (animacy neuron) [1] which takes input from the processing blocks in the two pathways and gives the animacy response which is the measure of animacy of an object in the scene.

The form pathway performs two functions, first the orientation detection and then the object detection. The orientation detection involves the identification of the local edges and their orientation and is performed by the layers F1 and F2. The next functional block is the F3 layer which performs the object detection to identify the shape of the object and its anterior direction (if any).

The motion pathway performs its processing in parallel to the form pathway. The first layer of the motion pathway is the position detectors (layer M1) which performs a filtering to remove the visual information of the object and extract the position of the object in the space. The next functional block is the velocity detection (layer M2 and M3) which encodes the velocity and the direction of motion of the object. The next block is the acceleration detection block (layer M4) which encodes the change in the velocity of the object.

[1] Animacy neuron block used in this thesis does not imply that an animacy neuron exists in the brain.

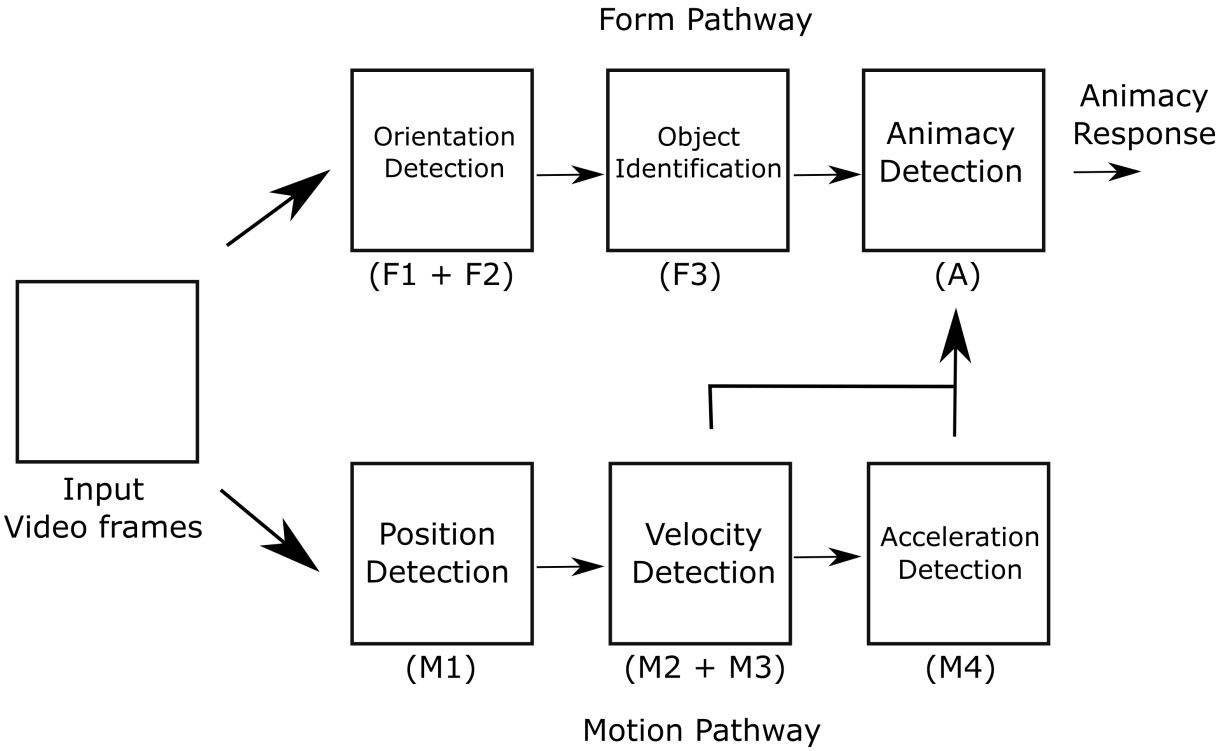


Figure 2.1: **The computational model of animacy perception**. The model consists of two pathways (form pathway and the motion pathway) each takes the video frames as input. The form pathway processes the visual features of the object and the motion pathways processes the spatial and temporal information of the object. The animacy detection module takes input from the object identification, velocity detection and the acceleration detection module to output the animacy rating.

After the feature extraction by the layers in the form pathway and the motion pathway, there is an animacy detection block (A) which takes input from the object detection, velocity detection and acceleration detection blocks and gives an animacy rating to the object in the scene.

2.1 Form Pathway

The form pathway is inspired from the ventral stream of the visual system (the "What" pathway) and consists of 3 consecutive layers. These layers process the input sequentially and are responsible for extracting the visual features and performing object identification and recognition. The first layer of the form pathway is the layer of local orientation detectors used to extract the information about the local edges and their orientation. The next layer is the layer of max pooling neurons used to introduce a local translational invariance. The third and the last layer is an RBF network for the object recognition. This layer is responsible for identifying the object type and its anterior direction. It needs to be trained beforehand with the images of the objects.

2.1.1 Orientation Detectors

The first layer of the form pathway consists of local orientation detector neurons. In this text, these neurons are termed as F1 neurons. These neurons model the simple cells present in the area V1,V2 of the visual cortex. These cells respond to an edge in their receptive field and are tuned to a particular orientation.

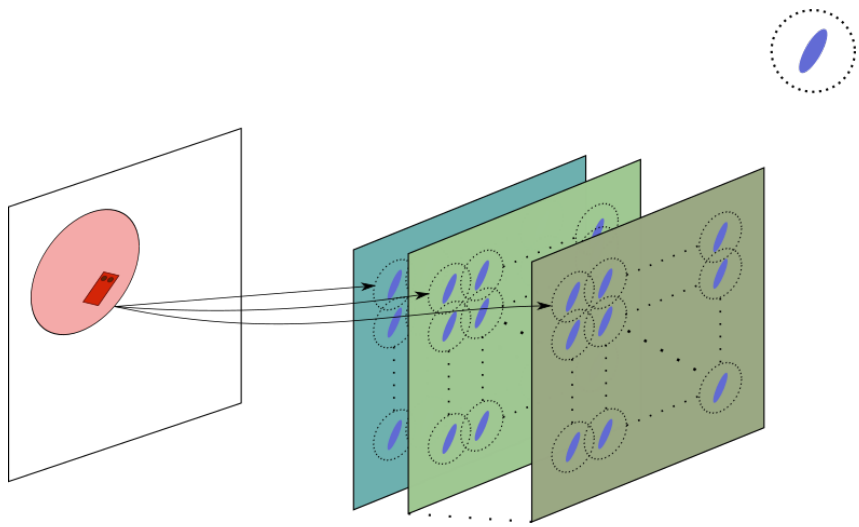


Figure 2.2: **The F1 layer.** It consists of 8 types of orientation detectors tuned to 8 different orientations

The F1 neurons are modeled using the gabor-like filters used in image processing. A gabor filter is a 2D filter obtained by multiplying a sinusoidal wave by a gaussian function. It is used to detect an edge of a particular orientation in its receptive field. Depending

on its parameters it is tuned to a particular orientation and responds only if an edge of that orientation is present in its receptive field.

The gabor filters used in this model can be represented by:

$$g(x, y, \lambda, \phi, \sigma, \gamma) = exp\left(- \frac{x' + \gamma^2 y'^2}{2\sigma^2} \right) \cos(\lambda x') \tag{2.1}$$

where,

$$x' = x \cos\phi + y \sin\phi, \qquad y' = y \cos\phi - x \sin\phi$$

x and y are the indices in the 2D space, λ is the wave number of the sinusoid, ϕ is the orientation angle, σ is the standard deviation of the gaussian function and γ is the spatial aspect ratio.

The first layer of the model has 8 different types of simple cells each tuned to a different orientation. Figure 2.3 shows the different orientation tuned simple cells present in the model.

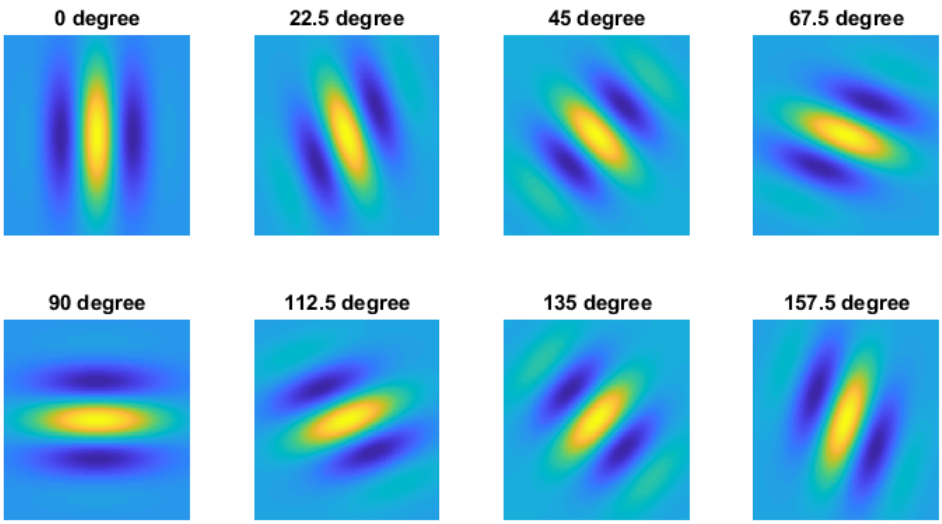


Figure 2.3: **Gabor Filters.** 8 types of gabor filters, each tuned to a different orientation

The orientation detector layer is an array of F1 neurons with overlapping receptive fields. The neurons in the layer are arranged such that the receptive fields of adjacent neurons overlaps and all of the neurons combined covers the entire visual field. Each receptive field is shared by 8 F1 neurons, each with different orientation tuning.

2.1.2 Pooling Neurons

The second layer of neurons is the pooling layer. The pooling neurons are termed as F2 neurons and the layer is termed as F2 layer in this text (figure 2.4). This layer is used to introduce the position and scale invariance to the output of the F1 neurons. F2 neurons receives the response of the F1 neurons in their receptive field and outputs the maximum response among them. The receptive field of each F2 neuron spreads across the space but

remains within the same orientation, i.e. each F2 neuron takes input from the F1 neurons of similar orientation tuning.

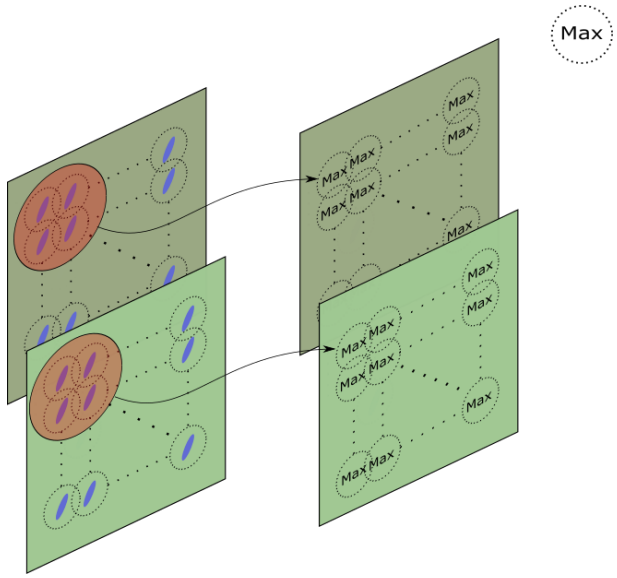


Figure 2.4: **The F2 layer.** This layer consists of the pooling neurons which pools over the F1 neurons in their receptive fields and outputs the maximum response among them.

2.1.3 Object Detectors

The next layer of the form pathway is the layer of the object detecting neurons. These neurons are termed as F3 neurons and the layer is termed as F3 layer in this text. They responds selectively to a particular object facing in a particular direction. These neurons are similar to the view-tuned neurons present in the IT area of the monkeys and gets tuned to complex shapes through learning. In the model, the F3 layer is modeled by the network of the Guassian Radial basis functions. The F3 neurons receive their input from the F2 neurons. Each radial basis function is centered to the F2 neuron output corresponding to the object it is supposed to encode. The radial basis functions are followed by a linear classifier to provide the score of each class. The weights of this classifier are learned in the training phase.

The RBF Network

The RBF network is made up of three layers (figure 2.5). The first layer is the layer of input vector. In the model, this input vector is the rearranged output of the F2 neurons. The input vector for a single F3 neuron is obtained by taking the output of the F2 neurons present in the receptive field of this F3 neuron and arranging it in the vector format. This vector is called as processed vector. The processed vector of an object is provided to the RBF neurons and then it's score for each class is obtained by the linear classifier. The score is similar to the probability measure and a larger score value implies the higher probability of the object belonging to that class. A gaussian radial basis function is given

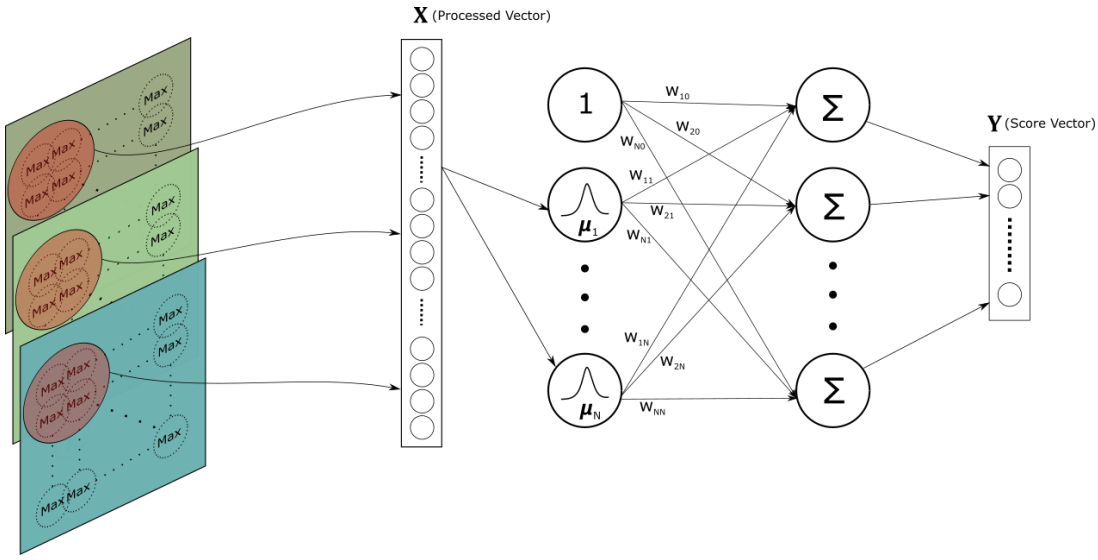


Figure 2.5: **The F3 layer.** This layer consists of the cells where each neuron is a RBF network. Each F3 neuron takes the vectorized input from the F2 neuron in its receptive field and outputs a score vector. The score value for each class represents the probability of the object in the receptive field belonging to that class.

as:

$$f(x) = e^{-\beta||x-\mu||^2} \tag{2.2}$$

Where, x is the input vector, μ is the center of the gaussian function and β is a parameter to control the width of the gaussian function. Small β value implies a wider curve.

$\mathbf{x}$ is the input vector to the RBF network. $\mathbf{f(x)} = [1, f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_N(\mathbf{x})]^T$ such that $f_i(\mathbf{x})$ is the value of the radial basis function with the input vector $\mathbf{x}$ and centered at the mean μ_i. $\mathbf{W}$ is the weight matrix of the classifier with elements $w_{ij}|i \in \{1, 2, ..., N\}$ and $j \in \{0, 1, ..., N\}$. The output $\mathbf{y}$ (score vector) of the network is given by:

$$\mathbf{y(x)} = \mathbf{Wf(x)} \tag{2.3}$$

Training the RBF network

Each RBF network is trained by the images of the different objects from different classes. The stimuli have two types of objects (one circular and one rectangular) and 36 different facing directions (0, 10, 20, . . . , 350). Hence, each RBF network outputs 2x36 neurons, each for a particular object type and facing direction. The circular object does not have any facing direction because of symmetry, hence, all the neurons tuned to the circular object have equal response irrespective of the facing direction.

The figure 2.6 shows 4 out of the 37 classes and one sample training image for each class. The training images for each class is generated by introducing a small translational shift within the receptive field of the F3 neuron. Additionally, for rectangular objects, rotation within the range of $-4°$ to $4°$ is introduced to get some more training images for the class. The processed vector of each training image in the class is calculated and

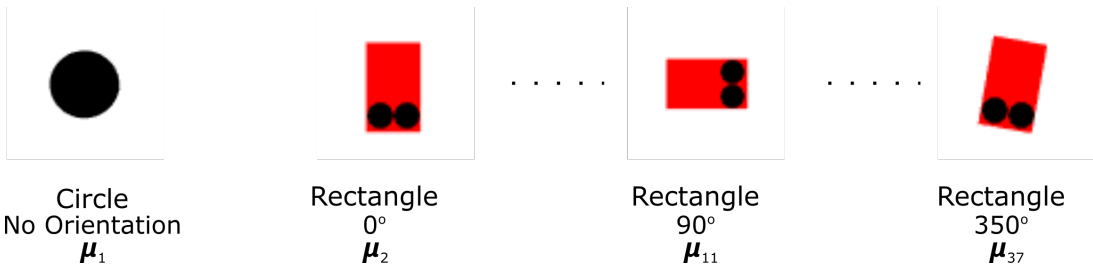


Figure 2.6: **Training images.** The RBF network is trained using these training images. The radial basis functions of the RBF network are centered over the processed vector of the training images.

averaged to get the RBF center for that class. Next, the weights of the classifier are learned using the Maximum likelihood solution for linear classification.

Say we have the training input matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}^T$ where $\mathbf{x}_j$ is the processed vector of the jth training image. $\mathbf{F(X)} = \{\mathbf{f(x}_1), ..., \mathbf{f(x}_M)\}^T$ is the output of RBF layer and is a matrix of dimensions $M \times (N+1)$. The output matrix is written as $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M\}^T$. Now the weight matrix is learned using the ML solution given by:

$$\mathbf{W} = \mathbf{Y}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^+ \tag{2.4}$$

where, A^+ represents the pseudo-inverse of a matrix A.

The output of the RBF network is taken and reorganized to get the output of the F3 layer. The F3 layer outputs 2x36 neurons (two kinds of objects with 36 orientations). $F_3(s, \phi)$ represents the neuron encoding the shape $s \in \{circle, rectange\}$ and the object orientation $\phi \in \{0, 10, 20, ..., 350\}$. The output of all the F3 neurons encoding circle are equal irrespective of the orientation. The output of the other neurons encoding a rectangle and a particular orientation would be equal to the RBF score of the rectangular object with the same orientation.

2.2 Motion Pathway

Inspired from the dorsal stream of the visual system (the "Where" pathway), the model has a motion pathway. This pathway includes layers which are responsible for the spatio-temporal processing to extract the information like the position, velocity, acceleration etc. The motion pathway has 4 layers. The first layer is the layer of position detectors which performs a frame by frame processing and replace the object by a gaussian blob. The objects in the input videos are restricted to the non-deformable objects such that the relative motion of different body parts is not possible. For the motion pathway, the information about the shape of the object can be dropped and the object can be replaced by a blob for simplifying the further processing. The next layer is the layer of motion detectors which encodes the velocity and the direction of the local motion in the visual field. This layer is followed by a layer of pooling neurons which pool over the local motion

and their output is the velocity and direction of motion of the object.The next and the final layer of motion pathway uses the output of the pooling layer and process it to encode the acceleration of the object.

2.2.1 Position Detectors

The first layer of motion pathway is the layer of the shape agnostic position detectors. These neurons are termed as M1 neurons and the layer as M1 layer. The visual features are already encoded in the form pathway. The M1 neurons are responsible for the position detection of individual objects. These are spread across the 2D space and the neuron corresponds to the middle of the object has the maximum response. Each M1 neuron can be modeled by a normalized gaussian filter with center at $(0, 0)$ and standard deviation σ in both the spatial dimensions. The activation of the M1 layer is given by a 2D convolution of the stimulus and the filter:

$$M_1 = p * \mathcal{N}(0, \sigma) \tag{2.5}$$

Where p is the stimulus in 2D space and $\mathcal{N}(0, \sigma)$ is the normalized gaussian filter centered at $(0, 0)$ and standard deviation σ.

With this filtering the local information about the object is lost and we are left with the global position. This replaces the object with a blob at the spatial position of the object which facilitates the process of extracting the global motion parameters of the object in a robust manner.

2.2.2 Motion Detectors

The second layer of motion pathway is the layer of the local motion detectors. These neurons are termed as M2 neurons and the layer is termed as M2 layer. The M2 neurons are similar to the motion direction selective neurons present in the V1,V2 and the MT area. These neurons provides the optical flow patterns in the stimulus and are modeled by the Reichardt detector. The motion pathway is interested in the global motion of the object and the local information of the object is already dropped at the M1 layer. The M2 neurons are used here to encode the translation motion of the blob provided by the M1 layer. A Reichardt detector is a simple spatio-temporal filter inspired by the motion detectors as shown in figure 2.7.

A polar coordinate system (r, θ) is used here to represent the 2D space. A single M2 neuron encodes for the speed and the direction of motion of each point in the 2D space. If we have M1 activations at two points in the space-time continuum represented by $M_1(r, \theta, t)$ and $M_1(r + \Delta r, \theta + \Delta\theta, t + 1)$, then the Reichardt detector for the position (r, θ) encoding the displacement of $(\Delta r, \Delta\theta)$ is given as:

$$M_2(r, \theta, \Delta r, \Delta\theta, t) = M_1(r, \theta, t) M_1(r + \Delta r, \theta + \Delta\theta, t + 1) \tag{2.6}$$

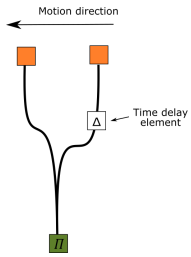


Figure 2.7: **The Reichardt detector.** A Reichardt detector is tuned to a particular displacement and time value. It measures the correlation between a neuron and another neuron at some displacement from the first one after a particular time delay.

Δr is a measure of velocity and $\Delta\theta$ is a measure of velocity direction, and hence, we replace these by v and $\dot{\theta}$ respectively.

2.2.3 Pooling Neurons

The next layer of the motion pathway is the layer of the pooling neurons. These neurons are named as M3 neurons and the layer as the M3 layer. The M3 neurons pool out the output of the M2 neurons present in their receptive field to encode the collective motion behavior. The pooling performed is the summation type pooling.

Each M3 neuron encoding the velocity of $(v, \dot{\theta})$ pools from the M2 output in its receptive field which is a patch in the 2D space (r, θ). This receptive field is typically of the order of the size of the object and hence equal to that of the receptive field of F3 neurons. Since, the stimulus contains a single object, the pooling is done over the entire 2D space for a simple implementation. The M3 layer output is then given by:

$$M_3(v, \dot{\theta}, t) = \sum_{r,\theta} M_2(r, \theta, \Delta r, \Delta\theta, t) \tag{2.7}$$

The neuron $M_3(v, \dot{\theta}, t)$ encodes the object velocity of magnitude v and direction $\dot{\theta}$.

2.2.4 Acceleration Detectors

The next layer of neurons is the M4 layer which encodes the change in the magnitude and the direction of the object velocity. It takes its input from the M3 layer. These neurons are also modeled by the Reichardt detector and takes in the M3 layer output at consecutive time steps.

$$M_4(\dot{v}, \ddot{\theta}, t) = M_3(v, \dot{\theta}, t) M_3(v, \dot{\theta}, t+1) \tag{2.8}$$

The neuron $M_4(\dot{v}, \ddot{\theta}, t)$ encodes if the velocity magnitude and the velocity direction of the object changes with $\dot{v}$ and $\ddot{\theta}$ respectively.

2.3 Animacy Detector

This section introduce an animacy detector which encodes the degree of perceived animacy of a stimulus. It is the final output of the model and takes its input from both the F3 and the M3 layers. This animacy detector is modeled based on the findings of Tremoulet and feldman Tremoulet and Feldman [2000] on the animacy perception in case of a single object. They studied the perception of the animacy based on the visual and motion features of the object and published the findings which are summarized below:

- Based on the facing direction and motion direction
  - An object with some facing direction (e.g. rectangle) is perceived as non-animate if the facing direction is not aligned with the direction of its motion.
  - A symmetric object without any facing direction (e.g. circle) is perceived as more animate than the object with facing direction not aligned with the motion direction.
  - If the facing direction of the object is aligned with the direction of its motion then it is perceived as most animate.
- Based on the speed and the motion direction
  - The perceived animacy increases with the increased speed deviation.
  - The perceived animacy increases with the increased deviation in the motion direction.

The animacy neuron is modeled with two parts, taking the above discussed two bases into consideration. First part encodes the animacy due to aligned facing and motion direction, and the second one due to the speed and the motion direction parameters.

The first part takes inputs from both the M3 layer and the F3 layer. This neuron is named as A1 neuron. It takes input from the orientation and motion direction encoding neurons from the F3 and the M3 layers respectively. The response of the orientation and motion direction neurons tuned to the same direction are multiplied, sum-pooled and passed to a sigmoid function to get the response of the A1 neuron. The sigmoid function is used to provide robustness to the alignment detection to the model. An asymmetric object with a clear anteroposterior orientation (for example the rectangle used in the input videos) achieves its maximum response when in the aligned condition. In case its facing direction is not aligned with the motion direction, the angle of misalignment does not matter for the animacy. A small misalignment is same as the large misalignment. To achieve this behavior of the A1 neuron, the sigmoid function is used. Although, the sigmoid function needs to be calibrated such that the aligned response falls in the 'ON' part, the misaligned response falls in the 'OFF' part and the response for the symmetric objects should fall in the linear region of the sigmoid function. The A1 neuron encodes the alignment of the objects facing direction and its direction of motion. It responds only

if the object's motion is in its facing direction. The output of the A1 neuron is then given as:

$$A_1(t) = S\left(\sum_i F_3(\phi_i) M_3(\dot{\theta}_i)\right) \tag{2.9}$$

where, $F_3(\phi_i)$ is the response of the F3 neuron encoding the orientation ϕ_i and $M_3(\dot{\theta}_i)$ is the response of the M3 neuron encoding the motion in the direction $\dot{\theta}_i$. S is the sigmoid function.

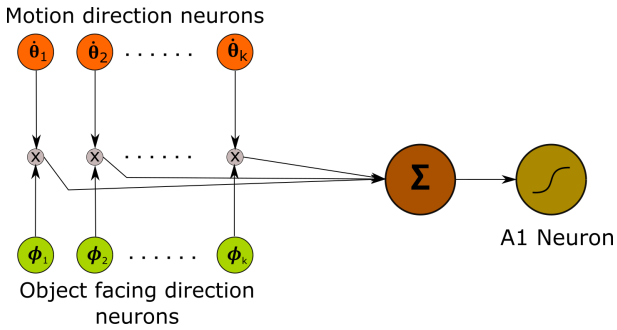


Figure 2.8: **The A1 neuron.** It encodes the aspect of animacy where the object's anteroposterior axis should be aligned to its direction of motion in order to be perceived as animate. The A1 neuron multiplies the motion direction and the object facing direction neurons tuned to the same angle and pools over all of the angles, followed by a sigmoid function.

The second part is termed as the A2 neuron and it takes inputs from the M3 and the M4 layer. To encode the degree of animacy due to the speed or motion direction deviation, it performs the weighted pooling of the velocity encoding neurons from the M3 layer, the acceleration encoding neurons and acceleration direction encoding neurons from the M4 layer. The weights are arranged such that the M3 neurons encoding the larger velocity, larger acceleration, and larger acceleration direction are given the larger weight. The simplest way for such weight distribution is to assign weight which are increasing linearly with the encoding parameter of the neuron. The A2 neuron then takes these weight pooled outputs and again perform a weighted summation to give its response. The output of the A2 neuron is given by:

$$A_2(t) = \sum_i w_i^v M_3(v_i) + \sum_j w_j^{\ddot{\theta}} M_4(\ddot{\theta}) + \sum_k w_k^{\dot{v}} M_4(\dot{v}) \tag{2.10}$$

The animacy neuron (A) is then the combination of A1 and A2 neurons. If the motion of the object is not aligned to its orientation then there is no perception of animacy irrespective of the type of motion. If there is no response from A1, there is no animacy. To achieve this, the response of the A1 and the A2 neuron are multiplied to get an 'AND' type of operation. The response of the animacy neuron at any time is the accumulated sum of the past responses and the present response normalized by the time duration.

$$A(t) = \frac{\sum_{i=0}^{t} A_1(i) A_2(i)}{t} \tag{2.11}$$

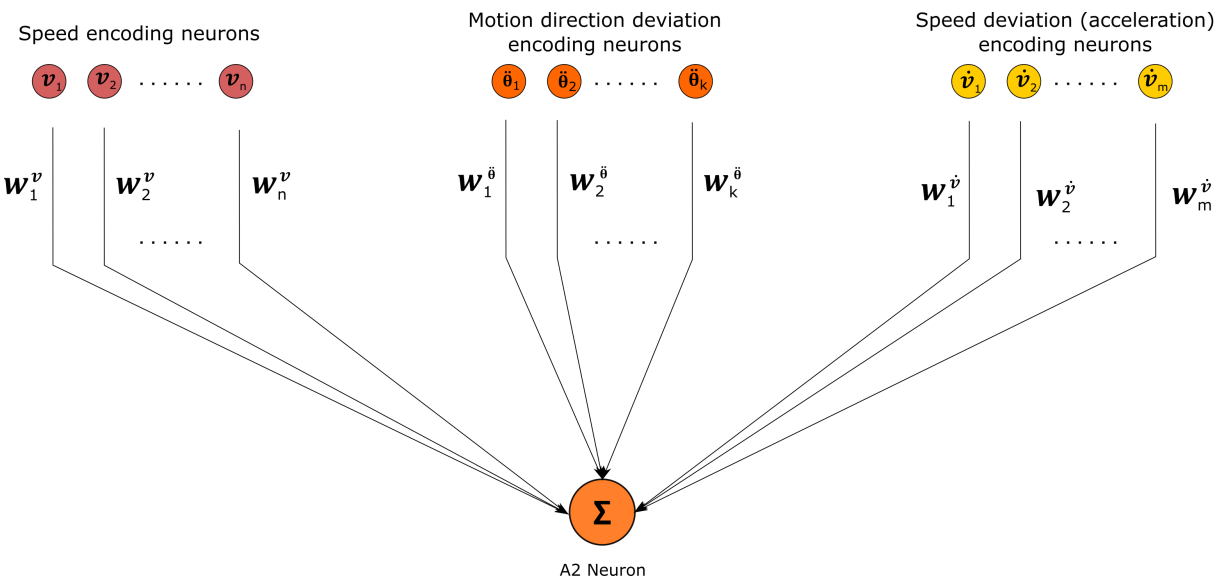


Figure 2.9: **The A2 neuron.** The output of the A2 neuron is the weighted pooling of the speed encoding, motion direction deviation and the speed deviation neurons. The weights for each type of neurons are linearly increasing with the encoding parameter.

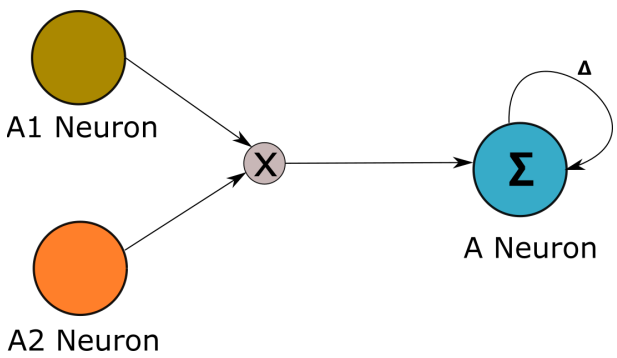


Figure 2.10: **The animacy neuron.** The output of the A1 and the A2 neurons are multiplied and averaged over the time to get the output of the animacy neuron.

Chapter 3

Results

In this chapter, we present various stimuli videos which are used to test the model and the results given by the model. This chapter is divided into two sections. The first section includes the details of the different stimuli used to evaluate the performance of the model. This section also discusses the expected output of the model for each stimulus based on the Tremoulet and Feldman [2000]. The second section discusses the processed output at the intermediate layers of the model and the final animacy results obtained from each stimulus video at the final stage of the model.

3.1 Stimuli Videos

This section discusses the stimuli videos that are passed on to the model to get the animacy response.

A total of 41 videos of a moving single object are used. In 20 (5*4) of them, the moving object is a circle and in the rest of the 21 (5*4+1) videos the object is a rectangle with eye like features. In each of the video, the object enters into the frame in straight line from the bottom of the frame with some velocity (say V) and then changes its direction of motion and the velocity. The 20 (5*4) videos for each object type are obtained by the combination of the 5 direction deviations $0°, 20°, 40°, 60°, 80°$ and 4 velocity changes $0.5V, 1V, 2V, 4V$. In the 20 videos for the rectangular object, the object's facing direction is always aligned with it's direction of motion and is called as the 'Aligned condition'. The facing direction of the rectangular object is towards its eyes. The one extra video in case of the rectangular object is for the 'Misaligned condition'. In this video the object changes it's direction of motion by 90° with no change in velocity and the facing direction is not aligned with the direction of motion (figure 3.2). Images corresponding to four videos for each object type are shown in the figure 3.11b and the figure 3.11a. These four images corresponds to the combination of extreme velocity change and direction deviation *i.e.* direction deviation $= 0°, 80°$ and velocity change $= 0.5V, 4V$.

The videos used are similar to the ones used by Tremoulet and Feldman [2000] in their study on the animacy perception in the case of a single moving object. However,

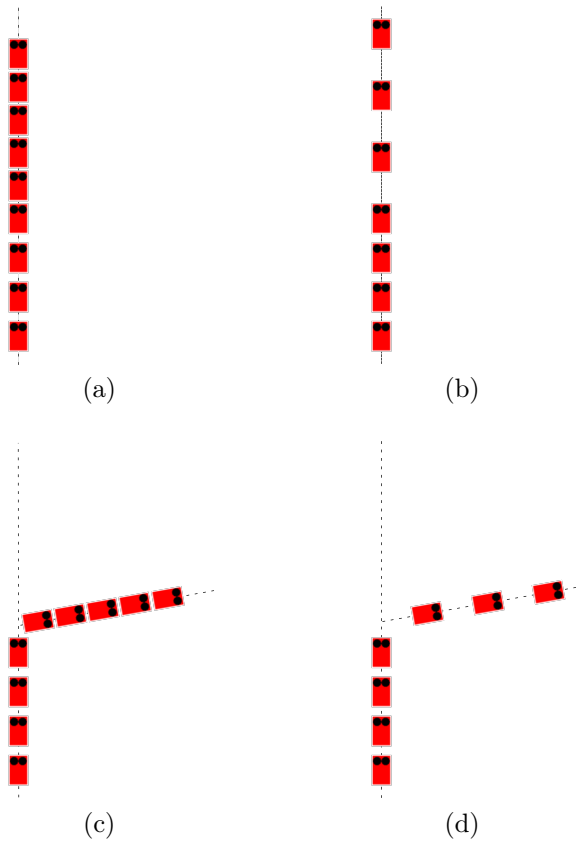


Figure 3.1: **The stimuli videos (aligned condition).** The figure shows the stimuli videos with the rectangular object in the four extreme configurations. a) Velocity change : 0.5V, direction deviation : 0° b) Velocity change : 4V, direction deviation : 0° c) Velocity change : 0.5V, direction deviation : 80° d) Velocity change : 4V, direction deviation : 80°

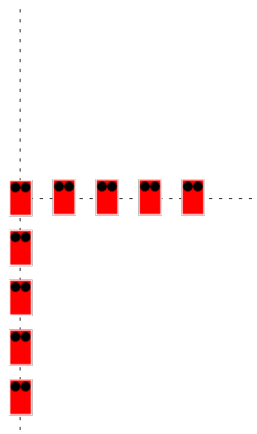


Figure 3.2: **The misaligned condition.** In this video, the rectangular object changes the direction of motion (direction deviation : 90°) but it's orientation remains the same. This kind of scenario is found out to be the least animate.

they used a plain rectangle and a plain circle as the two objects, while I use a plain circle and a rectangle with eye like features. A plain circle is completely symmetric and has no facing direction, but, a plain rectangle gives a sense of two possible facing directions. The purpose of adding eyes to the rectangle is to give a definite visual cue for facing direction (i.e. towards the eyes). The Tremoulet and Feldman [2000] findings suggests that the animacy perception increases both with the increasing direction deviation and the velocity change. In the case of the 'Misaligned condition' the object does not look animate and has a very low animacy rating.

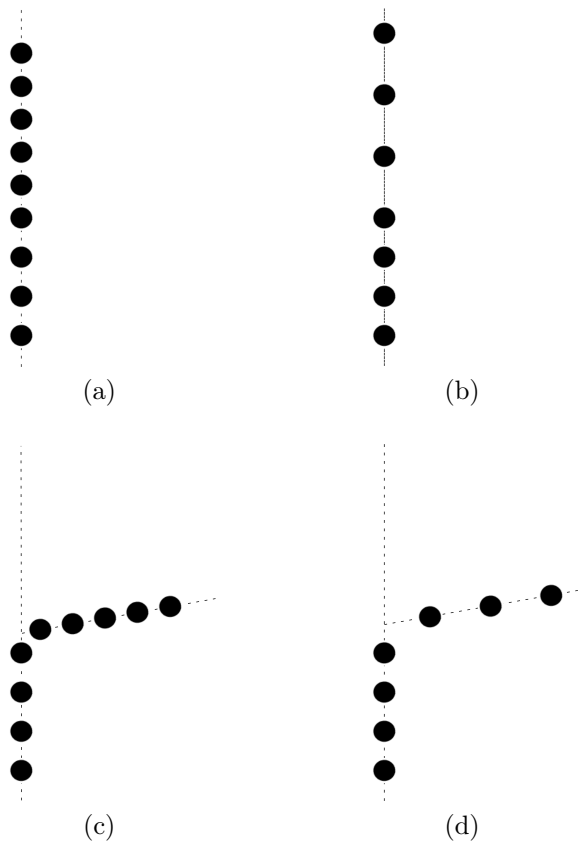


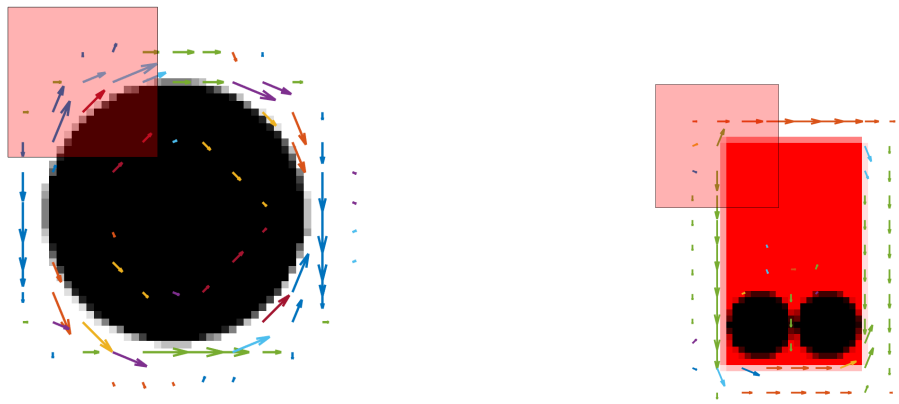
Figure 3.3: **The stimuli videos.** The figure shows the stimuli videos with the circular object in the four extreme configurations. a)velocity change : 0.5V, direction deviation : 0° b) velocity change : 4V, direction deviation : 0° c) velocity change : 0.5V, direction deviation : 80° d) velocity change : 4V, direction deviation : 80°

3.2 Results

In this section, the results obtained at the intermediate layers of the model after the processing of the stimuli and the final output of the model for each stimulus are discussed.

3.2.1 Form Pathway

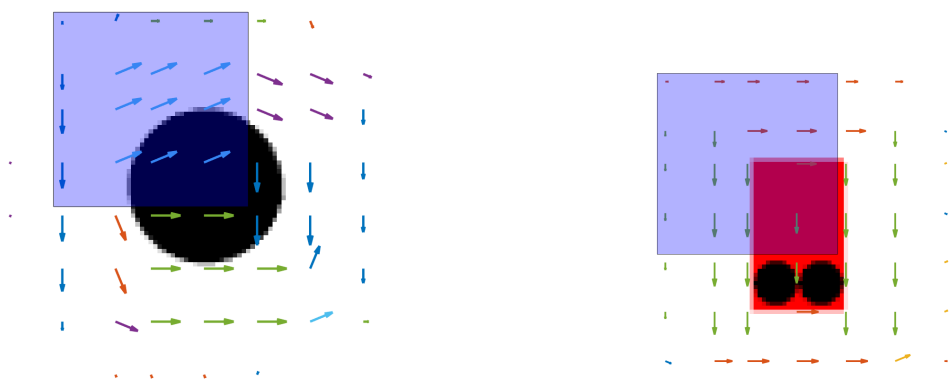
The first layer of form pathway is the F1 layer of edge detectors. Figure 3.4 shows a part of the single frame of the input video and the output of the F1 neurons on top of it. It



(a) (b)

Figure 3.4: **The F1 response.** The arrows shown represent the response of the F1 neurons. At any position, the F1 neuron with the maximum response is shown. The length of the arrow is proportional to the response of the F1 neuron at that position and the orientation of the arrow represents the angle at which the neuron is tuned. The red translucent rectangle shows the receptive field of an F1 neuron compared to the object. a) The circular object b) The rectangular object.

shows the F1 response for both types of objects in the stimulus. The F1 neurons are spread over the visual field. The translucent red square in the figure shows the receptive field area of any randomly selected F1 neuron. Each receptive field is shared by neurons of different orientation tuning. The arrows in the figure shows the response of the F1 neurons, where the orientation of the arrow represents the tuned orientation of the F1 neuron with maximum response and the length of the arrow is proportional to the response of the neuron.



(a) (b)

Figure 3.5: **The F2 response.** The arrows shown represents the response of the F2 neurons. A F2 neuron performs the max pooling over the F1 neurons in its receptive field. The blue translucent rectangle shows the receptive field of a F2 neuron compared to the object. a) The circular object b) The rectangular object.

The next layer (F2) is the pooling layer to provide local translational invariance. It pools over the F1 neurons in its receptive field by performing maximum type pooling.

Figure 3.5 shows the response of the pooling neurons, and the blue square shows the receptive field of the F2 neurons.

The next layer (F3) is the layer of object detector neurons. It consists of $36 * 2 = 72$ neurons (36 for each facing direction in {0,10,20,...,350} and 2 for each object type $\{Circle, Rectangle\}$). Figure 3.6 shows the response of the F3 layer. The yellow square in the subfigures 3.6a and 3.6c shows the receptive field of F3 neurons and the arrows shows the marginalized response of facing direction neurons. The length of the arrow is proportional to the response in that direction. In the subfigure 3.6a all the arrows are of equal length capturing the symmetry of the object. In the subfigure 3.6c the arrow towards the eyes is of the longest and hence captured the facing direction of the object nicely.

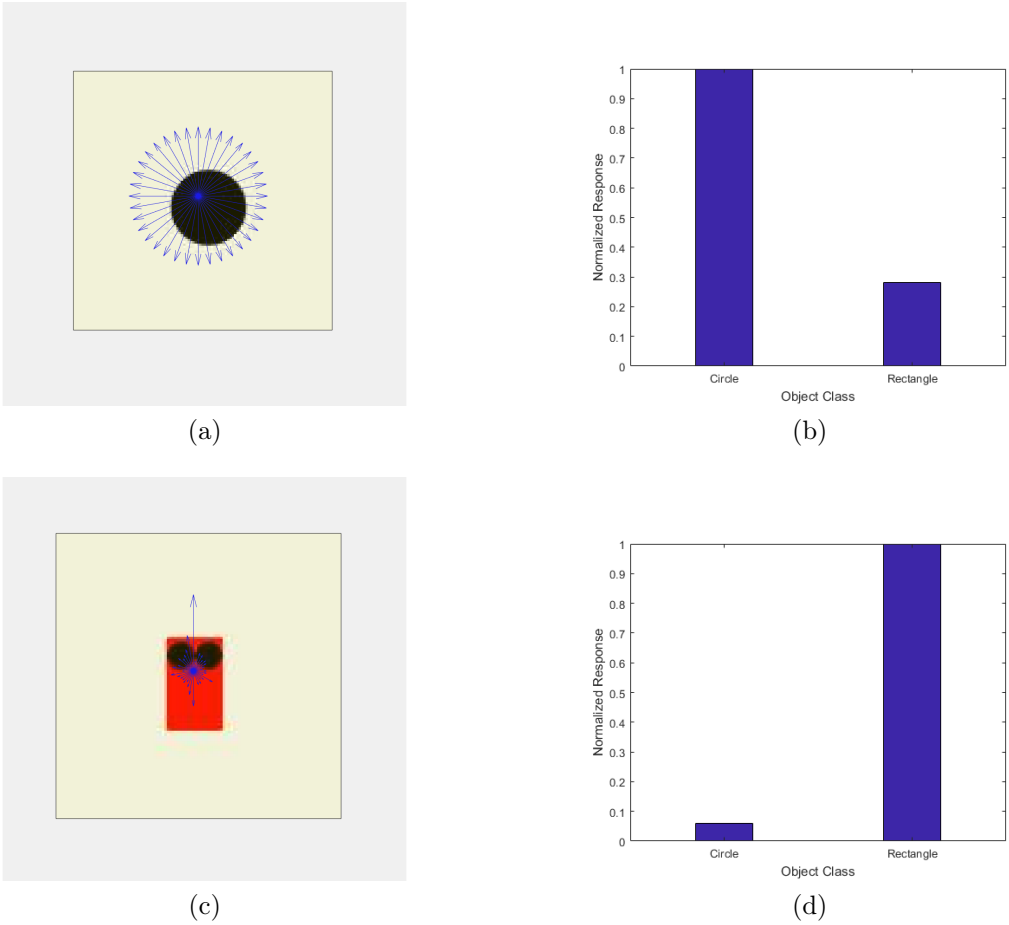


Figure 3.6: **The F3 response.** a) The response of all the neurons tuned to the different directions are equal in case of the circular object since the object is symmetric and there is no specific facing direction. The translucent yellow square shows the size of the receptive field of the F3 neurons. b) The response of the object identification F3 neurons in case of the circular object. The response of the neuron tuned to the circular object is significantly higher than the neuron tuned to the rectangular object. c) The response of the F3 neurons for a rectangular object. The facing direction is correctly detected as the maximum response neuron is tuned towards the object's frontal direction (the frontal direction is towards the eyes). d) The response of the object identification F3 neurons in case of the rectangular object. The response of the neuron tuned to the rectangular object is significantly higher than the neuron tuned to the circular object.

The subfigures 3.6b and 3.6d shows the marginalized response of the object type neurons for the *Circle* and the *Rectangle* respectively. We can see from the response that the object type detector neurons can detect the object robustly.

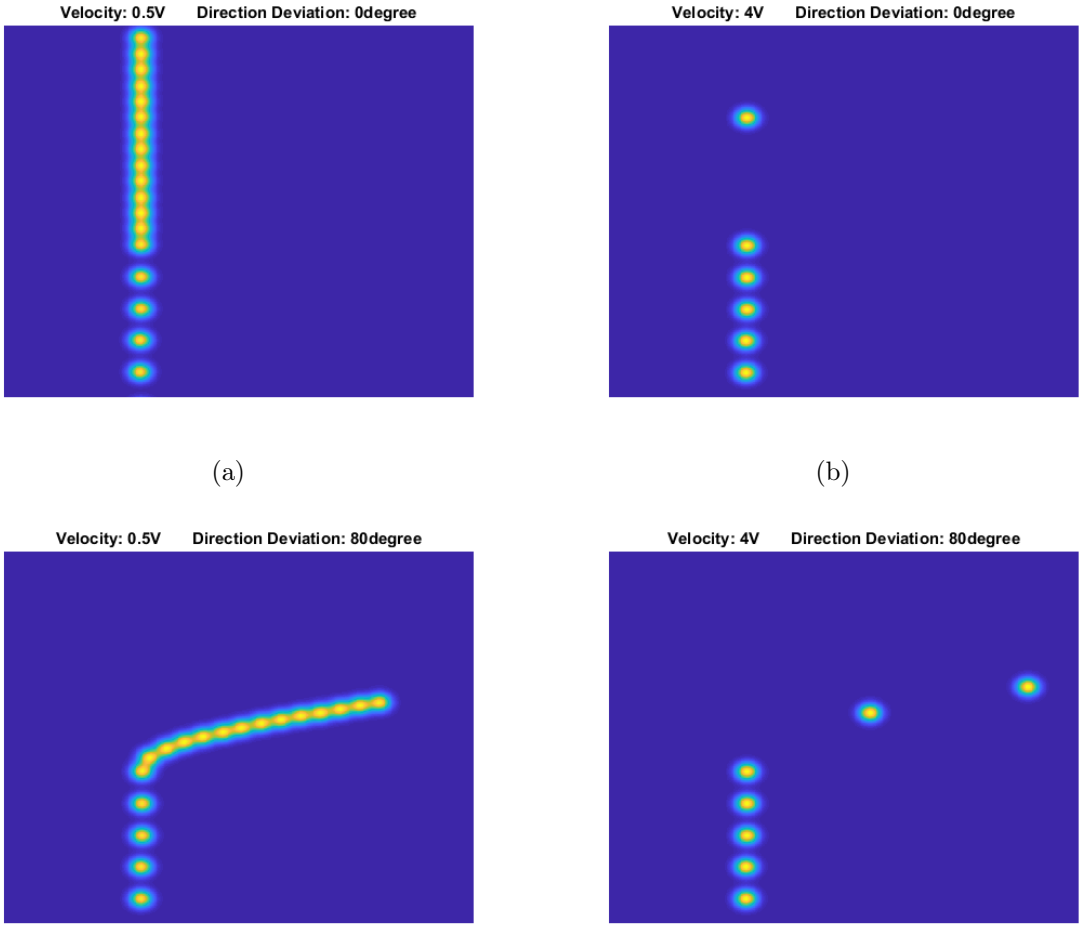
3.2.2 Motion Pathway

The first layer of the motion pathway is the position detection layer (M1) consists of the gaussian filters. The filtering removes the local visual features of the object and replace it by a gaussian blob indicating the position of the object in the global space.The motion of this blob is then processed by the further layers of the motion pathway. Figure 3.7 shows the response of the M1 layer for the four stimuli videos a) Velocity : 0.5V direction deviation : 0 degree b) Velocity : 4V direction deviation : 0 degree c) Velocity : 0.5V direction deviation : 80 degree d) Velocity : 4V direction deviation : 80 degree. Every image in the figure represents all the processed frames of the input video. All the processed frames are added together to form these images. Since the video is sampled uniformly, the distance between two consecutive blobs is the measure of the velocity. The processed output is same for the rectangle video and the circle video of the same velocity and direction change.

The next layer is the M3 layer that consists of the direction deviation and velocity neurons. Figures 3.8 and 3.9 shows the output of the M3 layer for four input videos with a) velocity : 0.5V direction deviation : 0 degree b) velocity : 4V direction deviation : 0 degree c) velocity : 0.5V direction deviation : 80 degree d) velocity : 4V direction deviation : 80 degree.

In figure 3.8, the x axis shows the time in frames and the y axis shows the direction neurons tuned to different motion directions. For the stimuli with the deviation of 0 degrees, there is no change in the neuronal activations throughout the video duration, while the change in the activations in cases where the direction deviation is 80 degrees can be seen. Note that the input videos with the velocity change of 4V has a shorter duration. This is because the on-screen time of the object is lesser in these cases, as the object with 4V moves quickly out of the frame.

Similarly, the figure 3.9 shows the response of the velocity neurons for the same set of input videos. For the videos with velocity change to 0.5V, the activation of the neurons encoding lower velocity increases and for the videos with the velocity change to 4V, the activation of the higher velocity encoding neurons gets increased.



(a) (b)

(c) (d)

Figure 3.7: **The M1 response.** The M1 neurons captures the position of the object. At the M1 layer, the object is replaced by a gaussian blob. The figure shows the M1 layer response for the 4 videos shown in figure 3.1.

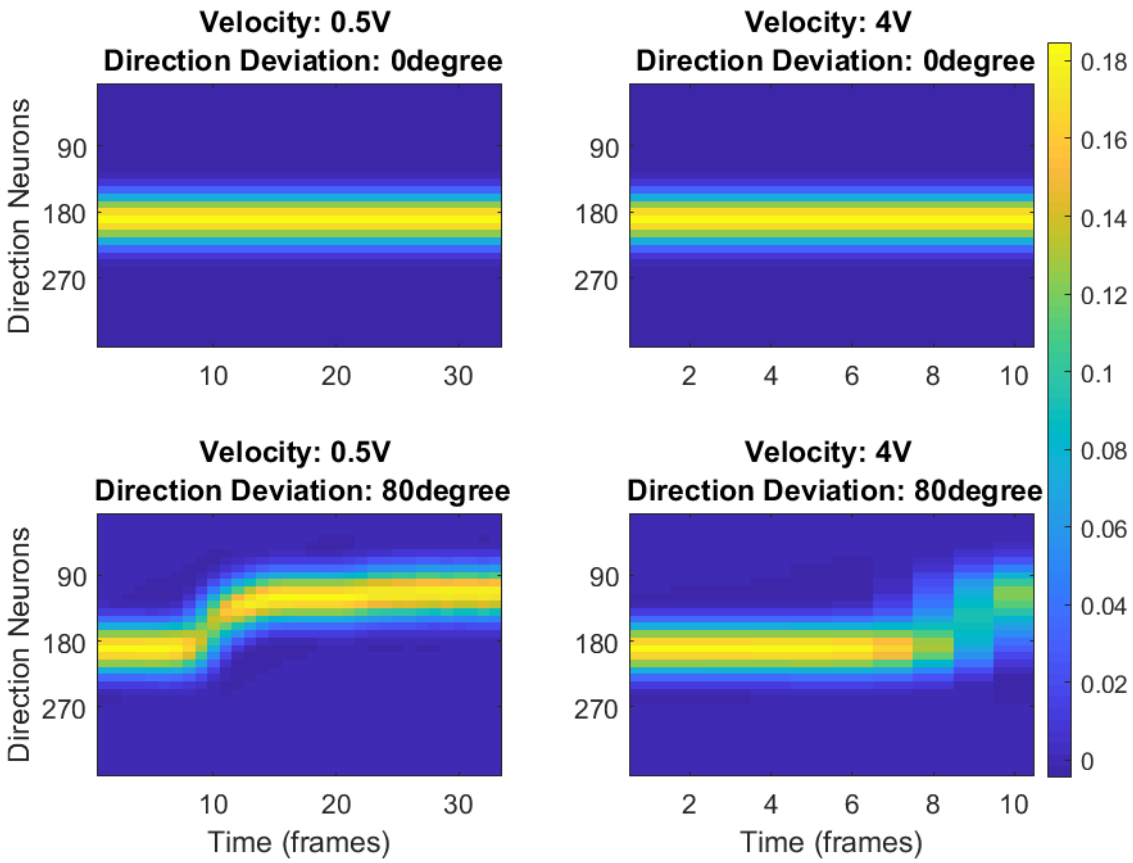


Figure 3.8: **The response of the direction deviation neurons from the M3 layer.** The response shown here is for the videos in the figure 3.1. We can see the shift in the activation for the videos with non zero direction deviation, however these neurons do not respond to the velocity change.

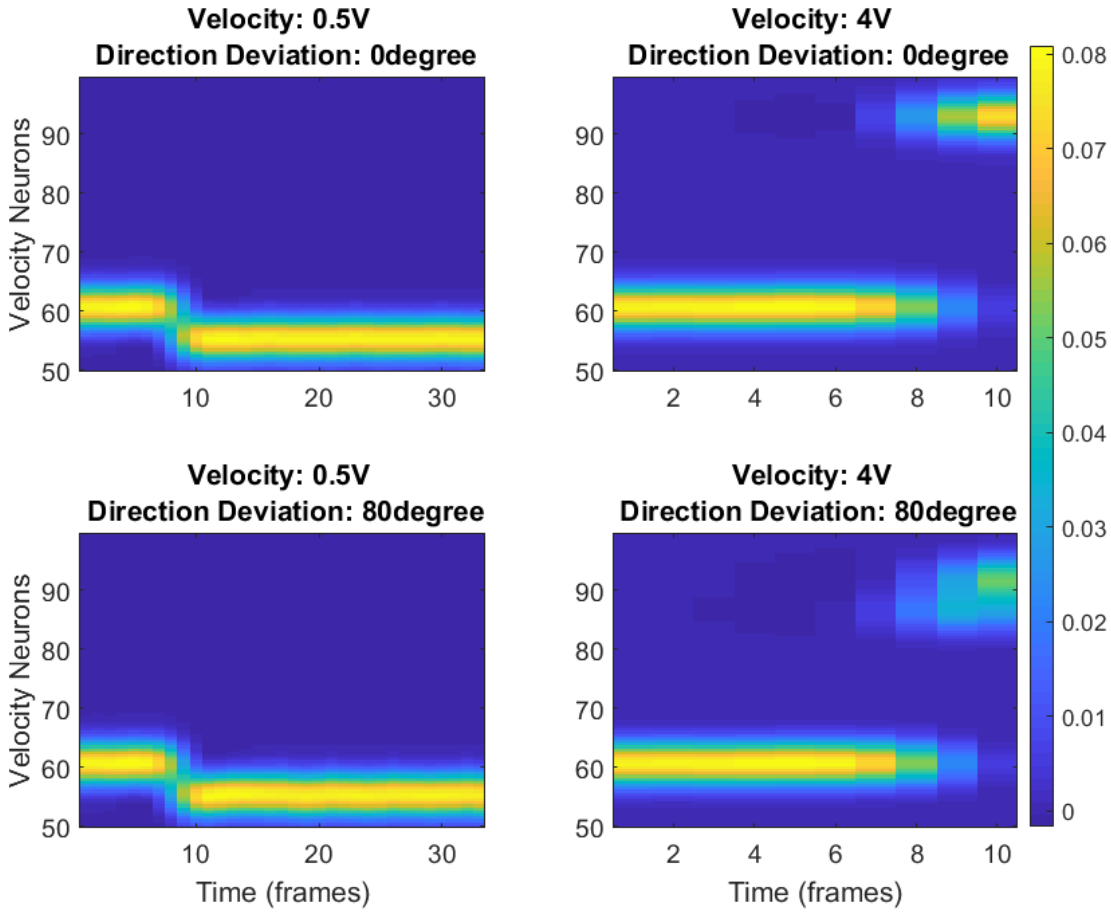


Figure 3.9: **The response of the velocity neurons from the M3 layer.** The response shown here is for the videos in the figure 3.1. The activation shifts from the neurons encoding initial velocity to the neurons encoding the final velocity as the velocity of the object changes. These neurons do not respond to the change in the direction of motion.

3.2.3 Animacy Neuron

The animacy neuron takes the input from the F3, M3 and the M4 layers and gives the output as the animacy rating of the object in the video. Figure 3.10 shows the response of the animacy neurons in 3 cases, 1) the misaligned condition, 2) the aligned condition with no velocity change and no direction deviation (i.e. changed velocity = $1V$ and direction deviation = $0°$), 3) the circle with no velocity change and no direction deviation. It can be seen that the aligned rectangle has the highest animacy rating followed by the circle and the misaligned condition has the lowest animacy rating. This result is similar to the findings of Tremoulet and Feldman [2000] where they stated that if an object's motion is not aligned with the anticipated anterior direction of the object then it does not look animate at all.

For further experiments, they used the videos with aligned rectangle and the circle (discussed in the input videos section) to study the effects of velocity change and the direction deviation on the animacy perception. The response of the model for such videos are shown in the figure 3.11. Figure 3.11a shows the animacy response of the model for the videos with the rectangle and the figure 3.11b shows the model response for the videos with the circle. In both the figures we can see that within the videos of same object type (rectangle or circle), the animacy response increases with the increment in the velocity change and with the direction deviation. The animacy response is shown using a color coding. Note that the color bar is differently encoded for both the object types and the color indicating the response encodes the difference within the object category. The response comparison between the object category is visible in the figure 3.13, where the mean marginalized animacy response is shown for the velocity change and the direction deviation separately. Figure 3.13b shows the mean animacy response of the rectangle videos and the circle videos with the change in the velocity. Figure 3.13a shows the animacy response of these videos with direction deviation. It can be seen that the mean animacy response increases with the increment in both the velocity change and with the direction deviation. The animacy response for the circle videos are lesser than the rectangle videos for the exactly same motion configuration (velocity change and direction deviation).

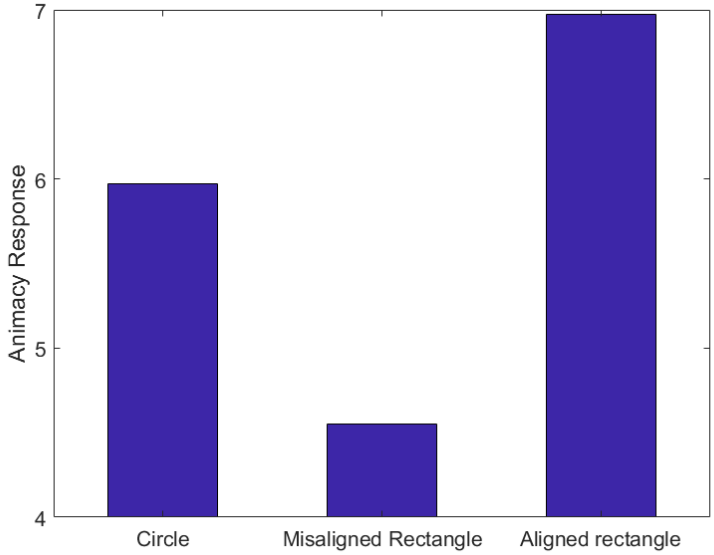
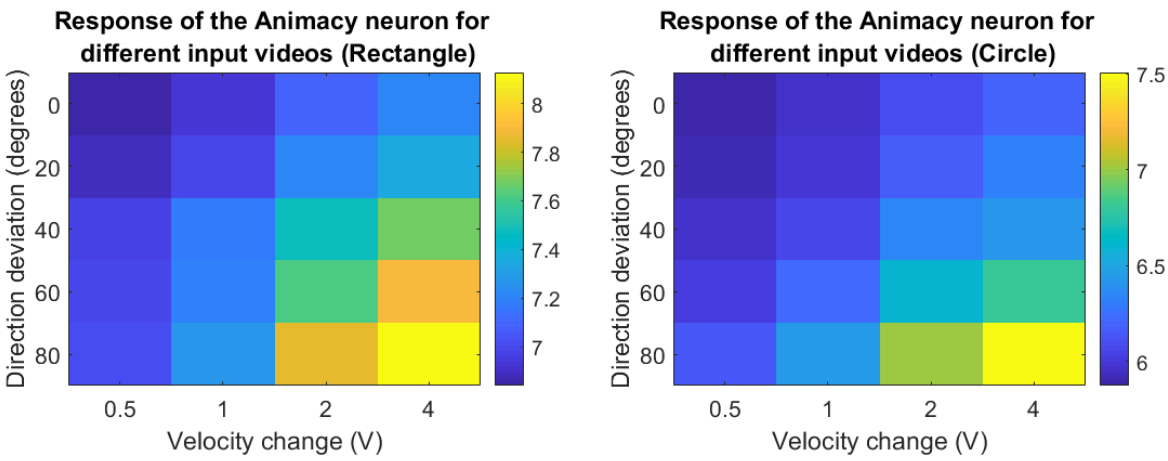
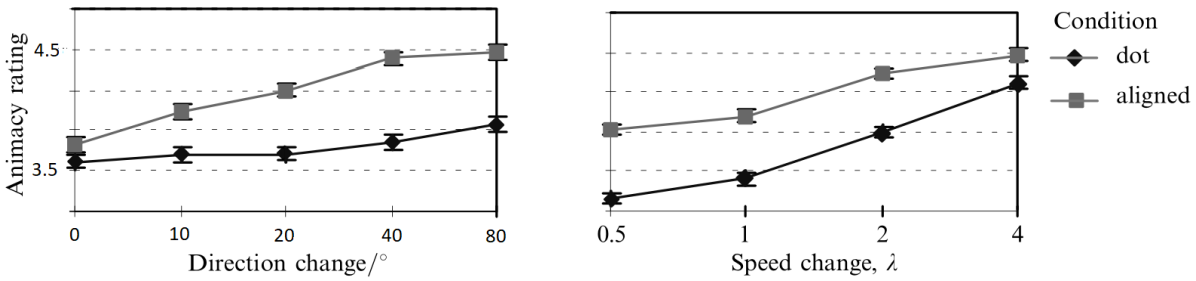


Figure 3.10: **The animacy response for the three conditions (The circle, the aligned rectangle and the misaligned rectangle).** The circle and the aligned rectangle condition has no velocity and the direction deviation. The misaligned condition has the 90° of motion direction deviation but the facing direction does not change. Even if there is a change in the motion direction in the misaligned condition, the animacy rating is the lowest. The aligned condition has the maximum response followed by the video with circular object.



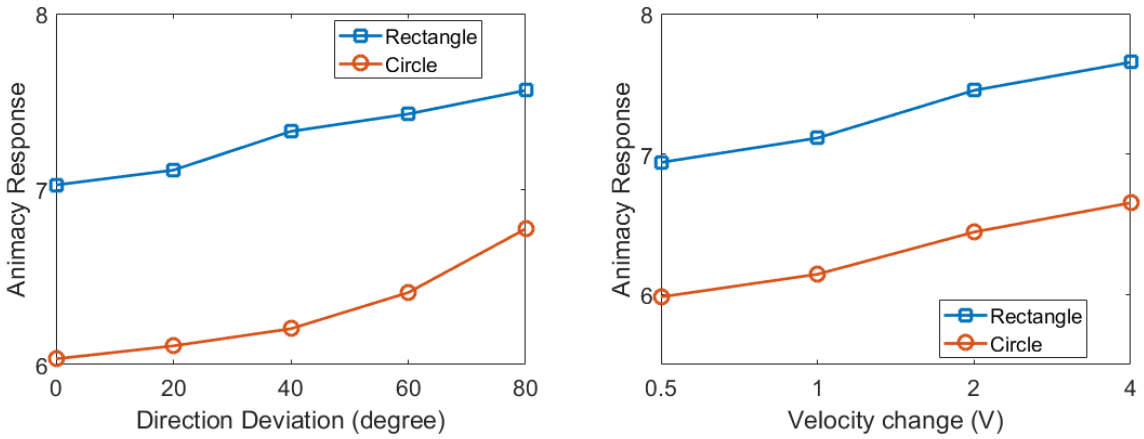
(a) (b)

Figure 3.11: **The animacy response of the input videos with the circular object and the aligned rectangular object.** For each object type, the animacy rating increases with the increasing velocity and the increasing direction deviation. The animacy rating of the videos with circular object is always less than that of the rectangular object for the same velocity and the direction deviation configuration (notice the difference in the color bar).



(a) (b)

Figure 3.12: The results from the Tremoulet and Feldman study on the animacy perception in case of a single object.The perceived animacy increases with the increasing change in the direction as well as the increasing change in the speed. The dot condition here represents the circle condition and the aligned condition represents the aligned rectangle condition in our case. The rating in the dot condition and the aligned condition is same in nature (monotonically increasing) but the dot condition has less response. **(from Tremoulet and Feldman [2000])**



(a) (b)

Figure 3.13: **The mean marginalized animacy response to show the change in animacy with change in velocity and the direction deviation.** The animacy response increases with the increasing final velocity and the increasing direction deviation. For the same configuration, the videos with circular object is perceived as less animate than the video with rectangular object.

Chapter 4

Discussion

In the first chapter we disccussed the animacy perception and the various cues behind it. We have also discussed the study by Tremoulet and Feldman Tremoulet and Feldman [2000] about the animacy perception in case of a single object. In the second chapter, we introduced a computational model for animacy perception in case of a single object. The model is a convolutional neural network with two pathways inspired by the ventral and dorsal pathways in the human visual system. In the third chapter, we discussed the input videos for the model which are similar to the ones used by Tremoulet and Feldman. Further, the output of the intermediate layers and the final animacy rating of the videos given by the model is presented.

The model can successfully replicate the qualitative results of Tremoulet and Feldman's study. The model's output are in compliance with the findings of the study, which are:

- If the object's anteroposterior axis is not aligned with its direction of motion, then the object is not perceived as animate.
- The animacy perception increases with the increasing angle deviation and the increasing velocity change.
- The object with no defined anteroposterior axis (a symmetrical object) is perceived as less animate than the object with a defined anteroposterior axis aligned with the direction of motion.

The model is capable of extracting the visual features of multiple objects and the motion features of a single object. It can combine these features to give the animacy ratings in the case of a single object. In order to give the animacy rating in the case of multiple objects, it needs to extract the motion features of multiple objects and the animacy neuron needs to be extended to combine all the extracted features properly. In the next section we discuss what the model can not achieve right now and how we can improve it.

4.1 Challenges and future work

Although the model can give animacy rating to the simple stimuli videos used by the Tremoulet and Feldman, it can not incorporate the complex cues for animacy perception. The animacy neuron takes in the simple motion cues (the absolute speed, the speed change and the direction deviation) and the simple feature (the object's shape and the facing direction) to give the animacy rating. The animacy perception is a complex process. It involves several motion cues and visual features. Additionally, if there are multiple objects then the same motion cues and visual features are not sufficient and the interaction between the objects also plays the role. In general, an object is perceived as animate if it looks like that the object's motion is due to the internal energy present in the object instead of some external interaction. Based on previous research in the field, the animacy perception can be summarized with the following points:

- In case of a single object.
  - If the motion of the object is simple with few deviation in kinematics, like the one used by Tremoulet and Feldman, the animacy perception depends on the simple cues such as the ones included in the animacy neuron of the model.
  - If the motion of the object is complex including several deviations in kinematics Schultz and Bülthoff [2013], the animacy also depends on the observer's experience. The observer can map to the motion with some previously seen motion and perceive the object to be animate or inanimate. For example, a falling leaf's motion includes several changes in the velocity, trajectory etc. but it is perceived to be inanimate. Here, the leaf is interacting with the air/wind resulting in such a motion and one can tell this based on the experience.
- In case of multiple objects
  - The animacy in the case of multiple objects not only depends on the object's motion but the interaction between the objects as well. If a change occurs in the motion of an object in a single object scenario, it might look like an animate object. If the same motion occurs in a multiple object scenario where the motion change appears to be due to the interaction with another object, it'll no longer be perceived as an animate object.

The model is limited to capture only the case with the single object and simple motion. It needs further modifications to incorporate the scenarios with single object with complex motion and the scenario with multiple objects. Now I will discuss briefly about how the model can be extended to include the multiple objects.

It's easy to extend the model for multiple objects. The form pathway is already capable of processing information for multiple objects if the two objects are not in the receptive field of the same F3 neuron. For the motion pathway, the pooling at the M3

layer should be performed within the receptive field of the M3 neurons which are spatially located at the positions of F3 neurons and with the same receptive field. Right now this pooling is performed over the whole visual field as there is only a single object present. To include the multiple objects, same pooling for each object is to be performed in the receptive field of M3 neurons in which the object is present.

There are other implementation related limitations of the model. The video is provided to the model in the form of the frames sampled with some sampling rate. Also, the model can't take the input videos of high resolution. The video sampling rate and the input resolution put restriction on the precision with which the model can detect the kinematic parameters like the position, the velocity and the acceleration etc. Each layer of the motion pathway introduces some error during the processing. If the change of the kinematic parameter which the model is supposed to detect is of the same order as that of the error, it affects the further processing and outputs the wrong result. Thus, the model is not robust enough for detecting the small changes in time and space. Next, I will discuss the case of the input video with velocity change of 0.5V and the direction deviation of 10° to show how the model fails to detect the velocity direction of the object.

Figure 4.1a shows the image corresponding to the video with the velocity change of 0.5V and the direction deviation of 10°. Figure 4.1b shows the velocity direction detected by the M3 layer. We can see that initially the direction of motion is 180° which is correctly detected. After the diversion, the direction of motion is 170°, however, the maximum responding M3 neuron jumps between 180° and 170°. Due to this, the further layers detect as if the object is making several changes in the motion and this leads to a relatively higher animacy rating. In this case, the speed is $0.5V$ which is small and for the used sampling rate the time interval between the consecutive frame is δt. The displacement in the perpendicular direction is also very small (given by $0.5V\delta t \sin 10°$) which can't be detected by the model because of the limited precision in space. Figure 4.1d shows the detected velocity direction in case of no velocity change and direction deviation of 10°. Since the velocity is not very low in this case, the M3 layer successfully detects the 10° change.

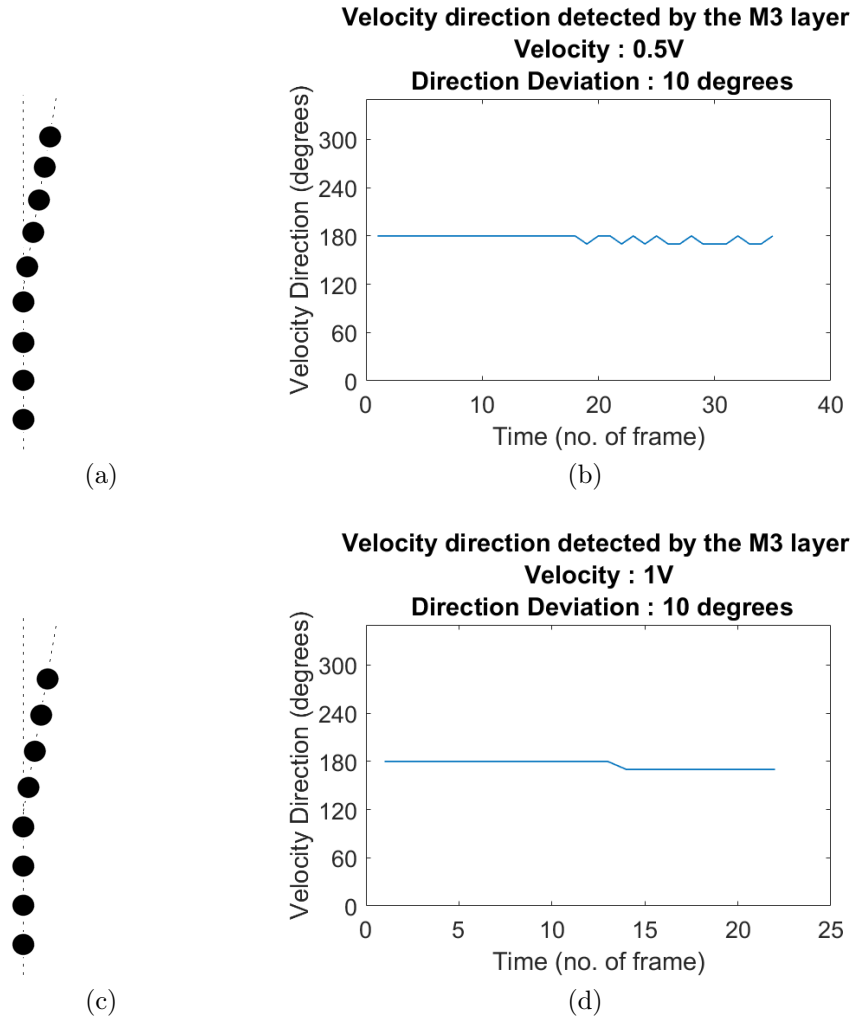


Figure 4.1: a) The video with the 10° deviation and the final velocity of 0.5V. b) Sometimes the M3 layer fails to detect the velocity direction of 10° in the case of the final velocity of 0.5V. c) The video with the 10° deviation and the final velocity of 1V. d) For the final velocity of 1V, the M3 layer successfully detects the change of 10°.

Bibliography

H. C. Barrett, P. M. Todd, G. F. Miller, and P. W. Blythe. Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, 26(4):313–331, 2005. ISSN 10905138. doi: 10.1016/j.evolhumbehav.2004.08.015. URL https://www.psychologytoday.com/files/attachments/95822/perceiving-animate-motion.pdf. 5

M. M. Chun, J. Golomb, and N. B. Turk-Browne. A Taxonomy of External and Internal Attention. *Dx.Doi.Org.Proxy.Lib.Ohio-State.Edu*, 62(1):73–101, 2010. ISSN 1545-2085. doi: 10.1146/annurev.psych.093008.100427. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.093008.100427{%}5Cnpapers3://publication/doi/10.1146/annurev.psych.093008.100427. 6

E. Di Giorgio, M. Lunghi, F. Simion, and G. Vallortigara. Visual cues of motion that trigger animacy perception at birth: The case of self-propulsion. *Developmental Science*, pages n/a–n/a, feb 2016. ISSN 14677687. doi: 10.1111/desc.12394. URL http://doi.wiley.com/10.1111/desc.12394. 3

W. H. Dittrich and S. E. Lea. Visual perception of intentional motion. *Perception*, 23(3):253–268, 1994. ISSN 03010066. doi: 10.1017/CBO9781107415324.004. 3, 6

T. Gao and B. J. Scholl. Perceiving animacy and intentionality. In M. D. Rutherford and V. A. Kuhlmeier, editors, *Social Perception*. The MIT Press, 2013. 6

T. Gao, G. E. Newman, and B. J. Scholl. The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, 59(2):154–179, 2009. ISSN 00100285. doi: 10.1016/j.cogpsych.2009.03.001. 8, 9

T. Gao, G. McCarthy, and B. J. Scholl. The Wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological science*, 21(12):1845–1853, dec 2010. ISSN 0956-7976. doi: 10.1177/0956797610388814. URL http://pss.sagepub.com/content/21/12/1845.short. 7, 8

T. Gao, B. J. Scholl, and G. McCarthy. Dissociating the Detection of Intentionality from Animacy in the Right Posterior Superior Temporal Sulcus. *The Journal of Neuroscience*, 32(41):14276–14280, 2012. ISSN 15378276. doi: 10.1016/j.biotechadv.2011.08.021.Secreted. 5

R. Gelman, F. Durgin, and L. Kaufman. Distinguishing between animates and inanimates: not by motion alone. *Causal Cognition: A Multidisciplinary Debate*, pages 151–184, 1996. doi: 10.1093/acprof:oso/9780198524021.003.0006. URL http://www.loc.gov/catdir/enhancements/fy0640/85013587-d.html. 5

M. A. Giese and T. Poggio. Cognitive neuroscience: Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4(3):179–192, 2003. ISSN 1471-003X. doi: 10.1038/nrn1057. URL http://www.nature.com/doifinder/10.1038/nrn1057. 13, 14

M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. URL http://www.cnbc.cmu.edu/braingroup/papers/goodale{_}milner{_}1992.pdf. 15

M. Hernik, P. Fearon, and G. Csibra. Action anticipation in human infants reveals assumptions about anteroposterior body-structure and action. *Proceedings. Biological sciences / The Royal Society*, 281(1781):20133205, 2014. ISSN 1471-2954. doi: 10.1098/rspb.2013.3205. URL http://www.ncbi.nlm.nih.gov/pubmed/24573853. 2

P. McAleer and S. A. Love. Perceiving intention in animacy displays. In M. D. Rutherford and V. A. Kuhlmeier, editors, *Social Perception*. The MIT Press, 2013. 11

A. Öhman and S. Mineka. The malicious serpent: Snakes as a prototypical stimulus for an evolved module of fear. *Current Directions in Psychological Science*, 12(1):5–9, 2003. ISSN 09637214. doi: 10.1111/1467-8721.01211. URL http://journals.sagepub.com/doi/pdf/10.1111/1467-8721.01211. 6

T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–266, 1990. ISSN 0028-0836. doi: 10.1038/343263a0. URL http://www.nature.com/doifinder/10.1038/343263a0. 13

D. H. Rakison and J. Derringer. Do infants possess an evolved spider-detection mechanism? *Cognition*, 107(1):381–393, 2008. ISSN 00100277. doi: 10.1016/j.cognition.2007.07.022. URL http://ac.els-cdn.com/S0010027707002090/1-s2.0-S0010027707002090-main.pdf?{_}tid=925dd48c-0412-11e7-b337-00000aab0f6b{&}acdnat=1488986527{_}dce0c5008523dcdd4d7627c744b15a87. 6

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–25, 1999. ISSN 1097-6256. doi: 10.1038/14819. URL http://www.ncbi.nlm.nih.gov/pubmed/10526343. 13

B. J. Scholl and P. D. Tremoulet. Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8):299–309, aug 2000. ISSN 13646613. doi: 10.1016/S1364-6613(00)01506-0. URL http://linkinghub.elsevier.com/retrieve/pii/S1364661300015060. 4

J. Schultz and H. H. Bülthoff. Parametric animacy percept evoked by a single moving dot mimicking natural stimuli. *Journal of vision*, 13(4):15, mar 2013. ISSN 1534-7362. doi: 10.1167/13.4.15. URL http://www.ncbi.nlm.nih.gov/pubmed/23525131. 40

T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:994–1000, 2005. ISSN 10636919. doi: 10.1109/CVPR.2005.254. URL http://ieeexplore.ieee.org/document/1467551/. 13

P. A. Szego and M. D. Rutherford. Actual and illusory differences in constant speed influence the perception of animacy similarly. *Journal of vision*, 7(12):5.1–7, sep 2007. ISSN 1534-7362. doi: 10.1167/7.12.5. URL http://jov.arvojournals.org/article.aspx?doi=10.1167/7.12.5. 3

A. Treisman. How the deployment of attention determines what we see. *Visual cognition*, 14(4-8):411–443, 2006. ISSN 1350-6285. doi: 10.1080/13506280500195250. URL http://www.tandfonline.com/doi/abs/10.1080/13506280500195250. 6

P. D. Tremoulet and J. Feldman. Perception of animacy from the motion of a single object. *Perception*, 29(8):943–951, 2000. ISSN 03010066. doi: 10.1068/p3101. 2, 3, 4, 5, 6, 10, 12, 13, 24, 27, 29, 36, 38, 39

P. D. Tremoulet and J. Feldman. The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception & psychophysics*, 68(6):1047–1058, 2006. ISSN 0031-5117. doi: 10.3758/BF03193364. 2, 5, 6

B. van Buren, S. Uddenberg, and B. J. Scholl. The automaticity of perceiving animacy: Goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. *Psychonomic Bulletin & Review*, pages 1–6, 2015. ISSN 1069-9384. doi: 10.3758/s13423-015-0966-5. URL http://link.springer.com/10.3758/s13423-015-0966-5. 3