Discriminating Animate from Inanimate Visual Stimuli

Brian Scassellati
MIT Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139
scaz@ai.mit.edu

Abstract

From as early as 6 months of age, human children distinguish between motion patterns generated by animate objects from patterns generated by moving inanimate objects, even when the only stimulus that the child observes is a single point of light moving against a blank background. The mechanisms by which the animate/inanimate distinction are made are unknown, but have been shown to rely only upon the spatial and temporal properties of the movement. In this paper, I present both a multi-agent architecture that performs this classification as well as detailed comparisons of the individual agent contributions against human baselines.

1 Introduction

One of the most basic visual skills is the ability to distinguish animate from inanimate objects. We can easily distinguish between the movement of a clock pendulum that swings back and forth on the wall from the movement of a mouse running back and forth across the floor. Michotte [1962] first documented that adults have a natural tendency to describe the movement of animate objects in terms of intent and desire, while the movements of inanimate objects are described in terms of the physical forces that act upon them and the physical laws that govern them. Furthermore, by using only single moving points of light on a blank background, Michotte showed that these perceptions can be guided by even simple visual motion without any additional context.

Leslie [1982] proposed that this distinction between animate and inanimate objects reflects a fundamental difference in how we reason about the causal properties of objects. According to Leslie, people effortlessly classify stimuli into three different categories based on the types of causal explanations that can be applied to those objects, and different modules in the brain have evolved to deal with each of these types of causation. Inanimate objects are described in terms of mechanical agency, that is, they can be explained by the rules of *mechanics*, and are processed by a special-purpose reasoning engine called the *Theory of Body* module (ToBY) which encapsulates the organism's intuitive knowledge about how objects move. This knowledge may not match the actual physical laws that govern the movement of objects, but rather is our intuitive understanding of physics. Animate objects are described either by their actions or by their attitudes, and are processed by the *Theory of Mind* module which has sometimes been called an "intuitive psychology." System 1 of the theory of mind module (ToMM-1) explains events in terms of the intent and goals of agents, that is, their *actions*. For example, if you see me approaching a glass of water you might assume that I want the water because I am thirsty. System 2 of the theory of mind module (ToMM-2) explains events in terms of the *attitudes* and beliefs of agents. If you see me approaching a glass of kerosene and lifting it to my lips, you might guess that I believe that the kerosene is actually water. Leslie further proposed that this sensitivity to the spatio-temporal properties of events is innate, but more recent work from Cohen and Amsel [1998] may show that it develops extremely rapidly in the first few months and is fully developed by 6-7 months.

Although many researchers have attempted to document the time course of the emergence of this skill, little effort has gone into identifying the mechanisms of how an adult or an infant performs this classification. This paper investigates a number of simple visual strategies that attempt to perform the classification of animate from inanimate stimuli based only on spatio-temporal properties without additional context. These strategies have been implemented on a humanoid robot called Cog as part of an on-going effort to establish basic social skills and to provide mechanisms for social learning [Scassellati, 2000]. A set of basic visual feature detectors and a context-sensitive attention system (described in section 2) select a sequence of visual targets (see Figure 1). The visual targets in each frame are linked together temporally to form spatio-temporal trajectories (section 3). These trajectories are then processed by a multi-agent representation that mimics Leslie's ToBY module by attempting to describe trajectories in terms of naive physical laws (section 4). The results of the implemented system on real-world environments are introduced, and a comparison against human performance on describing identical data is discussed in section 5.

2 Visual Precursors

Cog's visual system has been designed to mimic aspects of an infant's visual system. Human infants show a preference for stimuli that exhibit certain low-level feature properties. For example, a four-month-old infant is more likely to look at a

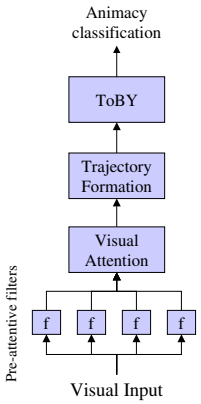

Figure 1: Overall architecture for distinguishing animate from inanimate stimuli. Visual input is processed by a set of simple feature detectors, each of which contributes to a visual attention process. Salient objects in each frame are linked together to form spatio-temporal trajectories, which are then classified by the "theory of body" (ToBY) module.

moving object than a static one, or a face-like object than one that has similar, but jumbled, features [Fagan, 1976]. Cog's perceptual system combines many low-level feature detectors that are ecologically relevant to an infant. Three of these features are used in this work: color saliency analysis, motion detection, and skin color detection. These low-level features are then filtered through an attentional mechanism which determines the most salient objects in each camera frame.

2.1 Pre-attentive visual routines

The color saturation filter is computed using an opponent-process model that identifies saturated areas of red, green, blue, and yellow [Itti *et al.*, 1998]. The color channels of the incoming video stream (r, g, and b) are normalized by the luminance l and transformed into four color-opponency channels (r', g', b', and y'):

$$r' = \frac{r}{l} - (\frac{g}{l} + \frac{b}{l})/2 \quad (1)$$

$$g' = \frac{g}{l} - (\frac{r}{l} + \frac{b}{l})/2 \quad (2)$$

$$b' = \frac{b}{l} - (\frac{r}{l} + \frac{g}{l})/2 \quad (3)$$

$$y' = \frac{\frac{r}{l} + \frac{g}{l}}{2} - \frac{b}{l} - \|\frac{r}{l} - \frac{g}{l}\| \quad (4)$$

The four opponent-color channels are thresholded and smoothed to produce the output color saliency feature map.

In parallel with the color saliency computations, The motion detection module uses temporal differencing and region growing to obtain bounding boxes of moving objects. The incoming image is converted to grayscale and placed into a ring of frame buffers. A raw motion map is computed by passing the absolute difference between consecutive images through a threshold function $\mathcal{T}$:

$$M_{raw} = \mathcal{T}(\|I_t - I_{t-1}\|) \quad (5)$$

This raw motion map is then smoothed to minimize point noise sources.

The third pre-attentive feature detector identifies regions that have color values that are within the range of skin tones [Breazeal *et al.*, 2000]. Incoming images are first filtered by a mask that identifies candidate areas as those that satisfy the following criteria on the red, green, and blue pixel components:

$$2g > r > 1.1g \qquad 2b > r > 0.9b \qquad 250 > r > 20 \quad (6)$$

The final weighting of each region is determined by a learned classification function that was trained on hand-classified image regions. The output is again median filtered with a small support area to minimize noise.

2.2 Visual attention

Low-level perceptual inputs are combined with high-level influences from motivations and habituation effects by the attention system. This system is based upon models of adult human visual search and attention [Wolfe, 1994], and has been reported previously [Breazeal and Scassellati, 1999]. The attention process constructs a linear combination of the input feature detectors and a time-decayed Gaussian field which represents habituation effects. High areas of activation in this composite generate a saccade to that location and compensatory neck movement. The weights of the feature detectors can be influenced by the motivational and emotional state of the robot to preferentially bias certain stimuli. For example, if the robot is searching for a playmate, the weight of the skin detector can be increased to cause the robot to show a preference for attending to faces. The output of the attention system is a labeled set of targets for each camera frame that indicate the positions (and feature properties) of the k most salient targets. For the experiments presented here, $k = 5$.

3 Computing Motion Trajectories

The attention system indicates the most salient objects at each time step, but does not give any indication of the temporal properties of those objects. Trajectories are formed using the multiple hypothesis tracking algorithm proposed by Reid [1979] and implemented by Cox and Hingorani [1996]. The centroids of the attention targets form a stream of target locations $\{P_t^1, P_t^2, ...P_t^k\}$ with a maximum of k targets present in each frame t. The objective is to produce a labeled trajectory which consists of a set of points, at most one from each frame, which identify a single object in the world as it moves through the field of view:

$$T = \{P_1^{i_1}, P_2^{i_2}, ...P_t^{i_n}\} \quad (7)$$

However, because the existence of a target from one frame to the next is uncertain, we must introduce a mechanism to compensate for objects that enter and leave the field of view and to compensate for irregularities in the earlier processing modules. To address these problems, we introduce phantom points that have undefined locations within the image plane but which can be used to complete trajectories for objects that enter, exit, or are occluded within the visual field. As each new point is introduced, a set of hypotheses linking that point

Figure 2: The last frame of a 30 frame sequence with five trajectories identified. Four nearly stationary trajectories were found (one on the person's head, one on the person's hand, one on the couch in the background, and one on the door in the background). The final trajectory resulted from the chair being pushed across the floor.

to prior trajectories are generated. These hypotheses include representations for false alarms, non-detection events, extensions of prior trajectories, and beginnings of new trajectories. The set of all hypotheses is pruned at each time step based on statistical models of the system noise levels and based on the similarity between detected targets. This similarity measurement is based on similarities of object features such as color content, size, and visual moments. At any point, the system maintains a small set of overlapping hypotheses so that future data may be used to disambiguate the scene. Of course, at any time step, the system can also produce the set of non-overlapping hypotheses that are statistically most likely. Figure 2 shows the last frame of a 30 frame sequence in which a chair was pushed across the floor and the five trajectories that were located.

4 The Theory of Body Module

To implement the variety of naive physical laws encompassed by the Theory of Body module, a simple agent-based approach was chosen. Each agent represents knowledge of a single theory about the behavior of inanimate physical objects. For every trajectory t, each agent a computes both an animacy vote α_{ta} and a certainty ρ_{ta}. The animacy votes range from $+1$ (indicating animacy) to -1 (indicating inanimacy), and the certainties range from 1 to 0. For these initial tests, five agents were constructed: an insufficient data agent, a static object agent, a straight line agent, an acceleration sign change agent, and an energy agent. These agents were chosen to handle simple, common motion trajectories observed in natural environments, and do not represent a complete set. Most notably missing is an agent to represent collisions, both elastic and inelastic.

At each time step, all current trajectories receive a current animacy vote V_t. Three different voting algorithms were tested to produce the final vote V_t for each trajectory t. The first voting method was a simple winner-take-all vote in which the winner was declared to be the agent with the greatest absolute value of the product: $V_t = \max_a \|\alpha_{ta} \times \rho_{ta}\|$ The second method was an average of all of the individual vote products: $V_t = \frac{1}{A}\sum_a(\alpha_{ta} \times \rho_{ta})$ where A is the number of agents voting. The third method was a weighted average of the products of the certainties and the animacy votes: $V_t = \frac{1}{A}\sum_a(w_a \times \alpha_{ta} \times \rho_{ta})$ where w_a is the weight for agent a. Weights were empirically chosen to maximize performance under normal, multi-object conditions in natural environments and were kept constant through out this experiment as 1.0 for all agents except the static object agent which had a weight of 2.0. The animacy vote at each time step is averaged with a time-decaying weight function to produce a sustained animacy measurement.

4.1 Insufficient Data Agent

The purpose of the insufficient data agent is to quickly eliminate trajectories that contain too few data points to properly compute statistical information against the noise background. Any trajectory with fewer than one-twentieth the maximum trajectory length or fewer than three data points is given an animacy vote $\alpha = 0.0$ with a certainty value of 1.0. In practice, maximum trajectory lengths of 60-120 were used (corresponding to trajectories spanning 2-4 seconds), so any trajectory of fewer than 3-6 data points was rejected.

4.2 Static Object Agent

Because the attention system still generates target points for objects that are stationary, there must be an agent that can classify objects that are not moving as inanimate. The static object agent rejects any trajectory that has an accumulated translation below a threshold value as inanimate. The certainty of the measurement is inversely proportional to the translated distance and is proportional to the length of the trajectory.

4.3 Straight Line Agent

The straight line agent looks for constant, sustained velocities. This agent computes the deviations of the velocity profile from the average velocity vector. If the sum of these deviations fall below a threshold, as would result from a straight linear movement, then the agent casts a vote for inanimacy. Below this threshold, the certainty is inversely proportional to the sum of the deviations. If the sum of the deviations is above a secondary threshold, indicating a trajectory with high curvature or multiple curvature changes, then the agent casts a vote for animacy. Above this threshold, the certainty is proportional to the sum of the deviations.

4.4 Acceleration Sign Change Agent

One proposal for finding animacy is to look for changes in the sign of the acceleration. According to this proposal, anything that can alter the direction of its acceleration must be operating under its own power (excluding contact with other objects). The acceleration sign change agent looks for zero-crossings in the acceleration profile of a trajectory. Anything with more than one zero-crossing is given an animacy vote with a certainty proportional to the number of zero crossings.

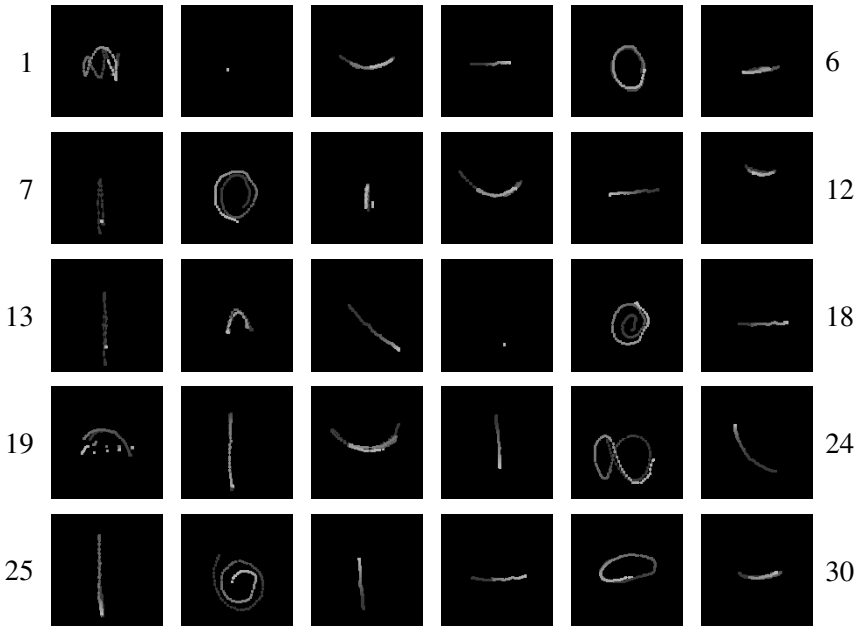

Figure 3: Thirty stimuli used in the evaluation of ToBY. Stimuli were collected by recording the position of the most salient object detected by the attention system when the robot observed natural scenes similar to the one shown in Figure 2. Each image shown here is the collapsed sequence of video frames, with more recent points being brighter than older points. Human subjects saw only a single bright point in each frame of the video sequence.

4.5 Energy Agent

Bingham, Schmidt, and Rosenblum [1995] have proposed that human adults judge animacy based on models of potential and kinetic energy. To explore their hypothesis, a simple energy model agent was implemented. The energy model agent judges an object that gains energy to be animate. The energy model computes the total energy of the system E based on a simple model of kinetic and potential energies:

$$E = \frac{1}{2}mv_y^2 + mgy \tag{8}$$

where m is the mass of the object, v_y the vertical velocity, g the gravity constant, and y the vertical position in the image. In practice, since the mass is a constant scale factor, it is not included in the calculations. This simple model assumes that an object higher in the image is further from the ground, and thus has more potential energy. The vertical distance and velocity are measured using the gravity vector from a three-axis inertial system as a guideline, allowing the robot to determine "up" even when its head is tilted. The certainty of the vote is proportional to the measured changes in energy.

5 Comparing ToBY's Performance to Human Performance

The performance of the individual agents was evaluated both on dynamic, real-world scenes at interactive rates and on more carefully controlled recorded video sequences.

For interactive video tasks, at each time step five attention targets were produced. Trajectories were allowed to grow to a length of sixty frames, but additional information on the long-term animacy scores for continuous trajectories were maintained as described in section 4. All three voting methods were tested. The winner-take-all and the weighted average voting methods produced extremely similar results, and eventually the winner-take-all strategy was employed for simplicity. The parameters of the ToBY module were tuned to match human judgments on long sequences of simple data structures (such as were produced by static objects or people moving back and forth throughout the room).

5.1 Motion Trajectory Stimuli

To further evaluate the individual ToBY agents on controlled data sequences, video from the robot's cameras were recorded and processed by the attention system to produce only a single salient object in each frame.[1] To remove all potential contextual cues, a new video sequence was created containing only a single moving dot representing the path taken by that

[1] This restriction on the number of targets was imposed following pilot experiments using multiple targets. Human subjects found the multiple target displays more difficult to observe and comprehend. Because each agent currently treats each trajectory independently, this restriction should not bias the comparison.

object set against a black background, which in essence is the only data available to the ToBY system. Thirty video segments of approximately 120 frames each were collected (see Figure 3). These trajectories included static objects (e.g. #2), swinging pendula (e.g. #3), objects that were thrown into the air (e.g. #7), as well as more complicated trajectories (e.g. #1). Figure 4 shows the trajectories grouped according to the category of movement, and can be matched to Figure 3 using the stimulus number in the second column. The third column of figure 4 shows whether or not the stimulus was animate or inanimate.

5.2 Human Animacy Judgments

Thirty-two adult, volunteer subjects were recruited for this study. Subjects ranged in age from 18 to 50, and included 14 women and 18 men. Subjects participated in a web-based questionnaire and were informed that they would be seeing video sequences containing only a single moving dot, and that this dot represented the movement of a real object. They were asked to rank each of the thirty trajectories shown in figure 3 on a scale of 1 (animate) to 10 (inanimate). Following initial pilot subjects (not included in this data), subjects were reminded that inanimate objects might still move (such as a boulder rolling down a hill) but should still be treated as inanimate. Subjects were allowed to review each video sequence as often as they liked, and no time limit was used.

The task facing subjects was inherently under-constrained, and the animacy judgments showed high variance (a typical variance for a single stimulus across all subjects was 2.15). Subjects tended to find multiple interpretations for a single stimulus, and there was never a case when all subjects agreed on the animacy/inanimacy of a trajectory. To simplify the analysis, and to remove some of the inter-subject variability, each response was re-coded from the 1-10 scale to a single animate (1-5) or inanimate (6-10) judgment. Subjects made an average of approximately 8 decisions that disagreed with the ground truth values. This overall performance measurement of 73% correct implies that the task is difficult, but not impossible. Column 4 of figure 4 shows the percentage of subjects who considered each stimulus to be animate. In two cases (stimuli #13 and #9), the majority of human subjects disagreed with the ground truth values. Stimulus #9 showed a dot moving alternately up and down, repeating a cycle approximately every 300 msec. Subjects reported seeing this movement as "too regular to be animate." Stimulus #13 may have been confusing to subjects in that it contained an inanimate trajectory (a ball being thrown and falling) that was obviously caused by an animate (but unseen) force.

5.3 ToBY Animacy Judgments

The identical video sequences shown to the human subjects were processed by the trajectory formation system and the ToBY system. Trajectory lengths were allowed to grow to 120 frames to take advantage of all of the information available in each short video clip. A winner-take-all selection method was imposed on the ToBY agents to simplify the reporting of the results, but subsequent processing with both other voting methods produced identical results. The final animacy judgment was determined to by the winning agent on the final time step. Columns 6 and 5 of figure 4 show the winning agent and that agent's animacy vote respectively.

Overall, ToBY agreed with the ground truth values on 23 of the 30 stimuli, and with the majority of human subjects on 21 of the 30 stimuli. On the static object categories, the circular movement stimuli, and the straight line movement stimuli, ToBY matched the ground truth values perfectly. This system also completely failed on all stimuli that had natural pendulum-like movements. While our original predictions indicated that the energy agent should be capable of dealing with this class of stimuli, human subjects seemed to be responding more to the repetitive nature of the stimulus rather than the transfer between kinetic and potential energy. ToBY also failed on one of the thrown objects (stimulus #20), which paused when it reached its apex, and on one other object (stimulus #19) which had a failure in the trajectory construction phase.

6 Conclusion

The distinction between animate and inanimate is a fundamental classification that humans as young as 6 months readily perform. Based on observations that humans can perform these judgments based purely on spatio-temporal signatures, this paper presented an implementation of a few simple naive rules for identifying animate objects. Using only the impoverished stimuli from the attentional system, and without any additional context, adults were quite capable of classifying animate and inanimate stimuli. While the set of agents explored in this paper is certainly insufficient to capture all classes of stimuli, as the pendulum example illustrates, these five simple rules are sufficient to explain a relatively broad class of motion profiles. These simple algorithms (like the agents presented here) may provide a quick first step, but do not begin to make the same kinds of contextual judgments that humans use.

In the future, we intend on extending this analysis to include comparisons against human performance for multi-target stimuli and for more complex object interactions including elastic and inelastic collisions.

Acknowledgments

Portions of this research were funded by DARPA/ITO under contract number DABT 63-99-1-0012, "Natural tasking of robots based on human interaction cues."

References

[Bingham *et al.*, 1995] Geoffrey P. Bingham, Richard C. Schmidt, and Lawrence D. Rosenblum. Dynamics and the orientation of kinematic forms in visual event recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 21(6):1473–1493, 1995.

[Breazeal and Scassellati, 1999] Cynthia Breazeal and Brian Scassellati. A context-dependent attention system for a social robot. In *1999 International Joint Conference on Artificial Intelligence*, 1999.

Stimulus Category	Stimulus Number	Ground Truth	Human Judgment	ToBY Judgment	ToBY Expert	Notes
Static Objects	2	Inanimate	3%	Inanimate	Static Object	
	16	Inanimate	6%	Inanimate	Static Object	
Thrown Objects	7	Inanimate	44%	Inanimate	Energy	
	13	Inanimate	**53%**	Inanimate	Energy	
	20	Animate	78%	**Inanimate**	Straight Line	Pause at apex
	25	Animate	81%	Animate	Energy	Velocity increases near apex
Circular Movements	5	Animate	59%	Animate	Energy	
	8	Animate	81%	Animate	Energy	
	17	Animate	81%	Animate	Straight Line	
	26	Animate	78%	Animate	Acc. Sign Change	
	29	Animate	56%	Animate	Energy	
Straight Line Movements	4	Inanimate	47%	Inanimate	Straight Line	
	11	Inanimate	36%	Inanimate	Straight Line	
	22	Inanimate	30%	Inanimate	Straight Line	
	27	Animate	53%	Animate	Energy	moving up
	15	Inanimate	37%	Inanimate	Straight Line	
	24	Animate	75%	Animate	Energy	moving up and left
Pendula	3	Inanimate	16%	**Animate**	Energy	
	10	Inanimate	12%	**Animate**	Acc. Sign Change	
	21	Inanimate	31%	**Animate**	Acc. Sign Change	
	30	Inanimate	19%	**Animate**	Acc. Sign Change	
	12	Inanimate	6%	**Animate**	Acc. Sign Change	
Erratic Movements	1	Animate	97%	Animate	Energy	random movements
	6	Animate	75%	Animate	Acc. Sign Change	left/right bouncing
	9	Animate	**31%**	Animate	Acc. Sign Change	up/down bouncing
	14	Animate	75%	Animate	Acc. Sign Change	repeated left/right hops
	18	Animate	87%	Animate	Straight Line	delay at center point
	19	Animate	93%	**Inanimate**	Little Data	failure to track
	23	Animate	81%	Animate	Energy	figure-8
	28	Animate	90%	Animate	Straight Line	delay at center point

Figure 4: Comparison of human animacy judgments with judgments produced by ToBY for each of the stimuli from figure 3. Column 3 is the ground truth, that is, whether the trajectory actually came from an animate or inanimate source. Column 4 shows the percentage of human subjects who considered the stimulus to be animate. Column 5 shows the animacy judgment of ToBY, and column 6 shows the agent that contributed that decision. Bold items in the human or ToBY judgment columns indicate a disagreement with the ground truth.

[Breazeal *et al.*, 2000] Cynthia Breazeal, Aaron Edsinger, Paul Fitzpatrick, Brian Scassellati, and Paulina Varchavskaia. Social constraints on animate vision. *IEEE Intelligent Systems*, July/August 2000. To appear.

[Cohen and Amsel, 1998] Leslie B. Cohen and Geoffrey Amsel. Precursors to infants' perception of the causality of a simple event. *Infant Behavior and Develoment*, 21(4):713–732, 1998.

[Cox and Hingorani, 1996] Ingemar J. Cox and Sunita L. Hingorani. An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 18(2):138–150, February 1996.

[Fagan, 1976] J. F. Fagan. Infants' recognition of invariant features of faces. *Child Development*, 47:627–638, 1976.

[Itti *et al.*, 1998] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[Leslie, 1982] Alan M. Leslie. The perception of causality in infants. *Perception*, 11:173–186, 1982.

[Michotte, 1962] A. Michotte. *The perception of causality*. Methuen, Andover, MA, 1962.

[Reid, 1979] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automated Control*, AC-24(6):843–854, December 1979.

[Scassellati, 2000] Brian Scassellati. Theory of mind for a humanoid robot. In *Proceedings of the First International IEEE/RSJ Conference on Humanoid Robotics*, 2000.

[Wolfe, 1994] Jeremy M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.