Which Motion Features Induce the Perception of Animacy?

Viksit Gaur
Department of Computer Science, Yale University
51 Prospect St., New Haven, CT 06511, USA
viksit@cs.yale.edu

Brian Scassellati
Department of Computer Science, Yale University
51 Prospect St., New Haven, CT 06511, USA
scaz@cs.yale.edu

***Abstract*— The perception of animacy is one of the fundamental social skills possessed by humans. The aim of this study is to determine whether such a skill can be accurately reproduced computationally, and if so, to describe the underlying factors which contribute to the animacy decision. We hypothesize a set of motion features which lead to such a distinction, construct a database of sample movements from both synthetic and natural stimuli, collect human judgments on the animacy of these stimuli, and evaluate the effectiveness of a computational system trained on this data to discriminate animacy.**

***Index Terms*— Animacy, Perceptual development, Causality, Naive physics**

I. Introduction

Objects around us may be classified as either having a controlling intelligence behind them or those dependent on purely physical circumstances. We can refer to these objects as animate or inanimate, although the distinction is not a clear one. What information do we need about an object, in order to classify it as animate? There are many artifacts which remind us of animacy, foremost of them being live biological motion. However, there are often moments where we classify abstract or non-living objects as animate, for example a complicated object trajectory generated by a mathematical equation, or a battery powered toy. The nuances that discriminate animate from inanimate motion have not been well described. In this work, we construct a system that performs this discrimination and we analyze the components of visual motion that contribute to this decision.

The phenomenon behind animate and inanimate motion was first studied in the 1900s, and shot to attention with the publication of Michotte's book, "The Perception of Causality" [2]. This was followed by Heider and Simmel's [3] classic 1944 article on "An experimental study of apparent behavior" which described the attribution of animacy and causality as having the characteristics of a perceptual judgment (being "fairly fast, automatic, irresistible and highly stimulus-driven") even though they were typically thought of as cognitive processes [4]. This suggests that the visual system infers properties like causality and animacy in the same way it does for physical properties like object motion. There is some evidence that other animals (including frogs [5]) operate in this same manner.

The efforts of many to determine the motion cues which mediate perceptual animacy have met with reasonable success. However, no absolute metric has been found or developed yet. Bassilli [6] developed 5 computer controlled displays, each showing 2 circles moving on a black background for his experiments. Dittrich and Lea [7] followed by presenting adult subjects with displays containing several randomly chosen, moving letters (called distractors), and a target letter whose motion was designed to simulate biologically meaningful and intentional motion - a predator chasing a prey, for example. They concluded that the perception of any motion as animate depended not only on the degree of interaction between the objects, but also on the degree of intentionality conveyed by it.

Stewart [8] set out to investigate how motion influences the perception of animacy. She hypothesized that observers should describe objects which violate the laws of physics as animate, since it would require them to have access to hidden energy sources. Stewart used this energy-violation hypothesis to predict that three types of motion were sufficient to identify animacy: starts from rest, changes in direction to avoid a collision and direct movement towards a goal. Gellman et al [9] suggested that the ability to classify animacy is not based solely on perceptual information, but also draws upon innate or early-developing knowledge of causal principles in humans. Blythe and his colleagues [10] on the other hand, have argued that a small set of motion cues can be sufficient not only to determine whether or not a moving object is animate, but also to determine what intention motivated the object's movement. They go on to present an algorithm which uses seven motion cues to predict observers' responses when asked to classify the motion into categories. Tremoulet and Feldman [1] however hypothesized that animacy could be perceived from point light movements, and formulated their experiments with extremely simple stimuli - showing a single white particle moving across a dark, featureless background. In cases where the trajectories of two objects were the same, the orientation and alignment of the object itself contributed to the perception of animacy, thus refuting Stewart's earlier claim of energy violations being the only contributing factor to such a decision.

Often, discussions of animacy occur within a larger conceptual framework that accounts for intentional and goal-directed behavior. These frameworks are often known as ”theory of mind” models because they require that the agent have some ability to interpret the hidden mental states (of goal and intention) of another entity. Based on the models of Leslie [13] and Baron-Cohen [20], Scassellati [11], [14] developed and implemented the basic components of a theory of mind for a humanoid robot. This implementation used Leslie's model of a "Theory of Body" (ToBY) to determine animacy of a visual target based on a set of naive physical laws. Objects which obeyed these hand-constructed physical laws were determined to be inanimate, while those that displayed some measure of self-propelled movement were seen as animate. This work serves as the basis for the work presented here. We propose a number of different visual cues to describe a particular motion, each of which is extracted from a given trajectory. The cues are based on physical properties such as the distance traveled by an object during its motion, the direction in which it did so, the energy it gained or lost during this time, and other such factors. Experiments with a model built using these features then allow us to conclude the importance of each of them in arriving at a judgment of whether a given motion is animate or not.

II. System Overview

It has been shown that a single moving object can create the subjective impression of being animate, based solely on its pattern of movement [1]. Some authors have also hypothesized that the visual system automatically detects events in which the observable kinetic energy increases [15], or is otherwise not conserved [8] - which might imply the presence of an animate entity with hidden energy sources. Our goal was to not only construct a system which could distinguish between animate and inanimate motions but also to analyze the relative contribution of various features of a physical movement that would contribute to a judgment of animacy. Our system consists of three main components: a visual preprocessor which converts visual data to the movement of an object to a single trajectory through space and time, a feature extractor which converts a motion trajectory into a lower-dimensional set of features, and a classification system that can be trained to discriminate animate from inanimate trajectories.

A. The Visual Pre-Processor

The visual pre-processor converts a sequence of visual images into a form suitable to submit to the next component, the feature extractor. The training data we consider consists of videos of point light sources generated using one of three methods (as shown in *Figure 1*). The first method (”hand-drawn”) extracted motion trajectories from hand-drawn trajectories obtained by recording the mouse movements of a human user engaged in a drawing task. The moving trajectory of the mouse pointer was used to generate a synthetic image that showed the position of the mouse pointer as a white dot moving over a black screen. In the second method (”real world”), an object with a small light source attached to it was moved through a motion trajectory in front of a camera system. The recorded video of this moving light source was used directly as data. The final method (”synthetic”), constructed video sequences synthetically by tracing a white point over a black background using a trajectory defined by hand-coded equations of motion.

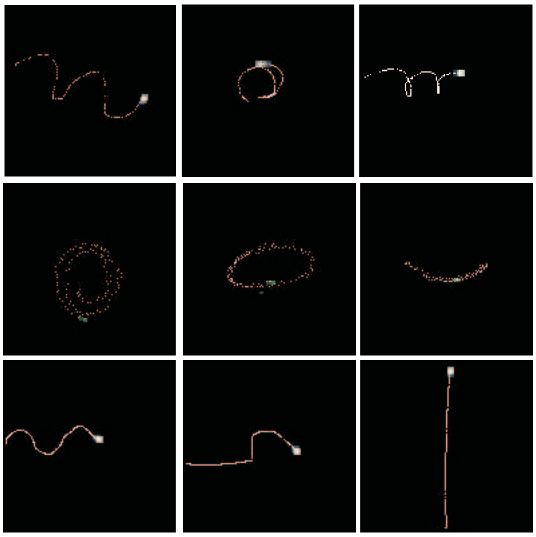

Fig. 1. Three types of test video sequences. Top row - ”Hand drawn” images; Middle row - ”Real world” images; and Bottom row - ”synthetic” images

B. The Feature extractor

Because the videos obtained earlier may be of arbitrary length, the first 120 frames of each trajectory is broken up into thirty subdivisions (breaks) each consisting of four spatial positions. The points thus obtained may however have noise, and contain jitter. The four points within each subdivision are averaged, and a piecewise polynomial cubic spline curve is fitted to these average coordinate points using a least squares approximation. *Figure 2* shows the generated spline curve of the object in motion in *Figure 1*. The smoothed data can now be used to calculate motion vectors, each represented by an *angle of movement* θ and ρ, the *distance between breaks*. Taken together, these two parameters constitute a set of motion vectors that describes the complete trajectory. These basic parameters are then used to derive features based on the principles of naive physics.

1) Velocity and Acceleration metrics: Tremoulet and Feldman [1] hypothesized that simultaneous changes in both speed and motion direction - occurring in a uniform, featureless environment - lead to animate interpretations. In the absence of any supporting context, these trajectories cannot normally be accounted for by inanimate motion sources common in the environment. One of the properties in our classification system can thus be through monitoring changes

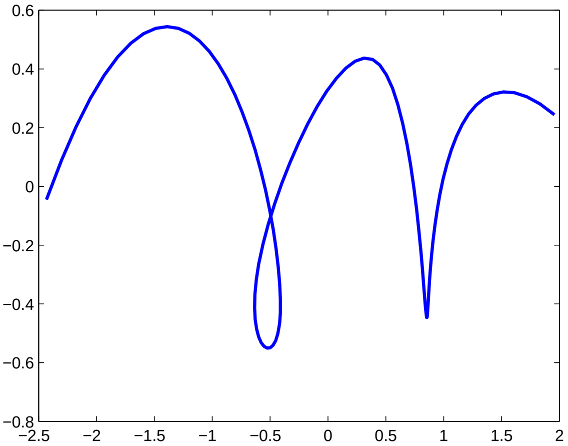

Fig. 2. The spline curve generated for the third hand-generated motion

in the direction, velocity and acceleration of the object. We calculate the mean, variance and range for both θ and ρ across the entire trajectory. The mean of distance ρ relates to the average velocity of the object, and its variance to the acceleration or deceleration endured. Directional information can be obtained from the angle θ, whose mean and variance give us directional changes.

2) Static metric: Certain objects can be easily be classified as inanimate, since they have a high chance of not undergoing any movement at all. This measure is proportional to the length of the trajectory of motion - the probability of it being inanimate rises higher as length of motion becomes shorter. From the above data, we have already obtained the distance metric as given by ρ. If this is near zero, the static feature flag is set to true.

3) Straight line metric: Although it is not a strict requirement, most animate movements follow an element of uncertainty, and the chances of them being geometrically perfect are minimal. We can thus include a measure of how closely the motion of an object approximates a straight line or a smooth curve by analyzing point locations with respect to time, and comparisons with the relevant trajectory equations. To calculate this, we formulate the equation of a straight line from the first two points available to us,

$$(y - y1) = ((y2 - y1)(x - x1))/(x2 - x1) \tag{1}$$

For successive values of (x, y) from the array, we substitute (x, y) and make sure that the equation is satisfied - which will only occur if the object moves in a straight line. We also check if the value of the initial angle θ of the object, and the range of angles it has moved through are close to each other, suggesting very slight deviation from a straight line. This method proves to be more efficient, since no object as analyzed in the real world moves exactly in a straight line.

4) Energy metric: We implement a simple energy monitoring system based on [15], which judges an object that gains energy to be animate. The total energy is calculated based on a simple model of kinetic and potential energies -

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

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. An object higher up in the image is assumed to have more potential energy. This is possible only if the object has some means of increasing its own energy - definitely the hallmark of an animate object. To determine a metric, we analyze the energy over the entire motion of the object (assuming that the mass of the object is a constant). We calculate if there has been any gain in the potential energy, since this would signify a force other than natural ones acting on the object.

Final attribute set	
Distance Metric (ρ)	Mean (ρ_V), variance (ρ_V) and range (ρ_R) sampled on each of the thirty trajectory breaks (3 features)
Direction Metric (θ)	Mean (θ_M), variance (θ_V), range(θ_R) sampled as above (3 features)
Spline coeff.	$1st$, $2nd$, $3rd$ and $4th$ order coefficients measure the sharpness of the mean, variance and range of the motion vectors (12 features)
Static Flag $\epsilon(0, 1)$	Static or not, depending on magnitude of distance travelled (1 feature)
Straight line Flag $\epsilon(0, 1)$	Straight line or not, depending on change in angle of motion (1 feature)
Energy agent (E)	Measures any increase in potential energy of the object (1 feature)

C. The Classification System

The underlying basis of the classification system is a naive Bayes classification algorithm. A repository of classification data about the motions was first generated after letting volunteers interact with the system. Their responses on whether a particular motion is animate or not, were integrated along with the extracted features, into a combined dataset which was used to train the classifier.

The process of classifying a movement as animate or inanimate is something inherently built into human beings, and there are no definitive means of quantifying it (yet). We thus need to consider the fact that none of the distinctions our system makes fall neatly into either of the two groups - there is bound to be an error in judgment at various points. Animacy is hence represented as a latent variable - one whose value can not be directly observed, but indirectly inferred from the values of other observable and measurable variables. In our scenario, the properties we extract from the image sequences play the role of the latter. One advantage of using latent variables is that it reduces the dimensionality of data. A large number of observable variables can be aggregated to represent an underlying concept, making it easier for human beings to understand and assimilate information - we can extrapolate this conclusion for machines.

The learning method we intend to follow is supervised learning [21] - that of creating a function from training data, whose aim will be to predict a class label of the input data, after having analysed a number of training examples. We choose a probabilistic classifier, specifically the naive Bayes classifier, also known as the independent feature model.

1) The Naive Bayes Classifier: We consider the function of properties (features) described earlier $(P_1, P_2, ..P_n)$, which describes the instance we wish to classify. The classifier can be thought of in a conditional model, where A is a dependent class variable with a small number of outcomes (a in A),

$$P(A|P_1, P_2....P_n) \tag{3}$$

where $a = 0$ if the motion is inanimate and $a = 1$ if the motion is animate. The above however, will not work for an arbitrarily large number of properties. Using Bayes theorem, we can thus get,

$$P(A|P_1, ...P_n) = \frac{P(A)P(P_1, ...P_n|A)}{P(P_1...P_n)} \tag{4}$$

The value of the denominator does not depend on A, and because the values of the properties once computed will remain constant, we can effectively ignore it. The numerator can be thought of as a joint probability model of the two terms,

$$P(A, P_1, ...P_n) = P(A)P(P_1|A)...P(P_n|A) \tag{5}$$

which is the same as,

$$P(A)\prod_{i=1}^{n} P(P_i|A) \tag{6}$$

The Bayes probability model can now be combined with a decision rule, in this case the *Maximum A Posteriori* (MAP) rule. The classification function for a particular instance to a class a, $classify(P_1, P_2...P_n)$ from a set of possible classes A thus becomes,

$$argmax_c P(A = a)\prod_{i=1}^{n} P(P_i = p_i|A = a). \tag{7}$$

We will also compute a vector of elements,

$$V_A = P(A)\prod_{i=1}^{n}(P_i|A) \tag{8}$$

After normalization, this vector denotes the class probabilities. (Both these and the conditional probabilities are estimated from training data). The class probability is the same as the relative class frequency. We can calculate the conditional probabilities by computing the number of instances which have their i_{th} attribute equal to P_i, and belong to A.

In this system however, we use the m-estimate mechanism of calculating probabilities due to the fact that the relative frequency method of calculation has problems when the instance size is small. Thus,

$$P = (n_c + mp)/(n + m) \tag{9}$$

where n is the total number of training examples, and n_c is the corresponding value when calculating the estimate. m is the equivalent sample size, which determines how heavily to weight p relative to the observed data. p is the prior estimate of probability we wish to estimate. We estimate by p by assuming uniform priors, and set $p = 1/k$.

III. Performance evaluation and testing

The classification system was built and tested with the Orange [18] framework. The dataset on which the testing was done consists of fifty videos each of which was classified by twelve human volunteers as animate or inanimate. For simplicity of analysis, the hand-drawn video trajectories and the equation-generated synthetic trajectories were lumped together. These videos, their extracted features, and the human-generated labels were taken as the training set. We started by testing the accuracy of the system using some common techniques. Analysis of the relation between the accuracy of the system in context of synthetic and real world motions was then done, followed by a knockout evaluation of the data to determine which one of the features was the most important in judging a motion as animate.

A. K-Fold Cross Validation results

The evaluation through k-fold cross validation method is a very common one in the machine learning community. The data set is here split into k equally sized subsets, and then in the i-th iteration, the $(i = 1..k)$ i_{th} subset is used for testing the classifier that has been built on all other remaining subsets. We see from Figure. 3 that as the number of datasets are increased, the accuracy of the system increases - the use of one training data set achieved 30% accuracy, which increased to 64.38% for seven samples, but decreased again for twelve samples, to 60.33%, which might highlight an aberration in the successive human perception of the test data.

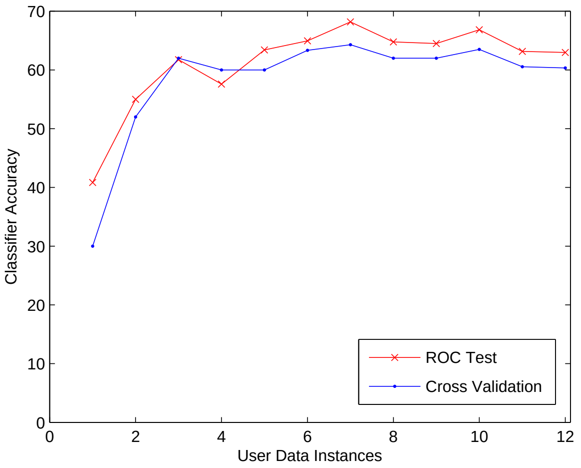

Fig. 3. Accuracy of the learning system

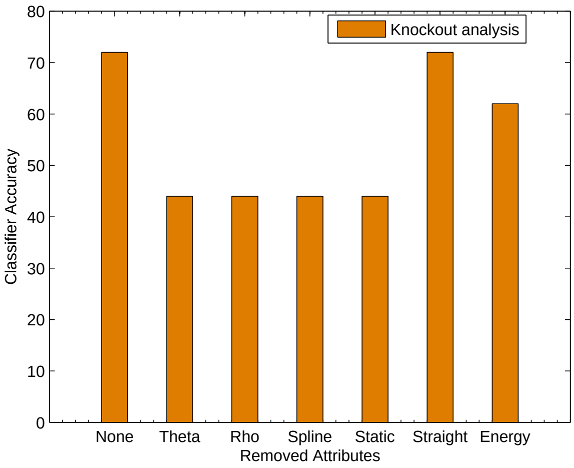

Fig. 4. Knockout analysis of the feature-set - Test I. Accuracy of the classifier when each of the mentioned features were removed from only the testing set

B. Area under ROC Curve results

Another method used is the area under ROC curve [19]. It is a discrimination measure that ranges from 0.5 (random guessing) to 1.0 (a clear margin exists in probability that divides the two classes). As we can see from Figure. 3, the results are quite consistent in terms of prediction of new data. Accuracy measures ranged from 40% in case of one data set, to about 68% for seven samples, and decreased again for twelve samples to about 63%. The discrimination measure remained at about 0.5, suggesting that there is no clear margin in the probability which divides the two classes - animate and inanimate, in accordance to our hypothesis.

C. Role of synthetic vs. real-world data

To determine if there are possible differences in interpretation of one type of movement compared to the other by humans, we analysed our system in two ways - by training it on the synthetic data set, and testing it on the real-world data, and vice versa. The accuracy of the system was found to be minimally different between these two conditions. In the first case, we obtain an accuracy value of 46.99%, and in the other, 48.7%. Compared to the accuracy levels in the region of 72% when the entire set is considered, this is significantly low, but it can be stated that there are no significant differences in their interpretation.

D. Knockout analysis of features

In order to answer our earlier question of determining which features contribute most to the animacy judgment, we ran knockout tests on the feature set. In the first test, a certain feature was nullified in the testing set, (but no changes made in the training set), and the system was run. The accompanying graph (Figure. 4) illusrates the results. The drop in accuracy to 44% remains constant when the velocity (ρ), direction (θ), staticness features, as well as the spline coefficients representing the sharpness of these changes are each nullified independently. There is no change in accuracy when the straight line vector is removed, which signifies its lesser contribution to the judgment process. The energy feature however, leads to a significant fall in accuracy, going down to 62%. When all features are retained, we get a high accuracy of 72%.

In the second test, both the training and testing sets were modified. For each step in the process, all attributes except one were consecutively removed, and the system was run. As is illustrated by the accompanying graph (Figure 5), the accuracy is a high 72% when all the attributes are retained. When testing only on direction (θ), we achieve an accuracy of 66%, which goes down to 58% when using only the velocity attribute set (ρ). On the other hand, considering only the sharpness factors for the above data leads to a higher accuracy in correctly classifying the motion, achieving 64%. The staticness metric, and the straight line metric perform on par, with average accuracy for both (taken independently) remaining 44%, signifying a lesser contribution to the judgment process. An interesting result of this was the relatively high accuracy obtained when retaining only the energy metric, which results in values near 68%.

We can thus conclude that the major features used in determining a particular motion as animate are a combination of the sharpness in change of velocity, and direction, with the energy metric playing a significant role in the process. The direction or velocity magnitude themselves contribute a much lesser amount to the final categorization of motion into either of the two classes. The staticness metric and the straight line metric are seen to contribute in a considerable,

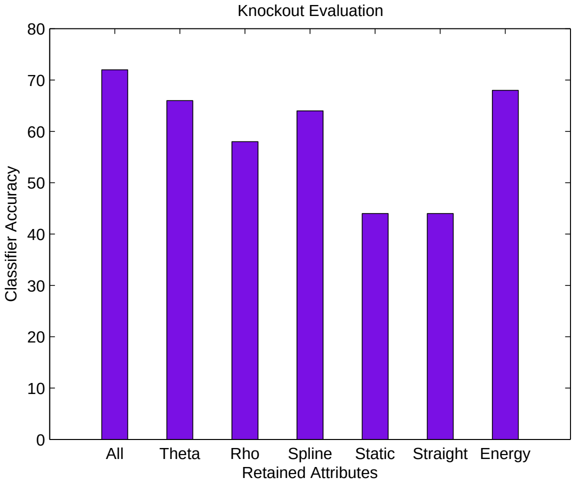

Fig. 5. Knockout analysis of the feature-set - Test II. Accuracy of the classifier when all but the mentioned attributes were removed. The major contributors to the judgment were theta, spline and the energy metrics

yet less substantial manner to the judgment.

IV. Discussion

Due to their innate complexity, a set of motions can not be discretely classified as being animate or inanimate. The fact that the 2 classes are not linearly separable is an obvious one (and substantiated by the AROC test) - mere physical representations of such motions give an insight into but one of the myriad features they possess. Humans have the ability to interpret complex features from such motion, the specifics of which are still unknown - it is definite however, that these relate to higher cognitive functions. Errors in judgment still abound, as is well illustrated by the samples of data which have been subjected to human analysis, and this fact highlights the blurred lines between animate and inanimate motion. After all, what would be the right classification for a movement in which a person falls down the stairs? If we were to approximate purely physical notions, then it could be argued that this is inanimate motion. However, when considering the randomness of movement, which is yet in sync with the various parts of the human body, this could be classified as animate. Our system of classification is a flexible one, able to adapt to changing data, rather than have extensive *a priori* knowledge about the system, and achieves our aim in developing a learning system which can learn through social interaction.

V. Acknowledgement

Support for this work was provided by a National Science Foundation CAREER award (#0238334). Some parts of the architecture used in this work was constructed under NSF grants #0205542 (ITR: A Framework for Rapid Development of Reliable Robotics Software) and #0209122 (ITR:Dance, a Programming Language for the Control of Humanoid Robots). This research was supported in part by a grant of computer software from QNX Software Systems Ltd. We would also like to thank all the participants of the experiments for their help.

References

[1] Tremoulet P.D. and Feldman J. *Perception of animacy from the motion of a single object*. 2000.

[2] Michotte A. *The Perception of Causality*. Basic Books (English Trans), 1962.

[3] Heider F. and Simmel M. *An experimental study of apparent behavior*. In Am. J. Psychol (57), pages 243 249, 1944.

[4] Scholl B. J. and Tremoulet P. D. *Perceptual causality and animacy*. In Trends in Cognitive Sciences 4(8), pages 299309, 2000.

[5] McCulloch W., Lettvin J., Maturana H. *What the frog's eye tells the brain*. The mind : Biological approaches to its functions, pages 233258, 1968.

[6] Bassilli J. *Temporal and spatial contingencies in the perception of social events*. In J. Pers. Soc. Psychol (33), pages 680685, 1976.

[7] Dittrich W. and Lea S. *Visual perception of intentional motion*. In Perception (23), pages 253268, 1994.

[8] Stewart J. A. *Perception of animacy*. Unpublished PhD Dissertation, Univ. of Pennsylvania.

[9] Gelman R. et al. *Distinguishing between animates and inanimates: not by motion alone*. In D. et al. Sperber, editor, In Causal Cognition: A Multidisciplinary Debate, pages 150184. Clarendon Press, 1995.

[10] Blythe P. et al. *How motion reveals intention: categorizing social interactions*. In Simple Heuristics That Make Us Smart, pages 257285. Oxford University Press, 1999.

[11] Scassellati B. *Foundations for a Theory of Mind for a Humanoid Robot*. PhD thesis, MIT EECS, 2001. PHd Thesis.

[12] Leslie A.M. *Pretense and representation: The origins of theory of mind*. Psychological Review, pages 412426, 1987.

[13] Baron-Cohen S. and Leslie A.M. *Does the autistic child have a theory of mind*. Cognition (Oct, 21(1)), pages 3746, 1985.

[14] Scassellati B. *Theory of mind for a humanoid robot*. Autonomous Robots, Vol 12, No 1, pages 1324, 2002.

[15] Rosenblum L.D., Bingham G.P. and Schmidt R.C. *Dynamics and the orientation of kinematic forms in visual event recognition*. Journal of Experimental Psychology, Dec 21(6), pages 1472 1493, 1995.

[16] Scassellati B. *Discriminating animate from inanimate from visual stimuli*. Unpublished paper, 2000.

[17] Brooks R.A., Breazeal C., Marjanovic M. and Scassellati B. *Investigating models of social development using a humanoid robot*. Analogy and Agents, page 52, 2001. 22

[18] Leban G. Demsar J, Zupan B. *Orange: From experimental machine learning to interactive data mining*. Technical report, 2004. Faculty of Computer and Information Science, University of Ljubljana. (www.ailab.si/orange).

[19] Beck J.R. and Schultz E.K. *The use of roc curves in test performance evaluation*. In Archives of Pathology and Laboratory Medicine, pages 1320, 1986.

[20] S Baron-Cohen and AM Leslie. *Does the autistic child have a Theory of mind*. Cognition (Oct, 21(1)), pages 37-46.

[21] *The Wikipedia collaborative encyclopedia*. http://www.wikipedia.org