Neural model for the visual recognition of animacy and social interaction

Mohammad Hovaidi-Ardestani[1,2] Nitin Saini[1,2] , Aleix M. Martinez[3], and Martin A. Giese[1]

[1] Section of Computational Sensomotorics, Department of Cognitive Neurology, CIN and HIH, University Clinic Tübingen,
Ottfried-Müller-Str. 25, 72076 Tübingen, Germany
[2] IMPRS for Cognitive and Systems Neuroscience, Tübingen, Germany
[3] Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA
Email: martin.giese@uni-tuebingen.de

Abstract. Humans reliably attribute social interpretations and agency to highly impoverished stimuli, such as interacting geometrical shapes. While it has been proposed that this capability is based on high-level cognitive processes, such as probabilistic reasoning, we demonstrate that it might be accounted for also by rather simple physiologically plausible neural mechanisms. Our model is a hierarchical neural network architecture with two pathways that analyze form and motion features. The highest hierarchy level contains neurons that have learned combinations of relative position-, motion-, and body-axis features. The model reproduces psychophysical results on the dependence of perceived animacy on motion smoothness and the orientation of the body axis. In addition, the model correctly classifies six categories of social interactions that have been frequently tested in the psychophysical literature. For the generation of training data we propose a novel algorithm that is derived from dynamic human navigation models, and which allows to generate arbitrary numbers of abstract social interaction stimuli by self-organization.

Keywords: hierarchy, neural network model, animacy, social interaction perception

1 Introduction

Humans spontaneously can decode animacy and social interactions from strongly impoverished stimuli. A classical study by Heider and Simmel [1] demonstrated that humans derived very consistently interpretations in terms of social interactions from simple geometrical figures that moved around in the two-dimensional plain. The figures were interpreted as living agents, to which even personality traits were attributed. More recent studies have characterized in more detail which critical features of simple stimuli affect the perception of animacy, that is whether the object is perceived as alive [2–4]. Furthermore, detailed studies have focused on the perception of social interactions between multiple moving

shapes, e.g. focusing on 'chasing' or 'fighting' [5, 6]. Six interaction types have been used in a number of studies [7–9], and McAleer & Pollick [9] showed that these categories can be reliably classified from stimuli showing moving circular disks whose movements were derived from real interactions.

Coarse neural substrates of the processing of such stimuli have been identified in fMRI studies. Animacy has been studied, modulating the movement parameters of individual moving shapes [10–12], and stimuli similar to the ones by Heider & Simmel have been frequently used in studies addressing Theory of Mind [13, 14]. In fMRI and monkey studies regions like the superior temporal sulcus (STS) and human area TPJ were found to be selective for these stimuli [15–18]. In spite of this localization of relevant cortical areas, the underlying exact neural circuits of this processing remain entirely unclear. Some theories have associated the processing of such abstract stimuli with probabilistic reasoning [19, 20], while others have linked them to lower-level visual processing [6]. So far no ideas exist how such functions could be accounted for by physiologically plausible neural circuits.

The goal of this paper is to present a simple neural model that reproduces some of the key observations in psychophysical experiments about the perception of animacy and social interactions from simple abstract stimuli. The model in its present form is simple, but in principle extendable for the processing of more complex stimuli that require also the processing of shape details or shapes in clutter. The model is an extension of classical models of the visual processing stream that account for the processing of object shape and actions [21–24]. However, such models never have been applied to account for the perception of animacy or social interaction. Our attempt to use these types of architectures is motivated by recent work that showed that models of this type for the recognition of hand actions also account for the perception of causality from simple stimulus displays that consist of moving disks [25]. This modeling work predicted also the existence of neurons in macaque cortex that are specifically involved in the visual perception of causality [26]. Here we show that a model based on similar principles accounts for the perception of animacy and social interactions.

In the following section, we first describe how we generated a stimulus set for training of the neural model, devising a generative model for social interaction stimuli that is based on a dynamical systems approach. We then describe the architecture of the model. The following section describes the results, followed by a brief discussion.

2 Stimulus synthesis

For the training of neural network models a sufficient set of stimuli is required. The problem is that from the classical psychophysical studies only a rather small set of stimuli is publicly available. For a meaningful application of learning-based neural networks approaches thus a sufficiently large training data set with similar properties needs to be generated. In our study we used movies showing individual

moving agents, and interaction of 2 agents (chasing, playing, following, flirting, guarding, fighting) described in psychophisical studies [7–9].

In order to model the interaction of two moving agents we exploited a dynamical systems approach, which before was used very successfully for the modeling of human navigation [27]. The underlying idea, originally derived from robotics [28], is to define a dynamical systems or differential equations for the heading directions ϕ_i and the instantaneous propagation speeds v_i of the interacting agents (in our case $i = 1, 2$). The specified movement is dependent on goal and obstacle points in the two dimensional plain, where the other agent can also act as goal or obstacle as well. We modified a model for human steering behavior during walking [29] to reproduce the movements during social interactions.

The resulting dynamics is given by the following differential equations for the heading direction:

$$\begin{aligned}\ddot{\phi}_i = &-b\dot{\phi}_i - k_g(\phi_i - \psi_{\mathrm{g},i})(e^{-c_1 d_{\mathrm{g},i}} + c_2) \\ &+k_o \sum_{n=1}^{N_{\mathrm{obst}}} (\phi_i - \psi_{\mathrm{o},ni})(e^{-c_3|\phi_i - \psi_{\mathrm{o},ni}|})(e^{-c_4 d_{\mathrm{o},ni}}).\end{aligned} \tag{1}$$

The variables $\psi_{\mathrm{g},i}$ and $d_{\mathrm{g},i}$ signify the absolute direction of the actual goal point and the distance of the goal from the agent in the 2D plain. Likewise, $\psi_{\mathrm{o},ni}$ and $d_{\mathrm{o},ni}$ signify the absolute direction and distance from obstacle number n from the agent, where N_{obst} is the number of relevant obstacles, and where k_m and c_m signify constants. The forward speed of the agents is specified by the two stochastic differential equations

$$\tau \dot{v}_i = -v_i + F_i(d_{\mathrm{g},i}) + k_\epsilon \epsilon_i(t), \tag{2}$$

where $\epsilon_i(t)$ is Gaussian white noise. The two functions F_i that specify the distance dependence of the speed dynamics are different for the two agents:

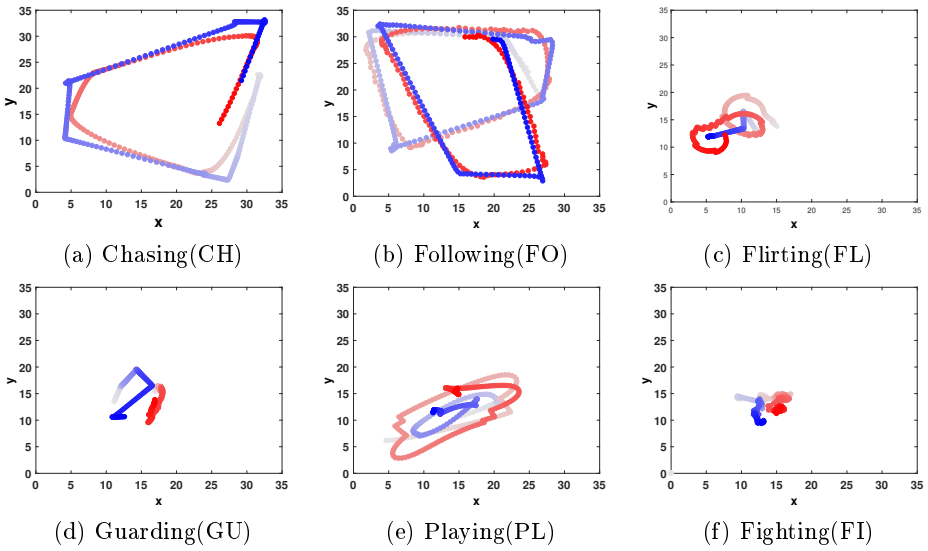
$$F_1(d) = \frac{1}{1 + e^{-c_5(d - c_6)}} - c_7 e^{-kd} \tag{3}$$

$$F_2(d) = \frac{c_8}{1 + e^{-c_9(d - c_{10})}} - c_{11} e^{-kd} + c_{12}. \tag{4}$$

Table 1: Parameters of simulation algorithm.

	Agent 1					Agent 2						
	k_ϵ	C_5	C_6	C_7	k	k_ϵ	C_8	C_9	C_{10}	C_{11}	C_{12}	k
Guarding (Gu)	0	1	5	0	0	0	1	1	3	0	0.5	0
Following (FO)	0	10	7	0	0	0	1	4	4	0	0	0
Fighting (FI)	1	1	3	1	0.1	1	1	1	3	1	0	0.1
Chasing (CH)	0	10	7	0	0	0	1	1	7	0	0	0
Flirting (FL)	0	1	5	0	0	1	0.6	1	2	1	0	0.5
Playing (PL)	0	1	5	0	0	1	1	1	10	0	0.5	0

The goal point of the second agent was typically the first agent. The goal points for the first agent was given by a sequence of fixed positions, which were randomly generated by uniformly sampling from the 2D plain and rejecting the samples that were closer than a fixed distance from the last sample. Since it turned out that the influence of the obstacle terms was rather low for the speed dynamics, we dropped the obstacle terms from the speed control dynamics. Table 1 provides an overview of the model parameters for the six simulated behaviors. We generated 50 stimuli for each interaction class. Figure 1 shows examples paths of the agents for the different behaviors for typical simulations.



(a) Chasing(CH) (b) Following(FO) (c) Flirting(FL)

(d) Guarding(GU) (e) Playing(PL) (f) Fighting(FI)

Fig. 1: Sample trajectories for 6 different social interactions. Colors indicate the positions of the two agents (agent 1: blue, agent 2: red). Color saturation indicates time, the color fading out after long times.

3 Model architecture

An overview of the model architecture is shown in Fig. 2. Building on classical biologically-inspired models for shape and action processing [21, 22], the model comprises a form and a motion pathway, each consisting of a hierarchy of feature detectors. Presently, these pathways were modelled following these classical papers, which was sufficient for the tested simple stimuli.

Form Pathway: The form pathway of the simple model implementation here comprises only three hierarchy layers. The first is composed from (even and uneven) Gabor filters with 8 different orientations (cf. [22]), whose centers were placed in a grid of 120 by 120 points across the pixel image. The outputs of this

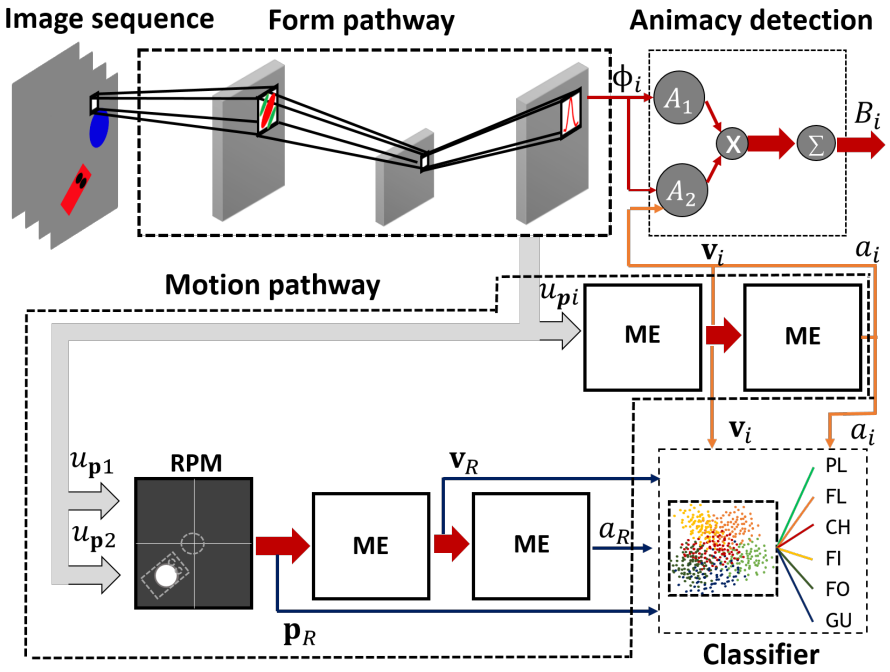


Fig. 2: Model consisting of a form and a motion pathway. ME signfies a layer of motion energy detectors, and RPM the relative position map. The top level of the model is formed by neural detectors for the perceived animacy, and a network that classifies six different types of interactions. (See text for details.)

Gabor filter array are pooled by the next layer using a maximum operation over a grid of 41 by 41 filters, separately for the different orientations, in order to increase the position-invariance of the representation. The highest layer of the form pathway is formed by Gaussian radial basis function, which are trained with the shapes of the agents in different 2D orientations. Opposed to many other object recognition architectures, these shape-selective neurons have receptive fields of limited size (about 20 percent of the width of the image), which is cosistent with neural data from area IT [30]. The outputs of this layer provide thus information about the identity of the agents, their positions, and their orientation in the image plain. The signal $u_k(\phi, x, y)$ is the output activity of the neural detectors detecting shape k at the 2D position (x, y). Summing this signal over all ϕ provides a neural activity distribution $u_{\mathbf{p}_k}(x, y)$ whose peak signals the position of agent k in the image. This signal is used to compute the velocity and the relative positions of the moving elements or animate objects. Similarly, by summing over the positions one obtains a activity distribution $u_{\phi_k}(\phi)$ over the directions with a peak at ϕ_k.

Motion Pathway: it analyzes the 2D motion and the relative motion of the moving agents. As input we use the time-dependent signals $u_{\mathbf{p}_k}(x, y)$ for each agent as input to a field of standard motion energy detectors (ME in Fig. 2), resulting in an output that encodes the motion energy in terms of a four-dimensional neural activity distribution (dropping the index k in the following) $u_{\mathbf{v}}(x, y, v_x, v_y, t)$, where $\mathbf{v} = (v_x, v_y)$ is the preferred velocity vector of the motion energy detector. Pooling this output activity distribution over all spatial positions using a maximum operation, a position-invariant neural representation of velocity is obtained. From this a neural representation of motion direction is obtained by pooling this activity distribution over all neurons with the same (similar) motion direction, resulting in a one-dimensional activity distribution $u_\theta(\theta, t)$ over the motion direction θ, from which the direction can be easily estimated by computing a population vector[1]. The same applies to the length of the velocity vector[2] $v = |\mathbf{v}|$. In order to compute also the acceleration of the agents, we transmit the position-invariant activity distribution $u_{\mathbf{v}}(v_x, v_y, t)$ as input to another field of motion energy detectors, which computes from this an energy distribution $u_{\mathbf{a}}(x, y, a_x, a_y, t)$ over the acceleration vectors $\mathbf{a} = (a_x, a_y)$. By pooling over directions, from this an activity distribution over the length of these vectors $a = |\mathbf{a}|$) is computed, and again this parameter can be estimated by a simple population vector. The population estimates of θ, $\mathbf{v}$ and a enter the animacy computation (s.b.).

For analyzing the relative motion of the two agents, following [22], the output distributions $u_{\mathbf{p}_k}(x, y)$ of the form pathway are also fed into a gain field network that computes a representation of the position of the second agent in a coordinate frame that is centered on the first. Its output is computed as convolution-like integral of the form $u_{\mathbf{p}_R}(x, y) = \int_{x',y'} u_{\mathbf{p}_1}(x', y') u_{\mathbf{p}_2}(x + x', y + y')\, \mathrm{d}x'\mathrm{d}y'$. This output defines a neural *relative position map* that represents the position of agent 2 as an activity peak in a coordinate frame that is centered on the first. The integral is taken over a finite region of shifts $|(x, y)| < D$, implying that situations where the agents have a distance substantially larger than D will not produce an output peak. This makes sense since agents that are too distant do not produce the percept of a social interaction. The activity distribution $u_{\mathbf{p}_R}(x, y, t)$ is again processed by a cascade of two levels of motion energy detectors in order to compute the relative speed and acceleration of the two agents. Population estimates of the relative distance $d_R = |\mathbf{p}_R|$, velocity $\mathbf{v}_R$, and the acceleration a_R enter the interaction classifier.

Recognition Level: the highest level of the model consists of a circuit that derives the perceived animacy of the two agents, and another one that classifies the perceived interaction class. The neurons detecting instantaneous animacy (dropping again the index k and time) multiply two input derived from the signal

[1] A simple estimate of the encoded angle is given by $\hat{\theta} = \arg\big((\sum_m \exp(i\theta_m) u_\theta(\theta_m, t))/(\sum_m u_\theta(\theta_m, t))\big)$, where the θ_m are the preferred directions of the neurons.

[2] Here the estimator is $\hat{v} = \arg\big((\sum_m v_m u_v(v_m, t))/(\sum_m u_v(v_m, t))\big)$, where the v_m are the preferred speeds of the neurons.

of both pathways signals $B = A_1 A_2$. The first signal measures the alignment of the body axis of the moving agent with its direction of its motion. It is just given by the scalar product of the activity distributions over the body axis of the agent $u_\phi(\phi)$ and the motion direction of the agent $u_\theta(\theta)$ in the form $A_1 = \sum_n u_\phi(\theta_n) u_\theta(\theta_n)$. The second signal A_2 linearly combines information about the speed, and the magnitude chages and angular changes of speed, which are given by a and the angular component of $\mathbf{a}$. The linear mixing weights of the animacy neurons were estimated by fitting the psychophysical results from [2]. Final animacy responses were computed as time averages over the whole trajectories.

The second circuit at the top level of the model classifies the different interaction types based on the following features: speeds $\mathbf{v}_i$ and acceleration a_i of the agents, and relative position $\mathbf{p}_R$, velocity $\mathbf{v}_R$, and acceleration a_R of the agents. These features served as inputs of different classifier models, We tested a multi-layer perceptron, linear and nonlinear discriminant analysis (see also [31]), k-nearest neighbor classification, and a linear and a nonlinear support vector machine.

4 Results

Results on animacy detection are shown in Fig. 3. The model reproduces at least qualitatively the dependence of animacy ratings on directions and speed changes [2]. In these experiments an agent shape moved along a straight line and then suddenly changed speed or direction by different amounts. In addition, the model reproduces the fact that a moving figure that has a body axis, like a rectangle, results in stronger perceived animacy than a circle if the movement, and that the rating is highest if the body axis is aligned with the motion than if it is not aligned [2].

Fig.4 shows example results from the application of the different classifier models for the 6 interaction behaviors in the study [9]. The classifiers were trained on movies generated with the stimulus generation algorithm described in section 2. The linear SVM classifier achieves 99% correct classifications on this data set. See Tab.2 for the results with the other classifiers. Most importantly, the model achieved also 100 % correct classifications on the example videos from [9], even though these movies were not used for training.

Classifier	Accuracy
Linear SVM	99.0%
Gaussian kernel SVM	96.3%
LDA	94.7%
KNN	94.7%
Nonlinear LDA	94.3%
Neural Network	94.0%

Table 2: Classification results with different classifiers (6 interaction types).

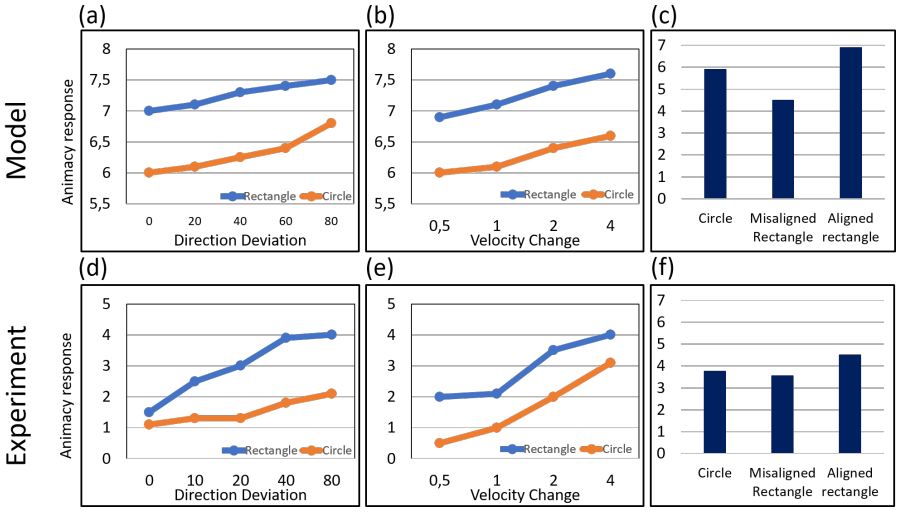
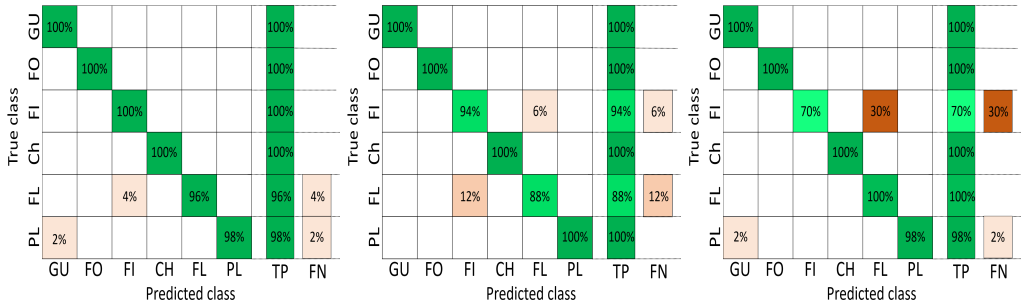


Fig. 3: Simulation results for animacy perception in comparison with experimental results. (a),(d): Dependence of animacy ratings on size of direction change. (b),(e): Dependence of animacy rating on size of speed change. (c),(f): Effect of alignment of body axis with motion direction, compared with moving circle (no body axis).

5 Conclusion

Our model accounts by combination of very elementary neural mechanisms for a number of classical results from animacy and social interaction perception from abstract figures. To our knowledge this is the first neural model that can account for such results. Evidently the model is only a proof-of-concept with many shortcomings, a major one being that the accuracy of the form and motion pathway that provide input to the animacy and interaction detection have to be improved. Since the model is in principle consistent with deep architectures for form and action recognition that can achieve high performance level it seems likely that it can be extended to the processing of much more challenging stimulus material. Even in its simple form the model proves that animacy and social interaction judgements partly might be derived by very elementary operations in hierarchical neural vision systems, without a need of sophisticated or accurate probabilistic inference.

Acknowledgments. This work was supported by: HFSP RGP0036/2016; the European Commission HBP FP7-ICT2013-FET-F/ 604102 and COGIMON H2020-644727, the DFG KA 1258/15-1, and BMBF CRNC FK: 01CQ1704.



(a) Linear SVM (one-vs-one) (b) Linear SVM (one-vs-all) (c) KNN

Fig. 4: Confusion matrices for the best(Linear SVM) and the worst(KNN) classifier; TP: true positive rate, FN stands for false negative rate. 50 videos per class.

References

1. Heider, F. and Simmel, M.: An Experimental Study of Apparent Behavior. The American Journal of Psychology (1944)
2. Tremoulet, P.D., Feldman, J.: Perception of animacy from the motion of a single object. Perception 29, 943–951 (2000)
3. Tremoulet, P. D. and Feldman, J.: The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. Perception and psychophysics (2006)
4. Hernik, M., Fearon, P., and Csibra, G.: Action anticipation in human infants reveals assumptions about anteroposterior body structure and action. Proceedings. Biological sciences (2014)
5. Scholl, B. J. and Tremoulet, P. D.: Perceptual causality and animacy. Trends in Cognitive Sciences, 4(8):299–309 (2000)
6. Gao, T. and Scholl, B. J.: Perceiving animacy and intentionality. In Rutherford, M. D. and Kuhlmeier, V. A., editors, Social Perception. The MIT Press (2013)
7. Blythe P, Miller GF, Todd PM.: How motion reveals intention: Categorizing social interactions. In: Gigerenzer G, Todd P (eds) Simple heuristics that make us smart. Oxford University Press, London, pp 257–285 (1999)
8. Barrett, H. C., Todd, P. M., Miller, G. F., Blythe, P. W.: Accurate judgments of intention from motion cues alone: A cross-cultural study. Evolution and Human Behavior, 26(4), 313–331 (2005)
9. McAleer, P., Pollick, F.E.: Understanding intention from minimal displays of human activity. Behavior Research Methods 40, 830–839 (2008)
10. Schultz, J., Friston, K. J., O'Doherty, J., Wolpert, D. M., Frith, C. D.: Activation in posterior superior temporal sulcus parallels parameter inducing the percept of animacy. Neuron, 45(4), 625–635 (2005)
11. Morito, Y., Tanabe, H. C., Kochiyama, T., Sadato, N.: Neural representation of animacy in the early visual areas: a functional MRI study. Brain Research Bulletin, 79(5), 271–280 (2009)

12. Shultz, S., McCarthy, G.: Perceived animacy influences the processing of human-like surface features in the fusiform gyrus. Neuropsychologia, 60, 115–120 (2014)
13. Blakemore, S.-J., Boyer, P., Pachot-Clouard, M., Meltzoff, A., Segebarth, C., Decety, J.: The Detection of Contingency and Animacy from Simple Animations in the Human Brain. Cerebral Cortex, 13(8), 837–844 (2003)
14. Yang, D. Y.-J., Rosenblau, G., Keifer, C., Pelphrey, K. A.: An integrative neural model of social perception, action observation, and theory of mind. Neuroscience and Biobehavioral Reviews, 51, 263–275 (2015)
15. Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., Nummenmaa, L.: Naturalistic FMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. Frontiers in Human Neuroscience, 6, 233 (2012)
16. Isik, L., Koldewyn, K., Beeler, D., Kanwisher, N.: Perceiving social interactions in the posterior superior temporal sulcus. PNAS 114, E9145–E9152 (2017)
17. Sliwa, J., Freiwald, W. A.: A dedicated network for social interaction processing in the primate brain. Science, 356(6339), 745–749 (2017)
18. Walbrin, J., Downing, P., Koldewyn, K.: Neural responses to visually observed social interactions. Neuropsychologia 112, 31–39 (2018)
19. Baker, C.L., Saxe, R., Tenenbaum, J.B.: Action understanding as inverse planning. Cognition, Reinforcement learning and higher cognition 113, 329–349 (2009)
20. Shu, T., Peng, Y., Fan, L., Lu, H., Zhu, S.-C.: Perception of Human Interaction Based on Motion Trajectories: From Aerial Videos to Decontextualized Animations. Topics in Cognitive Science, 10(1) (2018)
21. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nat. Neurosci. 2, 1019–1025 (1999)
22. Giese, M.A., Poggio, T.: Neural mechanisms for the recognition of biological movements. Nat Rev Neurosci 4, 179–192 (2003)
23. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A Biologically Inspired System for Action Recognition, in: IEEE 11th International Conference on Computer Vision (2007)
24. Fleischer, F., Caggiano, V., Thier, P., Giese, M. A.:Physiologically Inspired Model for the Visual Recognition of Transitive Hand Actions. The Journal of Neuroscience, 15(33), 6563-80 (2013)
25. Fleischer, F., Christensen, A., Caggiano, V., Thier, P., Giese, M. A.: Neural theory for the perception of causal actions. Psychological Research, 76(4), 476–493 (2012)
26. Caggiano, V., Fleischer, F., Pomper, J. K., Giese, M. A., Thier, P.: Mirror Neurons in Monkey Premotor Area F5 Show Tuning for Critical Features of Visual Causality Perception. Current Biology, 26(22), 3077–3082 (2016)
27. Warren, W. H.: The dynamics of perception and action. Psychological Review, 113(2), 358–389 (2006)
28. Schöner, G., Dose, M.: A dynamical systems approach to task-level system integration used to plan and control autonomous vehicle motion. Robotics and Autonomous Systems, 10(4), 253–267 (1992)
29. Fajen, B.R., Warren, W.H.: Behavioral dynamics of steering, obstacle avoidance, and route selection. Journal of Experimental Psychology: Human Perception Performance (2003)
30. diCarlo, J.J., Zoccolan, D., Rust, N.C.: How does the brain solve visual object recognition? Neuron, 73(3), 415–434 (2012)
31. You, D., Hamsici, O. C., Martinez, A. M.: Kernel optimization in discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(3), 631-638 (2011)