Modeling visual Affordances: The Selective Attention for Action Model (SAAM)

Christoph Böhme* and **Dietmar Heinke** [1]

Abstract.

Classically, visual attention is assumed to be influenced by visual properties of objects, e. g. as assessed in visual search tasks. However, recent experimental evidence suggests that visual attention is also guided by action-related properties of objects ("affordances")[1, 2], e. g. the handle of a cup *affords* grasping the cup; therefore attention is drawn towards the handle. In a first step towards modelling this interaction between attention and action, we implemented the Selective Attention for Action model (SAAM). The design of SAAM is based on the Selective Attention for Identification model (SAIM)[3]. For instance, we also followed a soft-constraint satisfaction approach in a connectionist framework. However, SAAM's selection process is guided by locations within objects suitable for grasping them whereas SAIM selects objects based on their visual properties. In order to implement SAAM's selection mechanism two sets of constraints were implemented. The first set of constraints took into account the anatomy of the hand, e. g. maximal possible distances between fingers. The second set of constraints (geometrical constraints) considered suitable contact points on objects by using simple edge detectors. We demonstrate here that SAAM can successfully mimic human behaviour by comparing simulated contact points with experimental data.

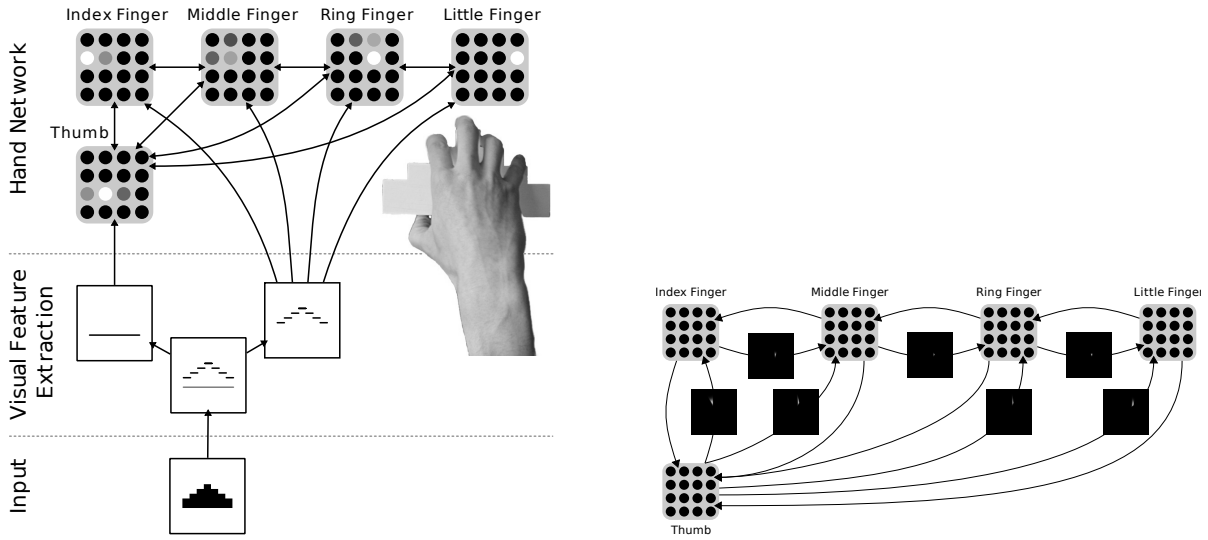
1 Introduction

Actions need to be tightly guided by vision in our daily interactions with our environment. To maintain such a direct guidance, J. J. Gibson postulated that the visual system automatically extract "affordances" of objects [2]. According to Gibson, affordance refers to parts or properties of visual objects that are directly linked to actions or motor performances. For instance, a handle of a cup *affords* directly a reaching and grasping action. Recently, experimental studies have produced empirical evidence in support for this theory. Neuroimaging studies showed that objects activate the premotor cortex even when no action has to be performed with the object [4, 5]. Behavioural studies indicated response interferences from affordances despite the fact that they were response-irrelevant [6, 7]. For instance, a recent study in Ref. [8] demonstrated that pictures of hand postures (precision or power grip) can influence subsequent categorisation of objects. In this study, participants had to categorise objects into either artefact or natural object. Additionally, and unknown to the participants, the objects could be manipulated with either a precision or a power grasp. The study showed that categorisation was faster when the hand postures were congruent with the grasp compared to hand postures being incongruent with the grasp. Hence, the participants' behaviour was influenced by action-related properties of objects irrelevant to the experimental task. This experiment together with earlier, similar studies can be interpreted as evidence for an automatic detection of affordances.

Interestingly, recent experimental evidence suggests that not only actions are triggered by affordances, but also that selective attention is guided towards action-relevant locations. Using event-related potentials (ERP) Handy *et al.* showed that spatial attention is more often directed towards the location of tools than non-tools [9]. Pellegrino *et al.* present similar evidence from two patients with visual extinction[10]. In general visual extinction is considered to be an attentional deficit in which patients, when confronted with several objects, fail to report objects on the left side of their body space. In contrast, when faced with only one object, patients can respond to the object irrespective of its location. This study demonstrated that this attentional deficit can be alleviated when the handle of a cup points to the left. Pellegrino *et al.* interpreted their results as evidence for automatically encoded affordance (without the patients' awareness) drawing the patients' attention into their "bad" visual field.

This paper aims to lay the foundations for a computational model of such affordance-based guidance of attention. We designed a connectionist model which determines contact points for a stable grasp of an object (see Fig. 1(a) for an illustration). The model extracts these contact points directly from the input image. Hence, such a model could be construed as an implementation of an automatic detection of object affordances for grasping. To realise the attentional guidance through affordances, we integrated the selection mechanisms employed in the Selective Attention for Identification Model (SAIM)[3]. Since this new model performs selection for action rather than identification, we termed the new model Selective Attention for Action Model (SAAM). There are only few computational models of affordance[11, 12]. However, Faggs *et al.* model does not process multiple-object scenes. On the other hand Ciseks model considers attentional processing with respect to pointing actions. However, in order to model crucial aspects of affordance-oriented processing, it is necessary to consider behaviours that require true physical interactions with objects, since this characteristic leads to an entirely different processing objective compared to classical perceptual processing, e. g. object recognition, where the physical environment is passively analysed. In this paper we will present first simulation results as well as an experimental verification of the model.

[1] School of Psychology, University of Birmingham Birmingham B15 2TT, United Kingdom, email: cxb632@bham.ac.uk



(a) Overall structure

(b) Excitatory connections between fingers.

Figure 1. The Selective Attention for Action model (SAAM)

2 The Selective Attention for Action Model (SAAM)

Figure 1(a) gives an overview of SAAM. The input consists of black&white images. The output of the model is generated in five "finger maps" of a "hand network". The finger maps encode the finger positions which are required for producing a stable grasp of the object in the input image. At the heart of SAAM's operation is the assumption that stable grasps are generated by taking into account two types of constraints, the geometrical constraints imposed from the object shape and the anatomical constraints given by the hand. In order to ensure that the hand network satisfies these constraints we followed an approach suggested in Ref. [13]. In this soft-constraint satisfaction approach, constraints define activity patterns in the finger maps that are permissible and others that are not. Then we defined an energy function for which the minimal values are generated by just these permissible activity values. To find these minima, a gradient descent procedure is applied resulting in a differential equation system. The differential equation system defines the topology of a biologically plausible network. The mathematical details of this energy minimisation approach are given in the next section. Here, we focus on a qualitative description of the two types of constraints and their implementation.

The geometrical constraints are extracted from the shape of the object in the visual feature extraction stage. To begin with, obviously, only edges constitute suitable contact points for grasps. Furthermore, edges have to be perpendicular to the direction of the forces exerted by the fingers. Hence, only edges with a horizontal orientation make up good contact points, since we consider only a horizontal hand orientation in this first version of the model (see Fig. 1(a)). We implemented horizontal edge detectors using Sobel filters [14]. Finally, to exert a stable grasp, thumb and fingers need to be located at opposing sides of an object. This requirement was realized by separating the output of the Sobel filters according to the direction of the gradient change at the edge. In fact, the algebraic sign of the response differs at the bottom of a 2D-shape compared to the top of a 2D-shape. Now, if one assumes the background colour to be white and the object colour to be black, the signs of the Sobel-filter responses indicate appropriate locations for the fingers and the thumb (see Fig. 1(a) for an illustration). The results of the separation feed into the corresponding finger maps providing the hand network with the geometrical constraints. Note that, of course, the assumptions about the object- and background-colours represent a strong simplification. On the other hand, this mechanism can be interpreted as mimicking the result of stereo vision. In such a resulting "depth image" real edges suitable for thumb or fingers could be easily identified.

The anatomical constraints implemented in the hand network take into account that the human hand cannot form every arbitrary finger configuration to perform grasps. For instance, the maximum grasp width is limited by the size of the hand and the arrangement of the fingers on the hand makes it impossible to place the index, middle, ring, and little finger in another order than this one. After applying the energy minimisation approach, these anatomical constraints are implemented by excitatory connections between the finger layers in the hand network (see Fig. 1(a) and 1(b)). Figure 1(b) also illustrates the weight matrices of the connections. Each weight matrix defines how every single neuron of one finger map projects onto another finger map. The direction of the projection is given by the arrows between the finger maps. For instance, neurons in the thumb map feed their activation along a narrow stretch into the index finger map, in fact, encoding possible grip sizes. Each neuron in the target map sums up all activation fed through the weight matrices. Note that all connections between the maps are bi-directional whereby the feedback path uses the transposed weight matrices of the

feedforward path. This is a direct result of the energy minimisation approach and ensures an overall consistency of the activity pattern in the hand network, since, for instance, the restriction in grip size between thumb and index finger applies in both directions. Finally, since a finger can be positioned at only one location, a winner-takes-all mechanism was implemented in all finger maps.

2.1 Mathematical Details

2.1.1 Visual Feature Extraction

The filter kernel K in the visual feature extraction process is a simple Sobel-filter [14]. In the response of the Sobel-filter the top edges of the object are marked with positive activation while the bottom edges are marked with negative activation. This characteristic of the filter is used to feed the correct input with the geometrical constraint applied into the finger maps and the thumb map. The finger maps receive the filter response with all negative activation set to zero. The thumb map, however, receives the negated filter response with all negative activation set to zero:

$$I_{ij}^{(\mathrm{f})} = \begin{cases} R_{ij} & \text{if } R_{ij} \geq 0, \\ 0 & \text{else.} \end{cases}$$

$$I_{ij}^{(\mathrm{t})} = \begin{cases} -R_{ij} & \text{if } -R_{ij} \geq 0, \\ 0 & \text{else.} \end{cases}$$

with $R_{ij} = I_{ij} * K$ whereby I_{ij} is the input image.

2.1.2 Hand Network

We used an energy function approach to satisfy the anatomical and geometrical constraints of grasping. In Ref. [13] an approach is suggested where minima in the energy function are introduced as a network state in which the constraints are satisfied. In the following derivation of the energy function, parts of the whole function are introduced, and each part relates to a particular constraint. At the end, the sum of all parts leads to the complete energy function, satisfying all constraints.

The units $y_{ij}^{(f)}$ of the hand network make up five fields. Each of these fields encodes the position of a finger. $y_{ij}^{(1)}$ encodes the thumb, $y_{ij}^{(2)}$ encodes the index finger, and so on to $y_{ij}^{(5)}$ for the little finger. For the anatomical constraint of possible finger positions the energy function is based on the Hopfield associative memory approach [15]:

$$E(y_i) = -\sum_{\substack{ij \\ i \neq j}} T_{ij} \cdot y_i \cdot y_j.$$

The minimum of the function is determined by the matrix T_{ij}. For T_{ij}s greater than zero, the corresponding y_is should either stay zero or become active in order to minimize the energy function. In the associative memory approach, T_{ij} is determined by a learning rule. Here, we chose the T_{ij} so that the hand network fulfils the anatomical constraints. These constraints are satisfied when units in the finger maps that encode finger positions of anatomically feasible postures are active at the same time. Hence, the T_{ij} for these units should be greater than zero, and for all other units, T_{ij} should be less than or equal to zero. This lead to the following equation:

$$E_{\mathrm{a}}(y_{ij}^{(g)}) = -\sum_{f=1}^{5} \sum_{\substack{g=1 \\ g \neq f}}^{5} \sum_{ij} \sum_{\substack{s=-L \\ s \neq 0}}^{L} \sum_{\substack{r=-L \\ r \neq 0}}^{L} T_{sr}^{(fg)} \cdot y_{ij}^{(g)} \cdot y_{i+s,j+r}^{(f)}.$$

In this equation $T_{ij}^{(fg)}$ denotes the weight matrix from finger f to finger g.

A further constraint is the fact that each finger map should encode only one position. The implementation of this constraint is based on the energy function proposed in Ref. [16]:

$$E_{\mathrm{WTA}}(y_i) = a \cdot (\sum_i y_i - 1)^2 - \sum_i y_i \cdot I_i.$$

This energy function defines a winner-takes-all (WTA) behaviour, where I_i is the input and y_i is the output of each unit. This energy function is minimal when all y_i are zero except one, and when the corresponding input I_i has the maximal value of all inputs. Applied to the hand network where each finger map requires a WTA-behaviour, the first part of the equation turns into:

$$E_{\mathrm{WTA}}^{\mathrm{a}}(y_{ij}^{(f)}) = \sum_{f=1}^{5} (\sum_{ij} y_{ij}^{(f)} - 1)^2.$$

The input part of the original WTA-equation was modified to take the geometrical constraints into account:

$$E_{\mathrm{f}}(y_{ij}^{(f)}) = -\sum_{f=2}^{5} \sum_{ij} w_f \cdot y_{ij}^{(f)} \cdot I_{ij}^{(\mathrm{f})},$$

$$E_{\mathrm{t}}(y_{ij}^{(1)}) = -\sum_{ij} w_1 \cdot y_{ij}^{(1)} \cdot I_{ij}^{(\mathrm{t})}.$$

These terms drive the finger maps towards choosing positions at the input object which are maximally convenient for a stable grasp. The w_f factors were introduced to compensate the effects of the different number of excitatory connections in each layer.

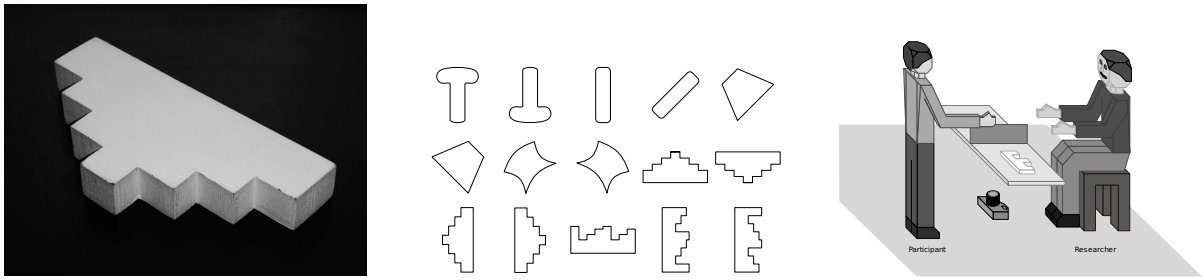
The Complete Model To consider all constraints, all energy functions need to be added, leading to the following complete energy function:

$$E(y_{ij}^{(f)}) = a_1 \cdot E_{\mathrm{WTA}}^{\mathrm{a}}(y_{ij}^{(f)}) + a_2 \cdot E_{\mathrm{f/t}}(y_{ij}^{(f)}) + a_3 \cdot E_{\mathrm{a}}(y_{ij}^{(f)}).$$

The parameters a_i weight the different constraints against each other. These parameters need to be chosen in a way that SAAM successfully selects contact points at objects in both conditions, single-object images and multiple-object images. The second condition is particularly important to demonstrate that SAAM can mimic affordance-based guidance of attention. Moreover, and importantly, SAAM has to mimic human-style contact points. Hereby, not only the parameters a_i are relevant, but also the weight matrices of the anatomical constraints strongly influence SAAM's behaviour.

Gradient Descent The energy function defines minima at certain values of y_i. To find these values, a gradient descent procedure can be used:

$$\tau \dot{x}_i = -\frac{\partial E(y_i)}{\partial y_i}.$$



(a) Object used in the study. (b) Conditions of the experiment. (c) Experimental set-up.

Figure 2. Material and procedure of the grasping experiment.

The factor τ is antiproportional to the speed of descent.

In the Hopfield approach, x_i and y_i are linked together by the sigmoid function:

$$y_i = \frac{1}{1 + e^{-m \cdot (x_i - s)}},$$

and the energy function includes a leaky integrator, so that the descent turns into

$$\tau \dot{x}_i = -x_i - \frac{\partial E(y_i)}{\partial y_i}.$$

Using these two assertions, the gradient descent is performed in a dynamic, neural-like network, where y_i can be related to the output activity of neurons, x_i the internal activity, and $\partial E(y_i)/\partial y_i$ gives the input to the neurons.

Applied to the energy function of SAAM, it leads to a dynamic unit (neuron) which forms the hand network:

$$\tau \dot{x}_{ij}^{(f)} = -x_{ij}^{(f)} - \frac{\partial E_{\text{total}}(y_{ij}^{(f)})}{\partial y_{ij}^{(f)}}.$$

To execute the gradient descent on a computer, a temporarily discrete version of the descent procedure was implemented.

3 Verification of the model

This study tested whether SAAM can generate expedient grasps in general and whether these grasps mimic human grasps. To accomplish this, simulations with single objects in the visual field were conducted. The results of the simulations were compared with experimental data on grasping these objects. In the following two sections we will at first present the experiment and its results and then compare its outcomes with the results from our simulations with SAAM.

3.1 Experiment

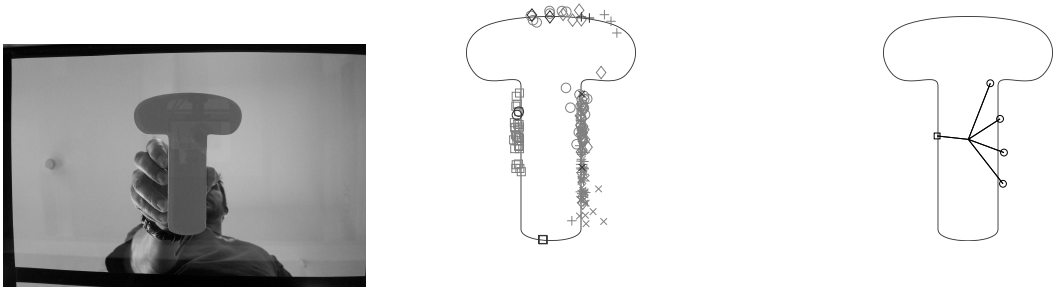
We conducted an experiment in which humans grasped objects. Interestingly, there are only very few published studies on this question. Most notably D. P. Carey *et al.* examined grasps of a stroke patient [17]. However, no studies with healthy participants can be found in the literature.

Participants We tested 18 school students visiting the psychology department on an open day. The mean age was 17.8 years. All participants but two were right-handed. The left-handed participants were excluded from further analysis because the objects had not always been mirrored correctly during the experiment.

Material For the experiment we designed six two-dimensional object shapes. The objects were made of 2.2 cm thick wood and were painted white. Their size was between 11.5×4 and 17.5×10 centimetres (see Fig. 2(a) for an example). By presenting the objects in different orientations we created fifteen conditions (see Fig. 2(b)). Note that the shapes are highly unfamiliar, non-usable. Hence, the influence of high-level object knowledge is limited in the experiment. We chose this set-up in order to be compatible with the simulations in which SAAM possesses no high-level knowledge either.

Procedure Figure 2(c) illustrates the experimental set-up. During the experiment participants and experimenter were situated on opposite sides of a glass table facing each other. The glass table was divided in two halves by a 15 cm high barrier. Participants were asked to position themselves so that their right hand was directly in front of the right half of the glass table. In each trial the experimenter placed one of the objects with both hands in the right half of the glass table. The participants were then asked to grasp the object, lift it and place it into the left half without releasing the grip. The experimenter took a picture with a camera from below the glass table (see Figure 3(a) for an example). After taking the photo, the participants were asked to return the object to the experimenter. The last step was introduced to ensure that the participants would not release their grasp before the photo was taken. As soon as the object was handed back to the experimenter, a new trial started by placing the next object in the right half of the glass table. Each participant took part in two blocks with fifteen trails each. The order of the trials was randomised.

Results To analyse the pictures taken in the experiment, we developed a software for marking the positions of the fingers in relation to the objects. In Figure 3(b) the resulting finger positions are shown for the first condition. Even



(a) Photo of grasp. (b) Extracted finger positions. (c) Mean finger positions.

Figure 3. Mean finger positions: Finger positions (b) are extracted from photos (a). The fingers are colour-coded (thumb marked in red; right-hand grasp). For each finger its mean position is calculated (c). The thumb position is highlighted by a square box.

though the grasps show some variability, in general, participants grasped the object in two ways: they either placed their thumb at the left side of the object and the fingers on the right side or they placed the thumb at the bottom of the object and the fingers on the top edges. These two sets of different grasping positions are indicated with two markers in Figure 3(b) (circle and square). Such sets different of grasping positions were observed in all conditions.

To determine a "typical" grip from the experimental data, averaging across these very different set of grasping positions would not make sense. Therefore, we calculated the mean finger positions for each set of grasping positions separately. The resulting mean positions for the first condition are shown in Figure 3(c). Sets of grasping positions containing only one or two samples were discarded as outliers. For the comparison with the simulation results we only considered the set of grasping positions for each object chosen in the majority of trials.

3.2 Simulations

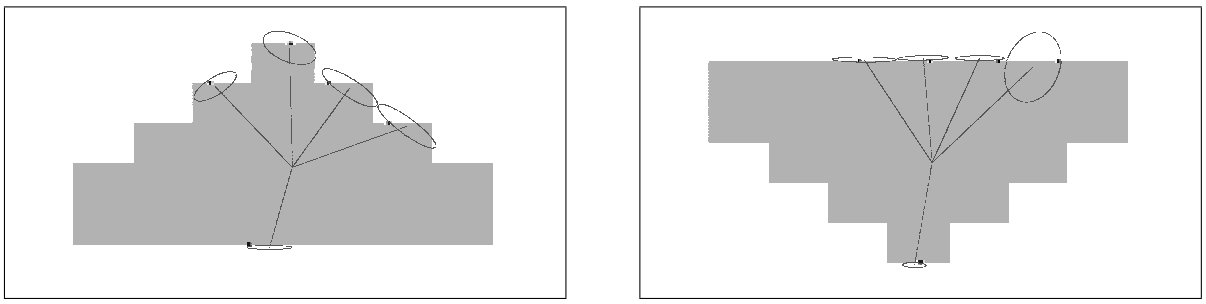
We conducted simulations with SAAM using the same objects as in the experiment. Figure 4 shows two examples of the simulation results. These illustrations also include the mean finger positions from the experimental results for a comparison with the simulation data. The ellipses around the mean finger positions illustrate the variations in the data. The comparison shows that most finger positions lie within the ellipses. Hence the theoretical assumptions behind SAAM that geometrical and anatomical constraints are sufficient to mimic human behaviour have been confirmed. Note that not all experimental conditions could be simulated with SAAM, since the model is currently only able to create horizontal grasps.

We also tested simulations with two objects in the visual field to test SAAM's ability to simulate attentional processes. The simulations were successful in the sense that contact points for only one object were selected and the second object was ignored (see Conclusion for further discussions).

4 Conclusion and Outlook

Recent experimental evidence indicates that visual attention is not only guided by visual properties of visual stimuli but also by affordances of visual objects. This paper sets out to develop a model of such affordance-based guidance of selective attention. As a case in point we chose to model grasping of objects and termed the model the Selective Attention for Action Model (SAAM). To detect the parts of an object which afford a stable grasp, SAAM performs a soft-constraint satisfaction approach by means of a Hopfield-style energy minimisation. The constraints were derived from the geometrical properties of the input object and the anatomical properties of the human hand. In a comparison between simulation results and experimental data from human participants we could show that these constraints are sufficient to simulate human grasps. We also tested whether SAAM cannot only extract object affordances but also implements the guidance of attention through affordances by using two-object images. Indeed, SAAM was able to select one of two objects based on their affordance. The interesting aspect here is that SAAM's performance is an emergent property from the interplay between the anatomical constraints. Especially, the competitive mechanism implemented in the finger maps is crucial for SAAM's attentional behaviour. This mechanism already proved important in SAIM [3] for simulating attentional effects of human object recognition. However, it should be noted that SAAM does not select whole objects as SAIM does. But, since SAAM and SAIM use similar mechanisms, it is conceivable that they can be combined to form one model. In such a model SAIM's selection mechanism of whole objects can be guided by the SAAM's selection of contact points. Hence, this new model could integrate both mechanisms, selection by visual-properties and by action-related properties, forming a more complete model of selective attention.

Despite the successes reported here, this work is still in its early stages. First, we will need to verify the priorities of object selection predicted by SAAM. We also plan to include grasps with a rotated hand to simulate a broader range of experimental data. Finally, there is a large amount of experimental data on the interaction between action knowledge



(a) Simulation 1 (b) Simulation 2

Figure 4. Comparison of experimental results and simulated grasps. The ellipses indicate the variation in the experimental data. The black dots mark the finger positions as generated by the simulations.

and attention (see Ref. [18] for a summary). Therefore, we aim to integrate action knowledge into SAAM, e. g. grasping a knife for cutting or stabbing. With these extensions SAAM will sufficiently contribute to the understanding of how humans determine object affordances and how these lead to a guidance of attention.

REFERENCES

[1] J. J. Gibson, *The senses considered as perceptual systems* (Houghton-Mifflin, Boston, 1966).

[2] J. J. Gibson, *The ecological approach to visual perception* (Houghton-Mifflin, Boston, 1979).

[3] D. Heinke and G. W. Humphreys, *Psychological Review* **110**, 29 (2003).

[4] S. T. Grafton, L. Fadiga, M. A. Arbib and G. Rizzolatti, *NeuroImage* **6**, 231 (1997).

[5] J. Grèzes and J. Decety, *Neuropsychologia* **40**, 212 (2002).

[6] M. Tucker and R. Ellis, *Journal of Experimental Psychology* **24**, 830 (1998).

[7] J. C. Phillips and R. Ward, *Visual Cognition* **9**, 540 (2002).

[8] A. M. Borghi, C. Bonfiglioli, L. Lugli, P. Ricciardelli, S. Rubichi and R. Nicoletti, *Neuroscience Letters* **411**, 17 (2007).

[9] T. C. Handy, S. T. Grafton, N. M. Shroff, S. Ketay and M. S. Gazzaniga, *Nature Neuroscience* **6**, 421 (2003).

[10] G. di Pellegrino, R. Rafal and S. P. Tipper, *Current Biology* **15**, 1469 (2005).

[11] A. H. Fagg and M. A. Arbib, *Neural Networks* **11**, 1277 (1998).

[12] P. Cisek, *Philosophical Transactions of the Royal Society* **362**, 1585 (2007).

[13] J. J. Hopfield and D. W. Tank, *Biological Cybernetics* **52**, 141 (1985).

[14] R. C. Gonzalez and R. E. Woods, *Digital Image Processing* (Addison-Wesley, 1993).

[15] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, in *Proc. of the Nat. Academy of Sciences*, (8)1982.

[16] E. Mjolsness and C. Garrett, *Neural Networks* **3**, 651 (1990).

[17] D. P. Carey, M. Harvey and A. D. Milner, *Neuropsychologia* **34**, 329 (1996).

[18] G. W. Humphreys and M. J. Riddoch, *Psychology of learning and motivation* **42**, 225 (2003).