Modeling the development of causality and occlusion perception in infants

Arthur Franz and Jochen Triesch
Frankfurt Institute for Advanced Studies, Johann Wolfgang Goethe University
Ruth-Moufang-Str. 1, 60438 Frankfurt am Main, Germany
{*franz, triesch*}*@fias.uni-frankfurt.de*

***Abstract*— Developmental researchers investigate many pieces of infants' physical knowledge, e.g. the perception of causality, occlusion or object permanence, but a theoretical framework that would unify all these pieces, account for the most basic phenomena and make testable predictions has not been provided yet. Here we make an attempt to unify and explain the emergence of causality and occlusion perception and its development in infancy using a simple artificial neural network that derives its representations from simplified motion detector and disparity cells as found in the primary visual cortex. The network accounts simultaneously for two experiments on causality and occlusion perception and develops a representation of object permanence during training. It also makes detailed testable predictions for the course of development and provides an account of *how* change occurs. We conclude that many aspects of physical knowledge can probably be learned from the statistical regularities of our environment while only few assumptions are needed.**

***Index Terms*— causality, launching, contact, solidity, occlusion, continuity, object permanence, model, Elman network, prediction**

I. Introduction

The question about the origin of knowledge is as old as humanity itself and has been one of the major questions pondered by philosophers. But since Jean Piaget it has also been tackled experimentally by developmental psychologists who investigate how human infants' knowledge develops in the course of their growth.

There are a few central phenomena that seem to constitute infants' most basic physical knowledge. Infants seem to perceive objects as moving on connected, unobstructed paths (continuity), as only affecting another object's motion if and only if they touch (contact) and that they normally do not pass through each other or through solid surfaces (solidity). These capacities have been suggested to be innate [13]. At some point infants learn about object permanence - the notion that objects continue to exist even though they are out of sight. Infants as young as 4 months have been shown to have reached this understanding [1].

As for causality the discussion can be traced back to Hume [3] in whose classical account the perception of causality in simple mechanical events is the result of repeated experiences of a constant conjunction between two events. Michotte [10] argued that causality could be perceived directly, for example, when one billiard ball collides with and launches another. He believed that in order to gain a "causal percept" infants would at least have to see enough "internal structure" to segregate a launching sequence into two movement components. Leslie [7] believed that an innate notion of force or pressure is needed and that the perception of cause and effect is performed by an innate motion analysis module. Mandler [8] suggested that seeing transfer of motion may provide the basis of infants' early interpretation of causal physical events and that no notions of force or pressure are necessary. We demonstrate that this is possible by constructing a computational model that learns to represent launching and occlusion events by merely observing them and detecting statistical regularities in them. We show that this model explains one of the fundamental experiments on the perception of causality in infants [6] while no innate force notions or modules are needed.

The model is an artificial neural network that is trained to predict its next inputs. We are going to model two experiments on causality and occlusion perception that rely on the so called habituation paradigm, e.g. experimenters repeatedly show a visual stimulus to an infant and measure the time it looks at it. As trials increase the looking time drops which is referred to as habituation, i.e. the infant gets "bored". After this phase usually two test stimuli are shown and the looking time is measured again. If the looking time increases again (dishabituation) then a stimulus is interpreted to be "novel" or "surprising" to the infant.

In our model the network's error in predicting its next input is used to model the infant looking time since novelty can be seen as a prediction failure. Prediction learning has been highlighted by a number of developmental theorists, e.g. [2], but was not referred to as a model for looking time. Schlesinger and Young [12] used a prediction network as a model of looking time but they did not pre-train their network which we consider essential in this kind of modeling (see Sect. II-B). [9] trained a network to predict occlusion events but the network is modularized and largely constructed by hand (including object recognition modules etc.) which makes it less parsimonious than our model. The work of [11] comes closest to our work but their model was trained exactly for the one task is was supposed to accomplish: representing an occluded object at a fixed position whereas in our model neither the position of the objects nor the sort

of task (occlusion or launching) is prespecified. Specifically, our model is the first to account for both occlusion and causality data.

II. Methods

A. General architecture

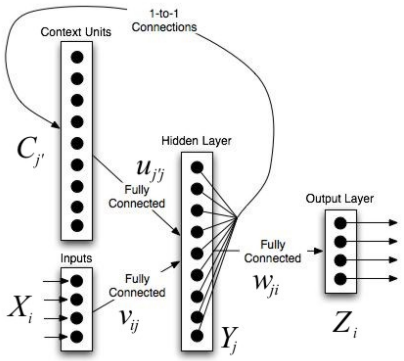

Fig. 1. The Elman network

We use a simple recurrent network, also known as the Elman network [2]. It consists of four layers of artificial neurons, named input, hidden, output and context layers, respectively (Fig. 1). The inputs to units in the hidden and output layers are weighted sums of the responses, X_i, C'_j, Y_j, from units at previous layers. The outputs Y_j and Z_i of the units are a Fermi function of the input:

$$Y_j = \left[1 + \exp\left(-\sum_{i=0}^{M} v_{ij} X_i + \sum_{j'=1}^{N} u_{j'j} C_{j'}\right)\right]^{-1} \quad (1)$$

and

$$Z_i = \left[1 + \exp\left(-\sum_{j=0}^{N} w_{ji} Y_j\right)\right]^{-1} \quad (2)$$

where v_{ij}, $u_{j'j}$ and w_{ji} are the weights. Every hidden and output unit has an additional constant input X_0 and Y_0 equal to 1. The weights v_{0j} and w_{0i} of these supplementary inputs act as threshold values for each unit and are also learned. The context layer derives its activity from the hidden layer by copying its activity at each sweep of calculation: $C_j := Y_j$. The Elman network is presented a temporal series of inputs $X_i(t), X_i(t+1), X_i(t+2), ...$ and its task is to learn from this sequence and predict the next input $X_i(t+1)$. We trained the network with the standard backpropagation algorithm minimizing the sum of the squares of the difference between the output $Z_i(t)$ and the next input $X_i(t+1)$. In our model we relate the prediction error

$$E(t) \equiv \sum_{i=1}^{M} |Z_i(t-1) - X_i(t)| \quad (3)$$

to the looking time in experiments with infants.

Our model is constructed to predict occlusion and launching events. Therefore, we need to represent motion and depth. In order to do so we split up the input layer into three maps, the "motion detectors" (first 7 units), "disparity" units (next 7 units) and "novelty" units (next 14 units) that represent the novelty of the environment. Fig. 2a) shows the inputs to the network at a launching event. The motion detector map is only active for a moving object (pixel). A unit is set to 1 if motion is present (in any direction) and to 0 otherwise. In order to predict occlusion events successfully one needs to distinguish merely three depth relations: farther away, same distance and closer than the object participating in the occlusion event. Therefore, units in the depth map can have three values, 0.0, 0.5 and 1.0, respectively. In Fig. 2g) we see an example of an occlusion event. The idea of the novelty map is that everything is new to an infant when it comes to the laboratory which leads to large looking times at first trials (see Sect. II-B for details).

B. Training

The network was trained in three phases: the pre-training, the habituation and the test phase. In real habituation experiments the infant is habituated to some repeating stimulus until the looking time drops and the habituation is terminated. Then the infant is presented test stimuli. Our model was trained in a similar way except that a pre-training phase is needed. The pre-training models the visual experience of the infant with the world before coming to the laboratory. Without pre-training the network would only learn what is presented during the habituation phase from which no interesting results can be expected.

We pre-trained the network in the following way. A freely moving pixel moved back and forth at depth 0.5 as in Fig. 2h). This motion was halted and reinitiated with probability of 5% at each time step. Another (but non-moving) pixel was added or removed with probability of 1% (at each time step) at a random position. When this second pixel was present, its depth was chosen to be 0.5 or 1.0 with equal probability, so that the pixel either became an obstacle or an occluder. In the occluder case the stimuli were like in Fig. 2g). In the obstacle case a launching event occured when the first (moving) pixel collided with the second one (see Fig. 2a)). During pre-training the occluder or obstacle could be at any position in the visual field while during the habituation and test phases the stimuli were exaclty those shown in Fig. 2. The stimuli during the pre-training were constrained to "possible" ones, i.e. to stimuli that we would expect to occur naturally like direct launching, occlusion or just free motion. By doing this we model the infant's pre-experimental experience with the world. Varying the pre-training time allows us to look inside the development of the model - therefore, the pre-training time corresponds to the age of the infant. The novelty units were set to a random but constant binary vector. The network was pre-trained for 10^6 time steps and the weights were saved every 1000 time steps. Then we performed the experiments described below using these saved weights that represent the developmental progress of the network.

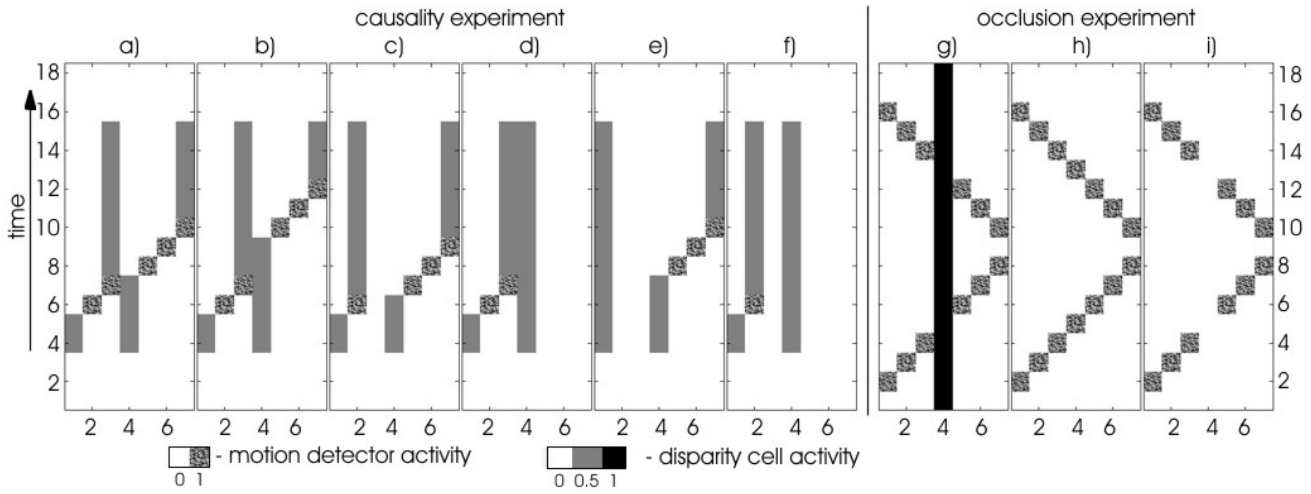

Fig. 2. Habituation (a,b,c,g) and test (d,e,f,h,i) stimuli. Motion detector and disparity layers are displayed on top of each other, textured pixels indicating motion detector activity and gray scale values indicating disparity cell activity. Note that whenever a motion detector is active the corresponding disparity cell's activity is 0.5. a) *direct launching* - first pixel launches the second pixel. b) *delayed launching* - second pixel moves off 2 time steps after collision. c) *launching-without-collision* - second pixel moves off without being touched. d) *no-reaction* - first pixel collides with second pixel which fails to move. e) *no-prior-movement* - second pixel moves without prior movement of first pixel. f) *no-reaction-no-collision* - first pixel stops before touching the second pixel which remains inert. g) *occluded trajectory* - pixel moves back and forth behind an occluder. h) *continuous trajectory* - pixel moves freely without occlusions or collisions. i) *discontinuous trajectory* - interrupted pixel motion

III. Perception of Causality: Modeling Experiment 1 in [6]

A. Description of the original experiment

Leslie [6] tested how 4.5- and 8-month-old infants perceive launching events. Infants were habituated to a cube starting to move and launching another cube which starts to move with the same speed as the first one while the first one stops moving after the collision (direct launching). Another group of infants was habituated to the same stimuli except that the start of the motion of the second cube was delayed (delayed launching). A third group of infants was presented a launching event without the first cube touching the second one. It stopped at some distance before but the second cube started to move off immediately just as if it had been launched (launching without collision). All infants were tested with basically two kinds of events: first cube moving, colliding with the second, stopping but without any reaction of the second cube (no-reaction). Alternatively, the second cube just started to move by itself without prior motion of the first cube (no-prior-movement). In Fig. 2a) - f) we see how these events have been presented to the neural network.

Leslie wanted to know whether the infant perceives the first cube as causing the second cube to move. The idea was that introducing a temporal or spatial gap between the two cubes would make the infants perceive two independent motions: one cube starting, moving, and stopping and then the second one doing the same - which are basically the test stimuli. Therefore, Leslie hypothesized that infants who were habituated with the delayed launching or launching-without-collision sequence would dishabituate less to the test stimuli than the infants exposed to the direct launching sequence.

B. Modeling procedure

After the pre-training the network was trained repeatedly with the direct launching stimulus in Fig. 2a) (alternately b) or c)) for 1000 time steps (habituation phase). Then, the total prediction error which is the sum of the prediction error over the 18 time steps of a stimulus (see Fig. 2) was calculated. After habituation, the test stimuli were presented once and the prediction error was calculated again. The direct and delayed launching habituations were tested with the no-reaction and no-prior-movement stimuli, Fig. 2d) and e), respectively, whereas the launching-without-collision habituation was tested with the no-reaction-no-collision and the no-prior-movement stimuli, Fig. 2f) and e), respectively. During the habituation and test trials the novelty units were switched to a different binary random but constant vector indicating the novelty of the laboratory environment (the novelty units don't play any role in this experiment but are important for the control case in the occlusion experiment below).

C. Results

The whole simulation was run 30 times. In Fig. 3 the results together with the results of the original experiment are shown.

Experimental result:
Looking times declined significantly from first to last habituation trials.
Model account:
During the habituation phase the stimulus was repeated over and over (1000 / 18 $\approx$ 56 trials). Thus, the network learned to predict the stimulus better, i.e. its prediction error dropped with time. As we relate the prediction error to the looking

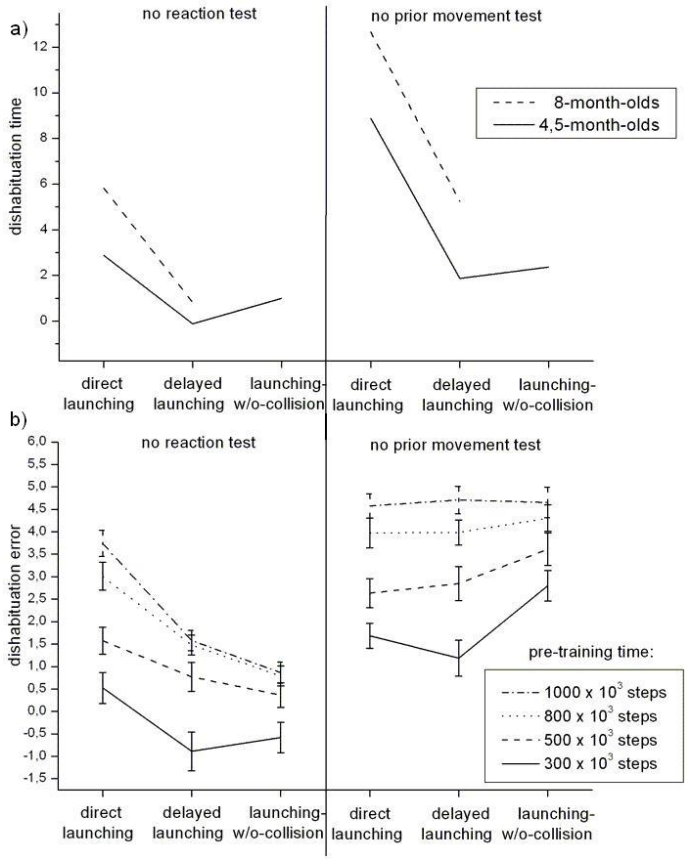

Fig. 3. Dishabituation time and error in a) experiment and b) model. Note that it can take negative values since it denotes a difference between looking times/errors.

time of infants this accounts for this result.

Experimental result:

The group habituated with the direct launching stimulus increased its looking time significantly more than the group habituated with delayed launching or launching-without-collision.

Model account:

Since the network was exposed to direct launching stimuli during pre-training already, there was not much to learn during the direct launching habituation. On the contrary, the other two habituation stimuli were more difficult to learn. Thus, the prediction errors of the last habituation trials were lowest in the direct launching case. Therefore, the networks ”dishabituation” to the test stimuli was higher after direct launching as compared to the other habituation cases. Unfortunately, this result doesn’t fit the data for the no-prior-movement test since the self-starting second object is equally surprising after each type of habituation.

Experimental result:

The no-prior-movement stimulus attracted significantly longer looking times than the no-reaction stimulus, regardless of the group.

Model account:

During habituation the network learned to expect the first pixel to start moving which happens in the no-reaction test but does not happen in the no-prior-movement test. Thus, in the no-prior-movement test, the network keeps predicting the first pixel to start moving which does not happen and yields a prediction error.

D. Model predictions

Although Leslie did not find any significant age effects, we see in his results, Fig. 3a), that the mean dishabituation time is higher for older infants (which could be random of course). But as we see in Fig. 3b) our model predicts that overall dishabituation times increase with age which is due to the fact that the habituation stimuli can be learned quicker and better after a long pre-training. Another prediction is that the direct launching results should be more similar to the delayed launching and launching-without-collision results for younger infants. This is due to the fact that the dishabituation errors are higher in the direct launching condition only because of prior exposure to direct launching stimuli during pre-training. But if the pre-training is short (young infants) then this effect vanishes as can be seen in Fig. 3b).

IV. Perception of occlusion: modeling Experiment 1 in [4]

A. Description of the original experiment

The experimenters tested 4- and 6-month-old infants and discovered an interesting effect in occlusion perception. They habituated the infants with a ball oscillating back and forth behind an occluder. Then two kinds of test displays were presented, both without any occluder: the first test display showed the ball continuously oscillating and the second test display showed the ball oscillating discontinuously, i.e. they used the same display as in the habituation but they removed the occluder such that it appeared that the ball oscillates but disappears behind an invisible occluder and reappears at its other end again. A separate control group was shown the same test stimuli but without prior habituation in order to control for some baseline preference for one of the test displays. In Fig. 2 we see the corresponding habituation, g), and test stimuli, h) and i), that we used in our model.

If infants perceive the ball as continuing to move behind the occluder during habituation then they should generalize their habituation to the continuously moving ball and show increased looking time at the discontinuous test display. However, if infants just learn the motion of the ball ”by heart” then they should generalize this perception to the discontinuous case which is identical to the ball’s motion in the habituation display. Thus, they should dishabituate more to the continuous display. The experimenters found that 4-month-olds show a preference for the continuous display but 6-month-olds dishabituate more to the discontinuous display (Fig. 4a)).

B. Modeling procedure

After pre-training the network was habituated repeatedly with the stimulus in Fig. 2g) for 1000 time steps. After

habituation the network was tested again once with the two stimuli Fig. 2h) and i) and the respective prediction errors, $E^{\text{exp}}_{\text{cont}}$ and $E^{\text{exp}}_{\text{discont}}$ were calculated. The prediction error was also calculated for the last habituation trial, $E^{\text{exp}}_{\text{baseline}}$, in order to be able to calculate the dishabituation time later.

Just as in the real experiment we also modeled the situation of the control group, i.e. we took the pre-trained weights, tested them directly without prior habituation and calculated again the respective prediction errors $E^{\text{control}}_{\text{cont}}$ and $E^{\text{control}}_{\text{discont}}$. In order to assess the dishabituation errors we presented the occlusion display 2g) once to the pre-trained network having still the old novelty units (see Sect. IV-C for the role of the novelty units). This prediction error, $E^{\text{control}}_{\text{baseline}}$, reflects how the network had learned so far.

Finally, we calculated the "looking preferences", P, for the experimental and control conditions, respectively.

$$P \equiv \frac{E_{\text{discont}} - E_{\text{baseline}}}{(E_{\text{cont}} - E_{\text{baseline}}) + (E_{\text{discont}} - E_{\text{baseline}})} \quad (4)$$

The difference between the test error and the baseline error (last habituation error) is what we call the dishabituation error which is analogous to the dishabituation time in real experiments. The original experiment the researchers used the same formula for the preference apart from the baseline values, i.e. they took the raw looking times to the test stimuli.

C. Results

The whole simulation was run 30 times. In Fig. 4 the model as well as the experimental results are shown.

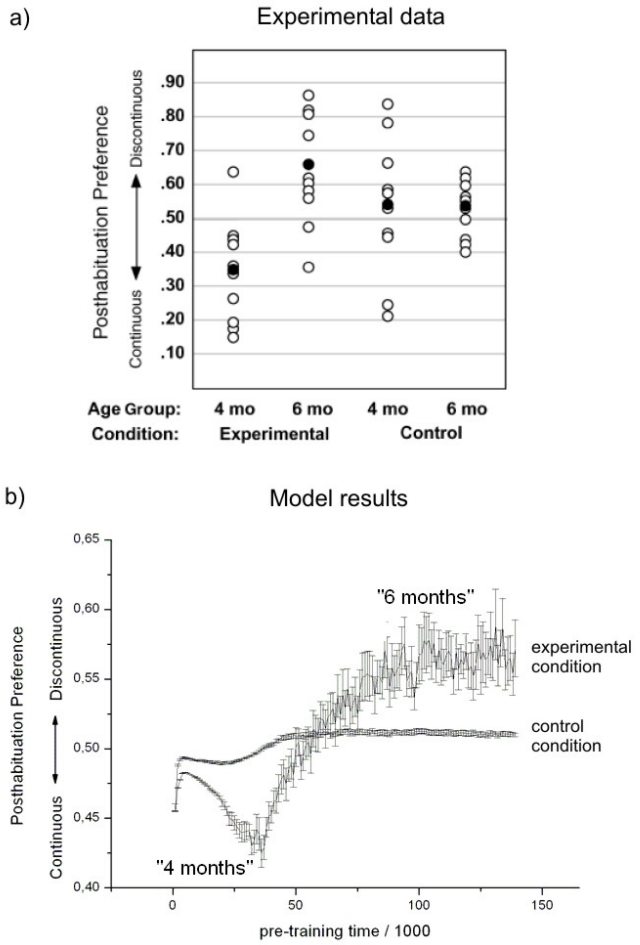

Fig. 4. Development of looking preferences for the continuous vs. discontinuous displays in a) experiment and b) model.

Experimental result:
4-month-old show a preference for the continuous display whereas 6-month-olds prefer the discontinuous display.
Model account:
In Fig. 4 we see that the preference first goes down to about 0.43 after 35000 trials and then increases until is saturates at 0.57 as a function of the pre-training time steps. Since the pre-training time corresponds to the infant age we get a similar result as observed experimentally. The preference curve can be explained in the following way: After 0 time steps the network did not predict any output at all. Therefore, the test displays are equal to the errors which are both large and to a preference around 0.5. After 35000 time steps the network learned to predict the trajectory of the pixel except at the occlusion position to some extent. It basically learned the pixel motion "by heart". This is the same case as before but only with smaller errors in total such that the failure to predict the continuous trajectory at the occlusion position gained more weight (preference for the continuous display increased). After the network has been exposed long enough to pre-training stimuli it learned to predict continuous trajectories and also that a trajectory is suppressed whenever there is an occluder. Therefore the network predicts the continuous display successfully but fails to predict the discontinuous one since it expects the pixel to move on in absence of an occluder (preference for the discontinuous display).
Experimental result:
The control group showed no preference for either test display.
Model account:
This is due to the general novelty of the laboratory environment. As the network was shown test stimuli without prior habituation it could not learn the value of the new novelty units which increases the total prediction error. Of course the network did have a baseline preference for the discontinuous ("unnatural") display after a long pre-training time. But this difference was small as compared to the prediction errors of the novelty units and reduced the total preference to around 0.5. Therefore we suspect that there may be a baseline preference also in infants but the novelty of the laboratory environment makes any preference non-significant.

Object permanence
In contrast to real experiments with infants we can examine *how* the system achieves its performance. First it learns to predict freely moving pixels in either direction. But then, when an occluder is present, it simply learned to suppress

the output of the motion detector layer at the position of the occluder. This is done by a few hidden units that are mainly driven by the activity of the disparity layer cell that represents the existence of an occluder. Whenever these hidden units are active they suppress the activity of the motion detector output layer at the same position where the occluder occurred by feeding in strongly negative connection weights to them. In this way the other hidden layer units *do* predict a moving pixel to continue its motion at the occluder position but the actual output of this prediction is suppressed by the former hidden units. Therefore, the network continues to represent the motion of the pixel even though there is an occluder which is exactly what object permanence means. We did not foresee this capacity of the network to develop, it just emerged as a solution to the occlusion problem.

D. Model predictions

As we can see from Fig. 4b) the model predicts that infants' preference should be similar to the model curve if the experiment will be performed for other ages as well.

V. Discussion

Leslie tested how infants perceive two object movements, one causing the other to move. They found that the perception of "causality" is disrupted if a temporal or a spatial gap is introduced between cause and effect. A gap would supposedly lead to the perception of two separate, independent movements whereas a direct launching is supposed to be perceived as two conjoined movements.
In the model there are no two separate movements. There are just sequences of input vectors. But the model learned that whenever a pixel approaches and touches a resting pixel it stops and makes the latter move in the same direction. Specifically, it learned that the second one moves off immediately (direct launching). Therefore, the model "dishabituates" more when presented a no-reaction test as compared to the condition where it has been habituated to delayed launching. Thus, there is no need for innate force notions of motion analysis modules as has been proposed by Leslie [7] whose experiment we modeled. Our model accounts for his data while suggesting that causality can be learned by merely observing the visual environment, registering statistical regularities and trying to predict them.

In the occlusion experiment [4] the researchers wanted to know whether infants are able to perceive a continuous trajectory although partially occluded. They found that 6-month-olds seem to perceive the trajectory veridically but 4-month-olds do not - they rather seem to perceive two distinct sections of the trajectory. This makes sense with regard to our model. Before learning enough about the continuity of a trajectory the model/infant can not know that it must be continuing behind an occluder. Only after being able to successfully predict a free trajectory the road is free to follow an object behind an occluder with the "mind's eye". This is exactly what happened in the model and object permanence emerged as discussed above.

In summary, we presented a simple framework - a network trying to predict its future inputs - that was able to develop representations for causality and occlusion perception as well as object permanence. It has learned about the continuity of object motion, about solidity and reaction of objects to contact. Therefore, there is no need to postulate innate principles as had been suggested by [13]. In our model we show that all these properties can be simply derived from statistical properties of visual input sequences.

One drawback of the model is that it uses backpropagation of an error signal which is not biologically plausible if we want the framework to be a model of the infant brain. This is certainly a weakness but can be overcome by using a more plausible network, e.g. a dynamic reservoir network with spiking neurons that have been shown to be able to perform prediction tasks [5].

A major topic of future work will be to provide the model with an active representation of occluded scenes. Although this model shows object permanence, i.e. it represents occluded inputs, it cannot do so for more than one time step. Beyond that we also hope to extend our explanations to other important phenomena like object unity and perception of support and containment.

References

[1] R. Baillargeon, *Object permanence in 3 1/2- and 4 1/2-month-old infants*, Developmental Psychology, vol. 23, no. 5, pp. 655-664, 1987

[2] J.L. Elman, *Finding structure in time*, Cognitive Science, vol. 14, pp. 179-211, 1990

[3] D. Hume, *A treatise of human nature*, Ed. L A Selby-Bigge (Oxford: Clarendon Press), 1740/1978

[4] S.P. Johnson, J.G. Bremner, A. Slater, U. Mason, K. Foster, A. Cheshire, *Infants' perception of object trajectories*, Child Development, vol. 74, pp. 94-108, 2003

[5] A. Lazar, G. Pipa, J. Triesch, *Fading memory and time series prediction in recurrent networks with different forms of plasticity*, Neural Networks, vol. 20, pp. 312-322, 2007

[6] A.M. Leslie, *The perception of causality in infants*, Perception, vol. 11, pp. 173-186, 1982

[7] A.M. Leslie, *ToMM, ToBY, and Agency: Core architecture and domain specificity* In L.A. Hirschfeld & S.A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* New York: Cambridge University Press. 1994

[8] J.M. Mandler, *The foundations of mind. Origins of conceptual thought*, Oxford Series in Cognitive Development, 2004

[9] D. Mareschal, K. Plunkett, P. Harris, *A computational and neuropsychological account for object-oriented behaviors in infancy*, Developmental Science 2:3, pp. 306-317, 1999

[10] A. Michotte, *The perception of causality*, (New York: Basic Books), 1963

[11] Y. Munakata, J.L. McClelland, M. H. Johnson, R. S. Siegler, *Rethinking infant knowledge: Towards and adaptive process account of successes and failures in object permanence tasks*, Psychological Review, vol. 104, no. 4, pp. 686-713, 1997

[12] M. Schlesinger, M.E. Young, *Examining the role of prediction in infants' physical knowledge*, In Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society (Boston, MA, USA), pp. 1047-1052, 2003

[13] E. Spelke, *Initial knowledge: six suggestions*, Cognition, vol. 50, pp. 431-445, 1994