

Modeling the Transition from Bottom-up to Top-Down Gaze Control Strategies in the Context of Gaze Following

Hector Jasso¹, Jochen Triesch^{2,3}, Christof Teuscher⁴, Gedeon Deák²

¹ Department of Computer Science and Engineering, University of California, San Diego, CA, USA (hjasso@cs.ucsd.edu)

² Department of Cognitive Science, University of California, San Diego, CA, USA (triesch@cogsci.ucsd.edu, deak@cogsci.ucsd.edu)

³ Frankfurt Institute for Advanced Studies, Max-von-Laue Str. 1, 60437 Frankfurt am Main, Germany

⁴ Los Alamos National Laboratory; CCS-1, MS-B287, Los Alamos, NM USA (christof@teuscher.ch)

Introduction

Visual attention is commonly characterized as driven by a combination of bottom-up and top-down signals [1]: Bottom-up signals are those resulting from a basic neuronal analysis of the visual input, based on features such as color, orientation, and intensity. Top-down signals, in turn, are characterized as task-driven and involving higher neuronal processing. While models of visual attention have concentrated on either bottom-up or top-down signals, there is a growing interest in schemes that integrate both in a single architecture [2, 3].

We present a developmental model of gaze following that integrates both signals: Object saliencies are derived directly from the visual input and thus represent bottom-up signals. Top-down signals, on the other hand, come from an indexing of the other person's head and eyes direction to paths along their line of view. This indexing, which allows the infant to use the other person's gaze as an indication of possible object locations outside its own (limited) field of view, is gradually learned through reinforcement learning [4]. This results in a transition from bottom-up to top-down gaze control strategy.

A Model of Gaze Following

Modeling the Environment. Infant and caregiver are positioned facing each other (see Figure 1, left). Objects can be placed anywhere except in the same location as the infant or caregiver. Time is discretized into time steps, each one corresponding to about 1 second. This roughly corresponds to the time it takes to shift gaze between any two positions.

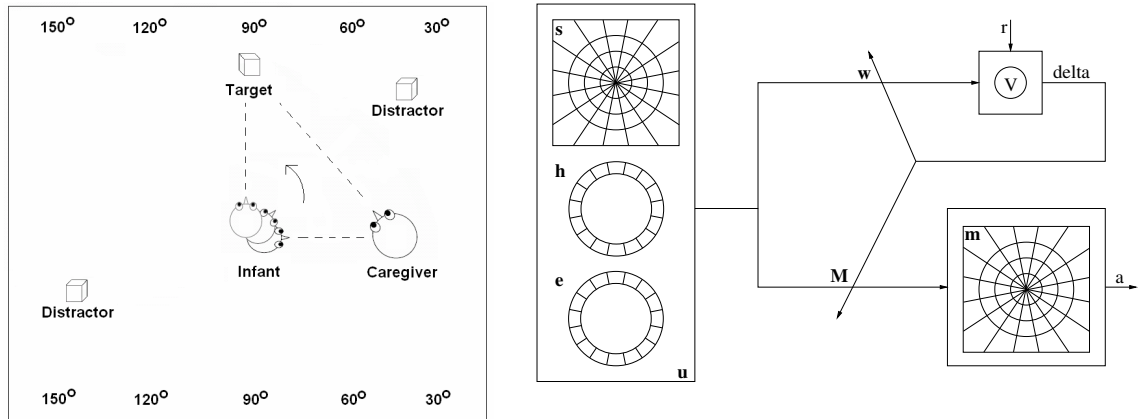


Figure 1. Left: Room setup. Infant following gaze in the presence of distracters. Right: Details of the actor-critic reinforcement learning model. Features calculated from the *Saliency Map* s , *Caregiver Head Direction* h , and *Caregiver Eyes Direction* e are weighted and added for each possible action. The action is selected using a softmax function.

Infant Visual System. The infant's visual input is processed by three different systems (see Figure 1, right): *Saliency map* ($s(t) = [s_1(t) \dots s_{64}(t)]$): Indicates the presence of visual saliency in a body-centered coordinate system with 64 different regions in space, along 16 heading ranges (slots) and 4 depth ranges. *Caregiver head direction* ($h(t) = [h_1(t) \dots h_{16}(t)]$): Indicates 16 possible caregiver head directions as perceived by the infant. *Caregiver eye direction* ($e(t) = [e_1(t) \dots e_{16}(t)]$): Indicates 16 possible caregiver eyed directions as perceived by the infant. The infant's visual input is calculated as $u(t) = [s(t) \ h(t) \ e(t)]^T$

Reinforcement Learning Model. The heart of the model is an actor-critic reinforcement learning algorithm [5, 4] (see Figure 1, right). The *critic* approximates the value of the current state using the infant's visual system and a set of weights $w(t) = [w_1(t) \dots w_{64}(t)]$, as $v(t) = w(t) u(t)$. The *actor* specifies the action to be taken, directing the infant's attention to one of 16 possible different headings and one of four possible different depths, for a total of 64 possible attention points. A particular action is chosen probabilistically according to a softmax formula. Action value parameters are calculated as $m(t) = M(t) u(t)$, where $M(t)$ is a weight matrix with as

many columns as there are input features and as many rows as there possible actions. The values of \mathbf{w} and \mathbf{M} are updated according to a delta rule.

Experiments and Results

Initialization: Before training, all weights (elements of \mathbf{w} and \mathbf{M}) are initialized to zero. **Training:** Then, throughout a number training trials, the infant acts and learns according to the reinforcement learning scheme described above. In each training trial, which lasts 10 time steps, several objects are positioned within a perimeter of 1.3 m from the infant, with the caregiver’s eyes directed to one of them, for all trial steps. The caregiver’s head direction is slightly offset from the eyes direction, to reflect a naturalistic setting. **Testing:** After every 5,000 training trials, 500 test trials are carried out (learning is disabled during testing): An object is positioned within the infant’s field of view, but the caregiver looks at an “imaginary object” positioned behind the infant, and at a different angle from the real one. Figure 2 shows the resulting emergence of top-down influences in visual attention.

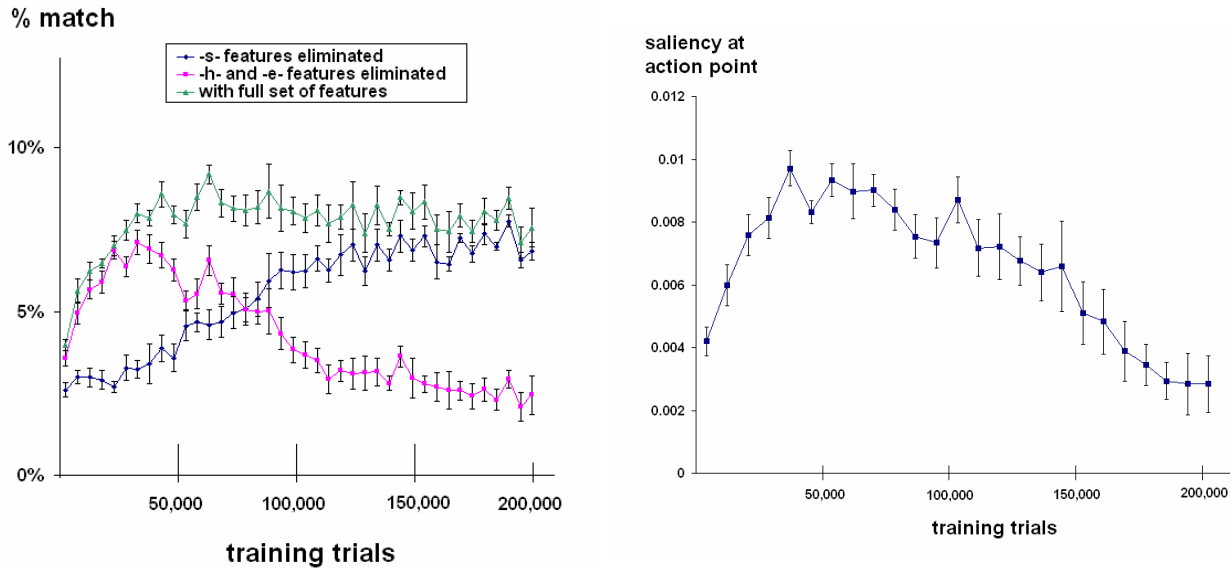


Figure 2. Emergence of top-down influences in visual attention. (Standard errors shown for 7 repetitions) Left: Percentage match between the selected action and the action with highest action value (from \mathbf{m}). At first, visual saliencies have a strong impact on the action selected. With learning, the influence is complemented by the top-down influences of the other person’s head and eye direction. Right: As the model learns to direct its attention to visual saliency, the saliency at the location where attention is directed (action point) grows. Then, as the model follows gaze instead to the “imaginary object” behind the infant (see text), this saliency decays.

Conclusions

The developmental model presented can help understand the transition from bottom-up influenced to progressively more top-down influenced looking behavior in humans. In particular, it shows how experience can be the driving force behind the transition. This model was developed within the MESA project at UCSD¹ [6]. We believe that such a developmental approach is well suited to explore further questions about the integration of top-down influences in visual attention such as imitation and theory of mind [7, 8].

References

- [1] Itti, L., & Koch, C. (2001) Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2:194-203.
- [2] Zelinsky, G., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2005) The role of top-down and bottom-up processes in guiding eye movements during visual search. *Nineteenth Annual Conference on Neural Information Processing Systems (NIPS 2005)*, Vancouver, B. C., Canada.
- [3] Navalpakkam, V. & Itti, L. (2005) Modeling the influence of task on attention. *Vision Research* 45:205-231.
- [4] Sutton, R.S., and Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.
- [5] Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA, USA: MIT Press.
- [6] Fasel, I., Deak, G. O., Triesch, J. and Movellan, J. R. (2002) Combining embodied models and empirical research for understanding the development of shared attention. *Second International Conference on Development and Learning, (ICDL'02)*, Cambridge, MA, USA.
- [7] Triesch, J., Jasso, H., Deák, G. O. (2006) Emergence of mirror neurons in a model of gaze following. *Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, USA.
- [8] Jasso, H., Triesch, J. (2006) Using eye direction cues for gaze following – A developmental model. *Fifth International Conference on Development and Learning (ICDL'06)*, Bloomington, IN, USA.

¹ <http://mesa.ucsd.edu>