Active Long Fixation Correlates with the Formation of Long-Term Memory

Jin-Hwa Kim (jhkim@bi.snu.ac.kr)
Interdisciplinary Program in Cognitive Science, Seoul National University
Seoul, 151-742, Republic of Korea

Byoung-Tak Zhang (btzhang@bi.snu.ac.kr)
School of Computer Science and Engineering, Seoul National University
Seoul, 151-742, Republic of Korea

Abstract

The application of eyewear accelerates the study on the eye movement, for the eye movement is a non-invasive and convenient indicator of the brain activities. We investigate the eye movements of the subjects watching the kids video. First, we conduct the long-term memory test of whether fixation duration correlates with long-term memory. Second, we classify its visual constraints into alert and no alert types. As a result of the test, the fixation duration itself is not decisive. However, the long fixations which are actively engaged with alert type movie clips statistically have higher scores, while the short fixations do not. Finally, we propose a simple computational model using the linear regression of two significant features, saliency scrutiny and fixation duration. It may provide an explanatory way to the efficient memory mechanism for the life-long sequences.

Keywords: Human Factors and Human-computer interaction; Cognitive Science; Memory; Statistics; Event cognition; Eye tracking

Introduction

The visual information is easily affected by the moment-to-moment environmental changes, heuristic but robust compensation strategies are required. Hence, how the brain processes the visual information provides the profound way to study the mechanism by which the brain precisely and efficiently processes the most dynamic and enormous sensory information.

There are a lot of studies on the computational modeling for the visual information, which include the visual fragment completion, the scene or object classification and recognition, and object tracking. These research topics often tend to focus on the objective for each task, not on the implementation of the method how the brain deals with the visual information. As a result, the computational approach to the modeling for the visual information processing of the brain is gradually changed to the optimization problem, which hinders the understandings of the human-level information processing abilities.

Particularly, the object tracking seems to describe how we pay attention to an interesting object, however, the eye movement, mostly controlled by the oculomotor system, complicates with how the brain works for the acquisition of the visual information (Henderson, 2003). For instance, in the fixation state, the human eyes only recognize the small portion of the whole sight. If you read this paper from an 8-inch distance fixing on one particular letter, you cannot read outside of next two words or about ten letters which are presented in the para-fovea. Since the brain is well-known for its parallel processing on the neural circuits, this sequential notion of eye movement for the visual system would be inefficient for information processing. Therefore, we should notice that the experimental results of covert attention and shifting receptive field are also the important subject to study (Zirnsak & Moore, 2014).

The studies on the reading eye movement, which are relatively well studied by psychologists and neuroscientists, reveal that fixation duration is related to the presence of the cognitive process, such as observing its correlation with linguistic attributes (Inhoff & Rayner, 1986; Rayner & Duffy, 1986).

We investigate the characteristics of eye movements on the video stimuli focusing on the formation of long-term memory using recognition test. In this study, we focus on the characteristics of the fixation duration as the evidence of the cognitive process to memorize. Moreover, as the movie clips which potentially induce emotional arousal are known to increase the recognition of the previously watched movie clips (Cahill et al., 1996; Cahill & McGaugh, 1998), we will see if the arousal effect is asserted by the duration of fixation. Finally, we propose a linear regression model based on the active features of eye movement, saliency scrutiny and fixation duration.

Materials and Methods

For this study, we prepared the video material *Pororo Season 3*, which is a famous kids video in Republic of Korea. In this video, there are artificial 3D-rendered characters and other objects who have distinctive traits, so visual features are easily captured by eye movements. *Pororo Season 3 DVD 1* contains 13 consecutive episodes, each with a different single storyline. The total playing time is 67 minutes and 50 seconds.

We recruited 11 participants with normal vision (6 males, 5 females; 23-31 age), who are graduate students in Seoul National University, voluntarily participated in the study. For the stability of collecting data and the eye gear requires appropriate head circumference, we chose the target participants for this study. All participants had not experienced a brain damage or a behavioral disorder. The participants were first time viewers of the video, *Pororo Season 3*. To prevent distraction, each participant took a set of tests for the two-split video, one

is about 32 minutes of the first half and the other is about 36 minutes of the later half, each on the other day. Later then, we merged two parts into one manually, in a way that the results do not overlap.

Participants watched the kids video in the room which has the experimental settings. The room is about 3 square meters surrounded by the opaque curtains. On the side of the room, a wide-screen HDTV (1920x1080 resolution, 885 mm x 500 mm, 16:9 ratio) was installed, and 2.1 channel speakers. Participants were guided to sit on a comfortable sofa in front of 1.7 m from the TV screen.

Concurrently, the eye movements and the user-perspective scenes were recorded by *Tobii Glasses eye tracker*, the corneal reflection based system with a sampling rate of 30 Hz. We used the *I-VT algorithm* as the fixation filter (system default), which classified fixations with the velocity threshold of 30 degree per second. Usually, the saccadic eye movements are discriminated with low velocities (less than 100 degree/second) and high velocities (higher than 300 degree/second), in which the velocity-based classification is a simple and reasonable approach (Salvucci & Goldberg, 2000).

We prepared the controlled memory test for each participant. We conducted this study 3-4 months after the first session of watching (the intervals are not consistent due to schedule conflicts). A memory test consisted of total 20 movie clips; 8 movie clips for long fixations, another 8 movie clips for short fixations, and the remaining 4 movie clips for the control, which were not seen in the previous study. In detail, the lengths of all movie clips were each 3-second long. *The long fixation sequences* were randomly picked from each participant's data containing the fixation longer than 1400 ms in the middle of the movie clip. *The short fixation sequences* were randomly picked from each participant's data containing the fixation shorter than 300 ms. The 4 control movie clips were randomly picked from the other season of *Pororo* series, *Pororo Season 2*.

Each participant identified 20 movie clips that were sorted randomly. Each participant gave a score for each movie clip between 1 and 5, which is an integer, depending on the assurance of whether he or she saw the movie clip before or not. Notice that, in this study, 1 is the lowest score, not 0, which means that the movie clip is surely not seen.

Fixation Duration

In the video watching task, we anticipated the characteristics of fixation duration to be different from those in the reading task. As expected, the fixation duration was changed more drastically, up to 10 seconds during watching the video. These changes were partly caused by the sequential changes of the visual stimuli and the fluctuated responses of the oculomotor system and cognitive processes.

In Figure 1 the sequences of frames which received more than 2 seconds of the fixation duration from at least 3 different participants are shown. We set the threshold to reduce the interpersonal variation. We got all 41 sequences across 1 hour 7 minutes 50 seconds length of the material. For the review, we chose 10 typical sequences. Each row represents an independent sequence and each column shows a single frame. The time interval between the frames is 0.5 seconds. The colored dots represent the fixation positions, whose durations are longer than 2 seconds. The same color means the same participant. Four different types of the sequences are listed as *alerted* (3), *successive* (3), *stationary* (3), and *unclassified* (1) for the review. The classification is conducted by one of authors and two other colleagues. Every sequence used by this study is classified by agreement. If there is a conflict or not fit to the major types, then the sequence is classified as *unclassified*.

Figure 1: The sequences of frames which received more than 2 seconds of fixation duration from at least 3 participants. First 3 sequences show the *alerted* type, then next 3 sequences show the *successive* type, then next 3 sequences show the *stationary* type, and then the last sequence shows the *unclassified* type of visual constraints for the review. For the description of those sequences, see the text in Section *Fixation Duration*.

The sequence of the *alerted* type was classified because the scene implied an unusual and, potentially dangerous or difficult situation, which may introduce a mental arousal. The sequence of the first row demonstrates an urgent moment that

a huge snowball is about to roll down the hill, which was previously rolled up by the robot, *Rody*. Second shows that *Pororo* has been fishing at the ice hole, but what he caught was *Shark*, a naughty character. Third shows that *Eddy* rolled his eyes to kick his ball avoiding the opponent *Pororo*.

The *successive* type shows that the fixation duration extends across more than 2 different scenes. Because the location of the target object is not changed or changed within the range of a foveal or central vision, 2-5°, the fixation holds its position (McMorris, 2014). The fourth sequence shows the closed-up characters are serially shown up in the center of the screen. The fixations duration of the fifth and sixth sequences extends across different scenes have different visual configurations.

The *stationary* type most clearly shows the characteristics that the indifferent scene maintains while the target object moves a little bit or even does not move. For there is not a particular event or a change of the scene, the participants tend to fixate their gazes. See the seventh through ninth sequences.

The sequence of the *unclassified* type takes various forms. The tenth sequence shows that *Pororo* and *Crong* just jumped out of the shoulder of the magician dragon *Tongtong*, who is flying in the sky. Two sunflowers spring *Pororo* and *Crong* into the sky in multiple times.

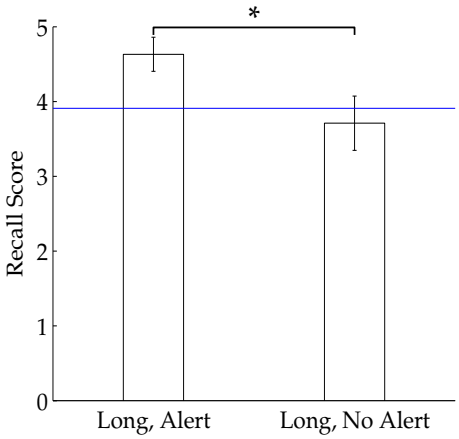

Figure 2: The memory test result for the long fixations which is on the *alert* movie clips or the *no alert* movie clips. The blue horizontal line indicates the mean scores of the long fixated movie clips. Error bars indicate ± 2 SEMs.

Long-Term Memory Formation

In the studies of reading eye movements, as noted before, the fixation duration is a good indicator for information processing. Even so, when we applied to a watching task, we have to be more cautious that the long fixation itself was not always induced when the internal state of a subject is affected. Furthermore, the sequences that received a long fixation did not decisively guarantee the quality of information processing nor its specificity. When we carefully look into the sequences in Figure 1, the fixations on *successive* or *stationary* type sequences can be interpreted as looking passively or even blankly.

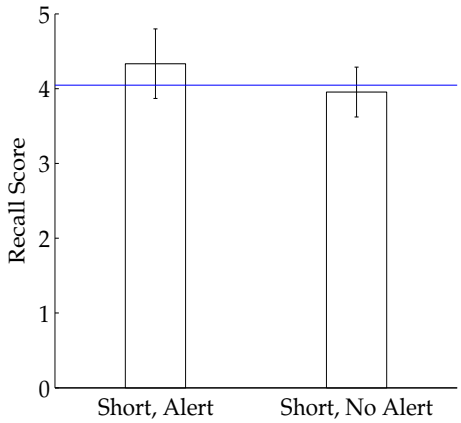

Figure 3: The memory test result for the short fixations which is on the *alert* movie clips or the *no alert* movie clips.

A stimulus is memorized by the constructive activities, which is a series of stimulations, giving attention, and acquisition. Yet the corresponding responses are not deterministic by the stimulus for the uncertainty of environmental perturbations or the complexity of internal states. Therefore, the formation of memory is the result of the cognitive process of response rather than the response itself. We need to discern the reliable indicator of the cognitive process with a sufficient care.

Recognition Test

We defined two types of fixation as long and short fixations. The long fixation was defined by the gaze holding its position after fixation filtering, which is longer than 1400 ms. The short fixation was defined by the one that is shorter than 300 ms. Each participant assessed in a recognition test rating 20 movie clips, containing 8 long fixated movie clips, 8 short fixated movie clips and 4 not-seen movie clips, which were not seen previously. Contrary to our expectations, the scores for the short fixated movie clips were not much different from the scores for the long fixated movie clips ($p = 0.5051$). This result makes sense considering that the long fixation is an attentive response, which can be actively or passively motivated by a reciprocal process in visual system.

The relationship between the emotional arousal and the formation of long-term memory was known for the studies in neuroscience (Cahill et al., 1996; Cahill & McGaugh, 1998). Although it is not included in this paper, the arousal effect on the recognition test was verified. *Alert* movie clips significantly got higher scores than *no alert* movie clips with the p-value of 0.0077 ($p < 0.01$). On top of this, more detailed analyses are conducted with regard to this.

Figure 2 shows the memory test result for the long fixation which was on the *alert* movie clip or the *no alert* movie clip. The *alert* movie clips included the emotional arousal events, which was previously defined. This classification is

rather definite because the content is an animation video for children. The numbers of ratings from 11 participants were 19 for the *alert* and 69 for *no alert*. The difference between the two mean scores for the long and short fixation types was significant. The p-value of two-sample t-test was 0.0104 ($<$ 0.05).

Figure 3 shows the memory test result for the short fixation which was on the *alert* movie clips or the *no alert* movie clips. The numbers of cases from 11 participants were 21 and 67, respectively. The p-value of two-sample t-test was 0.2484 ($>$ 0.05). In the short fixation cases, there was no significance between the two content types, *alert* type and *no alert* type.

Gaze Variations

The difference of the eye movements between the *alert* movie clips and the *no alert* movie clips for the long fixated was observed by the variation of gazes. Actually, the gazes keep moving in the fixation state, because eye balls are fixed by twitching extraocular muscles. However, the most contribution to that variation was a smooth pursuit. In the smooth pursuit, the eyes were following a slowly moving object or interesting features. Because our experimental settings did not and can not precisely define the smooth pursuit, all the eye movements slower than 30 degree per second were candidates.

To get the variation of the gazes excluding the variation from the smooth pursuit, we used the window sliding technique. Varying the size of a time window, we moved the time window on the time series by 1/30 second in every step, and took the median of the variations. Figure 4 shows the significant level changes according to the change of the window size. If the window size was smaller than 200 ms, p-values were lower than 0.05. In the long fixated group, the *alert* movie clips tended to have smaller gaze variations than the *no alert* movie clips have, whereas the short fixated group did not show these differences.

Computational Model

Based on the study of bottom-up attention (Koch & Ullman, 1985), the computational model was implemented by the saliency map-based approach (Itti, Koch, & Niebur, 1998). This model used the visual information, i.e., intensities, color opponencies and orientations, as a source to measure the conspicuity (Parkhurst, Law, & Niebur, 2002). While reflecting physiological evidences for the basis mechanisms of visual information processing, active selection to increase the information gain, which is depicted as "best question to ask" (Reinagel & Zador, 1999; Zetzsche et al., 1998), the implement had both processing efficiency and robustness to noises. It took an image as input, and it returns the saliency map.

Using the saliency map from the computational bottom-up attention model, we examined the association between the gaze fitness to the model and the score. The fitness function is defined as below. We used the SaliencyToolbox with default parameters for getting the saliency map to analyze, and aligned the gaze coordination to the area of the saliency

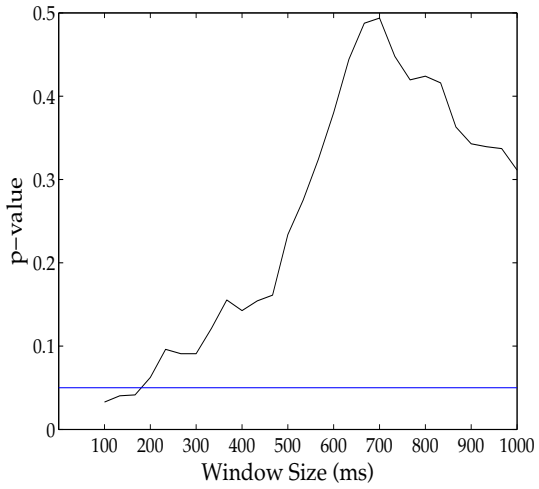

Figure 4: The significant levels of gaze variations with regard to the window sizes. The blue horizontal line shows the 5% significant level.

map, which had a smaller size after sub-sampling (Walther & Koch, 2006).

$$X_{saliency}^{(i,j)} = \sum_{t=1}^{T} \mathcal{S}_{x_t,y_t}^{(i,j,t)} \tag{1}$$

In Equation 1, x_t and y_t represent the mapped gaze coordination at t time in a fixation duration. $\mathcal{S}^{(i,j,t)}$ is a saliency map which is the output of the model for a given screenshot at t time in the duration of a movie clip, that was given to the participant i for the j-th movie clip of the recognition test.

We fit the parameters for the linear regression model for the scores, which was defined by Equation 2,

$$Y_{score} = \mathcal{B}_0 + \sum_{l \in \mathcal{L}} \mathcal{B}_l \cdot X_l \tag{2}$$

, $\mathcal{B}$s are the coefficients of the model, and $\mathcal{L}$ is a set of features as

$$\mathcal{L} = \{duration, saliency\}. \tag{3}$$

$X_{duration}$ is the fixation duration in second, and $X_{saliency}$ is defined by Equation 1. We used only the data of the long fixations, because the short fixated movie clips did not show the significant differences on the scores for the duration and the gaze fitness to the saliency map.

Table 1: Estimated coefficients of the linear regression model.

Coef.	Estimate	SE	tStat	pValue
$\mathcal{B}_0$	4.8241	0.48622	9.9217	7.39e-16
$\mathcal{B}_{duration}$	-0.42214	0.18627	-2.2663	0.025974
$\mathcal{B}_{saliency}$	0.032642	0.015957	2.0455	0.043893

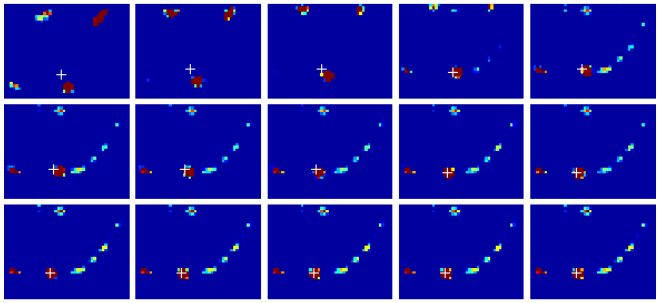

Figure 5: A series of saliency probing, one of those got the highest recognition score, is shown. From the left to the right, and from the top to the bottom, time goes with an interval of 100 ms (total length = 1500 ms). Each box represents a saliency map, the more reddish represents the higher saliency and the more bluish represents the lower saliency. The white cross represents the gaze position of the participant. As time goes, the eye gaze tracks down the highest saliency point (first row) and moves slowly in accordance with shifting of saliency map due to displacement of scene camera (second and third rows).

Table 1 shows the estimated coefficients of the linear regression model for the prediction. All shown parameters were statistically significant, though the gaze variation was excluded for the fitting due to its unexplainable for the scores. Whereas, the gaze fitness reasonably explained the score with the p-value 0.043893. The estimated value for the gaze fitness was positive, which means the score positively correlated with the gaze fitness in the model. Figure 5 shows a series of saliency probing for the gaze fitness. For details, please refer to the caption of that.

Interestingly, the fixation duration, which was longer than 1400 ms, negatively correlated with the score, though the standard error of that was relatively high.

The assessment of the model is shown in Table 2. The number of observation was 88 since each of 11 participants rates 8 long fixated movie clips, respectively. The other models, like logistic regression and non-linear regression, are also examined, but they did not explain better than the linear model.

Table 2: Assessment of the linear regression model.

Attribute	Value
# of observations	88
Error degree of freedom	85
RMSE	1.34
R^2	0.106
Adjusted R^2	0.0847
F-statistics vs. constant model	5.02 (p-value = 0.00866)

Discussions

After all, what is the meaning of the long fixation durations on the video stimuli? Despite the fact that it could be a latency time to process the cognitive information of the visual stimuli, in the other perspective, it could be waiting time to the potentially salient moment on that eye position. The waiting on the prospective location for a dramatic change or a new event is an efficient way of information processing.

Though the long fixation itself does not describe the presence of the cognitive process to memorize, in the condition of the long fixation, the arousal effect is remarkable compared to the others in the condition of the short fixation. The short fixation is relatively complex and vague, the arousal effect on the short fixation might be not observable.

The characteristics of the eye movement on the arousal effect are probed by the statistical method and the computational attention model. First of all, as we report in Section Gaze Variations, the gaze variation on the alerted movie clips is smaller than those on the other case, when the window size is shorter than 200 ms, to minimize the variation from the slow pursuit. Yet, the gaze variation does not have a statistical power to predict on the recognition score, which is the measurement for the long-term memory. We conclude that the gaze variation partly explains only for the arousal effect, not for the further cognitive process, memorization.

Second, we establish the linear model to predict the score using the computational attention model in addition to the fixation duration. For the short fixated movie clips, we cannot find the predicting variables, so, only the observed data of the long fixated movie clips are used. This is the backward study of Itti (2006)s work (Itti, 2006), which investigates the fitness of the model to human gaze, whereas our study uses the model to evaluate the attentiveness of human. In machine learning, Zou et al. (2012) reported that stimulated fixations help to capture the useful invariant features in the image recognition task (Zou, Zhu, Ng, & Yu, 2012). This gives a hint that how it works from a computational perspective.

Although we formulate the simple method via summation of saliency scores, the method shows the significant level for the prediction. The representation of saliency map is also found in the neuronal structures (Fecteau & Munoz, 2006), which have a peak activity in the physically salient position. And, for the fixation durations longer than 1400 ms, the longer fixated movie clips are the lesser recognized. This suggests that the power of saliency probing tends to diminish as the fixation duration is behind time.

The value of the adjusted R^2 for the model reminds that the information of eye movement should be carefully used to estimate the cognitive process. There is the limitation of the analysis coverage, because the majority of the eye movement is the short fixated ones.

Therefore, the investigation is not complete. We look forward to having more distinct features. A temporal modeling using the salient detectors (Marr & Hildreth, 1980; Canny, 1986) and the optic flow (Koenderink, 1986) may be

a promising option. Moreover, the cognitive modeling for the scene comprehension, which is related to the emotion-based reaction, guides us into the diffent level of a methodological stage.

As discussed, the long fixations are constrained by the visual content of the video stimuli, and with regard to that, the eye movements are planned in a reciprocal manner (Zhang, 2013). The estimation of the eye movement is viable on top of a reciprocally anticipatory model (Robert, 1985). The serial information of the fixation positions can be used to parsimoniously select the portion of visual features on the scene for an application using the features, and, it can be a methodological breakthrough for the cognitive modeling on the endless stream of the visual information.

Conclusions

The visual constraints, which induce long fixations, are summed up as three distinctive types: alerted, successive and stationary. We found the arousal effect only for the long fixated movie clips of the alerted type. The small gaze variation for the long fixation indicates an active response to the arousal stimuli. Though, the gaze variation does not help to estimate for the recognition score, among long fixations, fixation duration and saliency-probing activity are significant for that. We can appreciate this computational model within the embodied cognitive framework with the perception-action cycling.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2010-0017734-Videome), supported in part by ICT R&D program funded by the Korea government (MSIP/IITP) (10035348-mLife, 14-824-09-014, 10044009-HRI.MESSI).

References

Cahill, L., Haier, R. J., Fallon, J., Alkire, M. T., Tang, C., Keator, D., et al. (1996). Amygdala activity at encoding correlated with long-term, free recall of emotional information. *Proceedings of the National Academy of Sciences*, *93*(15), 8016-8021.

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences*, *21*(7), 294 - 299.

Canny, J. (1986). A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*(6), 679–698.

Fecteau, J. H., & Munoz, D. P. (2006). Salience, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences*, *10*(8), 382–390.

Henderson, J. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, *40*(6), 431–439.

Itti, L. (2006). Quantitative modelling of perceptual salience at human eye position. *Visual Cognition*, *14*(4-8), 959–984.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, *4*(4), 219–227.

Koenderink, J. J. (1986). Optic flow. *Vision research*, *26*(1), 161–179.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *207*(1167), 187–217.

McMorris, T. (2014). *Acquisition and performance of sports skills*. John Wiley & Sons.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201.

Reinagel, P., & Zador, A. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, *10*(4), 341–350.

Robert, R. (1985). Anticipatory systems: Philosophical, mathematical and methodological foundations. *Pergamon Press*.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the symposium on eye tracking research & applications* (pp. 71–78). New York, New York, USA: ACM Press.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural networks : the official journal of the International Neural Network Society*, *19*(9), 1395–407.

Zetzsche, C., Schill, K., Deubel, H., Krieger, G., Umkehrer, E., & Beinlich, S. (1998). Investigation of a sensorimotor system for saccadic scene analysis: An integrated approach. In *Proceedings of the fifth international conference on simulation of adaptive behavior on from animals to animats* (Vol. 5, pp. 120–126).

Zhang, B.-T. (2013). Information-theoretic objective functions for lifelong learning. In *Aaai spring symposium: Lifelong machine learning*.

Zirnsak, M., & Moore, T. (2014). Saccades and shifting receptive fields: Anticipating consequences or selecting targets? *Trends in Cognitive Sciences*.

Zou, W. Y., Zhu, S., Ng, A. Y., & Yu, K. (2012). Deep learning of invariant features via simulated fixations in video. *Advances in Neural Information Processing Systems*, 3212–3220.