The synergy of top-down and bottom-up attention in complex task: going beyond saliency models.

Enkhbold Nyamsuren (e.nyamsuren@rug.nl)
Niels A. Taatgen (n.a.taatgen@rug.nl)
Department of Artificial Intelligence, University of Groningen,
Nijenborgh 9, 9747 AG Groningen, Netherlands

Abstract

This paper studies how visual perception of a scene is affected by cognitive processes beyond the scene's bottom-up saliency. The game of SET is taken as an example where contrast-based salient parts of a scene are ignored in favor of a larger group of similar elements. Using results from a laboratory experiment and a model simulation we explain how three cognitive mechanisms, differential acuity, visual iconic memory and declarative retrieval, considered together help to explain player's visual perception in SET.

Introduction

Many studies describe how perception of a visual scene is governed by visual bottom-up mechanisms (Rayner, 1998). The conclusions derived in those studies are often based on results from relatively simple tasks involving free scanning or target search. It is widely accepted that visual attention is drawn toward a scene's salient parts (Egeth & Yantis, 1997). This bottom-up saliency is commonly used to explain *pop-out effect* of items that are increasingly different from its surroundings (Theeuwes, 1992). However, these findings alone may lead to incorrect conclusions if used within a context of more complex problem-solving tasks. It is important to consider a relationship between salience and other cognitive mechanisms to properly understand the inner workings of human mind in such tasks. We use the game of SET[1] as an example of a problem-solving task that gives results that can be interpreted initially as contradictory to the visual pop-out effect. Next, we describe how the same results can be explained within a framework that combines bottom-up saliency with top-down goal-directed attention.

The deck in SET consists of 81 cards. Each card is uniquely defined by a combination of four attributes: color, shape, shading and number of shapes. Each attribute can have one of three distinct values: red, green, and blue for the color; open, solid and textured for the shading; one, two and three for the number; oval, rectangle and squiggle for the shape. At any moment in the game, 12 cards are dealt face up (Figure 1). From 12 cards, players should find any combination of three cards, referred to as a *set*, satisfying a rule stating that in the three cards the values for each particular attribute should be all the same or all different.

Jacob and Hochstein (2008) studied how bottom-up components of the game, such as attribute value distribution among cards, influences player's strategy. They concluded that players prefer to search for a set inside the largest group of cards that share at least one common value. They referred to a common value as the Most Abundant Value (MAV) and the group of cards that contained it as a MAV group. Sets that were inside MAV group were found sooner than sets outside of the group with an observed probability being significantly higher than a chance probability.

According to the bottom-up saliency mechanism it is expected that players should start a search with visually unique, hence most salient, cards. However, Jacob and Hochstein's finding suggests that player's visual attention is drawn toward larger group of cards that are visually similar. From a perspective of a bottom-up saliency, this is a highly counterintuitive result. Furthermore, another study by Nyamsuren and Taatgen (2013b) revealed that a similarity along particular attribute dimension plays more important role in players' strategy than the saliency of any individual card. Players are more likely to search for a set among larger group of cards with the same color than to attend a card, for example, with a unique shape.

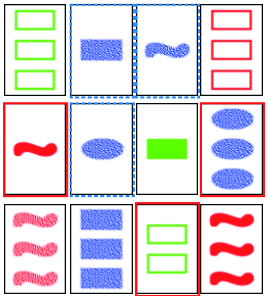

Figure 1: An example array of 12 cards. The cards with solid and dashed borders represent two valid sets.

In this paper, we describe a more controlled experiment with set cards with an aim of more in-depth exploration of underlying cognitive processes. In order to use the MAV strategy, subjects must be able to recognize very quickly, which attribute values are most common. The goal of the study is to focus on this particular aspect of SET: to answer the question what cognitive processes facilitate such quick recognition in players. Based on experimental results and model simulations, we describe how three cognitive mechanisms that include visual acuity, visual memory and declarative memory retrieval help to explain MAV effect and bias toward similarity in color attribute.

[1] SET is a game by Set Enterprises (www.setgame.com)

Experiment

Design and Procedure

14 subjects participated in the experiment. All subjects were students of University of Groningen. Subjects' age ranged from 18 to 27 (*M=22*). Subjects started each trial by looking at the center of a computer screen. Next, they were shown a 3×4 array of SET cards for a predetermined duration of time. After image of cards disappeared, subject was prompted to select one of 12 possible attribute values subject perceived as being the most abundant. The experiment consisted of 336 unique trials generated semi-randomly. Trials were divided into a short and a long condition block. The array of cards was shown to subjects for 600 and 2000 ms in the short and long conditions respectively. For half of the subjects, blocks were presented in a reverse order. Within a block, trials were presented in a random sequence. In each block, the MAV group size varied from 6 to 12. There were six trials in each combination of MAV group size and attribute type. Prior to experiment, subjects were asked to do eight, four from each block, trials to let them get familiar with an experiment setup. Results from those trials were not included in the analysis. In addition, subjects' eye movements were recorded. We used the EyeLink 1000, a desktop-mounted remote eye tracker with monocular sampling rate of 500Hz and spatial resolution of < 0.01° RMS. Exactly the same experiment setup and stimulus sizes as in Nyamsuren and Taatgen (2013b) were used in this study.

Experiment Results

Scanpaths The difference in trial durations also results in quite clear difference in scanpaths. Subjects on average make 8.8 (*SE=0.38*) fixations in the long condition compared to 2.9 (*SE=0.17*) fixations in the short condition. Figure 2 provides a more detailed look on the trials' fixation counts. There is an 87% probability that subject will make from seven to 11 fixations in the long condition. In contrast, subjects are likely to make only 2 to 4 fixations in 94% of all trials in the short condition.

Figure 3a shows mean durations of fixations in a trial. All durations are measured in milliseconds. The last fixations are excluded from the calculation of these means since it is likely that those fixations were interrupted when the time limit was reached. The first two fixations do not show much difference between the short and long conditions. The durations for consecutive fixations in the long condition does not change much. In contrast, durations of third and fourth fixations in the short condition gradually become lower. There can two explanations to this. It may be an artifact of averaging. Smaller number of trials with three or four fixations may be resulting in lower mean. On the other hand, it is possible that shorter durations are deliberate. To test this hypothesis we have also calculated the average duration of fixations in the short condition trials with exactly four fixations. As we have expected, fixations in these trials have much shorter durations than respective fixations in the long condition trials. Therefore, it is indeed possible that subjects were deliberately making shorter fixations in the short condition.

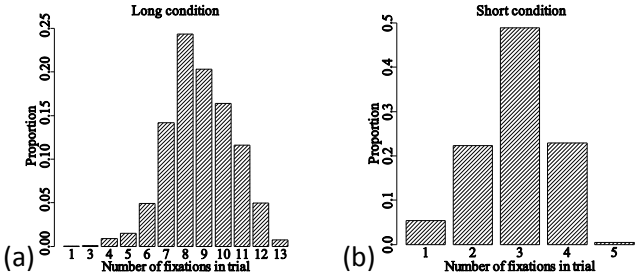

Figure 2: Frequencies of fixation counts subjects made during a trial. Frequencies are calculated separately for the (a) long and (b) short conditions.

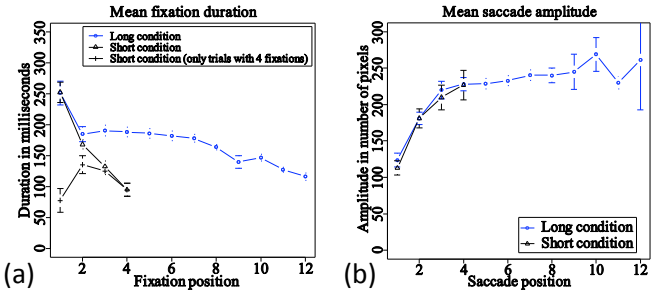

Figure 3: (a) Changes in mean fixation durations over course of a trial in the short and long conditions. (b) Changes in saccade amplitude over the course of a trial in the short and long condition.

Figure 3b shows how saccade amplitude changes over the course of a trial in both long and short conditions. Amplitude is measured in number of pixels that the saccade covers. There is not much difference between the two duration conditions. However, there is an obvious gradual rise in saccade amplitude as trial progresses. It suggests that there is a specific pattern in subjects' scanpaths.

Accuracy As Figure 4 shows, the overall accuracy increases as MAV group size increases. This is true for both short and long conditions. A test of proportions on pooled data indicate that subjects were more accurate in the long condition than in the short condition, $\chi^2(1, N=4704) = 35.63$, $p < 0.001$. However, as Figure 4 shows, there are remarkably small differences in accuracies with respect to group sizes in two duration conditions.

Figure 5 shows a boxplot of accuracy variations based on attribute type and duration. We did logistic mixed-effect regression analysis using the duration condition, attribute type and the interaction between the two as predictors. The intercept in the regression model reflects expected accuracy in a short condition trial where the MAV belongs to shading. Relative accuracy increased when MAV belonged to color ($z = 3.19$, $p = 0.001$) and decreased when MAV

belonged to either number ($z = -4.142$, $p < 0.001$) or shape ($z = -2.577$, $p = 0.01$). Overall performance in the long condition increased significantly ($z = 2.093$, $p < 0.036$). However, there were no significant interactions between duration conditions and attribute types.

Chi-square tests confirmed that subjects were significantly better at identifying the MAV with a color attribute than any other attribute type. Subjects showed little difference in accuracies in the short and long conditions with respect to color ($\chi^2(1, N=1176) = 2.91$, $p = 0.088$). It is surprising that, despite the significant difference in average number of fixations made, subjects are equally good at identifying color value in both duration conditions. In contrast, accuracies in the long condition were significantly higher for number ($\chi^2(1, N=1176) = 15.283$, $p < 0.001$), shape ($\chi^2(1, N=1176) = 16.94$, $p < 0.001$) and shading ($\chi^2(1, N=1176) = 4.12$, $p = 0.04$) than in the short condition.

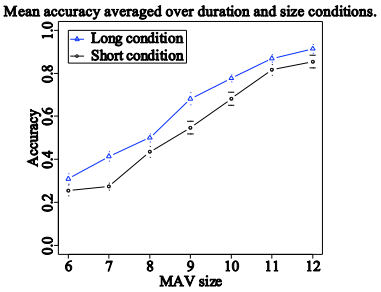

Figure 4: Mean accuracies averaged over all combinations of MAV group sizes and duration conditions.

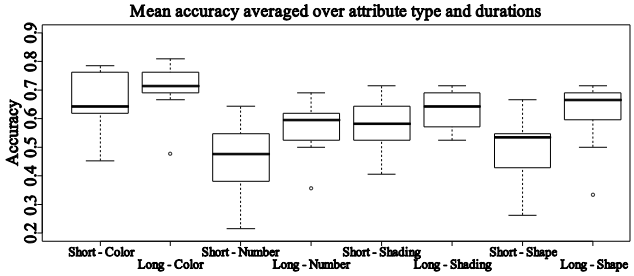

Figure 5: Mean accuracies averaged over all combinations of attribute types and duration conditions.

Experiment Discussion

Effect of MAV Group Size on Accuracy This effect can be explained by the priming of declarative memory by the visual system. There are several studies indicating that the human visual system has some form of iconic memory (Kieras, 2009). It is a low-resolution high-capacity memory where visual information is stored pre-attentively for a short duration of time. The process of gathering information is massively parallel and almost instantaneous. However, information about a visual object is stored as a collection of separate feature channels (such as color or shape) rather than single coherent object (Treisman & Gelade, 1980). Therefore, iconic memory has just enough resolution to guide further attention shifts and encoding.

There is evidence that visual perception can influence processes of memory retrieval (Wais, Rubens, Boccanfuso, & Gazzaley, 2010). It is reasonable to assume that visual stimuli can facilitate memory retrieval of items that are in some form related to the stimuli. Furthermore, we assume the same process applies to iconic and declarative memories. Items in iconic memory facilitate retrieval of similar or related items in declarative memory. In other words, items in declarative memory get activated by items in iconic memory. The strength of such activation depends on the number of items in iconic memory that are related to the item in declarative memory.

This interaction between iconic and declarative memories can explain why subjects find it easier to identify the MAV among larger group of cards. Subjects need to do two tasks: (1) gather visual information through attention shifts and (2) retrieve the MAV from memory when prompted. The second retrieval step is influenced by the content of iconic memory that was gathered during the first step. When MAV group size is large, more values enter iconic memory, and corresponding MAV value in declarative memory receives a higher activation during the retrieval.

Effect of Attribute Type and Duration on Accuracy The exchange of activations from iconic to declarative memories also helps to explain why subjects are better at identifying color values than values from any other attribute type.

However, there are studies showing that an ability to capture finer details of a visual scene becomes worse as the distance from a foveal region increases (Nelson & Loftus, 1998). This introduces limitations on what visual features can be gathered into iconic memory. As an object is further away from the foveal region it becomes more likely that some of its features will not enter iconic memory due to limitations of peripheral vision. A feature's acuity threshold defines the maximum distance from a foveal point at which the feature is still recognizable (Kieras, 2009). Compared to other features, color has a higher threshold making it easier to recognize in the peripherals. Thus, color values have a higher chance of entering iconic memory thereby spreading more activation to the same values in declarative memory.

When features, such as shape and shading, have a limited acuity, subjects need to fixate closer to respective visual objects to bring them within threshold distance. This explains why subjects perform better in the long condition trials. Subjects can make more fixations and gather a more complete gist of the visual scene in iconic memory, which then facilitates a more accurate declarative retrieval.

Scanpaths There are two interesting effects in subjects' scanpaths. Firstly, subjects seem to react to time pressure in the short condition by having shorter fixation durations. This behavior also supports our assumption that iconic memory and peripheral vision play an important role. It is possible that subjects compensate for a shorter duration by

making as many fixations as possible and accumulating in iconic memory as much visual information as possible. The pattern of increasing saccade amplitudes provides a clue about preferences of possible fixation locations. Subjects start by fixating on the cards closest to the center of the screen and gradually switch to the cards on the peripherals. These fixations from inwards toward outwards should result in increasing saccade amplitudes shown in Figure 3b. In addition to providing more clues about subjects' behavior, scanpaths provide additional measurements besides accuracy against which model fit can be evaluated.

Cognitive Model

Cognitive Architecture

We have used ACT-R cognitive architecture (Anderson, 2007) to develop the model. An additional module called Pre-attentive and Attentive Vision (Nyamsuren & Taatgen, 2013a) was used instead of ACT-R's default vision module. The PAAV module provides several extra functionalities that are otherwise not supported by ACT-R.

PAAV can pre-attentively capture the gist of a visual scene and store it in iconic memory. The content of iconic memory is updated before and after each saccade and before each time the memory is accessed. The update process is instantaneous from a perspective of model's timeframe. Iconic memory may contain complete information for some visual objects, such as an object's color, shape, shading and size. However, for most visual objects the iconic memory will contain incomplete information (e.g. color only) due to limited acuity. PAAV recognizes that not everything in a visual scene can be resolved by model's peripheral vision at any given moment. In PAAV two parameters, *a* and *b*, define differential acuities of color, shape, size and shading with color having the highest acuity. Fitness of these parameters was tested on models of three different visual search tasks and the updated model of game of SET (Nyamsuren & Taatgen, 2013a). An object's feature in iconic memory, although persisting through saccades, decays after a short period of time (currently 4 sec) if not recognizable in peripheral vision anymore.

The content of iconic memory is used to guide the model's visual attention. Visual objects with the highest saliency values are prioritized for visual attention and further encoding. In PAAV, the bottom-up saliency is a sum of saliency values calculated for each feature dimension as a function of contrast to its surrounding. For example, a single red card among green, otherwise similar, cards will be the most salient one and draw the model's attention. PAAV uses a binary measure of similarity: 1 for exact match and 0 otherwise. No adjustable parameters are used in calculation of bottom-up saliency (Nyamsuren & Taatgen, 2013a). It is a simplified version of Wolfe's (2007) saliency function.

In ACT-R knowledge chunks are stored in declarative memory. Each chunk has an activation value that usually reflects chunk's recency and frequency of use by a model. A chunk with the highest activation has the highest probability of retrieval. Besides frequency and recency, a chunk's activation can be increased by the content of iconic memory. Each visual object in iconic memory spreads activation to every declarative chunk with the same features. So depending on the content of iconic memory at the time the results of two same retrievals can differ. The model uses exactly the same set of parameters for declarative retrieval as in the original model of game of SET. Details of those parameters are described in Nyamsuren & Taatgen (2013b).

Model of MAV Task

Model Strategy Model performed 50 times the same two blocks of trials subjects did. Model starts each trial while fixating at the center of the screen. When cards are shown, models need some time to create a working memory before the first saccade is made. At the same time, model updates its iconic memory with representations of cards. Then model follows with free scanning using bottom-up saliency values to calculate consecutive fixation points. Each fixation is followed by encoding of an attended card. Free scanning stops when time limit is reached and representations of cards disappear. At this point model retrieves any one of 12 possible attribute values from declarative memory. Result of this retrieval depends on content of iconic memory the model has built up during the free scanning. The retrieved value is recorded as model's response for the trial.

Model Accuracy Model is quite good at replicating subjects' accuracy. Figure 6 shows that model's accuracy increases linearly as the MAV group size increases. This effect is present in both the short and long condition. However, just like subjects, the model shows a better performance in the long condition.

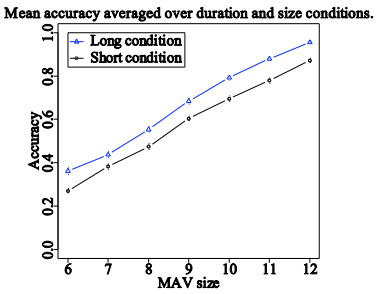

Figure 6: Mean accuracies averaged over all combinations of MAV group sizes and duration conditions.

The model is also good at reflecting subjects' accuracy depending on combination of attribute types and duration conditions. Firstly, as Figure 7, there is a general increase in model's accuracy in the long condition. Except in color, the model clearly benefits from additional time in all other three attributes. Next, Figure 7 shows that model is much better at identifying MAV belonging to color attribute than to any other attribute type. Similar to human performance, model's accuracy for color in the short condition is higher than the accuracies for other three attribute types in the longer trials.

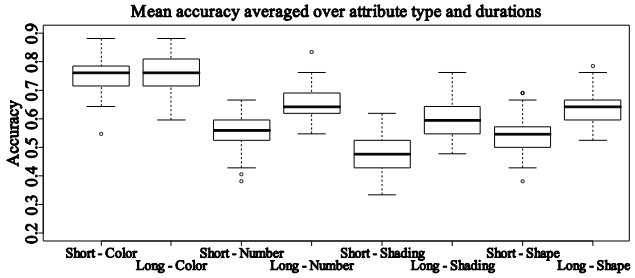

Figure 7: Mean accuracies averaged over all combinations of attribute types and duration conditions.

Model Scanpaths Comparison of model's scanpaths to that of subjects should give additional measure of how well the model fits human data at the level of raw eye movements. Figure 8 shows distributions of fixation counts the model made in the long and short conditions. In 99% of all long condition trials, the model made 9-10 fixations. It is within a range of 7-11 fixations subjects made. In the short condition, the model made either two or three fixations. It is also within a range of 2-4 fixations subjects made. As Figure 9a shows, model's fixation durations do not differ in the long and short conditions. The lower duration for the third fixation in the short condition is a result of interruption due to duration limit.

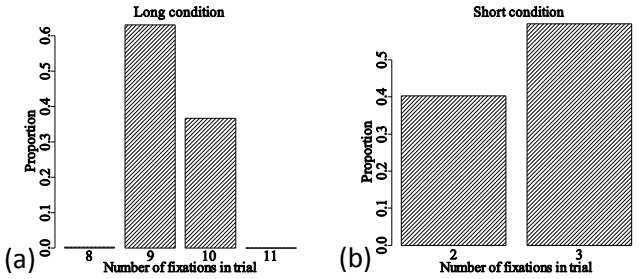

Figure 8: Frequencies of fixation counts model made during a trial. Frequencies are calculated separately for (a) long and (b) short conditions.

The model was able to reproduce a pattern of increasing saccade amplitudes in long condition trials, as it is shown in Figure 9b. It was not completely expected since we have not incorporated into the model any deliberate mechanisms to promote this behavior. Because the model makes only one or two saccades in a short condition trial, it is hard to make any conclusive statements about the pattern of amplitude changes. The same model is used in both duration conditions. Hence, there is no reason to expect the model to show different scanpath pattern in the short condition. The lower amplitude for the second saccade in the short condition is most likely due to smaller number of observations from which the mean is calculated. For exactly the same reason, amplitudes for the 9th and 10th saccades drop in the long condition since there are fewer trials that have more than 10 fixations.

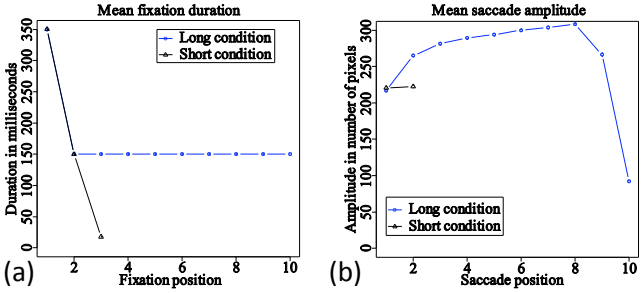

Figure 9: (a) Changes in model's mean fixation durations over course of the trial in the short and long conditions. (b) Changes in model's saccade amplitude over the course of the trial in the short and long condition.

Discussion on Model Results

The point at which model has to decide on a choice of MAV is the retrieval from a declarative memory. As model shows, the spreading activation from iconic memory is a major factor deciding the result of this retrieval. However, it is possible to counter-argue that spreading activation from iconic memory is not necessary, and items in declarative memory are activated directly through visual encoding of similar items. Such mechanism is possible and used in our model. Cards with the MAV have a higher chance probability of getting visual attention and being encoded. As a result, the MAV in declarative memory receives more activation and is retrieved. Although this argument would explain subjects' behavior in the long condition, it does not explain why there is a similar effect of MAV group size in the short condition. Neither subjects nor model can encode more than two cards in the short condition, and it is not enough to influence the retrieval. Instead, it is likely that subjects rely on visual information in peripheral regions for choosing MAV. Furthermore, the fact that subjects are quite good at identifying the MAV even within 600 ms implies that process of gathering information from peripherals is very efficient. The model simulation suggests that it may be massively parallel and instantaneous.

In the other side, acuity limitations of visual features in peripheral vision can result in incomplete and inaccurate iconic memory. This imperfect internal representation may explain why subjects fail to reach 100% accuracy. It also explains why subjects get better given opportunity to do more fixations in the long condition. More fixations negate the effect of low acuity and result in a more complete representation of the scene inside iconic memory. Furthermore, giving a higher acuity to color in model simulation increases model's accuracy in identifying the most abundant color values in both conditions. This result is similar to the result from the experiment, and, therefore, supports the assumption that human vision is affected significantly by different acuity properties of visual features.

The model produces the same pattern of increasing saccade durations in the long condition without any deliberate mechanisms. It suggests that spatial arrangement

and the bottom-up salient parts of the visual scene define the topology of fixation points, more specifically the characteristic fixations from inwards to outwards. In the model, cards around the edges of the screen are not fully visible due to limited acuity. Those cards have reduced bottom-up activation compared to cards around the center of the screen. As a result, the model prefers to fixate on cards closer to the screen center at the early stages of the trial. We were not able to simulate the deliberate reduction in fixation durations subjects have shown in the short condition. Visual processes currently implemented in ACT-R do not provide appropriate mechanisms to simulate this effect.

Discussion and Conclusion

The model fits subjects' accuracies and scanpaths well supporting the hypothesis that the same cognitive processes simulated in the model may also be used by human subjects. More specifically, a combined effect of differential acuity, pre-attentive visual iconic memory and implicit communication with declarative memory can influence our visual perception of the world.

The results from this study can explain player's behavior in game of SET. Player has to decide on a group of cards to be searched for a set. This choice is made through a declarative retrieval of an attribute value that is common among group cards. Similar to the experiment's task, this retrieval is influenced by a content of iconic memory introducing a bias toward a larger group of cards and cards with same color. The retrieved value is used to target attention to specific cards with that value. This top-down control over eye movements overrides the bottom-up saliency of the scene. It explains both why players are better at finding set within a group with many similar cards (Jacob & Hochstein, 2008) and the general preference toward cards with a similar color (Nyamsuren & Taatgen, 2013b). The model of SET player implemented on the same principles described here was able to simulate player's behavior (Nyamsuren & Taatgen, 2013a, 2013b). It is a good example of a case where cognitive mechanisms beyond bottom-up saliency can influence the behavior in a reasonably complex problem-solving task. It implies that not every eye movement pattern can be attributed to bottom-up salient components of the scene.

Subjects are far better in identifying the MAV even in the most difficult conditions. In 600 ms condition with smallest MAV group size, subjects show much higher accuracy than 8% chance probability of success. This result indicates that capabilities of human visual system may be higher than previously expected. The ability to capture a gist of a visual scene from first few fixations is known for a long time (Loftus & Mackworth, 1978). However, it is commonly viewed that functionality of such gist is limited to attentional guidance and providing early structural information for encoding, a preview effect (Rayner, 1998). On the other hand, our study suggests that gist, in form of iconic memory, may be involved in decision-making. It is possible through connections between memories in human brain. In this study, we talked about similarity-based cross-memory activations between iconic and declarative memories. However, it may be possible that similar cross activations exist between other forms of memory.

The model code and the data can be downloaded via following link: http://www.ai.rug.nl/~n_egii/models/.

Acknowledgement

We would like to express our gratitude to Eveline Broers for conducting a pilot experiment and giving a valuable insight into data analysis.

References

Anderson, J. R. (2007). *How Can Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

Egeth, H. E., & Yantis, S. (1997). Visual attention: control, representation, and time course. *Annual Review of Psychology , 48*, 269-297.

Jacob, M., & Hochstein, S. (2008). Set recognition as a window to perceptual and cognitive processes. *Perception & Psychophysics , 70* (7), 1165-1184.

Kieras, D. (2009). The Persistent Visual Store as the Locus of Fixation Memory in Visual Search Tasks. In A. Howes, D. Peebles, & R. Cooper (Ed.), *9th International Conference on Cognitive Modeling - ICCM2009*. Manchester, UK.

Loftus, g. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance , 4*, 565-572.

Nelson, W. W., & Loftus, G. R. (1980). The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning and Memory , 6*, 391-399.

Nyamsuren, E., & Taatgen, N. A. (2013a). Pre-attentive and Attentive Vision Module. *Cognitive Systems Research , 24*, 62-71.

Nyamsuren, E., & Taatgen, N. A. (2013b). Set as instance of a real-world visual-cognitive task. *Cognitive Science , 37* (1), 146-175.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin , 124*, 372–422.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics , 51*, 599-606.

Treisman, A., & Gelade, G. (1980). A Feature-integration Theory of Attention. *Cognitive Psychology , 12*, 97-136.

Wais, P. E., Rubens, M. T., Boccanfuso, J., & Gazzaley, A. (2010). Neural Mechanisms Underlying the Impact of Visual Distraction on Long-term Memory Retrieval. *Journal of Neuroscience , 30* (25), 8541-8550.

Wolfe, J. M. (2007). Guided Search 4.0: Current Progress With a Model of Visual Search. In W. Gray, *Integrated Models of Cognitive Systems* (pp. 99-119). New York: Oxford.