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

Jin-Hwa Kim (jhkim@bi.snu.ac.kr)
Cognitive Science Program, 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

As the eye movement is a non-invasive and convenient indicator of the brain activity, the application of eyewear accelerate the study on the eye movement, especially beyond the reading task. We investigate the eye movements of the subjects watching the kids video. The characteristics of the eye movements of watching are different from that of reading, having a long-tail and no rounded peak. The relatively undemanding load for the cognitive process of the video sequences may cause the most part of the short fixations, and the spatio-temporal constraints of given stimuli may attribute to the long fixations. We analyze the video sequences by classifying into two different sequence groups have the long and short fixation duration, respectively. 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 is not important, however, the long fixations which are voluntarily engaged with *Alert* type sequences associate with long-term memory than any other cases. Finally, we propose the computational modeling to predict those active long fixations, which may provide the way to effectively summarize the story of a given video.

Keywords: Eye movement; fixation; spatio-temporal; long-term memory, computational modeling.

Introduction

The brain is the most intelligent organ of the human body. It receives many different forms of the sensory information and processes these information appropriately with regard to its nature for doing well. Especially, the visual information takes a very special position among other kinds of sensory information as it does not need to touch or smell it directly, causing the chemical reaction inside the body, and is transfered to the remote target far freely than any other types of the sense. Furthermore it gives more chance to survive in the situation of being threaten by the predators using the eyes to detect the foes in the remote place before their approaching.

But this notable advantage is not freely given one as it requires more delicate and clever way of interpreting the visual information. Because the visual information can easily be affected by the moment-to-moment environmental changes, heuristic but robust compensation strategies are demanded. So, how the brain processes the visual information provides the profound way to study the mechanism that the brain precisely and efficiently processes the most dynamic and enormous sensory information.

There are a lot of studies on the computational modelling for the visual information, which include the visual fragment completion, the scene or object classification and recognition (Winn, Criminisi, & Minka, 2005; Lazebnik, Schmid, & Ponce, 2006), and object tracking (Wu, Lim, & Yang, 2013). These research topics often tend to focus on the objective for each tasks, not on the implementation of the method how the brain deals with. As a result, the computational approach to the modeling for the visual information processing of the brain is gradually changed to the optimization problem to effectively solve, which hinders the understanding of the human-level information processing abilities.

Particularly, the object tracking seems to describe how we keep the attention toward the interesting object, however, the eye movement, which is mostly controlled by the oculomotor system, is better to imply the way of how the brain works for the acquisition of the visual information (Henderson, 2003). For an instance, in the fixation state the human eyes only recognize the small portion of the whole sight. If you read this paper in a 8-inch distance fixing on one particular letter, you cannot read next two words or approximately about ten letters which are presented in the para-fovea. For the brain is well-known for its parallel processing on the neural circuits, this sequential notion of eye movement visual system does not suit to the efficiency of the brain.

In comparison with the reading eye movement, which are relatively well studied by psychologists and neuroscientists (Rayner, 1998; Reichle, Pollatsek, Fisher, & Rayner, 1998), the study of the eye movement on the video stimuli has been neglected or avoided due to the uncertainty of the research theme or its application (Tatler, Hayhoe, Land, & Ballard, 2011). However, the recent advancement of the sensory device, like the Google Glass and the mobile devices for the eye tracking, enables to study on the more natural experimental environment and to implement the research model or applications on those mobile device. Moreover, there is a different aspect of studying on the eye movement for video stimuli. How long the fixation sustains is more constrained by the video content rather than the reading materials do, and the selection of the next fixation and the saccadic movement direction tend to be more liberal than the dominance of horizontal searching of reading does.

In this study, we investigate the characteristics of eye movements toward the video stimuli. The basic elements of the eye movements can be segregated into the fixation du-

ration and the saccade vector, which consists of the saccade direction and the length of the saccadic movement (Findlay & Walker, 1999; Feng, 2006). In this study, we focus on the characteristics of the fixation duration as the evidence of the cognitive process.

Materials and Methods

Experiment 1

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 simple shaped artificial 3D-rendered characters which have marked individualities. *Pororo Season 3* contains the 13 consecutive episodes that each episode has a single storyline. Remarkably, 4 characters are intervene to the main plot among the existing characters, the robotic aliens *Popo* and *Pipi* in the first episode, the homemade robot *Rody* in the episode 3 and the magician Tongtong in the episode 6. Though the experiments and the analyses using these features are not fully considered in this study, we look forward to publish the results soon.

We recruit 18 participants with normal vision (11 males, 6 females; 23-31 years of age; median is 25), who are voluntarily participate in the study. All participants had not experienced a brain damage or a behavioral disorder. The participants are new to the video, *Pororo Season 3*. Experimental procedures were agreed by the Research Ethics of Seoul National University and the participants signed a consent form.

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

Eye movements and concurrently the user-perspective scenes are 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 classifies the fixations with the velocity threshold, 30 degree per second. Usually, the saccadic eye movements are discriminated between low velocities (less than 100 degree/second) and high velocities (higher than 300 degree/second), velocity-based classification is simple but reasonable (Salvucci & Goldberg, 2000).

We classified the event type of the eye movements into three categories; fixations, saccades and unclassified. The unclassified data is discarded and not used in this study.

Experiment 2

For 12 participants who have participated in *Experiment 1*, we prepared the controlled memory test for each participant. We conduct this experiment 3-4 months later after *Experiment 1* (the intervals are not consistent due to a schedule conflict). A memory test consists of total 20 video sequences; 8 sequences for long fixations, another 8 sequences for short fix-
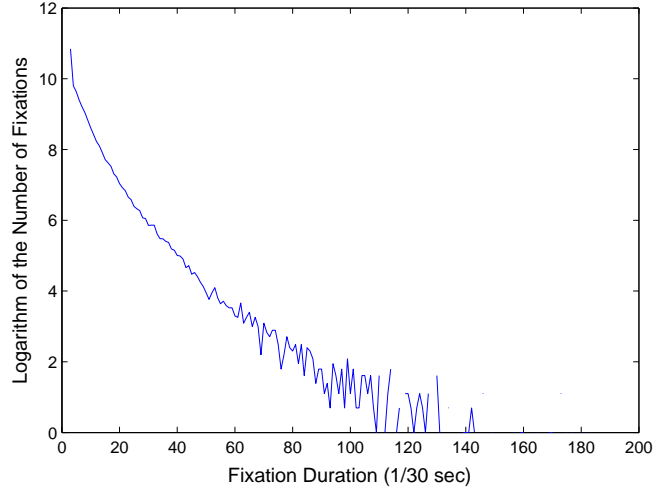

Figure 1: The marginal distribution of fixation durations. All of the 158,643 fixation durations of 18 participants is used. The x-axis represents the number of time unit, 1/30 sec., therefore, 30 indicates 1 second. The y-axis represents the logarithm of the number of fixation durations which are the same time length with regard to the time unit of x-axis.
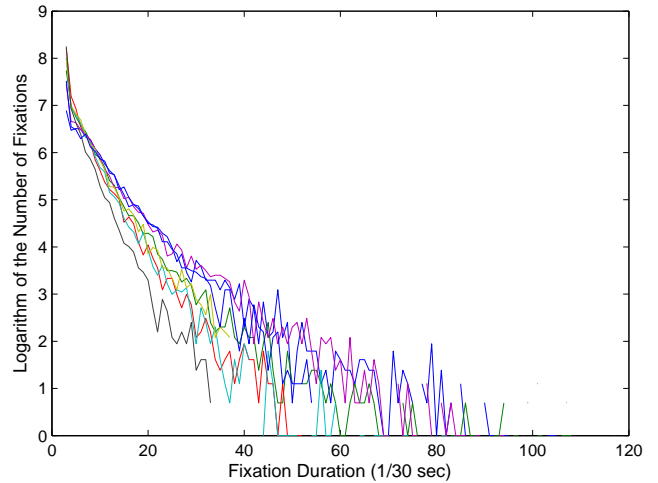

Figure 2: The distribution of fixation durations for the randomly selected 8 participants. We find the interpersonal differences among all 18 participants, the median of fixation durations varies from 133 ms to 267 ms (mean=183.3, std=41.7).

ations, and the remaining 4 sequences for the control, which are not seen in the previous experiment. The lengths of all video sequences are the same as 3 seconds. *The long fixation sequences* are randomly picked from each participant's data containing the longer than 2 seconds of fixation in the middle of the sequence. *The short fixation sequences* are randomly picked from each participant's data containing the shorter than 300 millisecond. The 4 control sequences are randomly picked from the other season of *Pororo* series, *Pororo Season 2*.

Each participant identifies 20 sequences randomly sorted. Each participant gives a sequence 1 to 5 score depending on the assurance of whether he or she saw the sequence before or not.

Fixation Duration

Fixation duration takes a important position in the reading literature as the duration of eye fixations seems to be constrained by the linguistic features of the fixated word (Rayner, 1998). In the video watching task, we anticipate the characteristics of fixation duration are different from that in the reading task. As we expect, the fixation duration changes more drastically, up to 10 seconds during watching the video. These changes are partly caused by the sequential changes of the visual stimuli and the fluctuated responses of the oculomotor system and cognitive processes. The length of fixation duration is not a deterministic property, because more than a single component dynamically contribute to the final motor command and the execution for the oculomotor system, we assumed the fixation duration as the random variable with probabilistic distribution (Rayner, 1998; Reichle, Rayner, & Pollatsek, 2004; Reichle, Pollatsek, & Rayner, 2006).

Marginal Distribution of Fixation Durations

Figure 1 shows the marginal distribution of fixation duration, which includes all of the 158,643 fixation durations of 18 participants. The x-axis represents the duration time, and the y-axis represents the log scale of the number of fixation durations across the whole data.

The shape of the marginal distribution of fixation duration (Figure 1) can be illustrated with the exponential function (see the logarithm of the y-axis and the approximation plotted with red dotted line). While the distribution of the *reading* fixation durations has a quiet different shape (Feng, 2006). The shape is roughly segmented into three parts, slow-rising ones for short fixations, fast-rising period until around 180 ms, and following a long tail for long fixation durations. These facts allow us to think that 180 ms of the fixation duration for reading is the most general case, but obviously not for watching the video.

Fixation durations which are longer than about 2 seconds are getting more unpredictable along with increasing the fixation duration. Though it is due to the logarithm increasing the sampling variance, occasionally the content of the video stimuli determines how long the eye gaze is fixated. In other words, the fixation more tends to maintain the gaze position
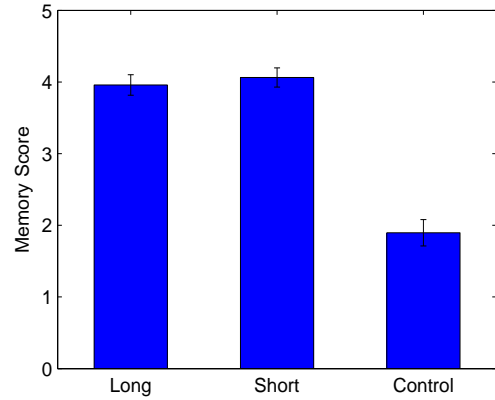

Figure 4: Long-term memory test result for two fixation types, one is that whose duration of fixation is longer than 2 seconds, and the other one is that whose duration of fixation is shorter than 300 ms. Each participant rates 8 long fixated sequences, 8 short fixated sequences and 4 control sequences which are not seen previously. Surprisingly, the mean of the memory scores of the short fixated sequences are slightly higher than the mean of the memory scores of the long fixated sequences, though it is not significant. Error bars indicate the standard error of the mean (SEM). For the detail of test, refer to *Materials and Methods*.

when the visual constraint imposes. Surprisingly, the visual constraints have more various forms than we expected.

Long Fixation Durations

In Figure 3 the sequences of frames which receive more than 2 seconds of the fixation duration from at least 3 different participants are shown. We set the threshold as 3 long fixations from the participants 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 means the independent sequence and each column means the single frame. The time interval between the frames is 0.5 seconds. The colored dots mean 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 *Alert* (3), *Successive* (3), *Stationary* (3), and *Unclassified* (1) for the review.

The sequences of the *Alert Focus* type are classified as the scene implies an unusual or potentially dangerous or difficult situation, which may introduce a mental arousal. The sequence of first row demonstrates an urgent moment that a huge snowball is about to roll down back the hill, which was previously rolled up by the robot *Rody*. Second shows that *Pororo* was 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 Focus* type shows that the fixation extends across more than 2 different scenes. Because the location of the target object is not changed or changed within the range of

Figure 3: The sequences of frames which receive more than 2 seconds of the fixation duration from at least 3 participants. Each row means the independent sequence and each column means the single frame. The time interval between the frames is 500 ms. The colored dots mean the fixation positions, whose durations are longer than 2 seconds. The same color means that which is from the same participant. First 3 sequences show the *Alert Focus* type, then next 3 sequences show the *Successive Focus* type, then next 3 sequences show the *Stationary Focus* 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 subsection *Long Fixation Durations*.

Table 1: Long Fixation Types.

Long fixation type	Description
Alert Focus	An urgent situation happens with an object, which can be easily targeted as a cause.
Successive Focus	Successive changes to keep attracting.
Stationary Focus	The scene is the same while the target object(s) moves a little or even does not move.
Unclassified	Cinematic techniques, which are tracking, tilt, zoom-in and out, and others.

a foveal or central vision, 2-5°, the fixation holds its position (McMorris, 2014). The fourth sequence shows the close up characters are serially shown up in the middle of the screen. The fixations of the fifth and sixth sequences extends across 2 different scenes have a different visual configuration.

The *Stationary Focus* type most clearly shows the characteristics that the indifferent scene maintains while the target object moves a little bit or even does not move. As 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 sequences of the *Unclassified* type have various forms. The tenth sequence shows that *Pororo* and *Crong* just jumped out of the shoulder of the magician dragon *Tongtong* who was flying in the sky. Two sunflowers spring *Pororo* and *Crong* into the sky in multiple times. But due to the perspective of the camera, the location of the two characters in the scene is almost fixed. Participants fixated their eye gaze on the middle of two characters while the background shifting up and down. In other cases, though not included in Figure 3, few cinematic techniques, tracking, tilt, zoom-in and zoom-out, are examples. But the number of these cases is so small compared to the other types, hence we classified all of them into the *Unclassified Focus*.

These various types of the long fixated sequences show that a naive approach to the predict the visual constraint which induces the long fixation using a single analytical method, like the salient feature detectors (Marr & Hildreth, 1980; Canny, 1986) or the optic flow (Koenderink, 1986) would fail. The *Successive Focus* and *Stationary Focus* types can be estimated by using the combination of the appropriate feature detectors and the optic flow, and the temporal modeling of the attention. However, the *Alert Focus* cannot be solved without the cognitive modeling for the scene comprehension, which is related to the emotion-based reaction.

Long-Term Memory Formation

Figure 4 shows the long-term memory test result for two fixation types, one is that whose duration of fixation is longer than 2 seconds, and the other one is that whose duration of fixation is shorter than 300 ms. Each participant rates 8 long fixated sequences, 8 short fixated sequences and 4 control sequences which are not seen previously. Surprisingly, the mean of the memory scores of the short fixated sequences are slightly higher than the mean of the memory scores of the long fixated sequences, though it is not significant. Error bars
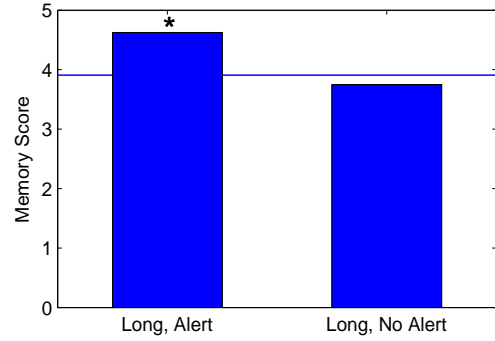

Figure 5: The memory test result for the *long* fixation types, Alert and No Alert. The number of cases from 11 participants are 16 and 72, respectively. The p-value of two-sample t-test for the *Alert* type is 0.0232 (< 0.05). The blue horizontal line indicates the mean of the memory scores of the long fixated sequences.
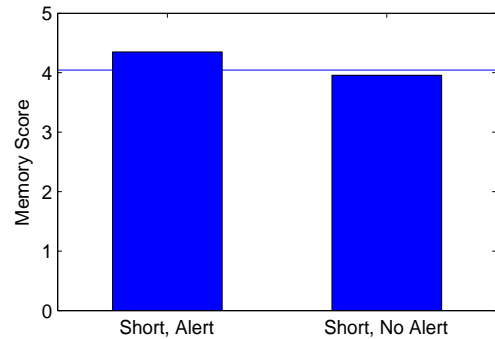

Figure 6: The memory test result for the *short* fixation types, Alert and No Alert. The number of cases from 11 participants are 20 and 68, respectively. The p-value of two-sample t-test for the *Alert* type is 0.2367 ($\gg 0.05$). The blue horizontal line indicates the mean of the memory scores of the short fixated sequences.

indicate the standard error of the mean (SEM).

Figure 5 shows the long-term memory test result for the long fixation types, Alert Focus, Successive Focus, Stationary Focus, and Unclassified. The number of cases are 18, 39, 30 and 9, respectively, from 12 participants. The significance test result of the *Alert Focus* type is 0.044 (< 0.05), but the others are far from that, 0.360, 0.187, and 0.073, respectively. The blue horizontal line indicates the mean of the memory scores of the long fixated sequences.

Discussions

The marginal distribution of the fixation duration shows the characteristics of the response toward the visual stimuli. The shape of the marginal distribution of fixation duration can be estimated as the exponential function though the marginal distribution of the reading fixation durations is illustrated as a left-skewed normal distribution, peaking at 180 ms. Then why reading and watching are so different from each other in this property? The first to think is the difference of the cognitive process during the fixation. Simply put, reading involves the visual processing in addition to the lexical processing. The visual process captures the letters through the retina, Lateral Geniculate Nucleus (LGN) and the primary visual cortex. Then the information of the letters is directed to the distributed lexical processing areas. Though skipping is occasionally occurred while reading, those serial processes spend some latency time. Since reading is an active task, the subject decides when and where to move to a next word, those latencies tend to be preserved. By the way, watching the video has a different condition. The lexical processes are typically not needed, which are heavy tasks in time. Also the video stimuli are passive in regard to the temporal aspect. Therefore the generation of the long fixation duration is sufficiently constrained by the duration of the stimulus, at the same time, the content of the stimulus.

We recap the visual constraints which trigger the long fixation durations as three types in Table 1. The *Contrast Focus* type seems that the contrast effect attracts the attention and marks the featured object. The second type, *Alert Focus*, is an ongoing urgent situation that makes the eyes fixates an object which is thought to be a cause or a factor. The last type is the *Successive Focus*. It looks like the successive appearing of the objects keeps the fixation longer, however, the interpretation that the successive absence of the other attracting elements just let the fixation persist is also possible. After all, what is the meaning of the long fixation durations on the video stimuli? Despite of the fact that it could be a latency time to process the cognitive information of the visual stimuli, in the other perspective, it is an waiting time to the potential salient moment on that eye position. All participants are grown-up persons, so familiar with almost all kinds of events in the episode of the video. Easily predicted, and waiting on the prospective location for a dramatic change or a new event is practical.

The computational modeling of the long fixation durations is worth to discuss. As we discussed, the long fixations are constrained by the visual content of the video stimuli, the probabilistic modeling of that would be unsuccessful to compatible with the unprecedented stimuli. The first possible attempt is the composition of the salient detectors (Marr & Hildreth, 1980; Canny, 1986) and the optic flow (Koenderink, 1986) accompanying the tentative cognitive modeling of that. As there are the types of the long fixation durations, the cognitive modeling can be projected to the lower dimension of the search space. Additionally, the estimation of the eye tracking is viable on top of the model. The serial information of the fixation positions enables us to intelligently select the portion of the visual features on the scene, like a human, it can be a breakthrough for the cognitive modeling on the endless stream of the visual information.

Conclusions

We studied the characteristics of the eye movements through the marginal distribution of fixation durations. We notice that the marginal distribution of fixation durations for the video stimuli has a form which is different from the marginal distribution of fixation durations for the reading materials. The behavioral basis for the difference may attribute to the lighter load for the cognitive process and the temporal and spatial constraints which are given by the video stimuli. Those constraints are summed up as three distinctive types, Contrast Focus, Alert Focus and Successive Focus. Also, we present the tentative suggestion for the computational modeling of the eye movements using the combination of the feature detectors and the optic flow in aid of the cognitive modeling.

Acknowledgments

References

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

Feng, G. (2006). Eye movements as time-series random variables: A stochastic model of eye movement control in reading. *Cognitive Systems Research*, 7(1), 70–95.

Findlay, J. M., & Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *The Behavioral and brain sciences*, 22(4), 661–674; discussion 674–721.

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

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

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural

scene categories. In *Computer vision and pattern recognition, 2006 ieee computer society conference on* (Vol. 2, pp. 2169–2178).

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.

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, *124*(3), 372–422.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a Model of Eye Movement Control in Reading. *Psychological Review*, *105*(1), 125–157.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006, March). EZ Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*(1), 4–22.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2004, March). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(04), 445–476.

Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the symposium on eye tracking research & applications - etra '00* (pp. 71–78). New York, New York, USA: ACM Press.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: reinterpreting salience. *Journal of vision*, *11*(5), 5.

Winn, J., Criminisi, A., & Minka, T. (2005). Object categorization by learned universal visual dictionary. In *Computer vision, 2005. iccv 2005. tenth ieee international conference on* (Vol. 2, pp. 1800–1807).

Wu, Y., Lim, J., & Yang, M.-H. (2013, June). Online object tracking: A benchmark. In *Computer vision and pattern recognition (cvpr), 2013 ieee conference on* (p. 2411-2418).