A Time-Dependent Model of Information Capacity of Visual Attention

Xiaodi Hou and Liqing Zhang

Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
filestorm@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract. What a human's eye tells a human's brain? In this paper, we analyze the information capacity of visual attention. Our hypothesis is that the limit of perceptible spatial frequency is related to observing time. Given more time, one can obtain higher resolution - that is, higher spatial frequency information, of the presented visual stimuli. We designed an experiment to simulate natural viewing conditions, in which time dependent characteristics of the attention can be evoked; and we recorded the temporal responses of 6 subjects. Based on the experiment results, we propose a person-independent model that characterizes the behavior of eyes, relating visual spatial resolution with the duration of attentional concentration time. This model suggests that the information capacity of visual attention is time-dependent.

1 Introduction

How much information is gained through one glimpse? There have been many attempts to answer this question[1]. To demonstrate the model in an information perspective, we consider the human visual perception pathway as an information channel. Any visual information whose spatial frequency is higher than the capacity of one's perception is unable to be transmitted through this channel. From this point of view, one can assert that what we "see" is the information that passes the band-limit filter of the visual channel[2].

1.1 Attention

Treisman and her colleagues in 1977 [3] classified the visual perception process into two categories, the pre-attentive process, and the attentive process. Generally speaking, the pre-attentive process is a parallel mechanism with coarse resolution and simple feature analysis. On the other hand, the attentive process is a serial process, much slower but with higher resolution. In tasks that require careful discriminations, our perception capacity is subjected to attention. An effective description of the behavior of attention is the "Zoom Lens" theory[4]. This theory proposed that the size of attentional focus can be concentrated to meet the requirement of successful perception. Recent researches even proved a physiological correlation of the "Zoom Lens" model[5]. In the "zoom lens" model,

only two factors are determinant to the information capacity of attention: the area that attention covers, and its spatial resolution.

1.2 Information capacity of attention

Previous studies indicated that the information capacity of attention is almost constant under uni-scaled visual stimuli [6]. However, when objects of different sizes are contained in one stimuli, the performance of attention varies[7]. To be specific, for more acute patterns, longer time is required to concentrate one's attention. It is easy for us to read several words of the headline of a newspaper in a short glimpse, but with the same observing time, it is hard even to see one single letter of the text font, which has a much higher spatial frequency than the headline. In an empirical level, this inequity in information capacity can be explained by zooming mechanism: with finer resolution, the observer needs more time to tune his/her attention.

In this paper, we aim at constructing a general model to quantify the information capacity of attention. More specifically, we try to (1) analyze the zooming mechanism of attention in respect of time and, (2) develop a quantitative formula that describes the viewing time duration and resolution of attention.

The result of our experiment shows a clear dependency of response time and spatial resolution of attention. We propose a model that describes response time of attention under stimuli of different spatial frequencies, and discussed our model in the context of cognitive science.

The introduction of temporal characteristics of attention enriches our understanding of the human vision, and may also advance current quantitative models. Currently, there is scant reference pertaining to the time-varying performance of our visual system. With the results of our experiments, it is possible for a researcher to deduct the frequency of incoming information given observation time. On the other hand, given the resolution of a stimulus, researchers may also predict the shortest viewing time that permits reliable perception.

2 The experiment

The goal of the experiment is to record the exact time duration required for successful attentional perception. In our experiment, we recorded the time duration of a *counting* task. In a counting task, a subject is told to enumerate the number of several identical items that are placed parallel to each other. Normally, the subject has to shift his/her attention continuously and sequentially like scanning. The serial counting task has a predominant advantage: it has a clear boundary over time, which opens opportunities for quantitative analysis. The stimuli on our experiment are strings of identical numeral characters, such as 0000, or 9999999. The length of each string is randomly chosen from 4 to 8. Numerals are also chosen in a random, so that the performance of a subject may not be hampered by particular features of certain numerals.

The spatial frequency is tuned by using different font sizes of the string. To quantify the response time of smaller stimuli is neither possible in the experiment nor valuable in applications. Since many researchers have suggested that in a counting task, the smallest interval should not smaller than 5 arcmin[10], which corresponds to 5px under the condition of our experiment, the possible sizes of a character in our experiment are 8px, 10px, 15px, 20px, 30px and 50px. The font sizes and the viewing distance are deliberately chosen so that there is no risk of over-pixelization[8], even at the smallest scale of 8px.

2.1 Evoking concentration of attention

The key in the experiment is to evoke the attention to concentrate in each trial. If the experimental task instead becomes a continuous process without interrupting, a subject would benefit by utilizing previous attentional status, and his/her performance would relate only weakly to the scale parameters[6]. In other words, aiming at quantifying the attention, we have to divert the attention from the status of being tuned to a particular position or a particular scale. Thus, we set each string a random appearing position, rather than making them pop up at the center of the monitor.

In the experiment, we also set an *anchor point*. An anchor point is located aside the screen. The subject is told to fix his/her attention on the anchor point just before and immediately after a counting trial. Naturally, when a new trial starts, the subject would abandon his/her attention at the anchor point, "zooming out" to search for the string, and re-concentrate his attention. In addition to the spatial disparity of the anchor point and stimuli, a difference in depth also required the subject to shift optical focus plane, thus further shuffles the subject's attention.

2.2 Experiment configurations

Subjects and environments: 6 subjects (all of which are college students with normal or corrected-to-normal vision) were included in the experiment; each was exposed to approximately 30 min stimuli. 5 of them were naïve to the purpose of the experiments. Subjects were given the instruction to "count the number of characters as fast as possible, and, when finished, look at the anchor point."

All visual stimuli were displayed on a calibrated 19-inch LCD monitor, with viewable size $376mm \times 301mm$, resolution $1280px \times 1024px$. The distance from the monitor to the subject is $1m$. The anchor point is located at $1.3m$ away from the subject, on the left to the screen. When concentrating attention on the anchor point, a subject moved his/her attention leftwards without turning head, so that the ocular muscle activities could be recorded by our apparatus.

Data recording: We used the NeuroScan system to collect electro-oculogram (EOG) at the sampling rate of 100Hz. EOG had been proved effective in tracking eye movements [9]. Our system has a temporal resolution high enough to

distinguish whether a subject is looking at the anchor point or looking at the stimuli. Fig.1 shows the recorded data from 3 of our subjects

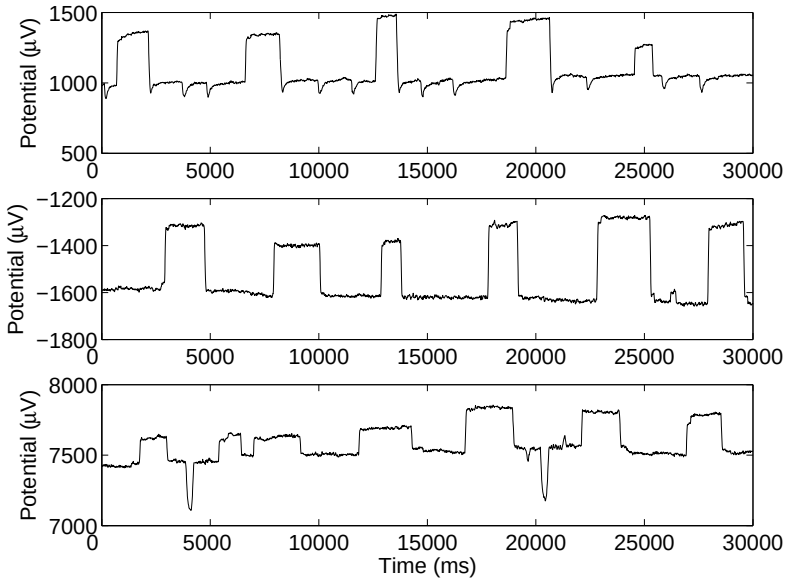

Fig. 1. This is a piece of horizontal EOG data. High potential corresponds to the activation of ocular muscle that makes the eye move rightwards. Although the potential is vulnerable to electrical activities of other muscles, the steep raises and falls of the curve are obvious. It is therefore easy to give a qualitative interpretation of the response time of attention.

3 Data Analysis

Since the viewing angle between the screen and the anchor point is 20°, shifting attention between stimuli and the anchor point results remarkable raises and falls in horizontal EOG signal. The rising edge of a EOG signal corresponds to the arrival of eyesight from anchor point to the screen, while the falling edge of a EOG signal corresponds to the departure of eyesight from the screen. Thus the "counting time" in each trial is the duration of the square wave. In processing, we tailored the periods overwhelmed by electrical activities of other muscles. At last, 1462 identifiable trials were included in our data set.

More generally, we prefer to interpret the data in frequency domain. Measured in c/deg, the frequency f corresponding to a particular font size s is given by

$$f = c/deg = 60/s.$$

The conversion from size domain to frequency domain is shown in Figure 2.

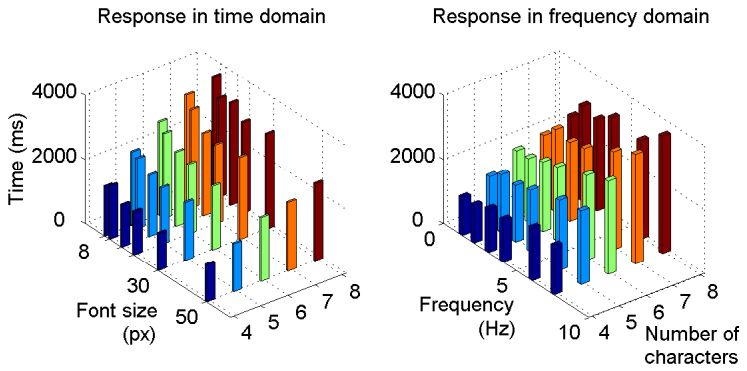

Fig. 2. A comparison of size domain and frequency domain representations

3.1 Two stages of response time

We analyzed the behavior of attention into 2 stages: *localizing* and *counting*. The localizing stage starts from the departure of attention from the anchor point to the moment when the subject detects the first numeral of the string. Note that this process is determined by the spatial frequency of the stimuli only, we denote this function as $L(f)$. The second part of the response time is the "counting" part. We denote this function as $C(f, n)$ One simplification can be made by implying $C(f, n)$ as a linear function of n, that is, $C(f, n) = n \cdot C(f)$. This linearity has been discussed by previous studies[10].

In sum, given frequency f and string length n of a trial, the response time $T(f, n)$ is

$$T(f, n) = L(f) + (n - k) \cdot C(f), \tag{1}$$

in which, $n-k$ denotes the times of jumping from one character to its propinquity.

It is possible for a subject to "apprehend" the length of a string from its shape without counting [10]. In that case, the actual numbers of counting may be less than the length of the string. For example, a subject may started counting from the third or the fourth numeral, or comprehended the number of last two or three numerals so the trial is finished in advance. A variable k is designed to take such "apprehensive counting" into consideration.

3.2 Normalization invariance

Although different people behaved differently in our experiment, the normalized responses along frequency axes all leaded to an identical shape. If we define the normalization as $N(f,n) = \frac{T(f,n)}{\max(T(f_i,n))}$, the normalized response time would be:

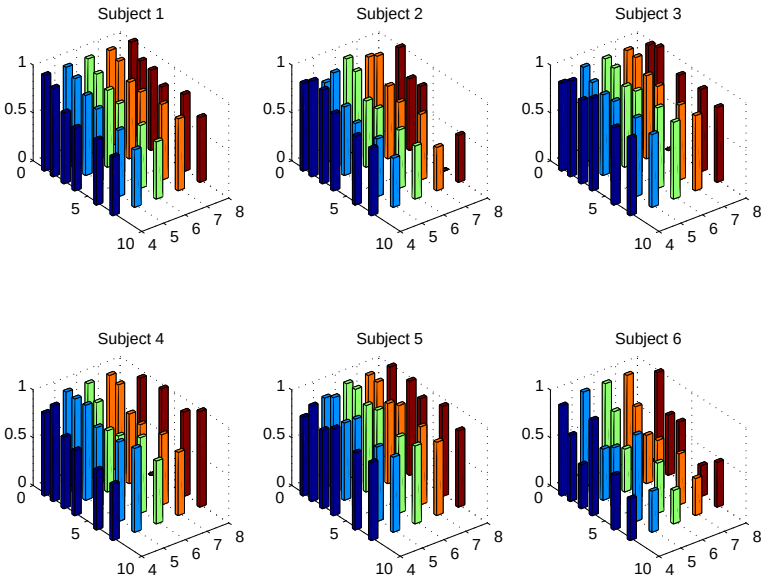

Fig. 3. Normalized data of different subjects

It is obvious from Figure 3 that the response time of different spatial frequency stimuli extend in a way that is irrespective to the string length. We call this property the *normalization invariance*.

This invariance can be defined as follows

$$\frac{T(f_i,n_i)}{T(f_j,n_i)} = \frac{T(f_i,n_j)}{T(f_j,n_j)}. \tag{2}$$

Substituting $T(f,n)$ into Equation 1, we obtain

$$L(f_i) \cdot C(f_j) = L(f_j) \cdot C(f_i), \tag{3}$$

or

$$\frac{L(f_i)}{L(f_j)} = \frac{C(f_i)}{C(f_j)}. \tag{4}$$

3.3 A computational model of attention

We adopt the exponential function to describe $L(f)$ and $C(f)$ as follows

$$L(f) = c_1 \cdot {c_2}^f, \tag{5}$$

$$C(f) = c_3 \cdot {c_2}^f, \tag{6}$$

where f is the spatial frequency of the stimuli, c_1, c_2, c_3 are parameters that distinguish the behavior of a particular person.

$L(f)$ and $C(f)$ satisfy the normalization invariance of Equation 4, since

$$L(f_1) \cdot C(f_2) = c_1 \cdot {c_2}^{f_1} \cdot c_3 \cdot {c_2}^{f_2} = c_1 \cdot c_3 \cdot {c_2}^{f_1+f_2} = L(f_2) \cdot C(f_1).$$

In this model, $L(f)$ and $C(f)$ share a common parameter c_2. In an empirical way, we interpret this parameter as a factor that summarizes the personal eye conditions and observing habits concerning to one subject in different spatial frequencies.

We define the error function as

$$e = \frac{\sum \left[T(f,n) - \hat{T}(f,n)\right]^2}{N_{f,n}}, \tag{7}$$

in which, $T(f,n)$ denotes the actual value of response, $\hat{T}(f,n)$ denotes the calculated value of our estimation, and $N_{f,n}$ denotes the number of trials to the corresponding frequency and character number.

In our experiment, the choice of k is not directly derived from our experimental results. However, since

$$T(f,n) = c_1 \cdot {c_2}^f + (n-k)c_3 \cdot {c_2}^f = (c_1 - kc_3) \cdot {c_2}^f + nc_3 \cdot {c_2}^f,$$

the values of c_2 and c_3 are independent to k. This fact legitimates us to propose an arbitrary k, and discuss c_2 and c_3 safely. Previous studies indicated that a subject could not enumerate more than 4 objects simultaneously in a counting task [10]. Accordingly, an acceptable choice of k in our framework is 4.

We plotted the response curve of each subject, and compare our prediction with actual records. From the figures shown below, we can see that the exponential function represents the characteristics of the responses of the subjects.

3.4 Speculations on the concentration process

We may interpret that c_1 and c_3 as the condition parameter of attentional concentration. Before the stimulus was detected, the attention had been concentrated at the anchor point. Once the subject detected the popped up string,

Table 1. parameters of different data sets

Data set	c_1	c_2	c_3	e
Subject 1	76.76	1.0691	31.87	841.6
Subject 2	61.16	1.0987	32.30	1311.8
Subject 3	157.12	1.0362	25.68	1006.6
Subject 4	118.94	1.0533	25.83	1986.3
Subject 5	219.76	1.0334	38.86	841.6
Subject 6	39.49	1.1552	32.09	1794.9
All	112.09	1.0633	33.57	4627.6

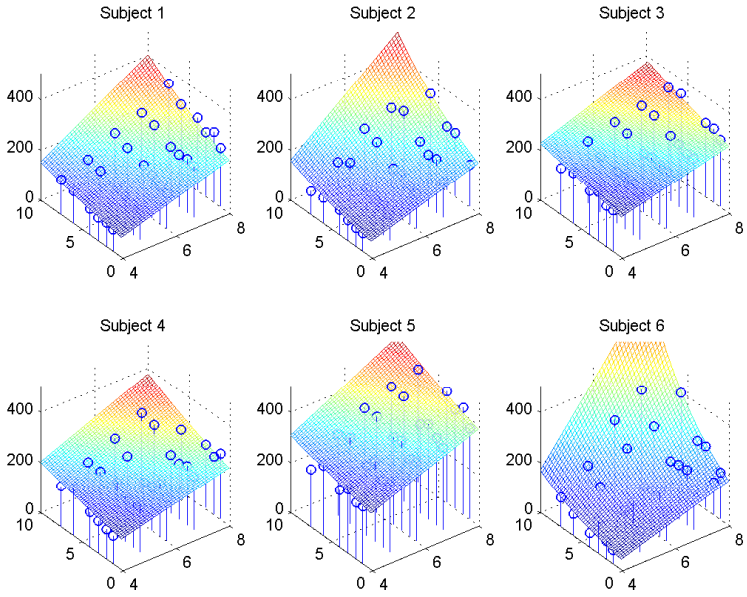

Fig. 4. Fitting experimental results with the model

he/she had to move his/her sight through a relatively long distance from the anchor point to the stimulus. In this process, the concentrated attention may not be maintained due to the rapid movement of eyes. In the stage of localization, the attention would thus be re-initialized from a totally diffused status to the desired scale. We denote such process of attentional concentration as $S(\infty) \rightarrow S(f)$, in which $S(\infty)$ describes the status of the worst condition of attentional concentration, $S(f)$ describes the degree of concentration after which the counting stage starts.

In the second stage of counting, a subject did not need to shuffle his attention because the eye moved with much smaller increments, and the concentrated attention, in some degree, may remained and might be reused in the counting of its neighboring numeral. Similarly, we denote this process as: $R(f') \rightarrow R(f)$.

4 Discussions

4.1 Of human vision

Inspired by Shannon's Theory, many scholars have analyzed human vision from a perspective of entropy[11]. These researches pioneered a new frontier to inspect the gap between human and machines. It is true that television programs are more attractive than a dead wall, but in order to calculate and compare the amount of information in different patterns, many studies oversimplified human to a camera. This camera prototype is questionable when we, for example investigate the human's attentional responses to grasslands and faces. Both images have similar amount of information in high frequency, but in general cases, human eyes are more inclined to attach to the latter. This phenomenon can be explained by our temporal analysis of attention. Faces are rich in both low and high frequency information, whereas a piece of well cultivated grasslands has extremely abundant in high frequency information but very scarce in low frequency information. According to our theory, the observer's attention has a very coarse resolution at a first look, and thus "blind" to the high frequency information of the grass, such as veins and shapes of the leaves. The low frequency component, however, catches the eyes and serves like an entrance. With the absence of low frequency information, one is less likely to initiate a careful observation at a certain region.

4.2 Of machine vision

As a counterpart to human vision, machine vision in many aspects aims at simulating the behavior of human. Nevertheless, before calculating in a humanoid way, we must be sure that the information we provide for an artificial processing system is identical to what our eyes provide to our brain. If not, it would be groundless to expect artificial processors to behave like human beings. We do not believe that a machine vision system should intentionally implement all the visual defects of human - electronic devices do not need to gazing on patterns

to enhance resolution. However, we should be aware of the complex and simple tasks for human vision system. In some daily tasks, a human brain may triumph over artificial devices, but this is not because we have sharper sensors, or faster processors, but because we have the wisdom to select proper data to process. The time-dependent capacity of attention can help us find out what kind of information is usually perceived by human vision system. Understanding the fact that with very limited information, a human brain still works in high performance, we can feed information of less quantity but more quality to an artificial system.

Acknowledgements

The work was supported by the National Basic Research Program of China (Grant No. 2005CB724301) and National Natural Science Foundation of China (Grant No.60375015).

References

1. Sperling, G.: The information available in brief visual presentation. Psychological Monographs, (1960)
2. Ruderman, D.:The statistics of natural images. Network: Computation in Neural Systems. **5** 517-548, (1994)
3. Treisman, A., Sykes, M., Gelade, D.: Selective attention and stimulus integration, In Donic, S.(Ed) Attention and performance VI 333-361 (1977)
4. Eriksen, CW., St. James, JD.: Visual attention within and around the field of focal attention: a zoom lens model. Perception and psychophysics. **(**40) 225-240 (1986)
5. Müller, N.G., Bartelt, O.A., Donner, T.H., Villringer, A., Brandt, S.A.:A physiological correlate of the "Zoom Lens" of visual attention. Journal of Neuroscience **23 (9)** 3561-3563 (2003)
6. Verghese, P., Pelli, D.G.: The information capacity of visual attention. Vision Research. **32 (5)** 983-995 (1992)
7. Verghese, P., Pelli, D.G.: The scale bandwidth of visual search. Vision Research. **34 (7)** 955-962 (1994)
8. Cha, K., Horch, K.W., Normann, R.A.: Reading speed with a pixelized vision system. Journal of the Optical Society of America A-Optics & Image Science, **9 (5)** 673-677 (1992)
9. Coughlin, M.J., Cutmore, T.R.H., Hine, T.J.: Automated eye tracking system calibration using artificial neural networks. **76** 207-220 (2004)
10. Intriligator, J., Cavanagh, P.: The spatial resolution of visual attention. Cognitive Psychology **43** 171-216 (2001)
11. Reinagel, P., Zador, A.M.: Natural scene statistics at the centre of gaze. Network: Computation in Neural System. **10** 1-10 (1999)