June 2005

Generating Sequence of Eye Fixations Using Decision-theoretic Attention Model

Erdan Gu
University of Pennsylvania

Jingbin Wang
Boston University

Norman I. Badler
University of Pennsylvania, badler@seas.upenn.edu

Generating Sequence of Eye Fixations Using Decision-theoretic Attention Model

Abstract
Human eyes scan images with serial eye fixations. We proposed a novel attention selectivity model for the automatic generation of eye fixations on 2D static scenes. An activation map was first computed by extracting primary visual features and detecting meaningful objects from the scene. An adaptable retinal filter was applied on this map to generate "Regions of Interest" (ROIs), whose locations corresponded to those of activation peaks and whose sizes were estimated by an iterative adjustment algorithm. The focus of attention was moved serially over the detected ROIs by a decision-theoretic mechanism. The generated sequence of eye fixations was determined from the perceptual benefit function based on perceptual costs and rewards, while the time distribution of different ROIs was estimated by a memory learning and decaying model. Finally, to demonstrate the effectiveness of the proposed attention model, the gaze tracking results of different human subjects and the simulated eye fixation shifting were compared.

Generating Sequence of Eye Fixations Using Decision-theoretic Attention Model

Erdan Gu[†] Jingbin Wang [‡] Norman I. Badler [†]
[†]Computer and Information Science Department, University of Pennsylvania, PA, 19104
[‡]Computer Science Department, Boston University, MA, 02215
[†]{erdan, badler}@seas.upenn.edu [‡]jingbinw@cs.bu.edu

Abstract

Human eyes scan images with serial eye fixations. We proposed a novel attention selectivity model for the automatic generation of eye fixations on 2D static scenes. An activation map was first computed by extracting primary visual features and detecting meaningful objects from the scene. An adaptable retinal filter was applied on this map to generate "Regions of Interest" (ROIs), whose locations corresponded to those of activation peaks and whose sizes were estimated by an iterative adjustment algorithm. The focus of attention was moved serially over the detected ROIs by a decision-theoretic mechanism. The generated sequence of eye fixations was determined from the perceptual benefit function based on perceptual costs and rewards, while the time distribution of different ROIs was estimated by a memory learning and decaying model. Finally, to demonstrate the effectiveness of the proposed attention model, the gaze tracking results of different human subjects and the simulated eye fixation shifting were compared.

1. Introduction

The human visual system is highly non-uniform in sampling, coding, processing and understanding. It is determined by the anatomical structure of the human retina, composed of a high-resolution central fovea and a low resolution periphery. The visual attention system directs the limited computational resources to a small subset of sensory information from environment stimuli for visual processing. Consequently, the visual system places the fovea on the interesting parts of the scene. How the visual attention system works efficiently will be decomposed into four subquestions:

- How does the visual system know what information is important enough to capture attention?

  The visual system usually employs two mechanisms to limit processing to important information of the world. They appear to be implemented in a rapid, bottom-up, conspicuous-driven manner or in a slower, top-down, task-prominent manner [6]. The bottom-up setting has been developed in many computer vision models [19] [8] to make use of "saliency" for directing attention. The saliency map is established by integrating all of the separated feature maps, which highlight certain parts of the scene that differ from their surroundings by specific feature extraction [3]. As described in Section 2, the saliency map in current work was a combination of primary feature maps on color, intensity and orientation [9]. Other than the feature saliency map, the final activation map also integrates objects (face) pop-out [2] [4] and the peak locations of the map became candidates for the "Regions of Interest" (ROIs). ROI, or fixation field, is the area of scene to be fixated upon.

- What kind of "mental image" results from the non-uniform coding of the world stimuli?

  Itti et al. [7] implemented the foveation filter through interpolation across levels of a Gaussian Pyramid to compute the "mental image". But Gaussian model is inconsistent with empirical data on the mapping from the primate retina to the visual cortex. The current method applied log-polar sampling [1] as an approximation to the foveated representation of the visual system. To speed up the computation, the method partitioned the log-polar retinal plane into receptive fields. Then, an adaptable division of receptive fields was applied iteratively to determine the appropriate size of the fixation field. The details of the above process is given in Section 3.

- How does the visual system know how to allocate the focus of attention to interpret the scene rather than doing it at random?

  Models of visual information acquisition are classified into two categories. Visual search [16] emphasizes on locating a single target and the search time required. Other models [21] focus on the eyes, regarded as a "single server queue", in visual scanning. The crucial concern is not target detection, but instead the scan

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 IEEE

order and the viewing time assigned to various ROIs. Many works [9] [17] present the sequence of the attended locations in the order of decreasing saliency. This strategy, however, conflicts with a fact of visual scanning people are not willing to move their gaze frequently. Therefore, considering perceptual cost and reward, Wickens et al. described an attentional expected value model, which was validated by the experiments of pilot task management [21]. But it was a descriptive model, which mainly stressed that the dynamic processing is under control. From a decision-theoretic perspective, we proposed a computational model to find an optimal selection strategy for 2D static scenes in Section 4.

- How do we assess the plausibility of our attention selectivity model?

  It is believed that eye movements are tightly linked to visual attention [22]. Thus, tracking eye movement is a suitable means for studying the simulated visual attention selectivity. In Sections 5 and 6, an empirical validation method was performed by comparing the performance of the computational model and human subjects. The actual eye fixation sequences and looking time can be obtained from gaze tracking results for eye movement video. Afterwards, an objective comparison criterion was defined to assess the plausibility of the model-simulated gaze shifting behavior.

2. Generation of Activation Map

One important mode of the attentional operation, bottom-up control, automatically performs independent extraction of features in parallel and processes them. The main criteria driving attention here is odd target pop-out, which generally falls into two categories: visual feature extraction at the lower level and object recognition at the higher level.

We applied the method of Itti et al [9] for the extraction of primary features (see Fig.1). First, a number of visual maps of multi-scale images are computed, for different image features, e.g., color, intensity and orientation, using center-surround difference. Then, the feature maps obtained on different scales are summed, respectively, into three saliency maps in a competitive way. Finally, a single scalar measure, which expresses the salience at each location of the scene image, can be determined by linear averaging the three saliency maps.

Besides the primary visual features, the current method also detected pop-out objects based on their social relevance, in particular, human faces by the method proposed by Paul Viola et al [20]. A learning process based on AdaBoost, as an efficient classifiers, was applied to select a small number of critical visual features from "Integral Image" windows. To reduce the computation cost, these classifiers were combined in a cascade manner, eliminating the need for further processing of background regions. The current application used the CMU face database for learning purposes. As a result, the final activation map was obtained by combining the scalar saliency map and the detected faces.

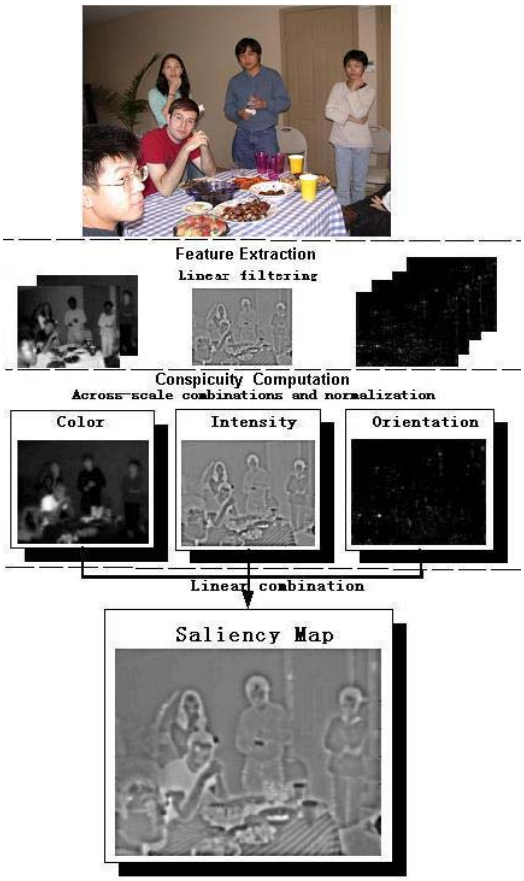

Figure 1: Generation of saliency map for "party" image: (1) feature extraction, (2) saliency computation for intensity, color and orientation, and saliency map generation.

3. Estimation of Regions of Interests

The human fovea has a much higher density of photoreceptor in the center than the periphery. As a result, people direct their fovea to the part of scene that they wish to attend. Given the computed activation map, the fixation points were defined as the peak locations of the activation map while fixation field sizes were estimated by an adaptable retinal filter centered on the fixation points.

3.1. Adaptable Retinal Filter

A fixation image is defined as the transformation of the world image by retinal filter. It is computed by a

complex-logarithmic fall-off function with eccentricity. The log-polar transformation, or so-called *logmap*, was studied as a good approximation to the retino-cortical mapping in the human visual system [18]. The logmap, $l(X)$, was defined as a conformal mapping from the cartesian plane $X = (x, y)^T$ to the log-polar retinal plane $z = (\xi, \eta)^T$:

$$l(X) = \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} log[x^2 + y^2] \\ arctan\frac{y}{x} \end{pmatrix} \quad I(z) = I(l^{-1}(X)) \tag{1}$$

To allow real-time computation of the logmap images $I(z)$, we partitioned the retinal plane into ten receptive fields, whose size and position correspond to a uniform partition of the plane. The innermost receptive field, defined as the fixation field, corresponds to the part of the scene falling onto fovea, being sampled with the highest resolution. Thus, the inner receptive field forms clear patch while the other fields represents the blurred patch of the retinal image. When an interesting location, e.g.,the "BBQ beef" in Fig. 2, was fixated on, the sampling rate of the fixation field rose, and consequently, the accuracy of the perceived information improved. On the other hand, as the size of the fixation field shrank, the blur patch became larger and lost more acuity due to the limited visual resources. Interpolation across different receptive fields was implemented to eliminate artifacts due to the sample discontinuities on the boundaries of receptive fields. The partition of receptive fields changes as a function of the fixation field size, illustrated by Fig. 2(a) and 2(b), which is in accordance with human dynamic sampling behavior.

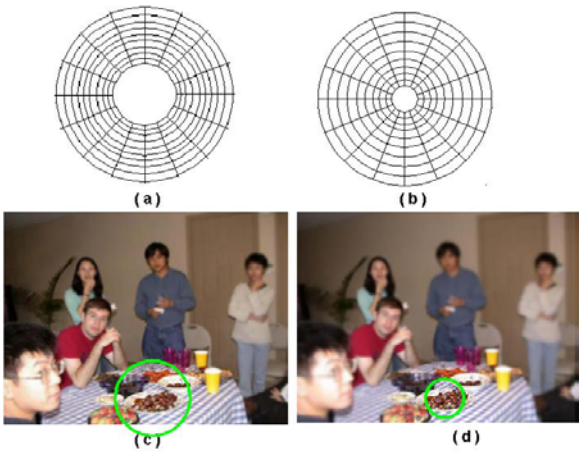

Figure 2: Adjustable retinal filter: (a-b): Partitions of receptive field with large fixation field and small one. (c-d): Resulting retinal images from the corresponding partition (a-b).

3.2. Estimation of Fixation Field Size

There are two traditional assumptions to account for the conspicuity-based attention strategy. The space-based attention theory processes everything within a spatial window while the object-based theory argues that attention is actually directed to objects rather than regions. In current implementation, a space-based setting was applied to facilitate the computation of the fixation field size. The fixation field was defined as a regular disk area with center position and radius. For the detected face area, the width of the sub-window used in face detection system was used as the approximated diameter of the fixation fields.

The candidates for fixation were always locations of the local maxima of the normalized activation map. The fixation field sizes, however, could vary in different scene images, and even for the same image, since they were dependent on the distance from the observer to the scene. Moreover, more activation peaks will merger into a single field when the larger size of fixation field offered. Therefore, as suggested by Pomplun's work [16], the method applied an iterative algorithm to adaptively adjust the fixation field sizes in order to achieve the desired fixation fields. The algorithm started using an initial fixation field with an arbitrary size, which agreed with human dynamic sampling behavior that starts from a random process. The computation process then counted the number of peaks of the activation map to determine the number of fixations. The size adjustment procedure stopped when the number of fixations approached an empirical threshold learned in the preliminary study.

In summary, the properties of the i-th detected ROI are represented as:

$$ROI_i = \{AV_i, A_i\}_{i=1..n} \qquad \sum_{i=1}^{n} AV_i = 1 \tag{2}$$

$$AV_i = W_i \sum_{x,y \in A_i} S_{x,y} P_{x,y} \tag{3}$$

$$A_i = \{x_{center}, y_{center}, radius\} \tag{4}$$

where AV_i is a weighted sum of the activation value within area A_i, which defines the geometrical properties of the area. S_{xy} represents the scalar value of the given location on the activation map. A position weight P_{xy} is assigned by a normalized Gaussian template centered at the image. Different weights W_i are applied to ROIs for the low level features and the face areas, respectively. The weights applied for the low-level features were chosen to be smaller than those applied for the faces, and the ratio of these two types of weights was about 0.5, which was empirically decided based on the preliminary study on a wide range of naturally occurring images.

4. Attention Selectivity

Attention selectivity attempts to optimally allocate the limited human visual resource to explore the perceived environment. It assures the retrieval of the necessary informa-

tion for interpreting the scene in a timely manner. To simulate the above procedure, the model should automatically decide where to look, when to look there and how long to stop there.

4.1. Decision-making Attention Shifting

In the current application, we assumed all observers are at a comfortable distance from the scene image so that eye movement with a fixed head pose suffices for acquiring the necessary information. The movement of eye is inexpensive but not "free". Thus, the fixation shifting behavior should be penalized. The current method took the above two aspects into account and modelled them via a designed benefit function, where the overall perceptual benefit was computed as the summation of penalties of gaze shifting and rewards of perceived information. The penalty was computed based on shifting distance between ROIs, while the perceptual reward was associated with the importance of the perceived information. Hence, we have:

$$B(k) = ReW(i_k) - C(i_{k-1}, i_k) + B(k-1) \quad (5)$$

$$ReW(i_k) = f(R(i_k)); \quad (6)$$

$$R(i_k) = K_r AV_{i_k}; \quad (7)$$

$$C(i_{k-1}, i_k) = K_c Dist(i_{k-1}, i_k) \quad (8)$$

where $B(k)$ represents the maximum gained benefit after the gaze fixation had shifted for k times and stopped on the i_k-th ROI[1]. Expected cost $C(i_{k-1}, i_k)$ represents the perceptual cost by shifting the attention from i_{k-1}-th ROI to the i_k-th ROI. Reward value $R(i_k)$ denotes the information importance of the i_k-th ROI relevant to other ROIs in the 2D scene and is dependent on the value AV_i. $ReW(i_k)$ is a rewardable variable, computed by function $f(R(i_k))$. $f(.)$ is a time dependent function and will be described in detail in next section. The function $Dist(i_{k-1}, i_k)$ is a L2 distance measurement between the i_{k-1}-th ROI to the i_k-th ROI. It is important to strike a careful balance between the influences from the perceptual costs and rewards values through constants K_c and K_r, so that the penalty value is usually smaller than the reward value to ensure the the profitability of gaze shifting. On the other hand, the cost can not be too small, otherwise it becomes negligible.

To maximize the benefit function, one basic strategy assumes people would normally shift their focus of attention to a nearby ROI than one farther away for comparable rewardable. But it allows the attention goes the farther ROI instead of the closer one when the former carries much more important information. This is different from the Greedy Heuristic in *scanpath* theory, which assumes that people is so lazy that they rather linger on many insignificant items close to the most salient object, instead of paying attention to other salient targets which are a little further away.

4.2. Memory Learning and Decay Model

The duration of the fixation affects the accuracy of the acquired information. Such Information accuracy is a key factor in computing the benefit of shifting attention because humans attempt to maintain a certain level of accuracy for the acquired information in practice.

There are two types of monitoring behaviors that people perform: overt monitoring and covert monitoring. During the overt monitoring, a person is actively attending to a target by placing their fovea on it, consequently, the information accuracy of the target becomes higher. If a high enough accuracy of ROI is reached, its reward is set to zero to simulate inhibition of return. Otherwise, people will continue attending to the same ROI since it has higher reward. During covert monitoring, the target is monitored from memory, and the accuracy decreases over time. When the accuracy drops below a certain threshold of tolerant accuracy, the value of $ReW(i_k)$ will set back to $R(i_k)$ and the target ROI will return to the competition pool as a candidate choice for the next location to be attended. This means that the fixation will often move back to targets already visited a long time ago. In the current implementation, the threshold of tolerant accuracy is 50%.

With respect to the above two types of monitoring behaviors, two models, a power law of learning model [14] and a model of memory decay [12], are applied to measure the accuracy level of the perceived information. These two models are respectively expressed as:

$$Learn : k_l * exp(b_l * t^l_{i_k}) = AV_{i_k} * \Delta P_{i_k} \quad (9)$$

$$\Rightarrow t^l_{i_k} = \frac{1}{b_l} \log(AV_{i_k} \Delta P_{i_k}) - \frac{1}{b_l} \log k_l \quad (10)$$

$$Decay : k_d * exp(b_d * t^d_{i_k}) = AV_{i_k} * \Delta P_{i_k} \quad (11)$$

$$\Rightarrow t^d_{i_k} = \frac{1}{b_d} \log(AV_{i_k} \Delta P_{i_k}) - \frac{1}{b_d} \log k_d \quad (12)$$

where the time $t^l_{i_k}$ denotes how long it takes to raise the accuracy level ΔP_{i_k} since the eye fixates on the i_k-th ROI. P_{i_k}, a percentage value, represents the information accuracy of the i_k-th ROI. ΔP_{i_k} represents how much accuracy is retained or forgotten for the i_k-th ROI, respectively, for the learning and decay models. For the first viewing i_k-th ROI , P_{i_k} goes up from 0 to 1, thus, ΔP_{i_k} equals 1.0. Similarly, $t^d_{i_k}$ is the time spent covertly monitoring the i_k-th ROI since the last overt fixation on it, which simulates the accuracy degradation process. The time function of decay is set to be much slower than the acquisition model although both of these process are exponential. k_l, b_l, b_d, and k_d are constants empirically decided, where the values of b_l and b_d decide the performance of the decay and acquirisition function. To determine the value of the key ratio $\beta = b_l/b_d$,

1 Subscript k indicates the times of eye fixation shiftings before the i_k-th ROI receieve the attention.

we used the magic number, 7 ± 2 slots for working memory, proposed by Miller [13]. In a summary, function $f(.)$ in EQ. 6 is defined as:

$$f(R(i_k)) = \begin{cases} R(i_k) & if \quad t^f_{i_k} < \frac{1}{b_l}\log(AV_{i_k}\Delta P_{i_k}) - \frac{\log k_l}{b_l} \\ R(i_k) & elseif \quad t^d_{i_k} > \frac{1}{b_d}\log(AV_{i_k}\Delta P_{i_k}) - \frac{\log k_d}{b_d} \\ 0 & otherwise \end{cases} \tag{13}$$

Given the list of ROIs for the synthetic "shape" image (Fig. 3), the activation array for six ROIs was calculated as $AV[6] = \{.41, .14, .23, .06, .14, .03\}$. The processing of acquisition and decay are plotted (Fig. 3:below), where the decay curves stop at the accuracy threshold for a tolerance 50%. For this example, information accuracies are maintained within the tolerant range until all ROIs have been scanned. Thus, with sufficient short-term memory, attention is not supposed to shift back the formerly attended ROI until all ROIs are viewed.

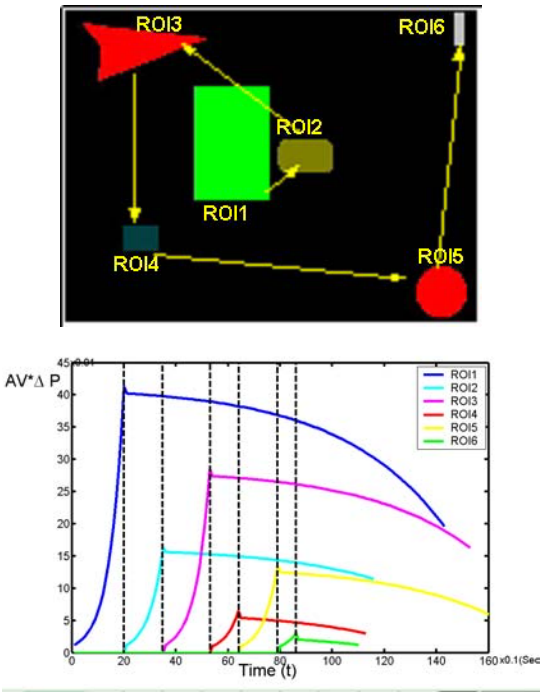

Figure 3: Top: Attention shifting path for a synthetic image "shape". Here, ROIs are defined as these arbitrary-shaped objects. Below: Memory acquisition and decay curves for the "shape" image. Dashed lines indicate the moment when the acquisition processing for currently attended ROI is completed. Afterward, attention shifts to the next attended ROI and currently attended ROI will be monitored from memory.

4.3. Finding Optimal Solution

Given the image marked with m ROIs, we can construct a complete graph with ROIs as the graph nodes. The edge weights in the graph are defined as the shifting benefit between nodes. The goal is to find an optimal path that passes through all ROIs and ends with a maximum value of $B(n)$ where $n \geq m$. Due to dynamically changing weights of edge in the graph, finding an optimal path can not be reduced to a shortest path problem. We solved the current problem by a dynamic programming mechanism, summarized as Algorithm 1. The algorithm takes m ROIs as input, and outputs the results as a transition path between ROIs, the corresponding time duration and the accuracy level for each ROI. Two examples of the generated eye fixation sequences are illustrated in Fig. 3 and Fig. 4, respectively.

Algorithm 1 MAXBENEFIT(ROI $j, j = 1 \ldots m$);

Step 1: $B(1)$=Max($R(j)$), $j \in 1..m, m \leq n$;
$i_1 = argmax_j(R(j))$;
$ReW(i_1)$=0;
Step 2: $B(2)$=$ReW(i_2)$-$C(i_2, i_1)$+$B(1)$;
$ReW(i_2)$=0;
$t(i_1) = \frac{1}{b_l}\log AV_{i_2} - \frac{1}{b_l}\log k_l$;
$P(i_1)$=1-$\frac{AV_{i_2}}{AV_{i_1}}$;
if $P(i_1)$ out of tolerance
$ReW(i_1)$=$R(i_1)$;
$t(i_1)$=0;
Step 3..$n-1$**:** ...
Step n**:** $B(n)$=$ReW(i_n)$-$C(i_n, i_{n-1})$+$B(n-1)$;
$ReW(i_n) = 0$;
for $k = i_1$ to i_{n-1}
$t(k)$=$t(k+1)$+...+$t(n)$;
$P(k)$=1-$\frac{1}{AV_k}$ $AV_{k+1} \ldots AV_n$;
if $P(k)$ out of tolerance
$ReW(k)$=$R(k)$;
$t(k)$=0;
Output: Max($B(n)$), and a path of ROI $i_1 \ldots i_n$ and a duration series $t(i_1) \ldots t(i_n)$ and accuracy level $P(i_1) \ldots P(i_n)$.

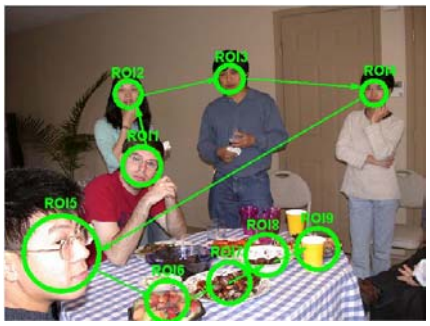

Figure 4: Attention shifting path for "party" image.

5. Validation of Attention Selectivity Model

The biological grounding of saliency-based attention, as reported in Itti's work, was empirically assessed by Ouerhani's work [15]. Thus, it is believed that the construction

Figure 5: Gaze tracking results on some frames

of the activation map and the subsequent estimation of ROIs are reliable. Moreover, recent works [17] have confirmed that fixation points correspond to the most salient parts of the scene. Thus, in the current application, we focused on validating the plausibility of the simulated attention selectivity mechanism. For this purpose, human subjects were asked to watch test images with detected ROIs marked on them. The eye movements of the subjects, actively scanning between the ROIs, were recorded. Then gaze shifting pattern were extracted by gazing tracking via *Active Appearance Model* (AAM) [5] as follows.

AAM applied PCA to model both the shape variations of image patches and their texture variations. The model established a compact parameterization of object variability for human eye as learned from a training set by estimating a set of latent variables. To model both the shape x and texture g of the eye movements, a parameterized shape and texture model can be represented as below:

$$x = \bar{x} + \phi_s b_s, \qquad g = \bar{g} + \phi_g b_g \tag{14}$$

where $\bar{x}$ and $\bar{g}$ denoted the mean shape and the mean normalized grey-level vector, respectively. Eigenvectors ϕ_s and ϕ_g were obtained from the training set covariances, and represented the variations of the shape and texture of eyes across the given training samples. A residual vector δI between the model and image, $\delta I = I_{image} - I_{model}$, was regressed against a known displacement vector δc, such that $\delta c = A\delta I$. By an iterative updating scheme, the robust gaze tracking was achieved even for low quality Web-Cam images. Fig. 5 shows some example frames for gaze tracking.

The obtained tracking sequences were parsed into the attention shifting path and fixation times were allocated for each ROI. These empirical results were then compared with the simulated results derived from the computational model for the same image data.

6. Experiments and Discussions

Different test images, seen many times in advance by all subjects, were used in the current experiments. It assures the subjects will use natural internal cognitive models to look at the images [17] with a fixed head pose, at a comfortable distance from the screen. The gaze-position accuracy needed was quite low so that simple calibration system can work out which ROI a gaze falls into, once the gaze tracking data is collected. We can then parse the tracking results into fixations, brief stops, saccade and rapid jumps, etc., using different parsing parameters. Before the objective comparison was performed, both model-simulated fixation times and actual duration times from human subjects were normalized since viewing times for various 2D scenes are heavily subject-dependent. The sequences and time distributions for "shape" (Fig. 3), "party" (Fig. 4)and "bench" (Fig. 6) images are shown in the Table 1, 2, 3, below.

To evaluate the coherence of the empirical and simulated results, we performed a quantitative measurement. A dissimilarity metric is defined as the minimal cost for transforming one distribution of the sequential representations onto the other precisely. Given n fixation shifting, let P be the computer-generated sequence, which consists of $P_i = (p_i; \rho_i^p), i \in 1..n$, where for the i-th fixation, p_i denotes the attended ROI and ρ_i^p is time distribution. While Q is the obtained empirical data with same representation. The editing distance $D_e(P, Q)$ was needed to minimize. The optimization problem was defined with the swapping operation, the only operation for editing, assigned to the unit cost $s(.)$, on the sequence.

$$s(Q_i, Q_j) = |i - j|(\rho_i^q + \rho_j^q) \tag{15}$$

$$N(s(Q_i, Q_j)) = \frac{s(q_i, q_j)}{\sum_{k=1}^{n}(s(q_i, q_k) + s(q_k, q_j))} \tag{16}$$

$$D_e(P, Q) = \min_{P,Q}(\sum_{i,j} N(s(P_i, P_j)), \sum_{i,j}(N(s(Q_i, Q_j)))) \tag{17}$$

We normalized the swap cost to make it comparable to the following L1 distance. Once the editing problem was solved, we got the distribution (P, Q') with the same sequence order, where Q' was the edited sequence of Q. Then, we used L1 distance, $D_d = \sum_{i=1}^{n} |\rho_i^p - \rho_i^q|$, to measure the dissimilarity. The total transformation cost D_t was obtained by $D_t = D_e + D_d$. The transformation cost denoted how closely two sets of data resembled each other in distribution for the sequential representations.

The results of the comparison demonstrated a correlation between the human eye movements and the computer modelled eye movements, for a synthetic image (Fig. 4), as shown in Table 1. For natural images, some human subjects presented a good correlation to the computer model while others were inharmonious with it. Besides the errors introduced by the simulation procedure of the proposed model, one possible reason for the result is the variation that exists between human subjects, causing each person to interpret the scene image differently. Moreover, inaccurate measurement of eye movements from the gaze tracking process could also introduce other type of errors. In summary, the above measurements demonstrate a preliminary but promising correlation between the attention of real humans and the proposed computer model.

Figure 7: Images sequences of fixations for "party" image: 1. the boy in red; 2.the girl on the left; 3.the boy in blue in the middle; 4.the girl on the right; 5. the boy with glasses on the left; 6.fruit salad; 7.BBQ beef; 8.meat ball; 9.yellow cup in the front.

Sequ	1	2	3	4	5	6	7	8	9	D_t
Comp	1: 0.19	2:0.13	3:0.15	4:0.11	5:0.13	6:0.08	7:0.09	8:0.07	9: 0.06	0%
Subj1	3: 0.19	2:0.15	1:0.16	4:0.13	5:0.12	6:0.06	7:0.09	8:0.06	9: 0.04	21.3%
Subj2	1: 0.18	2:0.12	3:0.19	4:0.12	5:0.11	6:0.10	7:0.10	8:0.05	9: 0.03	17.0%
Subj3	2: 0.15	1:0.16	3:0.16	4:0.13	5:0.10	6:0.08	7:0.10	8:0.07	9: 0.06	17.0%
Subj4	1: 0.18	2:0.12	3:0.18	4:0.13	5:0.14	7:0.08	6:0.10	8:0.05	9: 0.04	13.8%

Table. 3 Coherence results for "party" image

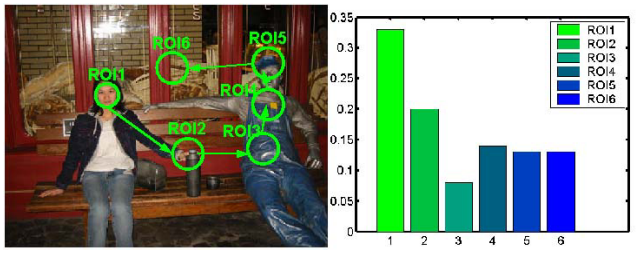

Figure 6: Computer-generated distribution of fixation for "bench" image was illustrated as histogram, and the sequence of the fixations was labelled on the image.

Sequ	1	2	3	4	5	6	D_t
Comp	1:0.25	2:0.17	3:0.21	4:0.12	5:0.17	6:0.08	0%
Subj1	1:0.18	2:0.19	3:0.25	4:0.12	5:0.15	6:0.11	19%
Subj2	1:0.16	2:0.16	3:0.25	4:0.13	5:0.15	6:0.15	24%
Subj3	1:0.20	2:0.17	3:0.22	4:0.14	5:0.15	6:0.12	14%

Table. 1 Coherence results for "shape" image: The model generated fixation sequence (row 1) and the sequences of three human subjects (rows 2-4) are shown. We use "3:0.19" for $p_1 = 3$, $\rho_1^p = 0.19$, meaning ROI3 was viewed as the first item in the sequence and the allocated time proportion was 19% over the whole sequence duration. The last column shows the computed transformation cost. The same notation applies for Table.2, 3.

ROIs	1	2	3	4	5	6	D_t
Sequ	1:0.33	2:0.20	3:0.08	4:0.14	5:0.13	6:0.13	0%
Subj1	1:0.35	6:0.12	2:0.14	3:0.06	4:0.09	5:0.23	44.7%
Subj2	1:0.42	2:0.12	3:0.05	4:0.06	5:0.29	6:0.06	42.0%
Subj3	1:0.38	2:0.25	3:0.04	4:0.05	5:0.20	6:0.08	35.0%

Table. 2 Coherence results for "bench" image

7. Conclusions

The human behavior of attention selectivity is elusive and still far from being well understood. A complete theory of top-down guidance remains unavailable for modelling the visual attention nowadays. This paper aims to present a decision-theoretic attention model, which allows automatic generation of the sequence of eye fixation and its time distributions on 2D scenes. The proposed attention model can be potentially useful in many applications, such as robot controlling, human-computer interaction, animation, or interactive games, etc. A full assessment of the model needs a large number of experiments that involve more human subjects and test images. In the future, we are planning to apply the current model for creating more human-like animation characters. Within this application, the performance of the proposed model will be better evaluated both in terms of the animated agents' behaviors and their interactions with the 3D virtual environment.

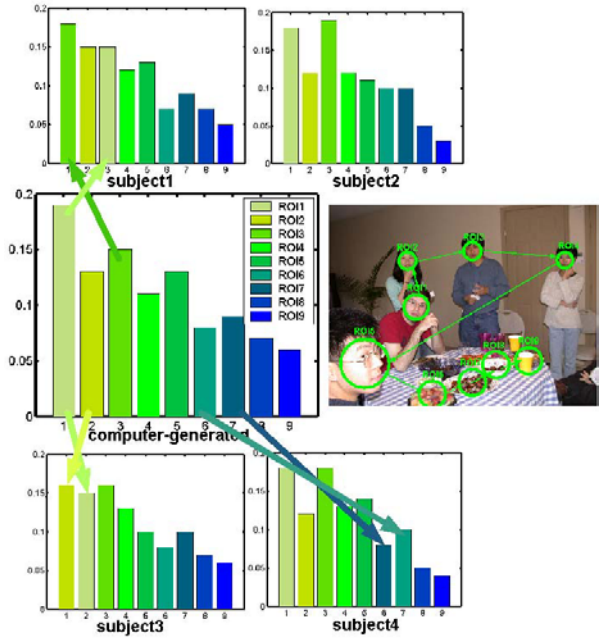

Figure 8: Middle: Computer-generated attention shifting sequence and time distribution results (left) for the "party" image (right). Top: Results from Subject 1 (left), where the fixation orders for the 1st and 3rd ROIs in the sequence are swapped; Results from Subject 2(right). Bottom: Results from Subject3 (left), where the order of the 1st and 2nd ROIs are swapped; Results from subject4 (right), where the orders of the 6th and 7th ROIs are swapped.

Acknowledgments

This work is partially supported by the ONR VITRE project under grant N000140410259. The authors are grateful to Catherine Stocker for her helpful editing work and Jan M. Allbeck, Amy Calhoun for their assistances. Also the authors thank all the voluntary participants in our experiments.

References

[1] A. Bernardino and J. Santos-Victor. A binocular stereo algorithm for log-polar foveated systems. In *Proc. 2nd International Workshop on Biologically Motivated Computer Vision*, pages 127–136. Springer-Verlag, 2002.

[2] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1146–1153, 1999.

[3] K. Brunnstrm, J. Eklundh, and T. Uhlin. Active fixation for scene exploration. *International Journal of Computer Vision*, 17:137–162, 1996.

[4] L. Chen, X. Xie, W. Ma, H. Zhang, and H. Zhou. Image adaptation based on attention model for small-form-factor device. In *Proc. of 9th International Conference on Multi-Media Modeling*, 2003.

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Lecture Notes in Computer Science*, 1407:484–499, 1998.

[6] L. Itti. Visual attention. *The Handbook of Brain Theory and Neural Networks*, pages 1196–1201, Jan. 2003.

[7] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proc. SPIE 48th Annual International Symposium on Optical*, pages 21–21, 2003.

[8] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–27, 2001.

[9] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[10] J.E.Hoffman and B. Subramaniam. Saccadic eye movements and visual selective attention. *Perception and Psychophysics*, 57:787–795, 1995.

[11] P. Majaranta and K.-J. Räihä. Twenty years of eye typing: systems and design issues. In *Proceedings of the symposium on Eye tracking research and applications*, pages 15–22. ACM Press, 2002.

[12] N. Moray. Designing for attention. *Attention: Slection, Awareness, and Motor Control*, 1993.

[13] A. Newell. Unified theories of cognition. 1990.

[14] A. Newell and P. Rosenbloom. Mechanisms of skill acqusition and the law of practice. *Cognitive skills and their acquistion*, pages 1–55, 1981.

[15] N. Ouerhani, R. von Wartburg, H. Hügli, and R. Müri. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision Image Anal.*, 3, no. 1:13–24, 2004.

[16] M. Pomplun, E. M. Reingold, and J. Shen. Area activation: A computational model of saccadic selectivity in visual search. *Cognitive Science*, 27:299–312, 2003.

[17] C. Privitera and L. Stark. Algorithms for defining visual regions-of-interest:comparison with eye fixations. *PAMI*, 22(9):970–982, 2000.

[18] E. Schwartz. Spatial mapping in primate sensory projection:analytic structure and relevance to perception. *Biological Cybernetics*, 25:181C194, 1977.

[19] Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligent*, 146:77–123, May 2003.

[20] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[21] C. D. Wickens, J. Helleberg, X. X. J. Goh, and W. J. Horrey. Pilot task management: Testing an attentional expected value model of visual scanning. In *Technical Report ARL-01-14/NASA-01-7*. NASA Ames Research Center Moffett Field, CA, 2001.

[22] A. Yarbus. Eye movements and vision. 1967.