An Approach for Preparing Groundtruth Data and Evaluating Visual Saliency Models

Rajarshi Pal, Jayanta Mukherjee, and Pabitra Mitra

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur, India
{rajarshi,jay,pabitra}@cse.iitkgp.ernet.in

Abstract. Evaluation is a key part while proposing a new model. To evaluate models of visual saliency, one needs to compare the model's output with salient locations in an image. This paper proposes an approach to find out the salient locations, i.e., groundtruth for experiments with visual saliency models. It is found that the proposed human hand-eye coordination based technique can be an alternative to costly human pupil-tracking based systems. Moreover, an evaluation metric is also proposed that suits the necessity of the saliency models.

Keywords: Evaluation, Visual saliency model, Groundtruth.

1 Introduction

Visual saliency models try to emulate human vision. Hence, the ideal way to evaluate these models is to estimate how similar (or dissimilar) the results obtained by these models are with the salient regions detected by human. Eye-tracking based systems are used to produce groundtruth data for the evaluation of saliency models.

The eye-tracking technology is costly and few research groups having access to it. The lack of technology of recording salient regions from human observers is evident as many of the works ([1], [2], [3]) have skipped the proper objective evaluation procedure. According to [4] the results obtained from the model was shown to a group of volunteers and they were asked to assess the result as either good, acceptable or failed. This evaluation is quite subjective. In [5], the model's performance is compared against randomly generated locations, not against salient locations reported by humans.

Therefore, technological bottleneck for collecting groundtruth data is a hindrance to the correct evaluation of visual saliency models. As an alternative to human eye-tracking based systems, in this paper, an approach to collect groundtruth data is discussed where volunteers' opinion will be recorded through hand-eye coordination.

Evaluation metric is equally important as collecting groundtruth data. Some of these metrics operate only if the groundtruth data is converted into a human attention map. An attention map (from human fixation) is formed by summing 2D Gaussian patches around each fixation point or using some variation of it.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 279–284, 2009.

Correlation coefficient between the human attention map and model predicted saliency map [6], difference between model predicted saliency map and human attention map [7], and area under ROC curve [8] are used as evaluation metrics. Another class of metrics may be formulated if the groundtruth is specified as a set of points. Average [9] and summation [10] of saliency values at these points are two such metrics. Though all these estimates are shown to work fine as a metric for saliency models, they are not tuned to the specific purpose of using visual saliency. They are very general in nature and are also applied in other fields of computer vision. In this paper, an alternative metric for evaluating saliency models is proposed. The proposed metric is very much tuned to the fact that real-time tasks process only a few salient locations. This metric discourages any model that selects less salient locations.

The outline of the remaining paper is as follows: Section 2 discusses the collection and compilation of groundtruth data. An evaluation strategy is proposed in section 3. Section 4 reports some experimental results to support the proposed approach and finally, section 5 draws the conclusion.

2 Preparation of Groundtruth Data

Let N be the number of image for which groundtruth data, i.e., the salient locations to be recorded. Let V be the number of volunteers recording the salient locations. Not to burden them with much load, each of them is shown n images. To ensure that opinions of a good number of volunteers are taken for each image, the values of these variable are set to satisfy $V \times n >> N$. The ratio $(V \times n)/N$ indicates the strength of evaluation. Typical values of N, V and n, for our experiments, are 50, 62 and 12, respectively.

2.1 A Session with an Individual

At first, the volunteer is shown a very prominent spot at a randomly chosen location within a white background. The spot remains for τ_1 (typically 100 ms in our experiments) time in the screen. The volunteer is asked to click the mouse to the location where the point appeared. The point disappears so quickly that the volunteer has to mark on the white background after its disappearance. The same process is repeated for m_1 (typically 12 in our experiments) times. The Euclidean distance between the actual position of appearance of the point and the position where the volunteer clicked is recorded for the last m_2 (typically 8 in our case) out of those m_1 instances. Mean μ (denoted as mean offset) and standard deviation σ (denoted as standard deviation of offset) of these recorded distances are estimated. The record for first (m_1 - m_2) cases are ignored to give the volunteer some time to get familiar with the procedure. One needs to be very much attentive to have a good hand-eye coordination. This phase serves two purposes. Firstly, it helps the volunteer to be attentive before the recording of salient locations for test images commences. Secondly, an estimate of how

far she clicks from the true position is found by taking the mean and standard deviation for the last m_2 instances.

As soon as the first phase completes, the second phase starts where the volunteer's opinion about the salient locations for n images are accumulated. Each image is shown to her for a very short period τ_2 (typically, 100 ms). Human vision has two components. When confronted to a scene, of which it has no prior clue, at first it will be guided to the visually salient locations in the scene. This is called *bottom-up component* of vision. With time the familiarity of the scene increases and the contents of the scene begin to be recognized/interpreted. Then this gradually enhanced understanding of the scene guides our vision. This is called *top-down component* of vision. As the objective here is to collect the groundtruth data to evaluate visual saliency model, i.e., bottom-up attention model, the time period τ_2 for which the image is shown to the volunteer needs to be very small. Smaller τ_2 indicates, lesser influence of top-down component of our vision and better capturing of salient locations. A white screen follows each image. The volunteer is asked to mark at the centers of each location that seems salient to her. Like the first phase, here too she marks on white screen that follows the image. This process is repeated for n images.

2.2 Combining Individual's Opinion to Form Salient Locations

Good hand-eye coordination is crucial for deriving groundtruth data. Therefore, the mean and standard deviation of both μ and σ (obtained in section 2.1) of all V volunteers are calculated. Let μ_μ and σ_μ be respectively mean and standard deviation for the mean distances μ and μ_σ and σ_σ are respectively mean and standard deviation for the standard deviations of distances σ of all the volunteers. The opinions of the volunteers, whose mean offset μ and variance of offsets σ are less than or equal to $(\mu_\mu + \sigma_\mu)$ and $(\mu_\sigma + \sigma_\sigma)$ respectively, are taken into consideration for groundtruth data preparation. Others opinions are discarded as their hand-eye coordination is poorer than the rest. Let S_V and S_v be the set of all volunteers and selected volunteers, respectively. Therefore,

$$S_v = \{i | (\mu_i \leq (\mu_\mu + \sigma_\mu)) \wedge (\sigma_i \leq (\mu_\sigma + \sigma_\sigma)) \wedge (i \in S_V)\} \tag{1}$$

Now, for each image the following procedure is applied to prepare the groundtruth data using the opinions of only the chosen set of volunteers S_v. Let, P_v be the set of points marked by volunteer v for an image I. Set of points P is found by taking together all the volunteers' responses for the image I.

$$P = \bigcup_{\forall v \in S_v} P_v \tag{2}$$

Let, N_v be the number of points marked by volunteer v for the image I. A set N is formed comprising of all N_v's.

$$N = \{N_v | \forall v\} \tag{3}$$

Next k-means clustering is performed on the set of points P. Number of clusters k is set to be the mode (i.e., the value that occurs most number of times) of the set N. The mean μ_C and standard deviation σ_C for each cluster C are also computed. As each volunteer marks at the approximate center of each location that seemed salient to her, a circular disk centering at μ_C of radius σ_C covers some portion around the center of each cluster. Collection of these circular disks (one or many for a particular image) constitute the groundtruth data for that image. It may be noted that all these circular regions are non-overlapping, as the clusters form a partitioning in the space. A binary image B (of same size as I) is constructed where all the pixels belonging to these disks (salient locations) are set to 1 and the remaining pixels are set to 0.

3 Evaluation Metric

It is checked whether there exist other locations which are more salient than the groundtruth indicated in B (obtained in previous section). Let S be the saliency map obtained for image I. Moreover, let m_i represent the maximum saliency value for each salient region R_i (in the form of a circle as discussed in previous section) in B.

$$m_i = max(S(R_i)) \tag{4}$$

where $S(R_i)$ represent the collection of values in the saliency map S corresponding to the region R_i.

It is checked whether there are other pixels that have saliency value greater than m_i and do not belong to any other salient region R_j in B. Let Γ is the set of such pixels. If no such pixel exists, then Γ becomes a null set.

$$\Gamma = \{x | \, (S(x) > m_i) \wedge (x \notin R_j) \wedge (\forall R_j \in B)\} \tag{5}$$

If Γ is not the null set, the error measure E_i for R_i is the sum of normalized distances for all pixels in Γ from R_i. Distances are normalized with respect to $\sqrt{(L^2 + M^2)}$, where image I is of size L-by-M.

$$E_i = \sum_{y \in \Gamma} mindist(y, R_i) / \sqrt{(L^2 + M^2)} \tag{6}$$

where mindist(a,A) is the minimum of all the Euclidean distances of a pixel a from a group of pixels A.

Error estimate E for the saliency model is the summation of all error estimates E_i for all values of i.

$$E = \sum_i E_i \tag{7}$$

In a nutshell, the summation of minimum distance of each pixel in the locations that are more salient (according to the computational model) from salient regions indicated in B is the proposed evaluation metric.

4 Experimental Validation

In our experiments, 50 images are chosen from a larger set of databases taken from iLab image database ([1], [11]), UCID([12]), Zurich natural image database ([13] and the Internet. A well-known saliency model [1] is used to validate the procedures described in above two sections. Groundtruth data is prepared using the procedure described in section 2. This data along with the input images is given in `http://www.facweb.iitkgp.ernet.in/~jay/VS/Groundtruth.html`. The groundtruth is represented in a binary image with regions filled with 1 belong to salient locations. Let for an image I, the set of salient locations obtained by the proposed procedure is denoted by R.

On the other hand, for each of the input images, a set of circular regions are chosen with randomly selected center positions. Let, for the image I the set of randomly chosen circular regions is denoted by T. For each image, the number of randomly selected circular regions, i.e., cardinality of T, is kept equal to the number of salient regions obtained in the groundtruth (cardinality of R, i.e., which is equal to k). Moreover, radius of each of the circular salient locations in R are maintained in T. The purpose, here, is to show that scores at locations obtained according to the proposed procedure (R) is higher than the scores at randomly selected locations (T). As in [9], average of saliency values at these locations are used as evaluation metric. Scatter diagram in figure 1 shows that average saliency value at the volunteer specified locations (average over 50 images is 140) is higher than average saliency value at randomly selected locations (average over 50 images is 72.16). It strengthens the fact that locations obtained by proposed procedure are salient.

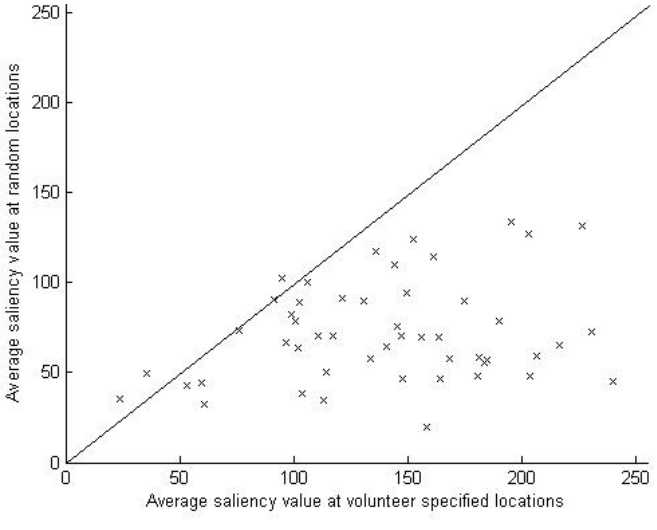

Fig. 1. Scatter Diagram of average saliency values at locations obtained by proposed procedure versus that of randomly chosen locations

It is also experimentally observed, as expected, that the average saliency value (measure of similarity) and the proposed metric (measure of dissimilarity) are negatively correlated. Their correlation coefficient is measured to be -0.58. This statistics justifies the proposed metric as a error measure.

5 Conclusion

In this paper, a hand-eye coordination based procedure is proposed to compile groundtruth data. Experimental results show that this can be an alternative to using eye-tracking systems. An evaluation metric is also proposed which is an error measure by definition. The fact that it negatively correlates with another evaluation metric (average saliency value) strengthens this proposed metric.

References

1. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
2. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision 45(2), 83–105 (2001)
3. Sun, Y., Fisher, R.: Object-based visual attention for computer vision. Artificial Intelligence 146, 77–123 (2003)
4. Yu, Z., Wong, H.S.: A rule based technique for extraction of visual attention regions based on real-time clustering. IEEE Transactions on Multimedia 9(4), 766–784 (2007)
5. Minut, S., Mahadevan, S.: A reinforcement learning model of selective visual attention. In: Proceedings of 15th International Conference on Autonomous Agents, pp. 457–464 (2001)
6. Meur, O.L., Callet, P.L., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(5), 802–817 (2006)
7. Bruce, N.D.B.: Features that draw visual attention: an information theoretic perspective. Neurocomputing 65-66, 125–133 (2005)
8. Gao, D., Vasconcelos, N.: Bottom-up saliency is a discriminant process. In: Proceedings of IEEE 11th International Conference on Computer Vision, pp. 1–6 (2007)
9. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vision Research 42, 107–123 (2002)
10. Meur, O.L., Thoreau, D., Callet, P.L., Barba, D.: A spatio-temporal model of the selective human visual attention. In: Proceedings of IEEE International Conference on Image Processing, pp. III–1188–1191 (2005)
11. Itti, L., Koch, C.: Feature combination strategies for saliency based visual attention systems. Journal of Electronic Imaging 10(1), 161–169 (2001)
12. Schaefer, G., Stich, M.: Ucid - an uncompressed color image database. In: SPIE Storage and Retrieval Methods and Applications for Multimedia, vol. 5307, pp. 472–480 (2004)
13. Frey, H.P., Konig, P., Einhauser, W.: The role of first- and second- order stimulus features for human overt attention. Perception and Psychophysics 69, 153–161 (2007)