# Investigating individual differences in learning-based visual saliency models

Binbin Ye[1] Yusuke Sugano[1] Yoichi Sato[1]

**Abstract:** Learning-based approaches for modeling visual saliency using a data set of human fixations are becoming increasingly popular in recent years. However, most of the prior studies do not consider individual differences in visual attention, which might potentially improve the fixation prediction performance of learned models. By taking the visual saliency model which incorporates visual field characteristics as an example, we investigate individual differences by statistically comparing different saliency models learned using person-dependent training data sets.

**Keywords:** Visual saliency, visual attention, individual difference, statistical hypoyhesis test

## 1. Introduction

Human's visual attention is attracted to salient stimuli, which means information captured by photoreceptor is not uniformly processed in the brain. Such mechanism is required for humans to rapidly discover important information. Some researches have shown that it is a challenging task even for human brains to simultaneously identify any and all interesting targets in one's visual field [16].

Modeling visual saliency has been a very active research field over the recent decade and many models of visual attention are available now [1]. Other than understanding and reasoning the mechanism behind visual attention, creating computational model that predicts interesting parts of images and videos has many meaningful real-life applications in the field of computer vision, such as image segmentation, thumbnailing, rendering, image compression, object detection and object recognition. Besides, it can also be found useful in robot control, advertisement design, and surveillance systems.

Itti *et al*. proposed the first complete implementation of computational model [9] by extracting low level visual features such as luminance contrast, color contrast, orientation and motion to predict interesting regions. Another representative method of visual attention model is Graph-based visual saliency, proposed by Harel *et al*. [7]. In this model, bottom-up features are extracted to form activation maps, then such activation maps will be treated as Markov chain to seek out nodes that have high dissimilarity with their surrounding nodes.

However, such conventional models often require a clear understanding of the biological visual systems in order to design the parameters, *e.g*., the type of visual features, shape and size of filters and normalization schemes. Since the mechanism of human visual system is not yet fully understood, designing such models in a biologically convincing way is not a trivial task.


**Fig. 1** Comparison of computational saliency map and actual fixation [10]

In Fig. 1, the first column shows input images, the second column shows saliency maps generated by Itti-Koch's model, and the last column shows comparisons between the saliency map and actual fixation locations. As can be seen, accurate prediction of fixation locations is not always easy for conventional bottom-up models.

Contrary to such rule-based approaches that try to formulate theories and assumptions behind visual attention, a data-driven approach of modeling visual attention is becoming increasingly popular in recent years. A data set of images and actual locations of human fixations is used as a training data to learn a computational model that accurately replicates the distribution of fixations. Judd *et al*. [10] proposed a method to use a linear support vector machine to train a model of saliency with predefine low, mid and high-level image features. Zhao *et al*. [18] proposed a method training

[1] Institute of Industrial Science, the University of Tokyo

a model of saliency by least square method with features defined in the Itti and Koch model [9]. Recently, Kubota *et al*. [12] improved learning-based model by introducing characteristics of human visual field. In this model, a visual field is divided into several regions according to the distance from the fixation center. Weight of features is trained separately in each field by least square method with graph-based visual saliency model [7] saliency map.

In these works, training data sets are usually collected from multiple test subjects and most prior researches do not take the consideration of individual difference when using these datasets. However, it is natural and instinctive to believe the existence of individual difference because people have different personal experience, gender and age, which might reflect in the habit of viewing a picture. Thus, considering individual difference in learning-based visual saliency model has a potential to improve its performance by, *e.g*., clustering people into a few major types and develop adoptive models for each type of people. The analysis on individual difference on visual attention is also helpful in the research of human biology systems. By studying the difference of viewing habit among infant, adult and the aged, for example, it will become possible to learn the growth and aging process of human brain and visual system.

In this work, as a prior study for future research, we investigate the existence of individual difference in Kubota *et al*.'s visual saliency model and their data set. We statistically compare models learned using individual data sets of each person, and the purpose of the experiment is two-fold:

(1) To assess if the personal models perform better on the fixation prediction task than a generic model
(2) To see how and where the difference between personal models arises

For 1), we compare NSS (normalized scanpath saliency) scores of personal and generic models via Wilcoxon signed rank test. For 2), we show a result of multivariate analysis of variance (MANOVA) of the feature weight vectors of the personal models. Throughout these tests, we provide a quantitative view on the individual difference of saliency models learned using personal training data.

## 2. Learning-based models of visual saliency

Learning-based visual saliency models generally consist of following steps [19]:

**Extraction of features**

Commonly used low-level features include color contrast, intensity, orientation and symmetry, as well as higher-level feature like face.

**Computation of individual feature maps**

Existing techniques are often used to generate feature maps with the extracted features, such as differentiation on image pyramids, Bayesian statistics, discriminant center-surround difference, entropy minimization and Markov chains.

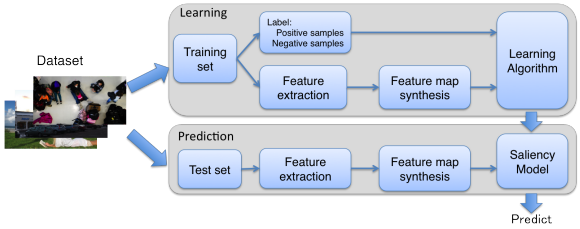



**Fig. 2** Flow chart of generating learning-based visual saliency models

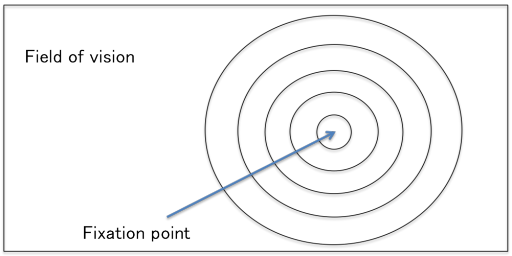



**Fig. 3** Divide the visual field into 6 regions according to fixation point

**Training of the feature integration model**

Given the feature maps and prerecorded fixation map, the task of learning-based model is to optimize the parameters of how to integrate them into the final saliency map. By overlaying fixation map over feature maps, positive training samples selected around the fixation locations on feature maps. Similarly, negative training samples are selected from the locations away from the fixations. Then postive samples and negative samples will be sent to a supervised learning technique to optimize the parameters of the model. Such techniques, typically, includes least square, artificial neural network, and support vector machines.Fig. 2 shows the flowchart of developing a learning-based model.

Kubota *et al*. [12] propounded that the visual field characteristics is an important effect that should be involved in optimizing model performance. Their baseline model is graph-based visual saliency (GBVS) model which is the bottom-up feature provider, as well as facial feature which is computed as a Gaussian distribution with respect to the center of the detected face, using a face detector from face.com [5]. In the learning phase, Kubota *et al*. divide the field of vision into 6 regions with the respect to the center of current fixation point in the saccade (Fig.3), guaranteeing each region has the same number of sacaade so that positive samples are equally distributed. Each region has 10 independent weights. They are color, intensity, orientation in 3 scale levels of 1/4, 1/8, 1/16 and face. Then a $i \times 60$ dimension matrix $\boldsymbol{F}$ of learning sample is made

$$\boldsymbol{F} = \begin{pmatrix} \text{Positive samples} \\ \text{Negative samples} \end{pmatrix},$$

with positive samples and negative samples. Each row is a traning sample and each column corresponds to a feature. On the other handd, elements in $i \times 1$ label matrix $\boldsymbol{l}$ is defined by

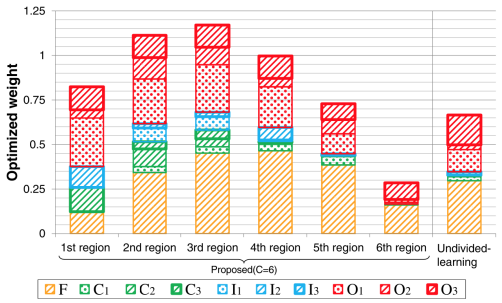



**Fig. 4** Comparison between divided model and undivided model. Feature name C=color, I=intensity, O=orientation, F=face. The following number $n$ is the scale level, which means the scale factor is $\frac{1}{2^{n+1}}$. e.g. C1 means Color contrast feature at the scale level of 1/4

**Fig. 5** Some samples of pictures in Kubota *et al.*'s dataset.

$$l_j = \begin{cases} 1 & l_j \text{ is positive sample} \\ 0 & l_j \text{ is negative sample} \end{cases}.$$

Finally, the model will be trained with non-negative least square technique,

$$\min_{w} \|\boldsymbol{FW} - \boldsymbol{l}\|^2, \text{ where } \forall w \in \boldsymbol{W}, w > 0,$$

gaining the feature weight vector $\boldsymbol{W}$ to minimize the error between $\boldsymbol{FW}$ and $\boldsymbol{l}$. Training result is shown in Fig. 4. It shows the weights and ratios of features in each region of visual field is significantly different from each other and they are also different from the weights in undivided model.

Another contribution of Kubota *et al.* is that they created a new data set with 57 ° of horizontal viewing angle, which is much wider than the data set applied in the prior works [2, 3, 10, 14, 14]

Fig. 5 shows examples of images contained in Kubota *et al.*'s data set. The 400 images in the dataset are randomly picked from Flickr Creative Commons. Gaze data of 15 test subjects is recorded by Tobii TX300 eye tracker at the rate of 60 HZ. Each image is shown to the subjects for 4 seconds. Eye movement faster than 22°/sec is considered a saccade and there are 5 saccade on each image on average.

## 3. Experiments

In this paper, we take Kubota *et al.*'s model as an example, investigate individual differences by statistically comparing different saliency models learned using person-dependent training data sets.

Fig. 6 shows training results using individual data of the 15 test subjects included in Kubota *et al.*'s data set. We can, subjectively, see that the shape of feature weight distribution of each person and ratio among the features in a each region is different. Our goal is to examine whether such an individual difference exists or not in a quantitative manner.

### 3.1 Effect of top-down feature in learning-based saliency

Before examining individual differences, we show additional experimental results on the effect of top-down feature used in Kubota *et al.*'s model. While human visual field characteristics can be correlated with bottom-up features such as color, intensity and orientation, its relationship with top-down features such as face is not obvious. Hence, incorporating face factor can even affect the learning results and it is not clear if their model can correctly reflect visual field characteristics.

In this section, we statistically examine if there exists an effect of using top-down feature in learning. More specifically, we compare models trained under following two conditions concerning face factor:

(1) All training samples are used, and all features are included.
(2) All training samples are used, but face factor is not considered.
(3) Samples in the images that include detectable faces are not used, and face factor is not considered

While the first condition corresponds to the original setting of Kubota *et al.*, face feature is simply excluded from their model in the second condition. In the third condition, images that include human faces are further excluded from the data set.

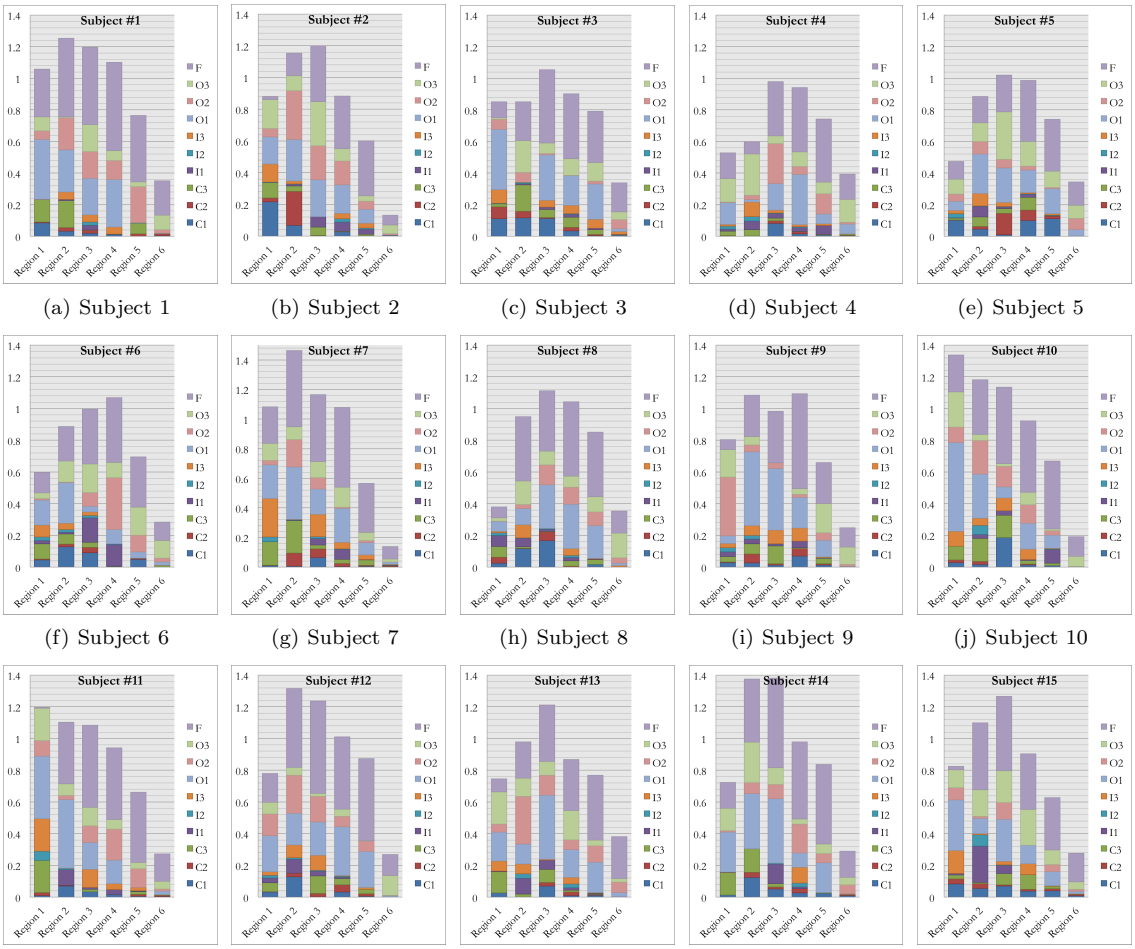
Examples of the learned feature weights under the three conditions are shown in Fig. 7. There are three figures for each subject. They are models trained under condition (1), (2) and (3) respectively and lined up in the order from left to right. The last grid shows a mean weight of the all subjects.

It can be seen that model *cond*.1 and model *cond*.2 are almost the same with each other and it suggests adding top-down feature does not affect the training results. In consideration of the fact that the amount of training samples is different that 146 out of 400 images are not used, it does not change a lot in the weights although some weights show dissimilarity in model *cond*.3.

To be more precise, we compare the Normalized Scanpath Saliency (NSS) [13] score over models in condition (1) and (2) to examine whether the perfromance are the same with each other.

The idea of NSS measure is to evaluate the pixel value on the saliency map along a subject's scanpath. To calculate NSS score, the first thing to do is normalize the model-predicted saliency map into a saliency map with a zero mean and unit standard deviation. Then the scanpath is overlaid on the normalized saliency map and the pixel values of saliency map on these fixation location are summed and averaged to get the NSS score.

We denote the model trained under condition ($i$) for sub-




(a) Subject 1 (b) Subject 2 (c) Subject 3 (d) Subject 4 (e) Subject 5

(f) Subject 6 (g) Subject 7 (h) Subject 8 (i) Subject 9 (j) Subject 10

(k) Subject 11 (l) Subject 12 (m) Subject 13 (n) Subject 14 (o) Subject 15

**Fig. 6** Feature weights in each region for each subject

ject $n$ by ${}^{cond.i}M_n$, where $n = 1, \cdots, 15$, We perform a two tailed Wilcoxon Signed-rank test [8] over the paired-data of ${}^{cond.1}\mathrm{NSS}_n$ and ${}^{cond.1}\mathrm{NSS}_n$, where $n = 1, \cdots, 15$. The null hypothesis is that ${}^{cond.1}\mathrm{NSS}_n - {}^{cond.2}\mathrm{NSS}_n$ comes from a distribution with zero median.

The $p$-value of hypothesis test is $p = 0.720 \gg 0.01$ which means we cannot reject the null hypothesis. Hence training with or without the face factor do not affect the weights of the rest 3 kinds of bottom-up feature and it is safe to compare the contribution in individual difference of the bottom-up features and face.

### 3.2 Performance comparison between personal and generic model

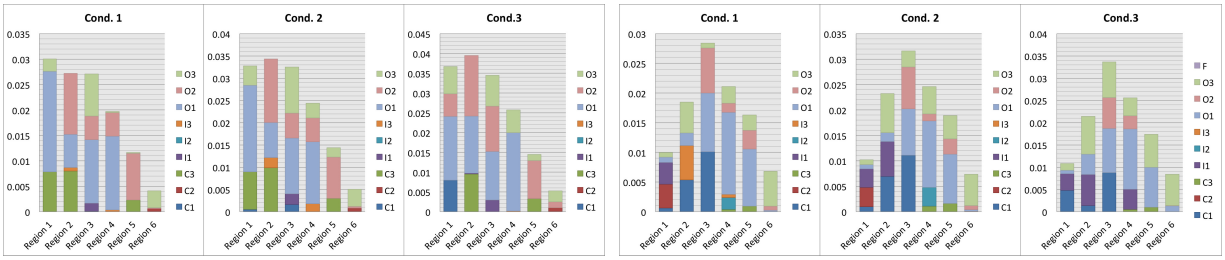
The first experiment about individual difference is comparing the score across personal and generic model. In this way, we first try to examine if personal models perform better than a generic model.

In this experiment, we divide the dataset into 100-image test set and 300-image training set. Training set is used for developing saliency models and the test set is for calculating NSS scores.

We denote the personal model for subject $n$ by ${}^{P}M_n$, where $n = 1, \cdots, 15$, and this is a model trained with the fixation data of subject $n$. Respectively, we denote the generic model without using training data from subject $n$ by ${}^{G}M_n$ and this model is trained with the data from rest of the test subjects. There are $10 * 6 = 60$ (10 features channels in 6 regions, 3 color channels, intensity channels, orientation channels and 1 face channel in each region) feature weights in one model.

The NSS score of ${}^{P}M_n$ and ${}^{G}M_n$, ${}^{P}\mathrm{NSS}_n$ and ${}^{G}\mathrm{NSS}_n$, which are tested with subject $n$'s fixation data should be different if the individual difference exists. In our experiment, ${}^{P}M_n$ is trained with all positive and negative samples in the training set, while ${}^{G}M_n$ is trained with 20% randomly selected samples from the training set. We perform a Wilcoxon Signed-rank test [8] over the paired-data of ${}^{P}\mathrm{NSS}_n$ and ${}^{G}\mathrm{NSS}_n$, where $n = 1, \cdots, 15$ (Fig. 8).

In our test, the null hypothesis is that ${}^{P}\mathrm{NSS}_n - {}^{G}\mathrm{NSS}_n$ comes from a distribution with zero median at the 1% significance level. As we assume that personal model has better performance than generic model, the alternate hypothesis states that ${}^{P}\mathrm{NSS}_n - {}^{G}\mathrm{NSS}_n$ come from a distribution with median greater than 0. The $p$-value of right-tailed hypothesis test is $p = 0.00018311 \ll 0.01$ so that it is reasonable to reject the null hypothesis at 1% significance level.




(a) Subject 1 (b) Subject 8

(c) Subject 15 (d) Mean of the feature of 15 subjects

**Fig. 7** Bottom-up feature weights learned with and without face factor in saliency model

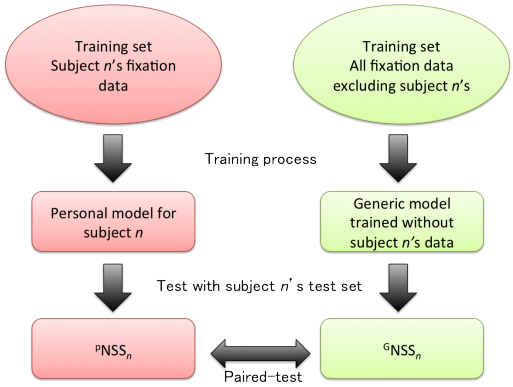



**Fig. 8** Data construction of experiment 1

### 3.3 Difference between personal models

While it is verified that personal models can perform better than generic models, more direct comparison of learned feature weights should be made to see individual difference. In this section, we carry out a one-way multivariate analysis of variance over the feature vectors (MANOVA) [11] to test the null hypothesis that the means of each personal model are the same n-dimensional multivariate vector at 5% significance level.

In this experiment, all images in the dataset are taken as training samples, in other words, there is no division of training set and test set. To obtain the feature vectors, we divide the dataset into 10 subsets as training sets, which makes every subset have 40 images. Then, for every test subject, 10 personal models will be trained with the data of 10 training sets as 10 observations. We denote the personal model for subject $n$ trained with subset $m$ by

$$^{P}M_{n,m} \quad \text{where} \quad n = 1, \cdots, 15 \text{ and } m = 1, \cdots, 10$$

The test results are listed in Table. 1. The first column stands for the type of feature vector that is put into the test. Second and third columns show how many different groups and how many observations per group are involved in the MANOVA test. The fourth column is the demension of feature vector. The last column is an estimate of the dimension of the space containing the group means. There is no evidence to reject the null hypothesis if d = 0. If d = 1, the null hypothesis can be rejected at the 5% significance level though the hypothesis that the mean lies on the same line.

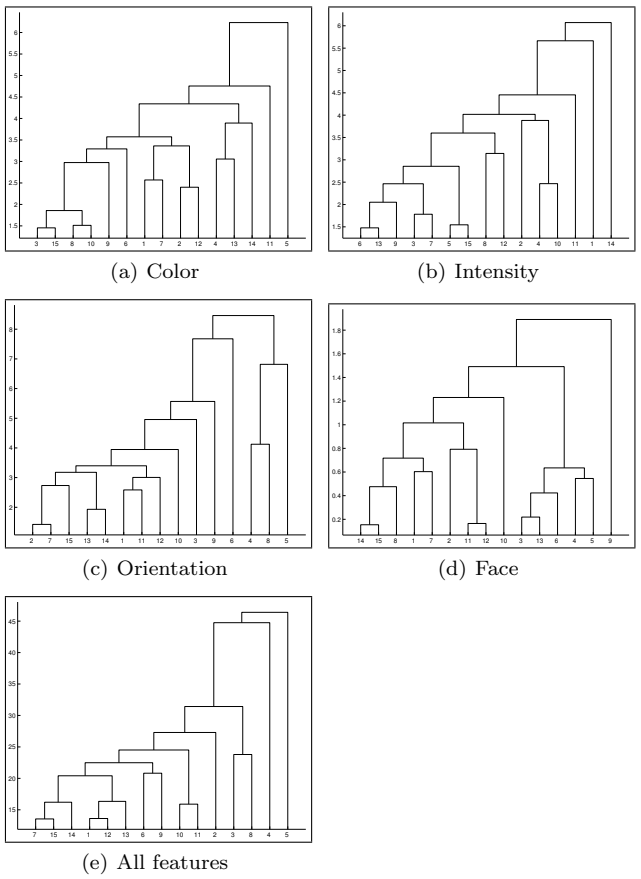
**Table 1** Result of multivariate analysis of variance

| Type of feature | Observations | Groups | Dimension | d |
|---|---|---|---|---|
| All | 10 | 15 | 60 | 2 |
| Color | 10 | 15 | 18 | 0 |
| Intensity | 10 | 15 | 18 | 0 |
| Orientation | 10 | 15 | 18 | 1 |
| Face | 10 | 15 | 6 | 0 |

When the test object is full feature vector of the models, we can reject the null hypothesis at 5% significance level although we cannot reject the hypothesis that the multivariate means may lie on the same plane in 60-dimensional space. This is another evidence indicating the existence of individual difference that the actual weight distribution has some sort of variation.

However, the difference seems to disappear when a single type of feature vectors is put into test. In the second row of Table 1, *e.g.*, $d = 0$ suggests that there are no difference in the distribution of color vector from test subject to test subject, so are distribution of intensity and face feature. Only the orientation feature shows the difference at 5% significance level.

We further apply a hierarchical clustering based on the result of MANOVA and visualize the clustering results as dendrogram in Fig. 9. We can see that color (Fig. 9(a)), intensity (Fig. 9(b)) and orientation (Fig. 9(c)) share a similar topology and subjects cannot be well grouped by any of them, although the d value of orientation is 1 and others are 0. On the other hand, subjects seem to be well grouped by



(a) Color (b) Intensity

(c) Orientation (d) Face

(e) All features

**Fig. 9** Hierarchical clustering using UPGMA (Unweighted Pair Group Method with Arithmetic Mean) based on different features

face (Fig. 9(d)) while the d value of face is 0.

When all features are considered, the topology of grouping becomes different from any of the above (Fig. 9(e)). Therefore, the individual difference is considered to be the result of the accumulation of the minor differences rather than caused by a single specific feature.

## 4. Conclusion

In this paper, we investigated the individual difference by two statistical experiments. We have shown that the NSS score of personal model is statistically higher than the NSS score of generic model. Under the premise, personal adaptive saliency model is possible to train to improve the accuracy of prediction.

We also attempted to explore where and how the difference between personal models arises in experiment 2, however we found there is no clear distinction between the test subjects until all features are used. There are two possible reasons for this phenomenon:

( 1 ) Different features are not perfectly independent from each other. Suppose there is an image contains a vertical white line over black background, orientation in the picture is represented by the intensity contrast.

( 2 ) Every single feature can be taken as a weak classifiers and the individual difference (strong classifier) is a linear combination of multiple weak classifiers, which is similar to the idea of AdaBoost.

It is an important task to examine these hypotheses with larger amount of data.

Even if personal adaptive saliency models can perform better than generic models, it is not practical to learn tailored models for each user. Hence, in future work, it is required to explore how such individual difference can be modeled efficiently without personal training. We are also planning to study differences that can arise from different categories, *e.g.*, ages of people.

## References

[1] Borji, A. and Itti, L.: State-of-the-Art in Visual Attention Modeling, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 35, No. 1, pp. 185–207 (online), DOI: 10.1109/TPAMI.2012.89 (2013).

[2] Bruce, N. D. and Tsotsos, J. K.: Saliency, attention, and visual search: An information theoretic approach, *Journal of vision*, Vol. 9, No. 3 (2009).

[3] Cerf, M., Frady, E. P. and Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model, *Journal of vision*, Vol. 9, No. 12 (2009).

[4] documentation center, M.: manova1, `http://www.mathworks.com/help/stats/manova1.html` (2013).

[5] Face.com: Face detection API, `http://face.com` (2009).

[6] Felzenszwalb, P., McAllester, D. and Ramanan, D.: A discriminatively trained, multiscale, deformable part model, *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8 (online), DOI: 10.1109/CVPR.2008.4587597 (2008).

[7] Harel, J., Koch, C. and Perona, P.: Graph-based visual saliency, *Advances in neural information processing systems*, pp. 545–552 (2006).

[8] Hollander, M. and Wolfe, D. A.: Nonparametric statistical methods (1973).

[9] Itti, L., Koch, C. and Niebur, E.: A model of saliency-based visual attention for rapid scene analysis, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 20, No. 11, pp. 1254–1259 (online), DOI: 10.1109/34.730558 (1998).

[10] Judd, T., Ehinger, K., Durand, F. and Torralba, A.: Learning to predict where humans look, *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2106–2113 (online), DOI: 10.1109/ICCV.2009.5459462 (2009).

[11] Krzanowski, W.: Principles of multivariate analysis: A user's perspective (1988).

[12] Kubota, H., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A. and Hiraki, K.: Incorporating visual field characteristics into a saliency map, *Proceedings of the Symposium on Eye Tracking Research and Applications*, ACM, pp. 333–336 (2012).

[13] Peters, R. J., Iyer, A., Itti, L. and Koch, C.: Components of bottom-up gaze allocation in natural images, *Vision research*, Vol. 45, No. 18, pp. 2397–2416 (2005).

[14] Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M. and Chua, T.-S.: An eye fixation database for saliency detection in images, *Computer Vision–ECCV 2010*, Springer, pp. 30–43 (2010).

[15] Simoncelli, E. and Freeman, W.: The steerable pyramid: a flexible architecture for multi-scale derivative computation, *Image Processing, 1995. Proceedings., International Conference on*, Vol. 3, pp. 444–447 vol.3 (online), DOI: 10.1109/ICIP.1995.537667 (1995).

[16] Tsotsos, J. K.: Is complexity theory appropriate for analyzing biological systems?, *Behavioral and Brain Sciences*, Vol. 14, pp. 770–773 (online), DOI: 10.1017/S0140525X00072484 (1991).

[17] Viola, P. and Jones, M.: Robust Real-Time Face Detection, *International Journal of Computer Vision*, Vol. 57, No. 2, pp. 137–154 (online), DOI: 10.1023/B:VISI.0000013087.49260.fb (2004).

[18] Zhao, Q. and Koch, C.: Learning a saliency map using fixated locations in natural scenes, *Journal of vision*, Vol. 11, No. 3 (2011).

[19] Zhao, Q. and Koch, C.: Learning saliency-based visual attention: A review, *Signal Processing*, Vol. 93, No. 6, pp. 1401–1407 (2013).