修　士　論　文

What affects learning-based visual saliency models?

(学習ベース視覚顕著性モデルに対する影響要因の解析)

東京大学大学院
情報理工学系研究科
電子情報学専攻

48-126456　叶　彬彬

指導教員　　佐藤 洋一　教授

平成26年2月

Abstract

Developing visual saliency models has been a very popular research field over the past decade. Many different models have been developed and successfully applied in other studies including computer vision, robotic, human cognition understanding, etc..

In recent years, learning-based approaches using actual human gaze data have been proven to be an efficient way to acquire accurate visual saliency models and attracted much interest. However, since visual saliency models are optimized in a supervised manner in these learning-based methods, the learned model is heavily influenced by the training data.

To explore the influence brought by training data, we first developed a new eye tracking dataset that recorded the eye fixations of young adults and elderlies watching images and video clips that contain various of scenes. Then we quantitatively investigated influence of individual difference, dataset characteristics, and personal distribution to a learning-based saliency model using statistical hypothesis test methods. Finally, we made a summary and discussed the importance of choosing appropriate experiment settings and the future works.

Acknowledgements

Much credit of this work must be given to those many people who gave me support, assistance and encouragement. First of all, I would like to express my sincere gratitude to my thesis advisor, Prof. Yoichi Sato, who is not only a mentor but also a dear friend. Prof. Yoichi Sato is the most curious person I have ever seen and the saying by Confucius that "He was of an active nature and yet fond of learning, and he was not ashamed to ask and learn of his inferiors" is the portrayal of his life. I learned how to think and analyze logically from him, which would benefit me for the rest of my life. I also would like to thank my parents, Younong Ye and Qiuchang Ye, and my fiancé, Xiaoying Bian, always have been supportive throughout my years in Japan.

I also would like to express thankfulness to all members of Sato Laboratory for their cooperation and assistance in my research, especially to Dr. Yusuke Sugano and Dr. Feng Lu for their technical support and useful advices. I would like to thank the secretaries, Sakie Suzuki, Yoko Imagawa, Chio Usui for their support and kindness.

I would like to give credits to the Bank of Communications, Tokyo branch, at where I did my 2-year internship. It provides me not only the financial support, but also the chance to improve my programming skills, and to know about finance systems. I would like to express special gratefulness to my department head, Mr. Jianbin Situ, from whom I learned and benefited a lot.

I also would like to thank all my friends. Without them, my life in Japan would be so boring.

Finally, I would like to express my love and gratitude to my parents and my fiancé, however such feelings are so strong that they are beyond my words. They are always there and will be always there for me, in my darkest hour, in my deepest despair, in my trials, and my tribulations, through our doubts, and frustrations, in my violence, in my turbulence, through my fear, and my confessions, in my anguish, and my pain, through my joy, and my sorrow. I would like to say to them that

You are always in my heart.

January 2014

Contents

Chapter 1

Introduction

1.1 Background

It is astonishing to know that every second, our eyes send a data stream to the brain at the rate of $875,0000$ bits/s [1] through the optic nerve. Actually, most of the information that represents high spatial resolution view in the data stream is provided by the most sensitive part of eyes, which is the small region on the retina called fovea, which corresponds to the central visual field. Increasingly peripheral locations are sampled at increasingly coarser spatial resolution, with a complex-logarithmic fall-off with eccentricity [2]. If photoreceptors density were the same everywhere else as fovea, the optic nerve would comprise on the order of one billion nerve fibers, which is approximately 1000 times as many as fibers in human visual system [3].

It could be seen that there is a mechanism in human visual system significantly cuts the uninterested information off to reduce the workload of brain, because the brain could not process huge amount of information sent by the optic nerves and this mechanism is called visual attention.

Studying human visual attention not only answers part of the question that how human evolves but also gives suggestions on advertisement and caution sign designing and even gives inspiration to computer vision scientist to reduce the computational complex in their researches.

However, analyzing visual attention implicate the knowledge of psychology, psychobiology, biology, neurophysiology and many other fields, which makes it not a trivial task. In recent 20 year, visual saliency model has been developed to model human's visual attention.

1.2 Visual saliency models

In 1980, Feature-Integration Theory [4] was proposed by Anne Treisman and Garry Gelade:

> "Features are registered early, automatically, and in parallel across the visual field, while objects are identified separately and only at a later stage, which requires focused attention. We assume that the visual scene is initially coded along a number of separable dimensions, such as color, orientation, spatial frequency, brightness, direction of movement."

This theory explained which features are important in forming visual attention and how visual attention comes into being.

Based on Feature-Integration Theory, Koch and Ullman [5] proposed a method to combine those features that draw visual attention and introduced the concept of saliency map, which is a topographic map that marks out conspicuous locations of a scene. Figure 1.1 shows an example of saliency map and actual human fixation. Figure 1.1(a) is the input image, and Figure 1.1(b) plots the fixation locations on the input image, while Figure 1.1(c) is the saliency map of the input image, where conspicuous locations are represent in white areas, and the brightness is proportional to the salient probability.

Itti and Koch [6] proposed the world's first complete implement of computational visual saliency model by combining low level features. After that, many visual saliency models using different approaches with various assumptions of visual attention were developed and applied in a variety of fields.

The resolution of camera is increasing rapidly in recent year, on the other hand, the computational complexity in image processing increases proportional to the resolution. Visual saliency models have been proved to be a cognition-driven, straightforward solution to this issue. It could efficiently reduce the search scope within a scene, which saves processing time and resources. Thus visual saliency models are widely used in computer vision [7], which includes image segmentation [8, 9], image thumbnailing [10, 11, 12], object detection [13, 14, 15, 16, 17] and recognition [18, 19, 20, 21, 22, 23, 24], visual tracking [25, 26], image and video compression [27, 28], etc.

Because visual attention bridges the gap between low level visual information and high-level cognition process, it is also useful in robotic, such as human-robot interaction [29], visual search [30], active vision [31].

(a) Input image (b) Fixation map (c) Saliency map

Figure 1.1: Example of saliency map, column (a) is the input images [32].

1.3 Purpose and approach

Generally speaking, approaches to develop visual attention models could be divided into two types. One is computational model, in which priori assumptions about feature selection and combination are needed. 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 attention selection is not yet well understood, designing such models in a biologically convincing way would be a challenging task.

The other one is machine learning approach that learns models from actual human gaze data without priori design. This approach is extremely data-dependent, although it has been proven to be an efficient way to acquire accurate visual saliency models. Thus, an investigation on the relationship between training data and learned model is indispensable, which means the training results could be very different when trained with two dataset that recorded in different setting, while there is no standard data collection protocol in dataset development. The purpose of this research is trying to investigate what elements are significantly affecting the training result and how they make a difference in a statistical manner, though which we would discuss the importance of choosing a proper parameter setting when developing a learning-based saliency model.

1.4 Thesis outline

In this chapter, we briefly introduce the background and history of visual attention model development. We show a sample of saliency map, make a list of applications of visual saliency models and state our research purpose.

In Chapter 2, we give more details on the foundation of the saliency model and two well-known computational models.

In Chapter 3, we introduce some related works, which include some popular eye tracking dataset and 5 recent popular learning-based visual saliency models.

In Chapter 4, we do a survey on some neurophysiology and psychology re-

searches that relate to human visual attention and define the factors we are to examine in the research. Then we describe the data collection protocol of our dataset and experiment settings.

In Chapter 5, we examine the potential influence brought by individual difference, viewing tasking difference, stimuli difference, and age difference in four independent experiments and list the results of experiments.

In Chapter 6, we discuss our findings, make a summary of the experiment and give a future direction of this work.

Chapter 2

Preliminaries of Visual Saliency Models

In the chapter, preliminaries of visual saliency models would be introduced. First of all, we bring up feature integration theory, the basis of many visual saliency models, which discusses which features are important and how they combine to direct human visual attention. Then we make a summary on general steps on developing computational visual saliency models and introduce two computational visual saliency models.

2.1 Feature Integration Theory

Feature integration theory is one of the most influential psychology theory modeling human visual attention discussing which features are important in forming attention and when attention is formed.

According to this theory, the perception processing could be divided into two serial stages. The first stage is called preattentive stage. In this stage, when given a stimuli, a person will focus on distinguishing features such as color, orientation, spatial frequency, brightness, direction of movement, which happens automatically and unconsciously and the idea of object will not form in this stage.

After such information is processed in the brain, perception processing stage moves to focused attention stage, where separated features of an object are combined to obtain the information of the whole object. While the theory assumes that focal attention is required to finish the feature integration.

Treisman *et al.* also indicated that human became aware of unitary objects in two different ways. One is through focal attention and the other one is through top-down processing. It is believed that these two ways usually work together to form our attention, although experiments could be designed to demonstrate that they are almost independent of the other. As a result of that, visual saliency

models could also be classified into two corresponding approaches – bottom-up and top-down approaches.

2.2 Computational visual saliency model

Broadly, a computational visual saliency model goes through the following steps:

1. Extracting visual features
2. Generating single channel saliency map from feature maps
3. Integrating multiple saliency maps into one final saliency map

In step 1, feature channels are determined and extracted into feature maps. Low-level features that are commonly used include color, intensity, orientation, flicker and motion. On the other hand, high-level features usually contain face and text. For a certain feature, different image processing techniques are accessible to extract it, which might bring affects to the final result.

In step 2, feature maps that extracted from an image or a scene are converted to saliency maps with some biologically plausible filters such as Gabor and Difference of Gaussian filters. Apart from that, other filters like Hidden Markov Model [33], Bayesian network [34], discriminant central-surround hypothesis [35], entropy minimization [36], and they are proved to have higher performance than Gabor or Difference of Gaussian filters. This processing could be seen as the preattentive stage in feature integration theory.

In the final step, saliency maps generated in step 2 are combined into one master saliency map, corresponding to focused attention stage. There have been many priori researches [37, 38] in psychology in support of linear summation, which is simple with moderate performance. Based on the linear combination, Itti and Koch [39] provided some strategies for normalizing the feature maps according to map distribution.

2.2.1 Itti-Koch visual saliency model

Itti *et al.*'s [6] work (see Figure 2.1) is a classic cognitive model to generate saliency map, in which three feature channels (color, intensity, and orientation) are used as it is suggested in feature integration theory.

In this model, when given an input image, it is firstly subsampled into a Gaussian pyramid and each pyramid level σ is decomposed into channels for Red (R), Green (G), Blue (B), Yellow (Y), Intensity (I), and local orientation (O_θ). For

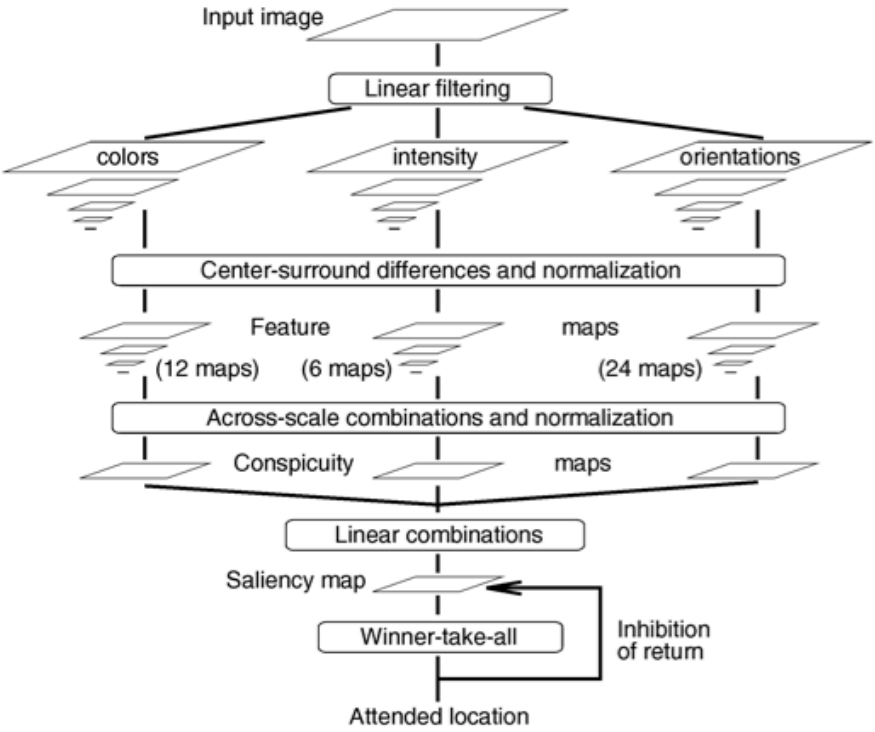


Figure 2.1: Work flow of model of Itti *et al.* [6]

each channel, center-surround "feature maps" f_l for different features l are constructed and normalized. Then in each channel, maps are summed across scale and normalized again.

$$f_l = \mathcal{N}\left(\sum_{c=2}^{4}\sum_{s=c+3}^{c+4} f_{l,c,s}\right), \forall l \in L_I \cup L_C \cup L_O$$
$$L_I = \{I\}, L_C = \{RG, BY\}, L_O = \{0°, 45°, 90°, 135°\} \tag{2.1}$$

where c and s means pyramid level for center and surrounding respectively.

Then these feature maps are linearly summed and normalized once more to compile the "conspicuity maps".

$$C_I = f_I, C_C = \mathcal{N}\left(\sum_{l\in L_C} f_l\right), C_O = \mathcal{N}\left(\sum_{l\in L_O} f_l\right) \tag{2.2}$$

Finally, conspicuity maps are linearly combined to generate the saliency map. For the result of this model please refer to Figure 2.2

$$S = \frac{1}{3}\sum_{k\in\{I,C,O\}} C_k \tag{2.3}$$

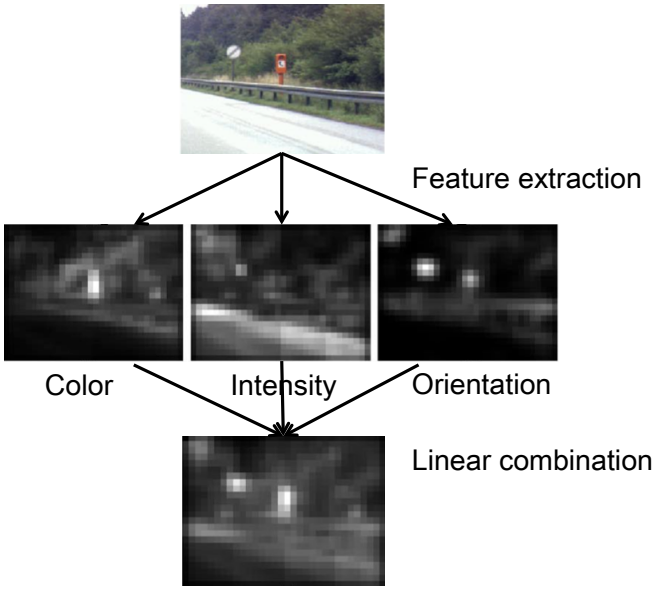


Figure 2.2: Saliency map created by the model of Itti *et al.* [6] The model used low level visual features such as luminance contrast, color contrast, orientation to predict interesting regions

This model is simple but with moderate performance, although it has the limitation that object is salient only if it represents with at least one predefined feature. While such problem may probably not avoid in feature oriented visual saliency model.

Additionally, Itti *et al.* [40] extend this model to spatio-temporal one by including the addition of flicker feature that detects temporal change (*e.g.*, onset and offset of lights), of a motion feature that detects objects moving in specific directions. Flicker is computed from the absolute difference between the luminance of the current frame and that of the previous frame and motion is computed from spatially-shifted differences between Gabor pyramids from the current and previous frames.

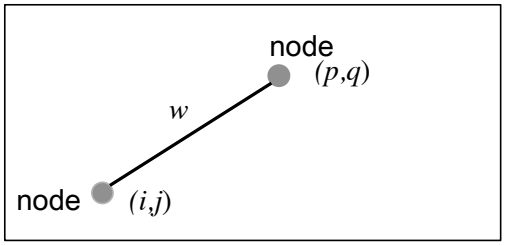
2.2.2 Graph-based visual saliency model

On 2007, Harel *et al.*propose a new bottom-up visual saliency model. It differs from Itti and Koch's basic visual saliency model in generating conspicuous map by drawing equivalence between graph and Markov chain.

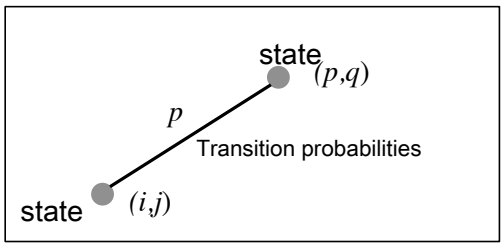
The basic idea of this model is computing activation map (corresponding to conspicuous map in Itti and Koch's model [6]) such that a location is unusual in its neighborhood will correspond to high value of activation.

At first, Harel defines the dissimilarity of given feature map between location (i,j) and (p,q) as (denoted by $M(i,j)$, $M(p,q)$)

$$d((i,j)\|(p,q)) \triangleq \left| \log \frac{M(i,j)}{M(p,q)} \right| \tag{2.4}$$



(a) Consider image as graph



(b) Consider image as Markov chain

Figure 2.3: Equivalence between graph and Markov chain

Take the pixels on image as nodes. By connecting node (i,j) with all the other nodes, a fully connected directed graph G_A can be obtained. The directed edge from node (i,j) and node (p,q) will be assigned a weight

$$w((i,j),(p,q)) \triangleq d((i,j)\|(p,q)) \cdot F(i-p, j-q) \tag{2.5}$$

where

$$F(a,b) \triangleq \exp\left(-\frac{a^2+b^2}{2\sigma^2}\right) \tag{2.6}$$

After that, a Markov chain is defined on G_A by normalizing the weights of the outbound edges of each node to 1, and drawing an equivalence between node & state, edge weight & state transition probabilities. (See Figure 2.3)

Then, the equilibrium distribution of this chain, reflecting the fraction of time a random walker would spend at each node/state if he were to walk forever, would naturally accumulate mass at nodes that have high dissimilarity with their surrounding nodes, which stands for the conspicuousness. By employing Markov chain to calculate the conspicuousness, the accuracy of this method improves greatly compared to the standard algorithm, where center-surround difference is used.

Chapter 3

Related Works

In this chapter, firstly we summarize the general steps of learning a visual saliency model and introduce five implementations of learning-based visual saliency model with detailed algorithm. Then we introduce public eye tracking dataset available on the Internet with their data collection settings.

3.1 Learning-based visual saliency model

As it is discussed in the section 2.2 that computational visual saliency model not only needs to select proper features for the model, but also requires precisely designing the shape and size of biologically plausible filters, and the normalization and integration of the feature map. While there are still a lot of unknowns waiting to explore on human visual attention system, it is too ambiguous to put all the trust on biological plausibility.

As an alternative to such conventional approaches, a data-driven learning approach of modeling visual attention has become popular. n the learning-based approach, human fixation data on natural images are used to learn a saliency model that can replicate the distribution of fixations accurately.

General implementation of learning-based visual saliency model is:

1. Collecting eye fixation data (optional).
2. Generating single channel saliency map from feature maps.
3. Labeling positive and negative learning sample.
4. Training visual saliency model with a learning algorithm.

Step 1 is collecting eye fixation data by showing stimulus to test subject under a certain data collecting protocol, however it is an optional step because there

are some eye tracking dataset released in the Internet. Respectively, it could be one of the most critical and challenging step in developing a learning-based visual saliency model because:

1. No standardized data collecting procedure: There is no guideline on the selection of stimulus, the resolution of the recording devices, the test subject attribute (age, gender, occupation, etc.) for creating fixation data set. Thus, all datasets are ad hoc and they could not always meet some specific requirement.

2. High labor intensive: Obtaining ground true for study costs time and effort. A dataset needs both stimulus and gaze data. For image stimulus, although it is easy to acquire from the giant image database such as Sun database [41], LableMe [42], ImageNet [43], recording the data when people looking is another story. Eye tracking devices must calibrate for every test subject and people get tired with one hour if they keep focusing on the stimuli, which limits the scale of fixation data set.

Learning-based visual saliency models become different from computational models after feature maps are created. Instead of applying linear summation or other preset integration, the weight of each feature is obtained through supervised learning with labeled samples. The followings are some examples of learning-based visual saliency models.

3.1.1 Nonparametric learning approach

As an early implementation, Kienzle *et al.* [44] developed a nonparametric visual saliency model. The feature of the model is that no feature maps are involved.

In this model, positive and negative samples are select as square patch cut out from the fixated location and background of images, accordingly. Pixel values in the patch is are stored in a feature vector $\boldsymbol{x}_i$ together with a label $y_i \in \{1; -1\}$. However it is not straightforward to choose an appropriate patch size and resolution. By making compromise between computational tractability and generality, the resolution of the patch is set to 13×13 and its size varies between $d = 0.47°$ and $d = 27°$ visual angle. Therefore, the feature vector $\boldsymbol{x}_i$ has 169 dimensions.

The saliency map is formulated as $f(x) : \mathbb{R}^{169} \to \mathbb{R}$ using a support vector machine (SVM):

$$f(x) = \Sigma_{i=1}^{m} \alpha_i y_i \exp(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|}{2\sigma^2}) \tag{3.1}$$

A experiment results suggest it has comparable performance with the conventional model of Itti and Koch[6] even with no cognitive feature, but it requires large amount of training samples due to the high dimensional feature vector.

3.1.2 Classic classification approaches

1. Judd *et al.*

Judd *et al.* [45] presented a model that applies a linear support vector machine to train a saliency model with predefined low, mid and high-level image features.

There is a wide range of low level features applied in this model, which include local energy of the steerable pyramid filters, contract of color, intensity and orientation, channel of red, blue and green, probability of each color, and features used in saliency models described by Oliva *et al.* [46] and Rosenholtz [47]. For mid-level feature, Judd used a horizon line detector form mid-level gist features. Viola Jones face detector [16] and the Felzenszwalb person detector are served as high-level features.

In the training phase, training sample are chosen from the top 20% salient locations and bottom 70% salient locations as positive sample and negative sample in the ratio of $1:1$ and they are trained in liblinear support vector machine.

2. Constrained linear regression approach (Zhao and Koch)

In contrary of using many features, Zhao and Koch [48] proposed a model using constrained linear regression with linear integration assumption as prior. The model has a simple structure and it is extremely easy to implement.

Zhao chose a bottom-up driven saliency model developed by Cerf [49] as the base features, which is a development based on Itti and Koch's work [6] by adding a face feature. Thus, the feature maps of Cerf *et al.*'s model are constructed over two color channels, one intensity channel, four orientation channels and a face channel.

Feature weight vectors are generated by overlaying image locations to the feature maps. That is, for example, for an image location $\boldsymbol{x}$, its feature vector is $\boldsymbol{v}(\boldsymbol{x}) = [C(\boldsymbol{x}), I(\boldsymbol{x}), O(\boldsymbol{x}), F(\boldsymbol{x})]^T$, where C, I, O, F are feature map of color, intensity, orientation and face of the image, respectively.

Training samples, on the other hand, are determined by convolving ground truth with an isotropic Gaussian Kernel.

Finally, the optimal weights of each feature are learned through linear least square regression with constraints, with $\boldsymbol{V}$ as the stack vector of feature vectors $\boldsymbol{v}(\boldsymbol{x}_i)$, $\boldsymbol{w} = [w_c, w_i, w_o, w_f]$ as the optimal weight of each feature and $\boldsymbol{M}_{fix}$as vectorized fixation map.

$$\arg\min_{\boldsymbol{w}} \|\boldsymbol{V} \times \boldsymbol{w} - \boldsymbol{M}_{fix}\|^2, \tag{3.2}$$

subject to,

$$\boldsymbol{w} \geq 0 \tag{3.3}$$

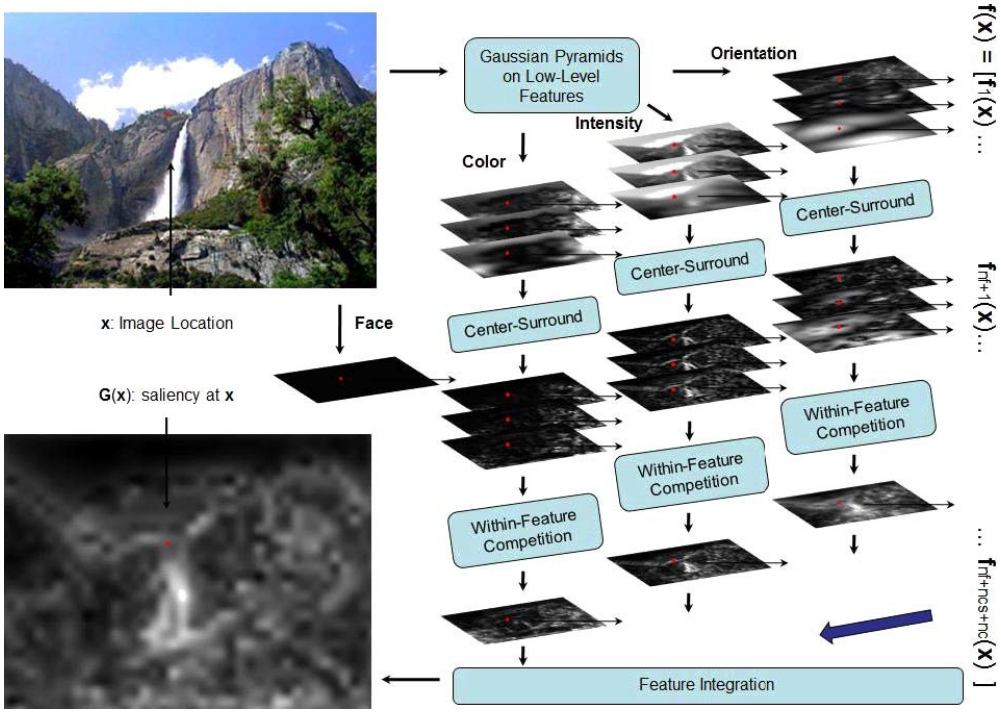


Figure 3.1: Feature map construction of Zhao *et al.*'s model

By solving the equation, optimal feature weight could be obtain. In this paper, Zhao *et al.*also tested the performance of the model using four public dataset.

AdaBoost approach (Zhao and Koch) To address the problem that features need to be selected at the first place. Zhao and Koch [50] proposed an AdaBoost based learning model, which is able to accomplish feature selection, thresholding , weight assignment and nonlinear integration automatically.

The choice of feature map is the same as the model Zhao proposed in 2011 [48]. It includes 4 feature channels: color, intensity, orientation and face. Upon the former three features, they are extracted in Gaussian Pyramids in different scales, and center-surround difference maps are constructed across pyramid scales to capture the difference. Face feature is extracted by Viola and Jones face detector [16]. (Figure 3.1)

Then many weak classifier are trained with features and fixations through iteration and combined to be a strong classifier (Figure 3.2(a)). For a new image, the feature vector is calculated and put into the strong classifier to obtain the saliency map(Figure 3.2(b)). Since all weak classifiers are trained for a single feature, AdaBoost selects features from the feature pool so that the best separation of positive sample and negative sample can be achieve. In this manner, AdaBoost can select the proper features and determined the threshold and optimal weight automatically.

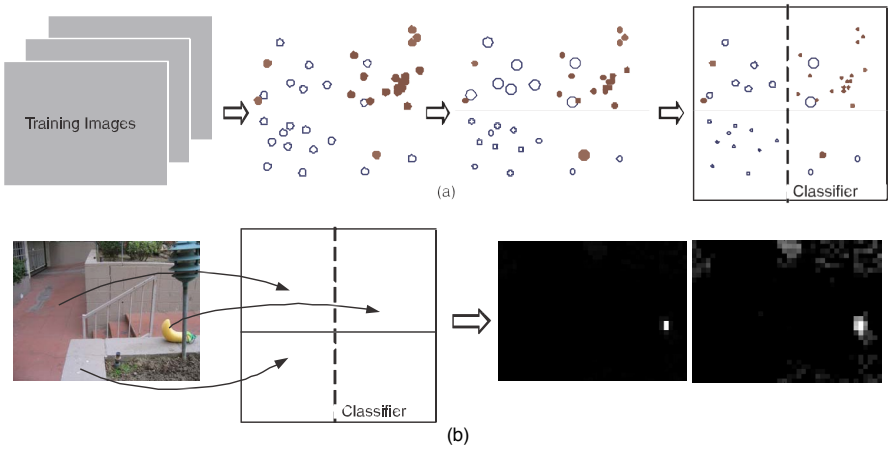


Figure 3.2: Training and testing process of Zhao *et al.*'s model

3.1.3 Biologically plausible learning approach

Unlike other learning-based models, Kubota *et al.*'s model [51] takes into account non-uniformity of sensitivity within a field of view by using different weights depending on the distance from a current fixation position. By learning the weights in a supervised manner, the model can predict the next fixation position more accurately. Hence, it reflects eye movement statistics more strongly than conventional models.

Figure 3.3 illustrates the basic concept of Kubota *et al.*'s model. Given input images, feature maps are generated first. Low-level features, including color, intensity and orientation are extracted using the graph-based visual saliency model [33]. A face feature is given as a Gaussian distribution around the center of a face detected by a face detector.

The visual field is then divided into c concentric regions around the current fixation position. The size of concentric regions are adjusted in such a way that each region contains the same number of fixation points. Nine low level features are used in the model: color, intensity, orientation in 3 scale levels, *i.e.*, 1/4, 1/8, 1/16 of the original image size, as well as a high level feature of face.

The output saliency map is computed by integrating these 60 (= 10 features $\times$6 regions) maps with individual weights, and these feature weights are learned using ground-truth saccade data from the training dataset. For each saccade s, a feature matrix (stacked feature maps) $\boldsymbol{F}_s$ and a target map $\boldsymbol{t}_s$ with a sharp peak around the next fixation location are obtained. A 60-dimensional weight vector $\boldsymbol{w}$ is then calculated by solving a nonnegative least squares problem:

$$\min_{w} \sum_{s} \|\boldsymbol{F}_s \boldsymbol{w} - \boldsymbol{t}_s\|^2, \text{ where } \forall w \in \boldsymbol{w}, w > 0. \tag{3.4}$$

In their experiment, they found that performance is positively correlated to the how many regions were divided and They showed that performance improvement

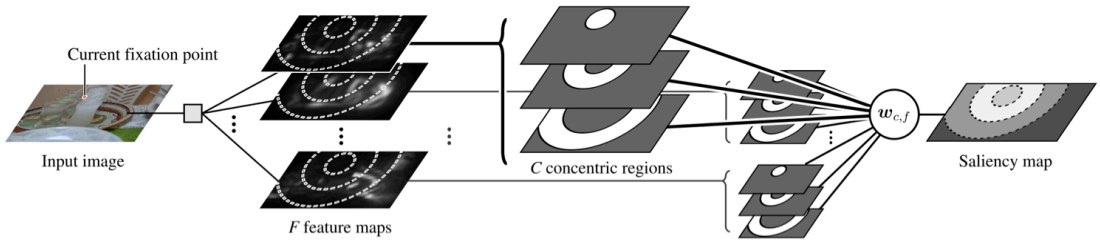


Figure 3.3: Flowchart of Kubota *et al.*'s model

became negligible after $c = 6$ (Figure A.1) but the computational complexity grew greatly.

3.2 Public eye tracking dataset

Since dataset is essential to a saliency model and decisive to the training result. We would go through some major datasets and show that they have very different settings in this section.

An fixation dataset usually includes visual stimulus (typically images) showing to subjects in the order of hundreds and eye fixation data recorded by eye tracking devices. Dataset parameters could be the amount of stimuli, number of test subject, viewing task and viewing angle.

Bruce dataset [52] contains 120 images viewed by 20 subjects. Images in this dataset consist of a variety of indoor and outdoor scenes, some with very salient items, others with no particular regions of interest. Images were presented in random order for 4 seconds each with a gray mask between each pair of images appearing for 2 seconds. Subjects were 0.75 meters away from the display, and no specific introduction was given when viewing the images.

The NUSEF dataset [53] includes gaze data from 75 subjects viewing 758 images containing faces, nudes, objects, human activities, everyday indoor and outdoor scenes, and some unpleasant concepts. Rather than showing the full image to all participants, a subset was presented and average number of viewer per image is 25.3. Test subjects were instructed to view freely 30 inch (76.2 cm) away from the monitor.

The MIT dataset [45] contains 1003 natural images from Flickr and LabelMe viewed by 15 test subjects and which outnumbers other datasets in terms of the number of images. Although subjects were instructed to view freely, a memory test was used to keep them from being absentminded. It is worth noticing that two of the viewers were researchers on the project.

Cerf *et al.* [49] used images containing faces, text, and complex man-made objects to create a dataset through 4 experiments with tasks of free viewing and search. However, the number of images and the number of subjects are not consistent in different experiments. The distance between the screen and the subject

Attribute / Dataset	Stimuli number	Subject Number	Stimuli length	Gaze reset
Bruce *et al.*	120	20	4s	no
NUSEF	400	75	5s	yes
MIT	1003	15	3s	no
FIFA	180	8	2s	yes
LeMeur *et al.*	27	40	15s	no
DOVES	101	29	5s	yes
Kubota *et al.*	400	15	4s	yes

Table 3.1: Comparison of public fixation datasets. (Part 1)

Attribute / Dataset	Viewing task	Viewing angle	Remark
Bruce *et al.*	Free view	$32° \times 24°$	\
NUSEF	Free view	$26° \times 19°$	Each image viewed by 25 subjects (average)
MIT	Memory test	$36° \times 27°$	\
FIFA	Multiple	$28° \times 21°$	Free viewing and search task
LeMeur *et al.*	Free view	Unknown	\
DOVES	Memory test	$17° \times 13°$	Grayscale image
Kubota *et al.*	Rating	$57° \times 34°$	\

Table 3.2: Comparison of public fixation datasets. (Part 2)

was 80 cm

Although the dataset created by LeMeur *et al.* [11] had 40 test subjects freely view images with strongly salient objects, only 46 images degraded from 10 nature images using different techniques were used.

DOVES dataset [54] included 101 images manually selected from a calibrated greyscale natural image database [55]. Eye movement data from 29 subjects were collected with a task of simple memory test.

Kubota *et al.* [51] created a dataset with wide view angle ($57° \times 34°$) in order to incorporate the visual field characteristics into visual saliency model. 400 nature images from Flickr were shown to 15 subjects in random order. To ensure that subjects were concentrate on the stimuli, they were asked to rate the viewed image before moving to the next one.

To make it more clear, Table 3.1 and Table 3.2 compare the settings of the datasets mentioned above.

Chapter 4

Problem Settings

4.1 Influential factor in visual saliency model

We have mentioned about our research purpose in section 1.3 that we intend to discover the relationship between the training data and the training result of learning-based visual saliency model. Therefore, it is necessary to consider what factors have influence on the training result with high probability.

1. Individual difference

Training datasets are 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.

2. Stimuli type

Apart from between-person difference, prior researches showed that the attention was stimuli-dependent and task-dependent. In Parkhurst *et al.*'s work [56], the ability to guide attention of three modeled stimulus features (color, intensity and orientation) was examined and found to vary with image type. In Parkhurst *et al.*'s experiment, four paid participants were instructed to view four kinds of images freely, which included fractals, nature landscapes, buildings and city scenes and home interiors (Figure 4.1). Later, saliency map of these images were generated through a computational model [6] and the average salience of the images

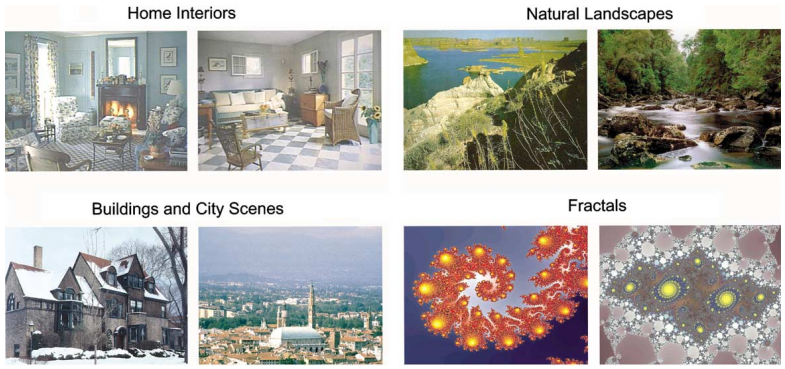
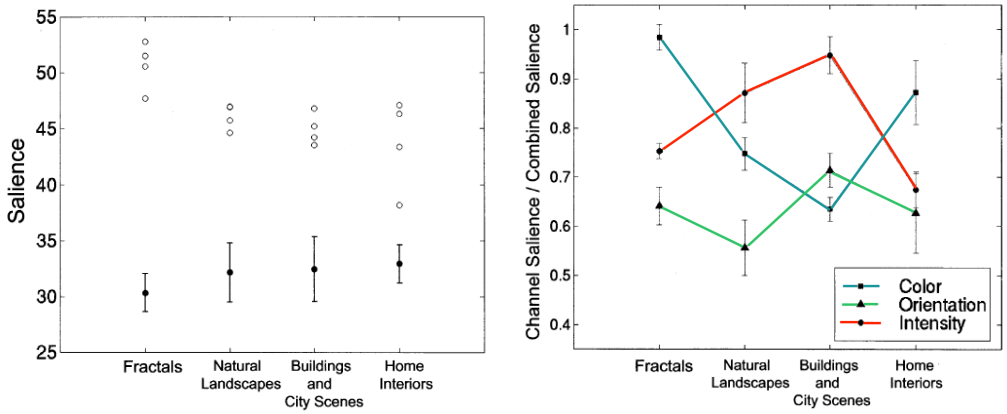


Figure 4.1: Examples of image dataset in Parkhurst *et al.*'s experiment



(a) The mean salience at the first fixation location is shown as an open circle for each participant within each database. The mean salience expected by chance for each database is shown as a closed circle with error-bars indicating plus/minus one standard error of the mean.

(b) The mean ratio of chance-adjusted salience for a single feature channel relative to the chance-adjusted salience for all channels

Figure 4.2: Results Parkhurst *et al.*'s experiment

were calculate by extracting the salience value at the fixations. By comparing the mean salience at the first fixation location (Figure 4.2(a)) and the contribution of each feature (Figure 4.2(b)), it could be seen that guide ability of different feature is stimuli-dependent.

3. Viewing task type

Human visual attention could be varied when different visual task is undertaken. Laar *et al.* [57] presented a neural networks can learn to focus its attention on important features depending on several visual search tasks, indicating even within a same type of visual task, minor changes (target selecting in the case of visual search) can cause a different attention flow change.

Also Peters *et al.* [58] incorporated task-dependent influences into a computational visual attention model. They divided the visual tasks into three type: passive viewing, active viewing and interactive viewing. Passive viewing is typical free viewing task; active viewing usually includes visual search, scene comprehension, reading and rating. Interactive viewing indicates a task in which the par-

ticipant's reaction serves as feedback to the revealing stimuli, *e.g.*, playing video games, driving and web browsing. Stimuli used in this research were contemporary three-dimensional video games, and the test subject's task was playing the video game. The result showed that it was efficient to improve the performance of the visual saliency model by considering the influence of visual task.

4. Visual ability drop due to aging

Human body function abilities decline as age grows. For human eyes, the useful field of view becomes narrow and dynamic visual acuity drops [59, 60, 61, 62, 63, 64, 65].

Ball *et al.* [59] designed an experiment to test the useful field of view of three subject group. Average age of young, middle-age and older group were 25, 45 and 69. The schematic diagram of the test stimulus is shown in Figure 4.3, a frowning face is shown in the fixation box and a peripheral target 10° from fixation is presented among 47 box distractors. Subjects viewed 28.5 cm away from a large monitor thus the visual angle was $60° \times 60°$. Subjects needed to answer three types of questions corresponding to three different task. Based on the answers to the questions that whether the peripheral target is present or not, whether the peripheral target is a smile or a frown face, whether the peripheral target is the same as the one in the fixation box, the error rate of each group was calculated.

Figure 4.4 shows two results from the experiment. Figure 4.4(a) presents effect of center task difficulty on peripheral localization performance as a function of age and eccentricity. The increase in error rates with increasing eccentricity for all age and center task combinations indicates a restriction of the useful field of vision. It is noticeable the difference between young and middle-age groups is marginally significant (the probability of p-value < 0.05 in Turkeys test is approximately 50%), while the difference between middle-age and older group is significant (Tested with Turkeys test at significant level 0.05).

The effect of distractor condition on peripheral localization as a function of age and eccentricity is shown in Figure 4.4(b). Young group and middle-age group differ only under the condition that distractor is presented. The older observers make significantly more errors than do the middle-aged and young observers in all distractor and eccentricity combinations (Tested with Turkeys test at significant level 0.05).

In addition to the size of useful field of vision, the dynamic visual acuity becomes worse when people grow old. Bennett *et al.* [62] investigated the effects of aging on motion detection and direction identification.

Stimuli used in their experiment were random dot cinematograms showing 300 black dots over a white background; black and white corresponded to luminance of approximately 5 and 95 cd/m^2, respectively. Every dot's direction was chosen randomly from a uniform distribution of directions. There were two kinds of cinematograms, designated Signal and Noise, which randomly appeared to the test subjects in the experiment.

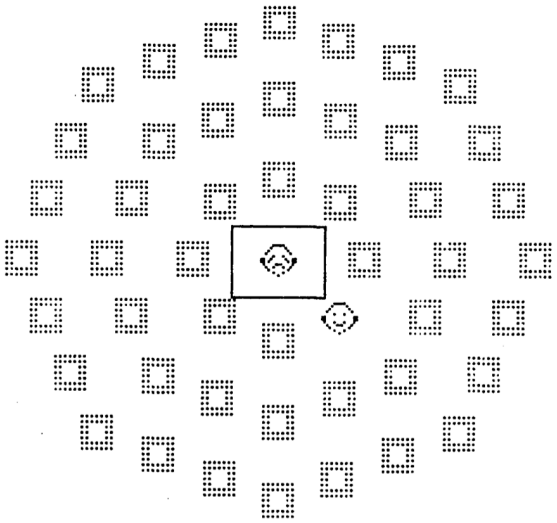
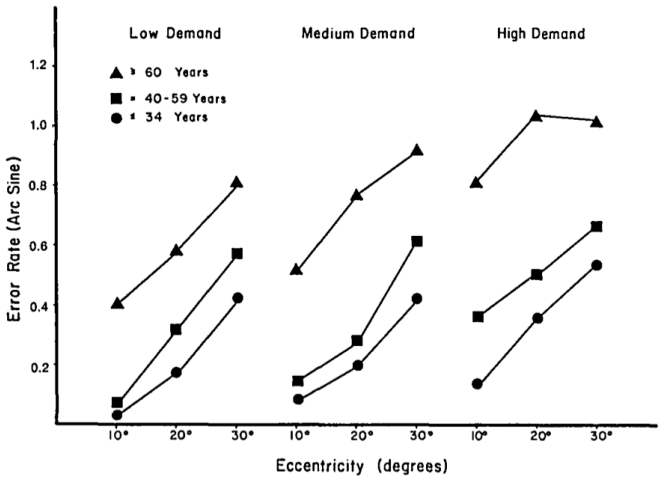
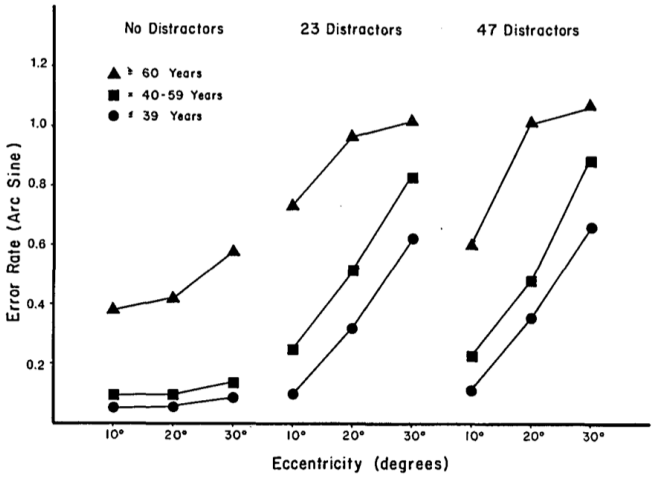


Figure 4.3: Schematic represent at the UFOV (Useful field of vision) task.

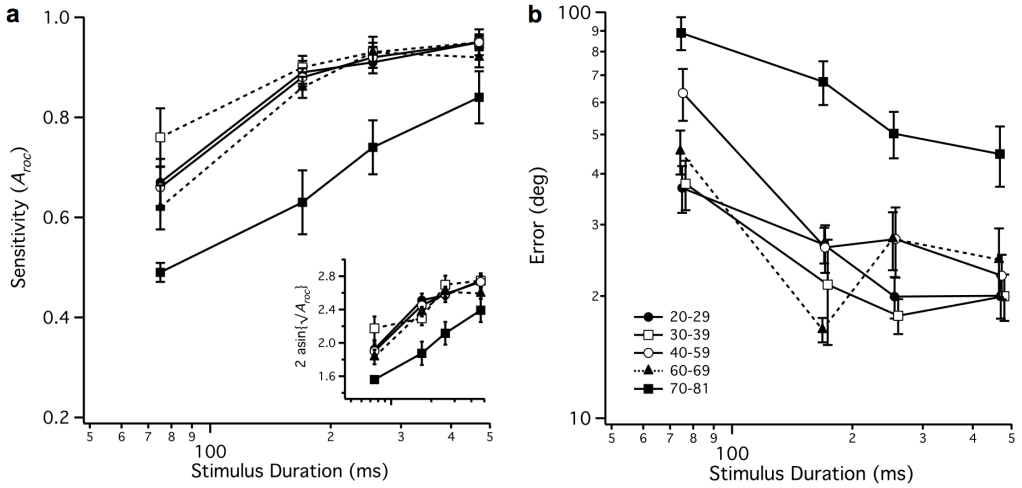


(a) Radial localization error rates for different center task demands as a function of age and eccentricity.



(b) Radial localization error rates for three different distractor conditions as functions of age and eccentricity.

Figure 4.4: Results of Ball *et al.*'s experiment



(a) Sensitivity to motion

(b) Moving direction judgment error rate

Figure 4.5: Results of Bennett *et al.*'s experiment

In Signal cinematograms, the directions of dots moved between frames were drawn from a uniform distribution spanning 250° around some mean direction. The signal cinematograms appears to young observers as a global flow in a direction corresponding to the mean of the underlying distribution. In Noise cinematograms, the directions in which dots moved between frames were drawn from a uniform distribution spanning 360° , producing a percept of incoherent, random motions.

39 healthy subjects aged from 23 to 81 were paid to participate in their experiment. Subjects had two task in the experiment: pointing out the direction in which in which the cinematogram's elements appeared to flow, and giving a confidence rating regarding whether the cinematogram is a signal one or a noise one.

From Figure 4.5, it can be read that the oldest group of observer had less sensitivity on motion and the error rate of their direction judgment was higher than younger observers.

4.2 Dataset

Through the discussion in the previous section, we narrowed the scope of influential factors of visual attention and decided to investigate the following influence on a visual saliency model.

1. Individual difference of the test subjects
2. Stimuli type
3. Viewing task

4. drops in visual ability cause by aging

It is obvious that none of the available public eye tracking datasets we have discussed in section 3.2 satisfy all our research purposes simultaneously. As a result of that, we decided to create a qualified dataset that includes:

1. Dynamic and static stimuli,
2. Stimuli include multiple contextual levels,
3. Stimuli viewed under different viewing task,
4. Young and old observers.

For the ease of understanding, we describe image and video dataset design scheme in separate manner.

1. Image dataset design scheme

We built a dataset with images of three levels of contextual complexity. The *low level* group contains abstract images without specific meaning, and we synthesized color fractal images to construct this group. The *mid level* and *high level* groups both consist of natural images, however the mid level group only contains indoor and outdoor still-life images without highly contextual meaning. The high level group contains a wider range of natural images such as sports scenes, human activities and paintings. Fixation data for each of the three groups were collected from the same set of participants under two different instructions. While no instruction was given under the *free viewing* condition, participants were asked to rate their preference of the images under the *preference rating* condition. In this way, we can quantitatively analyze differences between 6 training sets (3 contextual levels $\times 2$ tasks) using the same set of test subjects. To the best of our knowledge, ours is the first and largest dataset that contains fully controlled variations of both stimulus types and viewing tasks. The image dataset will be made available on-line for further studies.

2. Video dataset design scheme

Our video dataset only include mid and high level contents and such video clips can be classified as nature scenes. Mid level videos contains nature and city views with moving objects like cars, cloud, water, animals etc., still objects with camera motion, and so on. High level videos are those with specific human activities, such as sports, playing games, writing shot from first-person and third-person perspective. No specific instruction was given to the participants during their viewing.

Figure 4.6: Some examples of our image dataset, 3 rows from top to bottom are examples of low, mid and high context level images

4.3 Data Collection Protocol

4.3.1 Image Dataset

In this section, we describe details of the data collection protocol for image dataset. We first prepared 870 full HD images in total, *i.e.*, 290 images for each of the contextual levels. The number of images was decided by taking the balance between time cost of experiment and convergence of saliency model weight vector, as our preliminary experiment showed that the data collecting speed was approximately 280 images per hour and the weight vectors were converged when trained with over 200 images. Images in the *low level* group images are generated using a fractal image generator. Public domain images collected from pixabay [1], Google image search [2] and SUN database [41] are used to construct the mid level and high level groups. Fig. 4.6 shows some examples of images in each group.

Eye movements were recorded using Tobii TX300 eye tracker at a sampling rate of 120Hz upon 9-point gaze calibration. Eye movements faster than 18° per second are considered a saccadic movement. Test subjects were 12 novice volunteers who were not told the purpose of the experiment and they are aged from 22 - 30 years old ($\mu = 24.33$, $\sigma = 2.19$). Images were presented on a 23-inch LCD display with a resolution of 1920×1080 at the distance of approximately 60 cm. Each image was presented for 4 seconds and followed by a black mask image with a white cross at the center for 1 second. During experiments, a chin rest was used to fix the subjects' head position.

Eight hundred and seventy images were divided into *free viewing* and *preference rating* sets so that each contextual level group contains 145 images for each task. First, for the *free viewing* task set, 435 images from 3 groups were mixed and displayed in a randomized order, and test subjects were instructed to view the images freely. The recording session was divided into three parts, *i.e.*, test subjects

[1] http://pixabay.com/
[2] http://images.google.com/

Figure 4.7: Screenshots of our video dataset, containing nature landscapes, city scenes, plants, animals, human activity, etc..

took a rest after viewing 145 images. Then for the *preference rating* task set, test subjects were instructed to give seven-grade ratings of their preferences by pressing a keypad after each image is displayed. Unlike the *free viewing* task, each of the 3 groups was displayed separately, and the display order was randomized within each category.

Additionally, in order to compare the training results of young and aged for static scene, we paid for 12 participants to view the mid and high level images and give preference ratings under the exactly same settings with others.

4.3.2 Video Dataset

The amount of video clips is 434. They are uncut footage videos collected from videezy [3] and xstockvideo [4]. Videos from videezy are completely free to use and those from xstockvideo under regular license are free to use for educational purpose. Figure 4.7 shows some examples of the video clips. Originally clips are at least 5-second long with least resolution of 960×540 and above 24 frames per second. We extracted the first 5 second contents from the original video, downsampled them to have the same frame rate of 24 and scaled them to full HD resolution (1920×1080). All clips were re-encoded in XviD and packaged in AVI format.

Eye movements were recorded using Tobii TX 300 eye tracker at a sampling rate of 120Hz upon 5-point gaze calibration. The threshold judging qualified eye movement is 18° per second. Eye movements faster than the threshold are considered a saccadic movement. 2 groups of test subjects were involved in our data collection. One group was consist of 12 novice volunteers aged from 23 - 43 years old ($\mu = 28.6$, $\sigma = 6.27$), as young group. We hired 12 novice test subjects aged from 65 - 77 years old ($\mu = 68.3$, $\sigma = 3.75$) as elderly group. All elderlies in our experiment were in good health status. The uncorrected visual acuity of the elderlies were above 0.5.

[3] http://www.videezy.com/
[4] http://www.xstockvideo.com/

Videos were presented on a 23-inch LCD display with a resolution of 1920×1080 at the distance of approximately 60 cm in a random order. After each video was presented, a black mask image with a white cross at the center was shown and test subjects had to press any key on a keyboard to proceed to next video. During experiments, a chin rest was used to fix the subjects' head position.

4.4 Experiment Settings and Methodologies

We can now discuss the experiment settings and methodologies based on the survey of potentially influential factors to the learning-based visual saliency model and the newly developed dataset.

We designed three experiments and testings and examined the data with Kubota *et al.*'s visual saliency model [51] in our research. The parameters to generate saliency model (feature type, region division, stimuli number, test subject) were different in each experiment.

The first one is to examine the existence of individual difference, on which has no preceding studies, while it is worthy verifying because it is crucial for researcher to choose test subjects. As when the difference exceeds a certain level, the increase in training data will possibly cause the performance decrease of the trained model.

We examined the existence of individual difference by doing a performance comparison between personal and generic models which were calculated using data collected by Kubota *et al.* [51]. A personal model is a model generated with this person's fixation data, while the corresponding generic model to the personal model, on the contrary, is a model learned from all test subjects' except but this person's one. Wilcoxon Signed-rank test and one-way multivariate analysis of variance over the feature vectors (MANOVA) were performed within this experiment.

In the second experiment, we compared saliency models generated from the image dataset we created with controlled variables to analyze how stimuli type and viewing task type affected learning-based saliency model. Specifically, we compared models trained with the low, mid and high context image groups in the same viewing task and models trained with the data from free viewing and preference rating tasks in a same contextual level. The consistency of these models were examined by Paired Hotelling's T-square test.

Lastly, we inspected the effects of aging on learning-based visual saliency model. Scientific research on neurophysiology and psychology leads to a conclusion that performance degradation of visual ability of human being comes along with senescence. We compared the difference between static image and dynamic models for young and aged group by Paired Hotelling's T-square test.

Chapter 5

Experiments

5.1 Influence by Individual Difference

Models trained in this section were trained with data collected by Kubota *et al.* [51]. Region division number was set to 6 to model the peripheral view because of the wide viewing angle ($57° \times 34°$).

5.1.1 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

dataset. Examples of the learned feature weights under the three conditions are shown in Fig. A.2.

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 compared the Normalized Scanpath Saliency (NSS) [66] score over models in condition (1) and (2) to examine whether the performance 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 denoted the model trained under condition (i) for subject n by ${}^{cond.i}M_n$, where $n = 1, 2, \cdots, 15$, We performed a two tailed Wilcoxon Signed-rank test [67] over the paired-data of ${}^{cond.1}\mathrm{NSS}_n$ and ${}^{cond.1}\mathrm{NSS}_n$, where $n = 1, 2, \cdots, 15$. 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 does 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.

5.1.2 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 tried to examine if personal models perform better than a generic model.

In this experiment, we divided 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 $n = 1, \cdots, 15$,

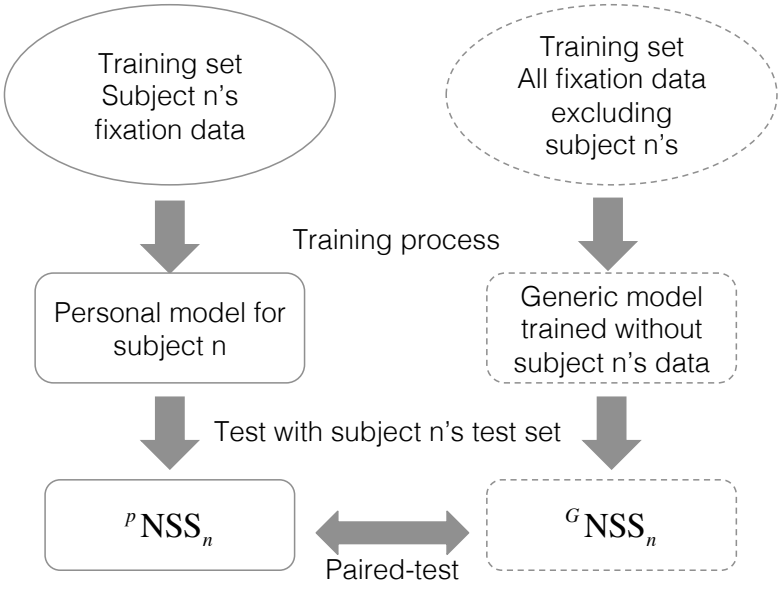


Figure 5.1: Data construction of personal and generic model

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 [67] over the paired-data of ${}^{P}\mathrm{NSS}_n$ and ${}^{G}\mathrm{NSS}_n$, where $n = 1, \cdots, 15$ (Fig. 5.1).

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 $p \ll 0.01$ significance level.

5.1.3 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) [68] to test the null hypothesis that the means of each personal model are the same n-dimensional multivariate vector at 5% significance level.

Table 5.1: 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

In this experiment, all images in the dataset were taken as training samples, in other words, there was no division of training set and test set. To obtain the feature vectors, we divided the dataset into 10 subsets as training sets, which made every subset have 40 images. Then, for every test subject, 10 personal models would 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

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

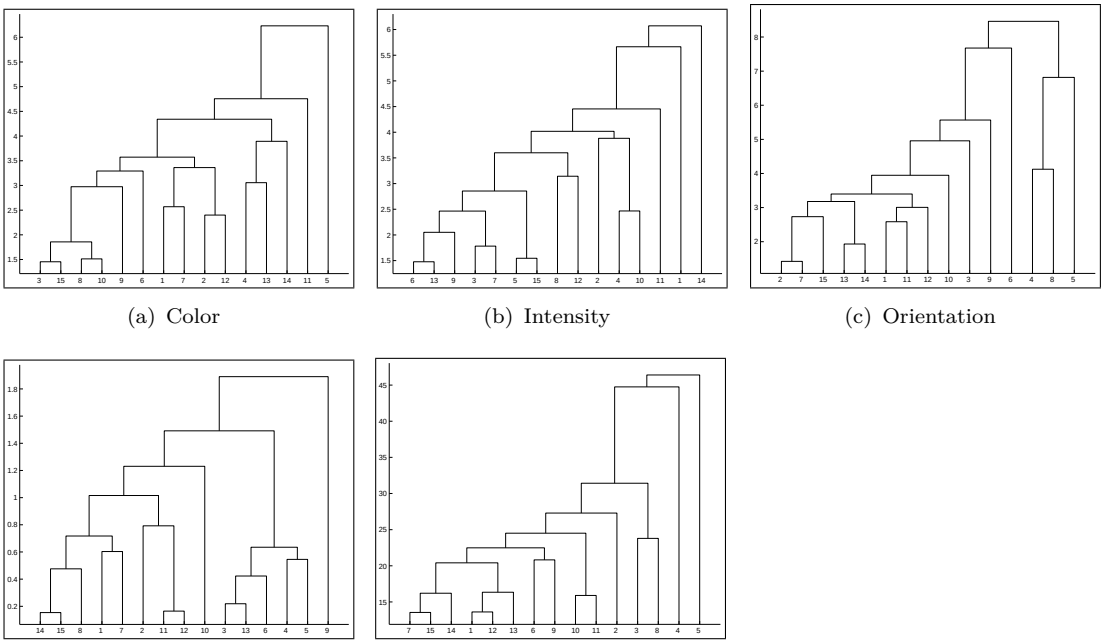
The test results are listed in Table. 5.1. The first column stands for the type of feature vector that is put into the test. Second and third columns show how many different groups and how many observations per group are involved in the MANOVA test. The fourth column is the dimension of feature vector. 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.

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 5.1, *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 applied a hierarchical clustering based on the result of MANOVA and visualize the clustering results as dendrogram in Fig. 5.2. We could see that color (Fig. 5.2(a)), intensity (Fig. 5.2(b)) and orientation (Fig. 5.2(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 face (Fig. 5.2(d)) while the d value of face is 0.

When all features are considered, the topology of grouping becomes different



(a) Color (b) Intensity (c) Orientation

(d) Face (e) All features

Figure 5.2: Hierarchical clustering using UPGMA (Unweighted Pair Group Method with Arithmetic Mean) based on different features

from any of the above (Fig. 5.2(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.

5.2 Influence by Viewing Task and Stimulus Types

In this section, the models were trained with the newly collected image dataset with a relatively narrow viewing angle ($45° \times 29°$)

Figure 5.3 shows the mean weight vectors trained with different stimulus types and viewing tasks. The legend in these figures indicates the name and scale level of the feature, in which C = Color, I = Intensity, O = Orientation and F = Face. For instance, C1 stands for color feature with scale level 1. The higher the scale level is, the less detail is considered. R1 to R3 in the horizontal axis means region 1 to region 3, respectively, representing from fovea view to peripheral view.

Based on these training results, we statistically examined the influence of stimulus and viewing task types on a learning-based visual saliency model. More specifically, we compared the weight vectors under the following two conditions:

1. Comparison between models trained with the low, mid and high context image groups in the same viewing task.
2. Comparison between models trained with the data from free viewing and preference rating tasks in a same contextual level.

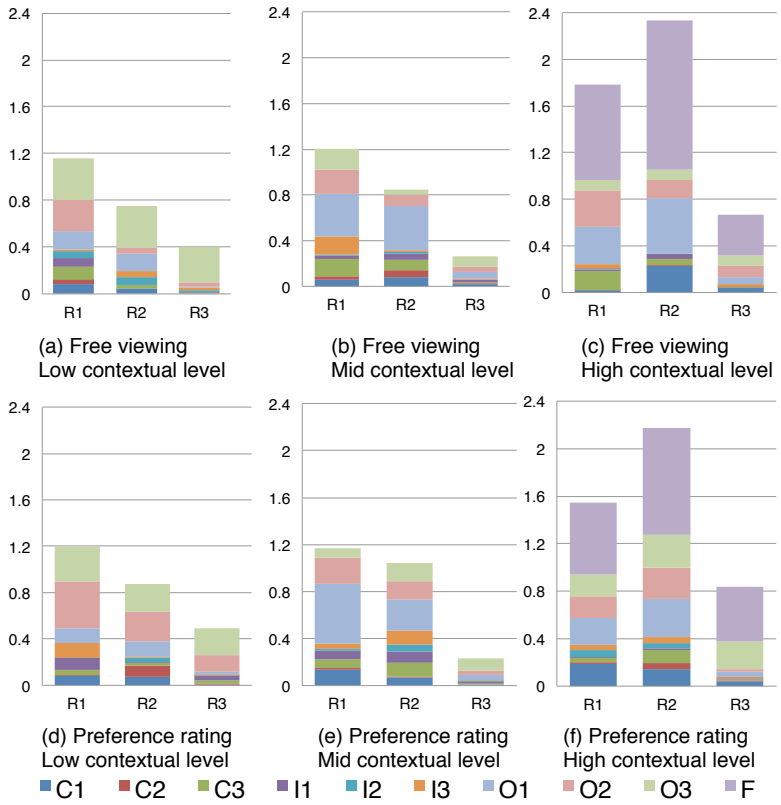


Figure 5.3: Mean weight vectors trained with different stimulus types and viewing tasks.

Given two sets of weight vectors from comparison pairs $W_1 = \{\boldsymbol{w}_{11}, \cdots, \boldsymbol{w}_{1n}\}$ and $W_2 = \{\boldsymbol{w}_{21}, \cdots, \boldsymbol{w}_{2n}\}$ ($n = 12$ in our case), we compared the weight vectors through the paired Hotelling's T^2 test [69]. Since the full dimension of the weight vector $d = 30$ is much higher than the number of observations $n = 12$, we compared each feature type independently. Accordingly, the dimension of the weight vector reduces as $d = 9$ for the color, intensity and orientation features, and $d = 3$ for the face feature.

Let $\mathcal{N}_p = (\boldsymbol{\mu}, \Sigma)$ denote p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Thus we have n observations in X_1: $\boldsymbol{x}_{11}, \cdots, \boldsymbol{x}_{1n} \sim \mathcal{N}_p = (\boldsymbol{\mu}_1, \Sigma_1)$ and X_2: $\boldsymbol{x}_{21}, \cdots, \boldsymbol{x}_{2n} \sim \mathcal{N}_p = (\boldsymbol{\mu}_2, \Sigma_2)$, where $n = 12$. Null hypothesis H_0 is

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \ . \tag{5.1}$$

Then we denote

$$\boldsymbol{y}_i = \boldsymbol{x}_{1i} - \boldsymbol{x}_{2i}, \tag{5.2}$$

So that $\boldsymbol{y}_i \sim \mathcal{N}_p = (\boldsymbol{\mu}_y, \Sigma_y)$ where $\boldsymbol{\mu}_y = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Null hypothesis H_0 becomes

$$\boldsymbol{\mu}_y = 0. \tag{5.3}$$

Sample mean vector for $\boldsymbol{y}_i$ is

	Low vs. Mid	Low vs. High	Mid vs. High
Color	> 0.50	0.09	0.43
Intensity	0.41	**0.02**	> 0.50
Orientation	0.30	**0.04**	> 0.50

Table 5.2: Differences brought by stimulus difference in free viewing.

	Low vs. Mid	Low vs. High	Mid vs. High
Color	0.11	> 0.50	> 0.50
Intensity	0.45	**0.02**	> 0.50
Orientation	0.15	0.28	**0.01**

Table 5.3: Differences brought by stimulus difference in preference rating.

$$\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \qquad , \tag{5.4}$$

So that we can calculate sample covariance matrix

$$S_y = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})' \qquad , \tag{5.5}$$

and Paired Hotelling's T^2 test statistic is given by

$$T^2 = n\bar{\boldsymbol{y}}' {S_y}^{-1} \bar{\boldsymbol{y}} \qquad , \tag{5.6}$$

With Paired Hotelling's T-square test statistic T^2, we have

$$F = \frac{n-p}{p(n-1)} T^2 \sim F_{p,n-p} \ , \tag{5.7}$$

where $F_{p,n-p}$ is the F-distribution with parameters p and $n-p$. We can reject H_0 at level α if $F > F_{p,n-p,\alpha}$, where $F_{p,n-p,\alpha}$ is the F-value with p and $n-p$ degrees of freedom, evaluated at level α. In our case, the significance level is set to be 0.05.

As with most saliency models, output saliency maps are always normalized to the fixed range in Kubota *et al.*'s model, and the global scale of the feature weight vector does not affect the output map. Hence, the weight vector is normalized as $\boldsymbol{w} = \boldsymbol{w}/|\boldsymbol{w}|$ before testing. Since low and mid contextual level images do not contain human faces, weights for the face feature always become 0 in these cases. Therefore, we normalized the full 30-dimensional vector only when comparing task differences with the high context group. In other cases including comparisons between low/mid and high context groups, 27-dimensional subvectors corresponding color, intensity and orientation features are normalized.

Tables 5.2, 5.3 and 5.4 show the p-values with the null hypothesis that two mean weight vectors are equal. p-values smaller than 0.05 are denoted in bold

	Low level	Mid level	High level
Color	**< 0.01**	> 0.50	0.21
Intensity	**0.02**	> 0.50	0.46
Orientation	> 0.50	> 0.50	**0.01**
Face	Uncompared	Uncompared	> 0.50

Table 5.4: Differences brought by viewing task difference (Free viewing and Preference rating) in different contextual levels.

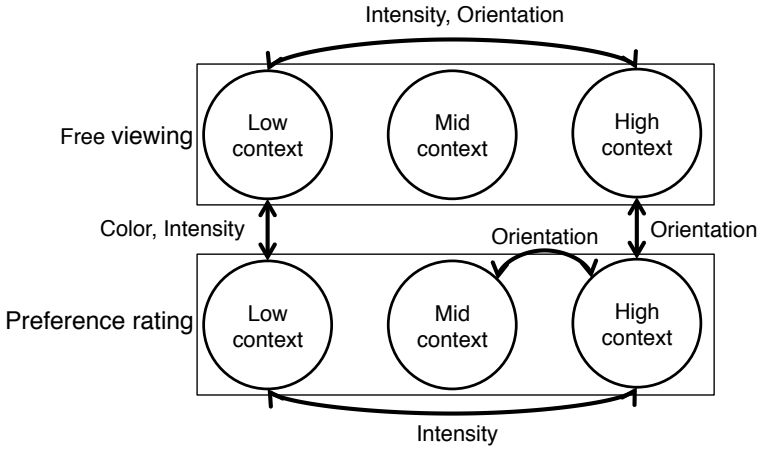


Figure 5.4: Differences in weight vectors between training sets. The double-arrow linking a comparison pair means that at least one of the feature weights is statistically different between the two groups at 0.05 significance level.

font. While the p-values are generally large, there are some comparison pairs with p-values lower than the 0.05 significance level, *i.e.*, significantly different feature weights.

The second column of Table 5.2 and 5.3 indicate that the feature weights learned using low and high level images are significantly different in both tasks. From the comparisons under the free viewing task (Table 5.2), it can be seen that each feature weight vector becomes gradually different in accordance with the contextual level, and the difference becomes significant between low and high level groups.

Interestingly, significant difference between feature weights learned using different tasks can only be seen in low and high level groups (the first and third column of Table 5.4). While the color and intensity feature weights became different for the low level group, the high level group shows significantly different orientation feature weights. This indicates that the choice of the task does influence the learned saliency model, and the effect is somewhat stimulus-dependent.

An overview of the differences is shown in Figure 5.4. It can be clearly seen that, among natural scene images (mid and high context groups), only the high context group viewed in preference rating task shows statistically significant difference from other groups.

5.3 Influence by aging

We trained models using data of young adults and elderlies viewing nature static and dynamic scenes. Number of region division in both cases was three. Features used in models for static and dynamic scene were different. For models of static image, color, intensity, orientation and face were used. As for the models for video, face feature was not counted, but flicker and motion feature for dynamic scene were considered. As we assumed that elderly came with a smaller useful field of view, the threshold of region division was determined by the data of young group. First saccade was used to generate the static model, and first five saccades were used for dynamic model.

Figure 5.5 shows the mean weight vectors trained by image dataset (Figure 5.5(a), 5.5(b)) and video dataset (Figure 5.5(c), 5.5(d)) of young and older observers. The legend in the figure stands for feature name and scale level of it. Similarly to Figure 5.3, C = Color, I = Intensity, O = Orientation and a newly added features M = Motion. It is worth noting that F in the static scene model stands for "Face", but it means "Flicker" in the dynamic scene model. Face feature only includes one scale level thus there is no number after "F". On the other hand, feature of flicker has three scales in the dynamic model. Scale level i represents the feature map is created at the size of $\frac{1}{2^{i+1}}$ of the original input image. R1 to R3 in the horizontal axis entails region 1 to region 3, from inner region to outer region of the visual field.

Based on the training results, we statistically tested the difference of feature weight vectors of young and older observers in following conditions:

1. Comparing overall difference of a single feature.
2. Comparing difference of a single feature in each region.

The test implementation was similar to that mentioned in section 5.2. For the ease of understanding, we denote young adult group as group 1 and elderly group as group 2. Given two sets of weight vectors from group 1, $W_1 = \{\boldsymbol{w}_{11}, \cdots, \boldsymbol{w}_{1n}\}$ and group 2, $W_2 = \{\boldsymbol{w}_{21}, \cdots, \boldsymbol{w}_{2n}\}$, we compared the weight vectors through the paired Hotelling's T^2 test [69]. Since the full dimension of the weight vector ($d = 30$ for static model and $d = 45$ for dynamic model) is much higher than the number of observations $n = 12$, we compared each feature type independently. Accordingly, the dimension of the weight vector reduces as $d = 9$ for the color, intensity, orientation, flicker and motion features, and $d = 3$ for the face feature. The weight vector is normalized as $\boldsymbol{w} = \boldsymbol{w}/|\boldsymbol{w}|$ before testing.

Here we repeated the definitions in section 5.2, let $\mathcal{N}_p = (\boldsymbol{\mu}, \Sigma)$ denote p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Thus we have n observations in X_1: $\boldsymbol{x}_{11}, \cdots, \boldsymbol{x}_{1n} \sim \mathcal{N}_p = (\boldsymbol{\mu}_1, \Sigma_1)$ and X_2: $\boldsymbol{x}_{21}, \cdots, \boldsymbol{x}_{2n} \sim \mathcal{N}_p = (\boldsymbol{\mu}_2, \Sigma_2)$, where $n = 12$. Null hypothesis H_0 is

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \ . \tag{5.8}$$

Then we denote

$$\boldsymbol{y}_i = \boldsymbol{x}_{1i} - \boldsymbol{x}_{2i}, \tag{5.9}$$

So that $\boldsymbol{y}_i \sim \mathcal{N}_p = (\boldsymbol{\mu}_y, \Sigma_y)$ where $\boldsymbol{\mu}_y = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Null hypothesis H_0 becomes

$$\boldsymbol{\mu}_y = 0. \tag{5.10}$$

Sample mean vector for $\boldsymbol{y}_i$ is

$$\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \qquad , \tag{5.11}$$

The difference in this test is that we used a shrinkage estimator of covariance matrix. The widely used standard covariance for two variables that

$$\operatorname{cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}),$$

is ill-suited for inferring large-scale covariance matrices from sparse data. We applied the shrinkage covariance estimator proposed by Schäfer and Strimmer [70], which is a statistically efficient and computationally fast alternative to that of Ledoit-Wolf shrinkage estimator [71], such that the covariance matrix can have guaranteed minimum mean squared error, be well-conditioned, and be always positive definite even for small sample sizes.

Consequently, the covariance matrix S_y in the test is:

$$s^*_{ij} = \begin{cases} s_{ij} & \text{if } i = j \\ r^*_{ij}\sqrt{s_{ii}s_{jj}} & \text{if } i \neq j \end{cases} \tag{5.12}$$

and

$$r^*_{ij} = \begin{cases} 1 & \text{if } i = j \\ r_{ij}\min(1, \max(0, 1 - \hat{\lambda}^*)) & \text{if } i \neq j \end{cases} \tag{5.13}$$

with

$$\hat{\lambda}^* = \frac{\sum_{i \neq j} \widehat{Var}(r_{ij})}{\sum_{i \neq j} r^2_{ij}} \tag{5.14}$$

Features / Model Type	Color	Intensity	Orientation	Face	Flicker	Motion
Static model	0.281	0.067	0.22	0.97	\	\
Dynamic model	NaN	NaN	0.082	\	0.358	0.115

Table 5.5: Differences brought by aging in static and dynamic visual saliency model.

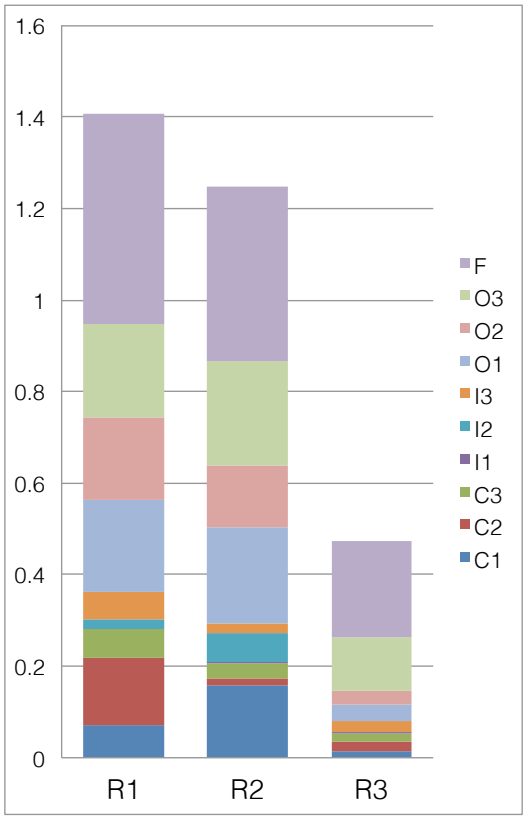
Then Paired Hotelling's T^2 test statistic is given by

$$T^2 = n\bar{\boldsymbol{y}}' {S_y}^{-1} \bar{\boldsymbol{y}} \qquad , \tag{5.15}$$

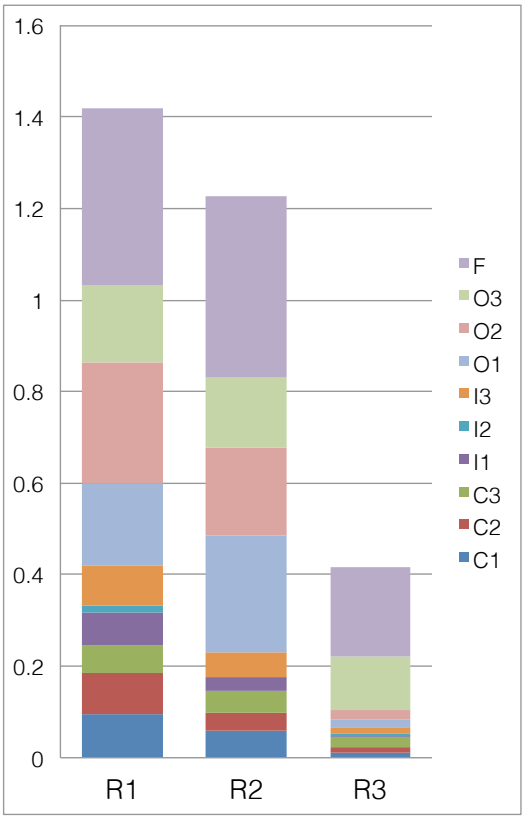
With Paired Hotelling's T-square test statistic T^2, we have

$$F = \frac{n-p}{p(n-1)} T^2 \sim F_{p,n-p} \ , \tag{5.16}$$

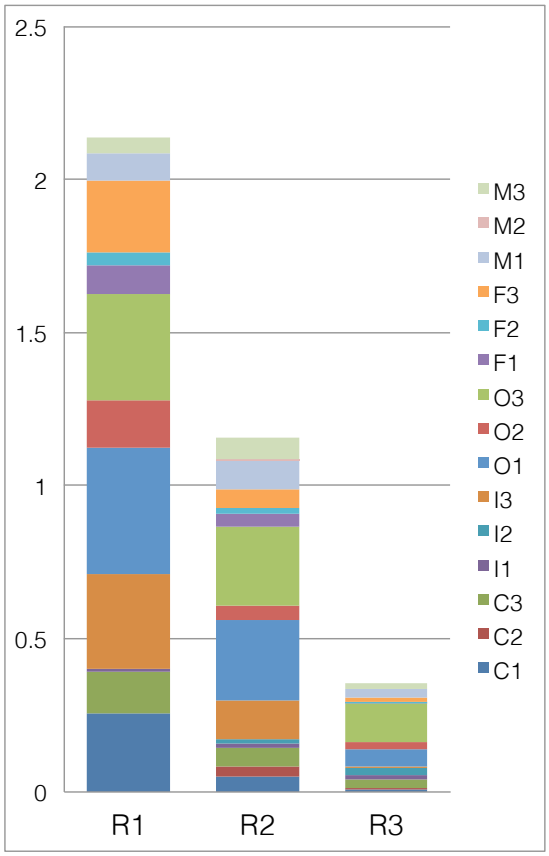
where $F_{p,n-p}$ is the F-distribution with parameters p and $n-p$. We reject H_0 at level α if $F > F_{p,n-p,\alpha}$, where $F_{p,n-p,\alpha}$ is the F-value with p and $n-p$ degrees of freedom, evaluated at level $\alpha = 0.05$. Table 5.5 shows the p-value with the null hypothesis that two mean weight vectors are equal. NaN in the table means the covariance matrix S_y is not reversible thus it is not able to calculate the T^2 test statistic for that test, and the backslash indicates the corresponding test has not been carried out. Interestingly, no p-values is smaller than 0.05, which indicates that the two groups of subjects, young adults and elderlies, have no difference in viewing images and videos.



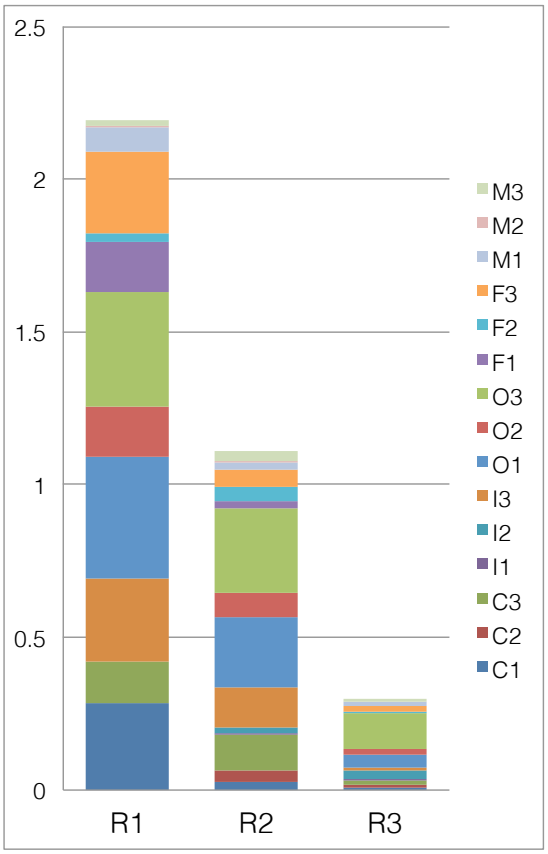
(a) Mean vector of static model, young observers



(b) Mean vector of static model, older observers



(c) Mean vector of dynamic model, young observers



(d) Mean vector of dynamic model, older observers

Figure 5.5: Differences brought by age difference.

Chapter 6

Summary

In this thesis, we presented a quantitative investigation on how individual difference and dataset characteristics affect learning-based a visual saliency model. We created a new image dataset that consists of low, mid, high contextual level images and fixation data collected during two different tasks, free viewing and preference rating, and a new video dataset that contains fixation data of young adults and elderlies. By taking Kubota *et al.*'s model as an example, we examined the influence of

1. Individual difference
2. Stimuli types and viewing tasks
3. Aging

by comparing learned feature weight vectors statistically.

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 section 5.1.3, however we found there is no clear distinction between the test subjects until all features are used. There are two possible reasons to explain 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.

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 were aware of that learned weight vectors become different according to the level of contextual complexity, and they become significantly different between low and high contextual groups. We also found that the influence brought by the task difference is heavily stimuli-dependent, *i.e.*, different tendencies were observed for each contextual group. This indicates that the relationship between visual saliency and task/stimuli is complex, and it is not a trivial task to investigate how we can build generic and/or task-specific models of visual saliency in a supervised manner. Although the mechanism behind visual attention has not been fully understood yet, our dataset and analysis can provide valuable information to establish a systematic and generic method for learning-based modeling of visual saliency.

Lastly, we found no significant difference in the training results of young adults and elderlies in our setting, either in the static or the dynamic scene, which is a little disappointing but not surprising because there were reports [62] showing that the age-related changes were apparent only in the oldest subjects that they tested (70–81 years of age). Another reason that only minor difference exists between two groups of subjects can be the narrow viewing angle ($45^\circ \times 29^\circ$) in the data collecting environment such that the participants was not viewing the stimulus with full visual field. Although prior researches suggested that the dynamic visual acuity of elderly is worse than young adults, the forming of attention might not need the acuity. That is, knowing there is something somewhere is enough to direct the visual attention, while acuity means knowing what is exactly there.

It is an important future direction to investigate if elderly visual attention model is different from that of young adults under extremer condition. Such as, recruiting older subjects, using a larger display device to show the stimuli, including stimuli that contain more fast motions. In addition, it might give interesting outcomes if we repeat our experiments on infants whose visual system is not fully developed.

List of Figures

List of Tables

Bibliography

[1] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. B. II, V. Balasubramanian, and P. Sterling, "How much the eye tells the brain," *Current Biology*, vol. 16, no. 14, pp. 1428 – 1434, 2006.

[2] L. Spillmann and J. S. Werner, *Visual perception: the neurophysiological foundations*. Academic Press, 1989.

[3] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, Inc., 1995.

[4] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[5] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence*. Springer, 1987, pp. 115–141.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[7] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 1, pp. 185–207, 2013.

[8] A. Mishra, Y. Aloimonos, and C. L. Fah, "Active segmentation with fixation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 468–475.

[9] A. Maki, P. Nordlund, and J.-O. Eklundh, "Attentional scene segmentation: integrating depth and motion," *Computer Vision and Image Understanding*, vol. 78, no. 3, pp. 351–373, 2000.

[10] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 2232–2239.

[11] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 5, pp. 802–817, 2006.

[12] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs, "Automatic thumbnail cropping and its effectiveness," in *Proceedings of the 16th annual ACM symposium on User interface software and technology*. ACM, 2003, pp. 95–104.

[13] S. Frintrop, *VOCUS: A visual attention system for object detection and goal-directed search*. Springer, 2006, vol. 3899.

[14] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2049–2056.

[15] N. J. Butko and J. R. Movellan, "Optimal scanning for faster object detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2751–2758.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[17] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modelling search for people in 900 scenes: A combined source model of eye guidance," *Visual Cognition*, vol. 17, no. 6-7, pp. 945–978, 2009.

[18] A. A. Salah, E. Alpaydin, and L. Akarun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 3, pp. 420–425, 2002.

[19] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Advances in neural information processing systems*, 2004, pp. 481–488.

[20] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 6, pp. 989–1005, 2009.

[21] S. Mitri, S. Frintrop, K. Pervolz, H. Surmann, and A. Nuchter, "Robust object detection at regions of interest with an application in ball recognition," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*. IEEE, 2005, pp. 125–130.

[22] G. Fritz, C. Seifert, L. Paletta, and H. Bischof, "Attentive object detection using an information theoretic saliency measure," in *Attention and Performance in Computational Vision*. Springer, 2005, pp. 29–41.

[23] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[24] S. Han and N. Vasconcelos, "Biologically plausible saliency mechanisms improve feedforward object recognition," *Vision research*, vol. 50, no. 22, pp. 2295–2307, 2010.

[25] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 1007–1013.

[26] S. Frintrop, "General object tracking with a component-based target descriptor," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on.* IEEE, 2010, pp. 4531–4536.

[27] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *Image Processing, IEEE Transactions on*, vol. 13, no. 10, pp. 1304–1318, 2004.

[28] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *Image Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 185–198, 2010.

[29] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," *rn*, vol. 255, p. 3, 1999.

[30] A. Belardinelli, "Salience features selection: Deriving a model from human evidence," Ph.D. dissertation, University of Rome La Sapienza, 2008.

[31] B. Mertsching, M.Bollmann, R. Hoischen, and S. Schmalz, "The neural active vision system," in *Handbook of Computer Vision and Applications*, B. Jähne, H. Haußecker, and P. Geißler, Eds. Academic Press, 1999.

[32] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Methods*, pp. 1–16, 2012.

[33] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS2006*, 2006, pp. 545–552.

[34] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Advances in neural information processing systems*, 2005, pp. 547–554.

[35] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of vision*, vol. 8, no. 7, 2008.

[36] R. Raj, W. S. Geisler, R. A. Frazor, and A. C. Bovik, "Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy," *JOSA A*, vol. 22, no. 10, pp. 2039–2049, 2005.

[37] S. Engmann, M. Bernard, T. Sieren, S. Onat, P. König, and W. Einhäuser, "Saliency on a natural scene background: Effects of color and luminance contrast add linearly," *Attention, Perception, & Psychophysics*, vol. 71, no. 6, pp. 1337–1352, 2009.

[38] S. Onat, K. Libertus, and P. König, "Integrating audiovisual information for the control of overt attention," *Journal of Vision*, vol. 7, no. 10, 2007.

[39] L. Itti and C. Koch, "A comparison of feature combination strategies for saliency-based visual attention systems," *SPIE human vision and electronic imaging IV (HVEI'99)*, vol. 3644, pp. 373–382, 1999.

[40] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Optical Science and Technology, SPIE's 48th Annual Meeting*. International Society for Optics and Photonics, 2004, pp. 64–78.

[41] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. CVPR2010*. IEEE, 2010, pp. 3485–3492.

[42] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[44] W. Kienzle, F. A. Wichmann, B. Schölkopf, and M. O. Franz, "A nonparametric approach to bottom-up visual saliency," in *Proc. NIPS2007*, 2007, pp. 689–696.

[45] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. ICCV2009*, 2009, pp. 2106–2113.

[46] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[47] R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," *Vision research*, vol. 39, no. 19, pp. 3157–3163, 1999.

[48] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of vision*, vol. 11, no. 3, 2011.

[49] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of vision*, vol. 9, no. 12, 2009.

[50] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using adaboost," *Journal of Vision*, vol. 12, no. 6, 2012.

[51] H. Kubota, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Incorporating visual field characteristics into a saliency map," in *Proc. ETRA2012*. ACM, 2012, pp. 333–336.

[52] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of vision*, vol. 9, no. 3, 2009.

[53] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Proc. ECCV2010*, Berlin, Heidelberg, 2010, pp. 30–43.

[54] I. Van Der Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Doves: a database of visual eye movements," *Spatial vision*, vol. 22, no. 2, pp. 161–177, 2009.

[55] J. H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 265, no. 1394, pp. 359–366, 1998.

[56] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision research*, vol. 42, no. 1, pp. 107–123, 2002.

[57] P. Van De Laar, T. Heskes, and S. Gielen, "Task-dependent learning of attention," *Neural networks*, vol. 10, no. 6, pp. 981–992, 1997.

[58] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on.* IEEE, 2007, pp. 1–8.

[59] K. K. Ball, B. L. Beard, D. L. Roenker, R. L. Miller, and D. S. Griggs, "Age and visual search: Expanding the useful field of view," *JOSA A*, vol. 5, no. 12, pp. 2210–2219, 1988.

[60] G. M. Long and R. F. Crambert, "The nature and basis of age-related changes in dynamic visual acuity." *Psychology and Aging*, vol. 5, no. 1, p. 138, 1990.

[61] H. Ishigaki and M. Miyao, "Implications for dynamic visual acuity with changes in age and sex," *Perceptual and motor skills*, vol. 78, no. 2, pp. 363–369, 1994.

[62] P. J. Bennett, R. Sekuler, and A. B. Sekuler, "The effects of aging on motion detection and direction identification," *Vision Research*, vol. 47, no. 6, pp. 799–809, Mar. 2007.

[63] J. M. Wood, "Aging, driving and vision," *Clinical and experimental optometry*, vol. 85, no. 4, pp. 214–220, 2002.

[64] P. J. B. M. M. A. B Sekuler, "Effects of aging on the useful field of view," *Experimental aging research*, vol. 26, no. 2, pp. 103–120, 2000.

[65] A. F. Kramer, D. G. Humphrey, J. F. Larish, and G. D. Logan, "Aging and inhibition: beyond a unitary view of inhibitory processing in attention." *Psychology and Aging*, vol. 9, no. 4, p. 491, 1994.

[66] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[67] D. A. Wolfe and M. Hollander, "Nonparametric statistical methods," *Nonparametric statistical methods*, 1973.

[68] W. Krzanowski, "Principles of multivariate analysis: A user's perspective," 1988.

[69] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*. Pearson, 2002.

[70] J. Schäfer, K. Strimmer *et al.*, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, p. 32, 2005.

[71] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, vol. 10, no. 5, pp. 603–621, 2003.

[72] J. K. Tsotsos, "Is complexity theory appropriate for analyzing biological systems?" *Behavioral and Brain Sciences*, vol. 14, pp. 770–773, 12 1991.

[73] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proc. ICIP2002*, 2002, pp. 900–903.

[74] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Adv Neural Inf Process Syst*, 2005, pp. 155–162.

[75] Q. Zhao and C. Koch, "Learning saliency-based visual attention: A review," *Signal Processing*, vol. 93, no. 6, pp. 1401–1407, 2013.

[76] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency modeling," in *Proc. ICCV2013*, Dec 2013, bu;mod;cv, pp. 921–928.

[77] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision research*, vol. 47, no. 19, pp. 2483–2498, 2007.

[78] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proc. ICIP2002*, 2002, pp. 900–903.

Appendix A

Figures

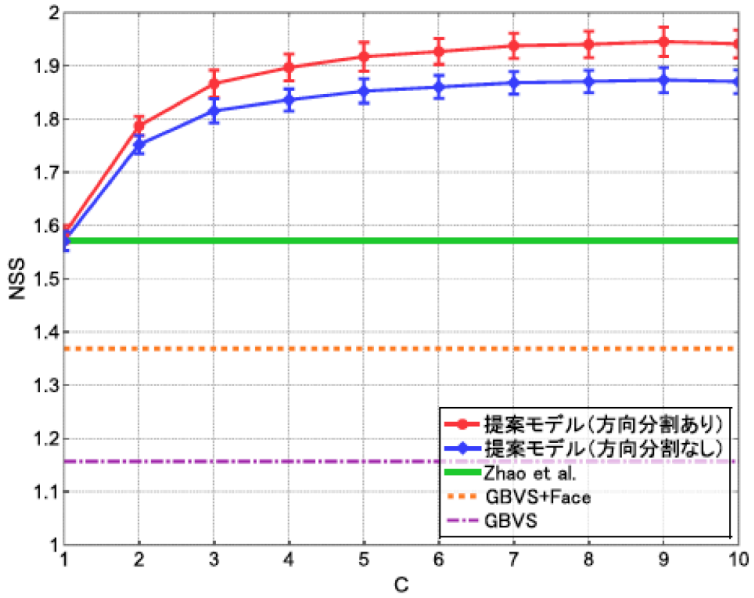
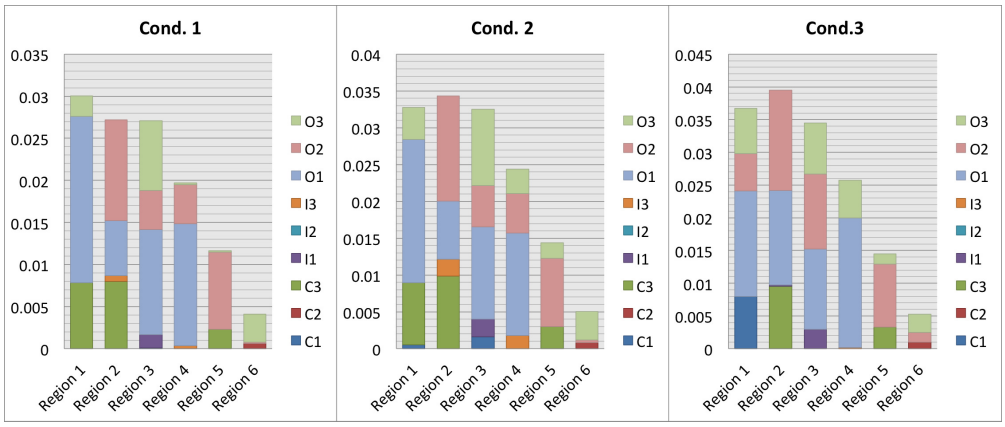
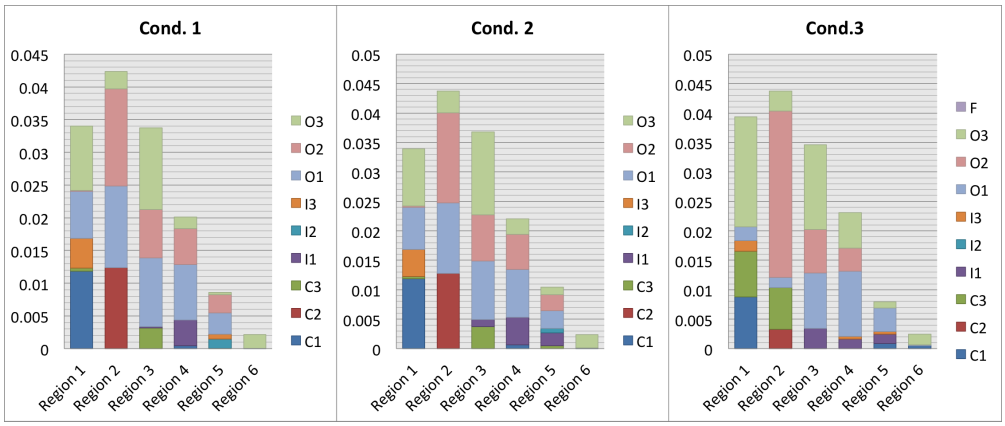


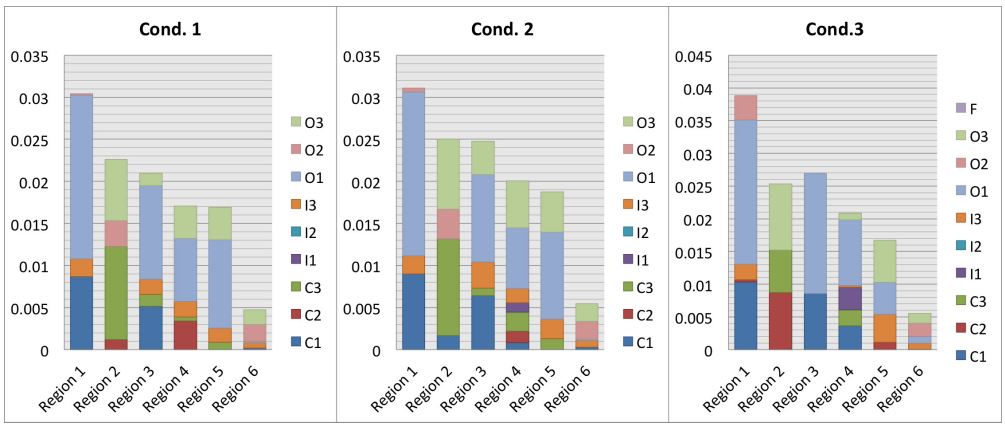
Figure A.1: Relation between performance and region division



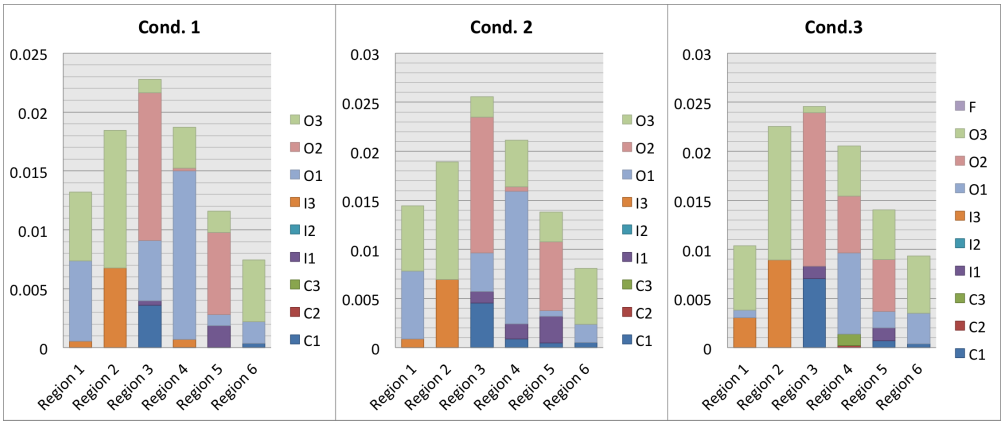
(a) Subject 1



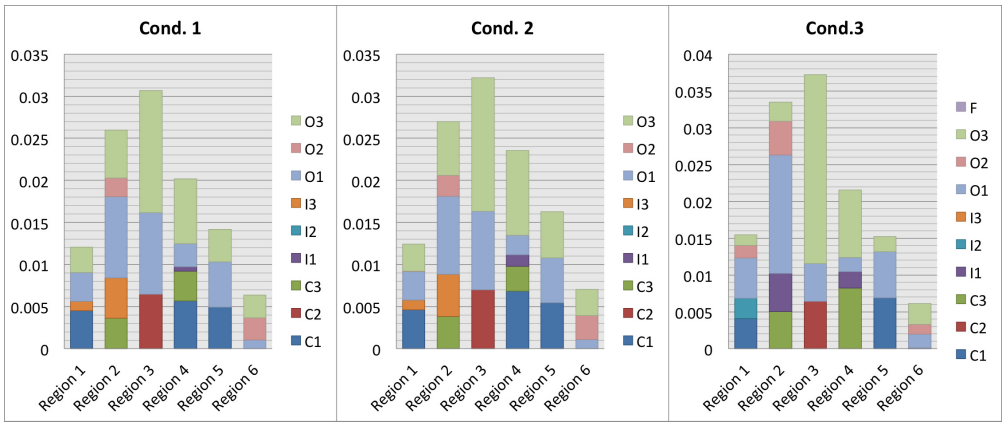
(b) Subject 2



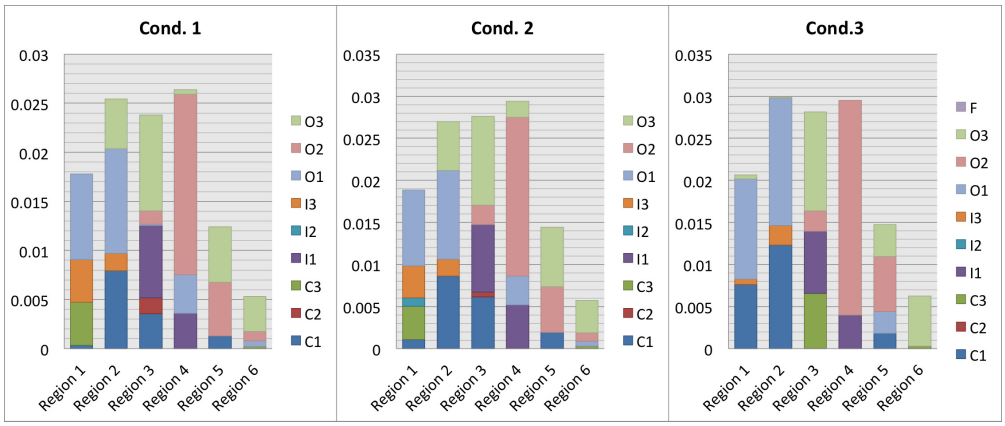
(c) Subject 3



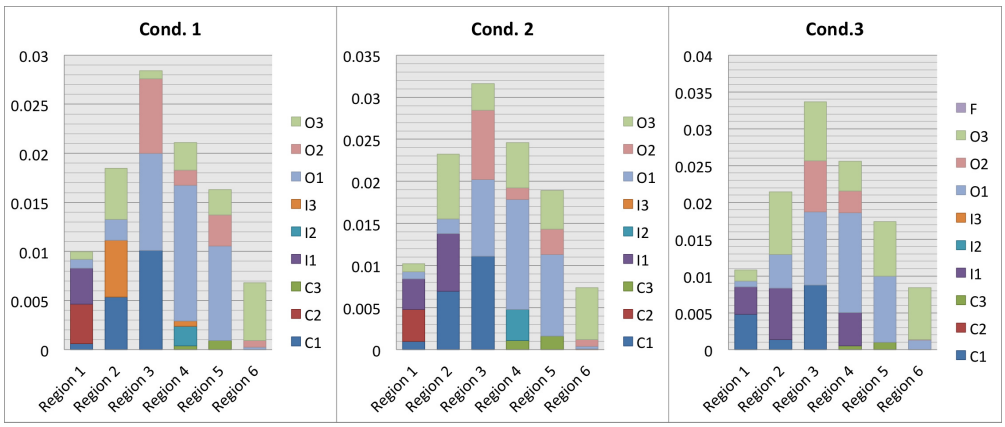
(d) Subject 4



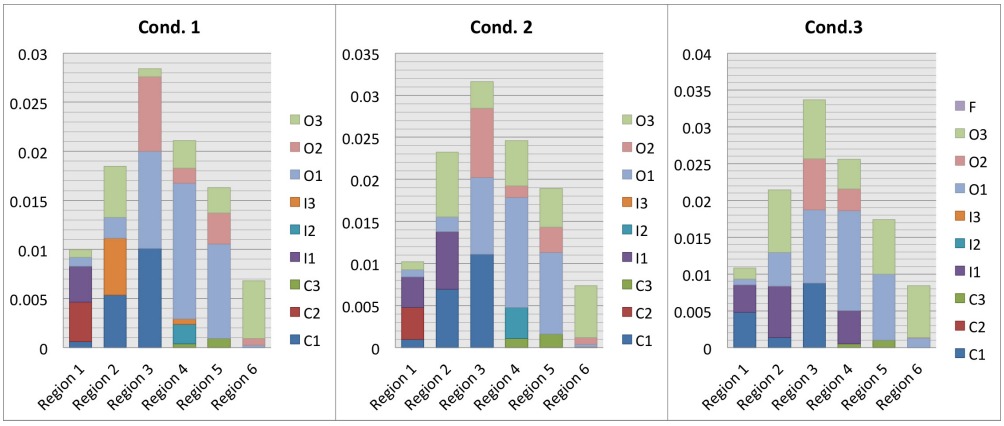
(e) Subject 5



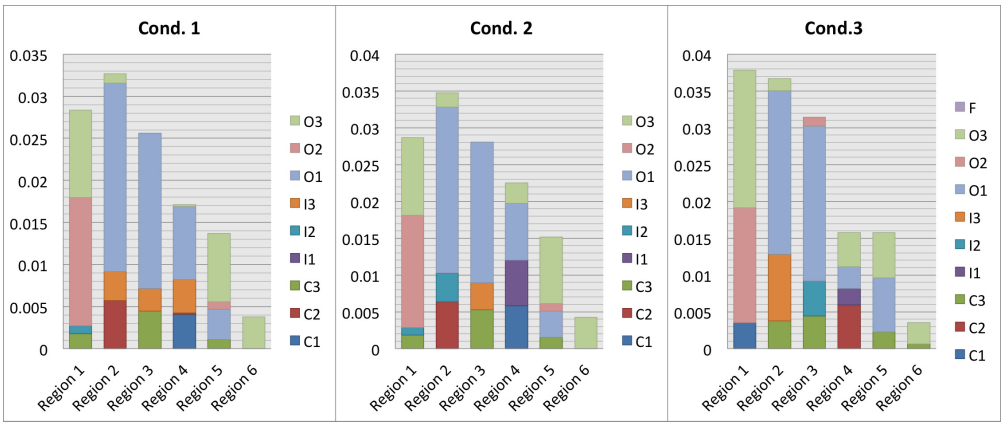
(f) Subject 6



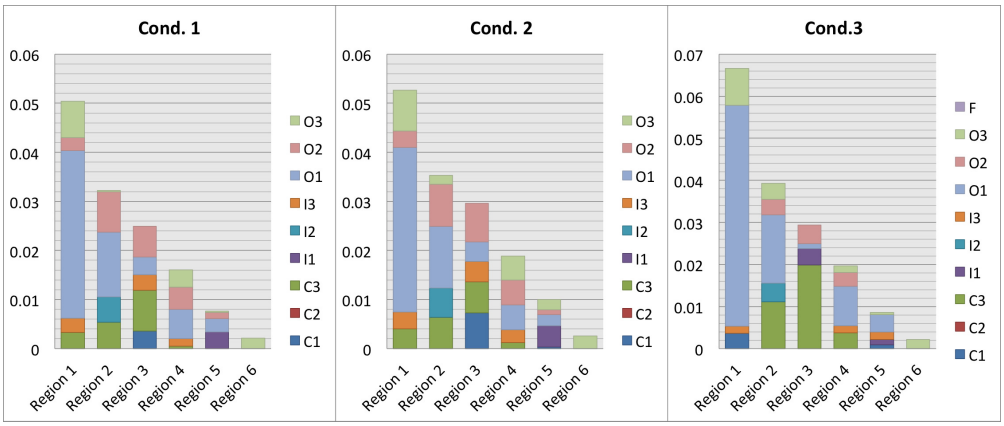
(g) Subject 9



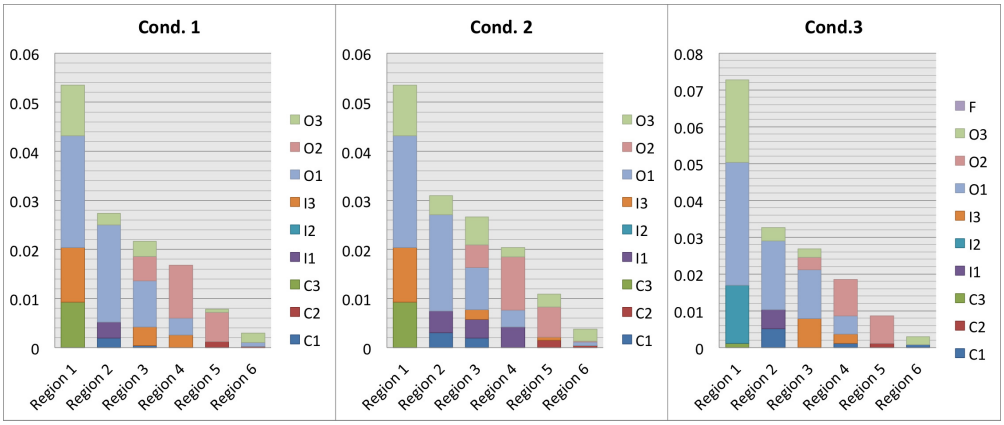
(h) Subject 8



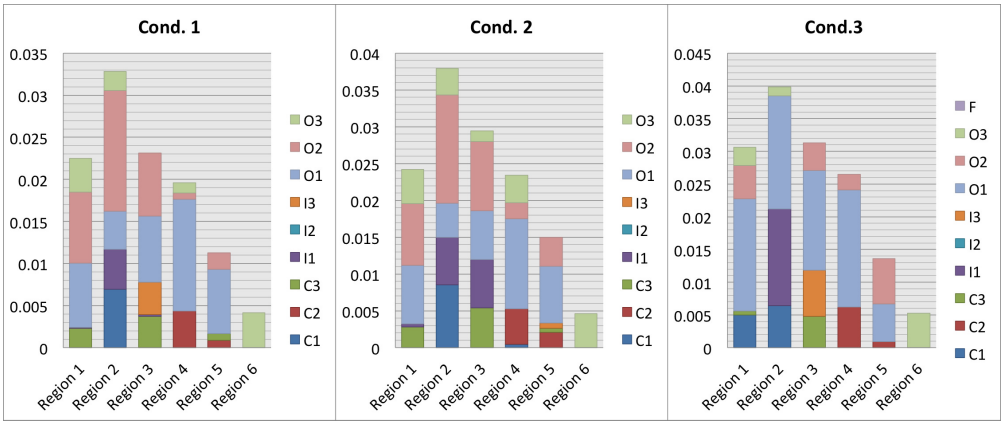
(i) Subject 9



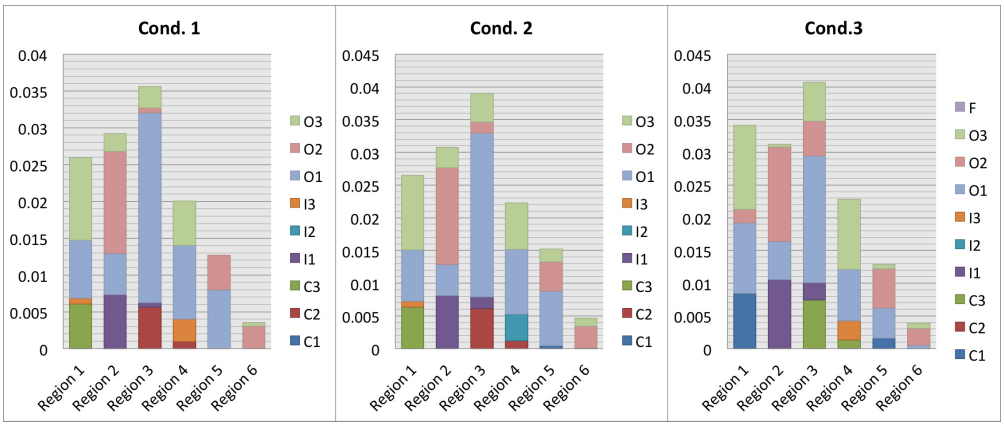
(j) Subject 10



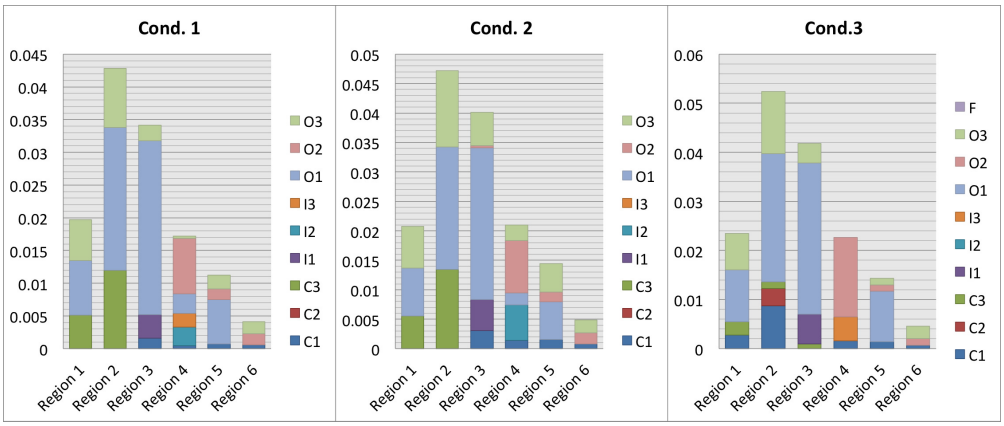
(k) Subject 11



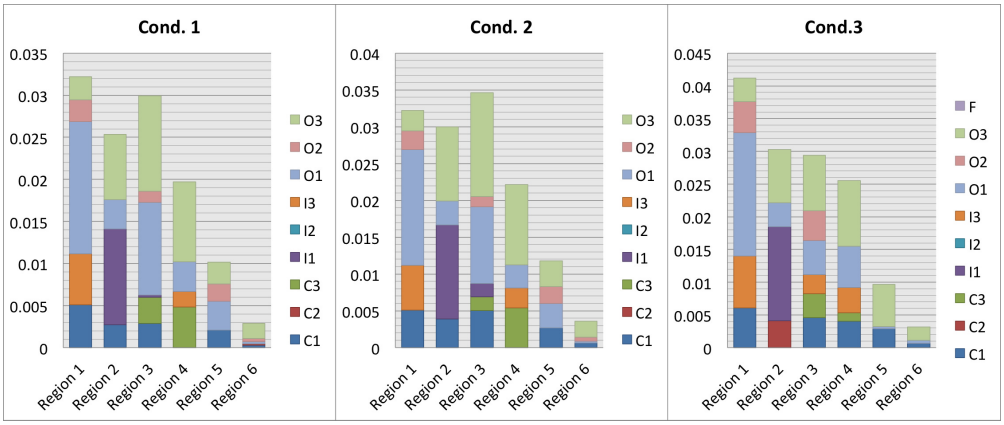
(l) Subject 12



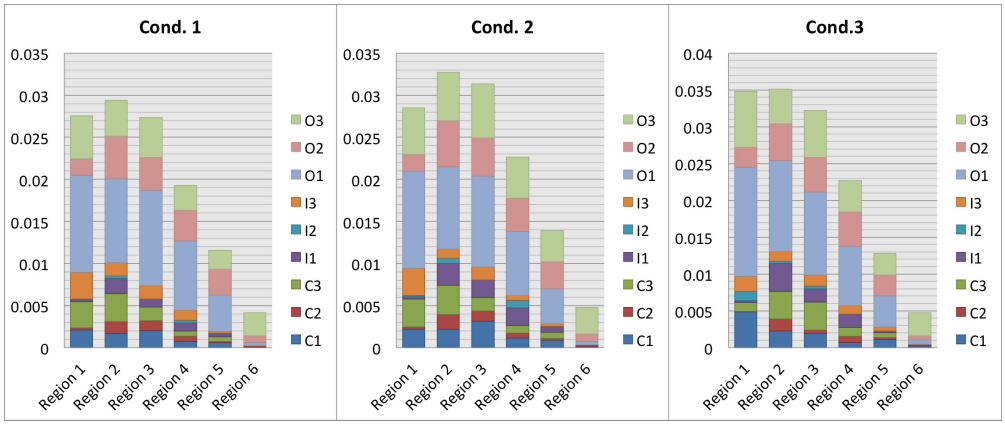
(m) Subject 13



(n) Subject 14



(o) Subject 15



(p) Mean of the feature of 15 subjects

Figure A.2: Bottom-up feature weights learned with and without face factor in saliency model

Publications

Binbin Ye, Yusuke Sugano, and Yoichi Sato, "Investigating individual differences in learning-based visual saliency model", in *JPSJ SIG: Computer Vision and Image Media (CVIM 2013)*, September 2013.

Binbin Ye, Yusuke Sugano, and Yoichi Sato, "Influence of stimulus and viewing task types on a learning-based visual saliency mode", in *Proc. Symposium on Eye Tracking Research and Applications (ETRA2014)*, March 2014.