論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Personalized Saliency in Non-immersive and Immersive Environments for Practical Applications
著者(和文)	Erum ZaibSumaira
Author(English)	Sumaira Erum Zaib
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12294号, 授与年月日:2022年12月31日, 学位の種別:課程博士, 審査員:山村 雅幸,瀧ノ上 正浩,小野 功,青西 亨,関嶋 政和
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12294号, Conferred date:2022/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Doctoral Dissertation

Personalized Saliency in Non-Immersive and Immersive Environments for Practical Applications

Sumaira Erum Zaib

Graduate Major in Artificial Intelligence
Department of Computer Science
School of Computing
Tokyo Institute of Technology

Supervisor: Professor Masayuki Yamamura

December, 2022

Abstract

Saliency is the ability of being important, noticeable, and standing out among others. Saliency that takes the heterogeneity of individuals into consideration is called personalized saliency. Currently personalized saliency models in non-immersive environments, such as smartphones and computers, use deep learning methods which are resource heavy on-device and pose network limitations and privacy concerns in cloud. In this thesis, we propose using gradient boosted tree regression and just the color spaces of user fixations to determine personalized saliency without collecting invasive personal data. In immersive environments such as virtual reality, we determine that the use of auditory and emotional saliency is much more useful for practical application purpose. Therefore, we propose using different machine learning algorithms along with heart rate from user's personal smartwatch to successfully induce and amplify emotions for personalized saliency. Using devices already owned by users and these techniques, we successfully bridge the gap between personalized saliency research and practical applications in both non-immersive and immersive environments.

Thesis Summary

Saliency is the ability of being important, noticeable, and standing out among others. Anything that attracts attention of the observer is salient in nature. Our brain has a magnificent ability to filter out what is important, according to the situation, from the constant incoming sensory information.

From visual perspective, properties such as color, contrast, hue, and brightness determine salient regions or objects in images. In videos the movement of the objects play an additional role in making the content salient or not. From auditory perspective, properties such as loudness, pitch and content determine saliency of the sound. Using these sensory based properties of images, videos, and sounds in determining saliency is called bottom-up saliency.

Another type of saliency is emotional saliency, where according to the situation at hand we can amplify or reduce the intensity of the emotions we are feeling. This can later effect how the event is stored in memory and remembered.

The task being performed also has a great impact on visual, auditory, and emotional saliency. What is salient in one situation might not be as salient in another. Calculating saliency using task-based information is called Top-down saliency. Both bottom-up and top-down saliency are used extensively in research to develop saliency models.

As we are moving towards a society that prefers personalized experience in everything, the need for personalized saliency also arises. Research has shown that saliency is personal. What object or sound grabs the attention of one person may be completely missed by another. What one person feels in a situation might be different for another. These differences arise due to differences in age, gender, culture, personality, and life experiences. Since traditional saliency models do not consider heterogeneity of individuals, special personalized saliency models are needed for individualistic experiences.

In non-immersive environments such as mobile phones and computers, visual saliency has important applications in image compression, website de-

signs, custom user experiences in applications etc. However, currently, almost all personalized saliency models use different deep learning methods to calculate personalized saliency. Even in general saliency the focus is on using deep learning technique to improve saliency. However, such methods cannot be used in actual devices used by consumers due to high computational resources required for deep learning. Even if model training was done in cloud, network bandwidth issues would reduce efficiency. Another important factor to take into consideration for practical applications is data privacy of the users. This limits the usage of cloud for model training and inference.

For non-immersive environments, in this thesis, we propose a personalized saliency prediction framework that only requires only the eye fixations of the individuals to compute personalized saliency using color spaces (RGB, CYMK, HSV, HSL) and a machine learning algorithm called gradient boosted tree regression. This eliminates the need to collect personal data from the users such has their likes, dislikes, gender, cultural background etc. Since on-device machine learning is well established, this framework can be used as a non-invasive way to offer services that require personalized saliency in mobile and computer applications.

In immersive environments such as Virtual Reality, though visual saliency is useful, but practical applications such as therapy for mental disorders, training of different sorts, gaming require more emotional saliency. Research has also shown that there is a disconnect in immersive environments between what people see and what they think about it. Therefore, a need beyond visual saliency arises.

Environment sounds and the physical condition of the user all play a role in what makes the situation emotionally salient. However, currently all research being done regarding emotion induction and personalization uses special gadgets and sensors to study the physiological parameters of the users such as heart rate, respiratory rate, and galvanic skin response in controlled laboratory environment.

What is not taken into consideration is the fact that everyday people do not possess such devices. These parameters then become impractical to application developers for the purpose of creating applications that require personalized saliency in immersive environments. Even if they are used, it would make these applications highly inaccessible to the public.

To tackle this problem, in this thesis, we propose a framework that uses machine learning and heart rate from everyday smartwatches used by people

for emotion induction in their own space and not in a laboratory. Using just the heart rate data and audio data we can induce and amplify emotion of the participants. Since fear is the most personal emotion, this research focused on inducing and amplifying fear. However, other emotions could also be used. We also present our discussion on how gender and environment impact heart rate. The results of control group and experimental group also confirmed that this research can be successfully applied in real life applications since all the parameters are accessible to the developers.

In conclusion, this study discusses why personalized saliency is important and proposes frameworks to calculate them successfully in both immersive and non-immersive environments for practical applications.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

With evolution, our brain has adapted itself to pay attention to only a subset of information, even though it receives constant information from all senses. We can divert our attention to any particular object or sound in an environment, we can control emotions and determine what is important according to the situation and task at hand. In living beings, with the passage of time, this ability to filter information needs constant improvement for survival of the species. [166]

Saliency is a property of anything that makes it stand out from others and grab attention of the observer. Anything that attracts attention, possess the property of saliency and is thus salient in nature. An object can be salient in any visual medium, a sound can be salient in an auditory medium, an emotion can be salient psychologically. Saliency can be based on properties such as colors or contrast in images. Properties such as pitch or loudness could impact the saliency of sound. Similarly, past experiences could impact the saliency of an emotion. Saliency is also effected when the observer is performing a specific task. Different tasks would require the observer to pay attention to different stimuli. Therefore, saliency of any particular stimuli could be based on its characteristics, on the task at hand or both.

1.1 Personalized Saliency

Since every individual is unique, what they pay attention to is also unique. What people pay attention to in any medium changes based on their age, personality, gender, culture and life experiences.

Even though extensive research has been done on saliency from visual, auditory and emotional perspective, personalized saliency is still a relatively

new field. Despite being a new field, it is an important one.

We are currently moving towards a society that wants personalized services and personalized consumer experience is at its peak. Every service wants to provide a unique user-tailored experience to the customer for their specific needs. This can be seen in social media content, media streaming services, shopping recommendations and even in gaming.

From a social perspective, personalized therapy for mental issues such as autism and different kinds of phobia using virtual reality is an important application that utilizes personalized emotional saliency. Emotional saliency utilizing virtual reality could also be used for training in medical and military fields. Students that have different learning needs that need further training to deal with Attention deficit hyperactivity disorder (ADHD) and anxiety could also use therapy with personalized emotional saliency to overcome the challenges that their other peers might not face.

1.2 Types of Saliency

With saliency, there are also different types of saliency that can be categorized based on how the information is being processed or in what medium saliency is being studied. To study personalized saliency in detail for non immersive and immersive environments, it is important for us to first understand these different categories of saliency.

When studying how the incoming sensory information is being processed, we can divide saliency into two categories:

- Bottom-up Saliency
- Top-down Saliency

When studying the medium in which saliency is to be detected, predicted or studied, we have multiple categories that can be further divided into subcategories. For the purpose of our research, we will be focusing on the following categories:

- Visual Saliency
- Auditory Saliency
- Emotional Saliency

1.2.1 Bottom-Up Saliency

Although usually associated with visual saliency, bottom-up saliency refers to how our senses are involved in determining what is important in our environment. This type of saliency is independent of the task at hand and is solely based on low level features. For images, these features can be colors, contrast, brightness, hues etc. For sound these features would be characteristics such as loudness or pitch.

Common practice to determine bottom-up saliency in visual medium is to let participants free view the images or videos, and then analyze what grabs their attention. Bottom-up saliency is therefore, where our focus goes involuntarily.

One of the earliest works that form the basis of most saliency models is done by Itti et al [84], Itti and Koch [86], and Koch and Ullman [107]. Based on works of [107], [137] and [18] Itti et al proposed a bottom-up saliency model in [84]. The model used "Feature Integration Theory" and was used to determine visual attention for rapid scene analysis.

Itti and Koch in [86] put forth the main five components that constitute bottom-up saliency models

- **Pre-attentive computation of visual features** Low level features of visuals are important in determining saliency but the context has to be taken into consideration.

- **Saliency** Creating a saliency map for visual stimuli is well structured and methodical way to depict saliency and reproducible human behavior regarding attention.

- **Attentional Selection** Eye movement and attention are highly interconnected. Therefore, references such as center bias need to be taken into consideration.

- **Inhibition of Return** To shift attention from one point to another, the most salient region needs to be suppressed. This will guide attention of the observer to the next salient location.

- **Attention and Recognition** Context in terms of scene and objects play a huge role in determining salient regions.

In more recent works, low-level features other than colors, intensity and orientation are being used for saliency detection. [228] proposed seven other

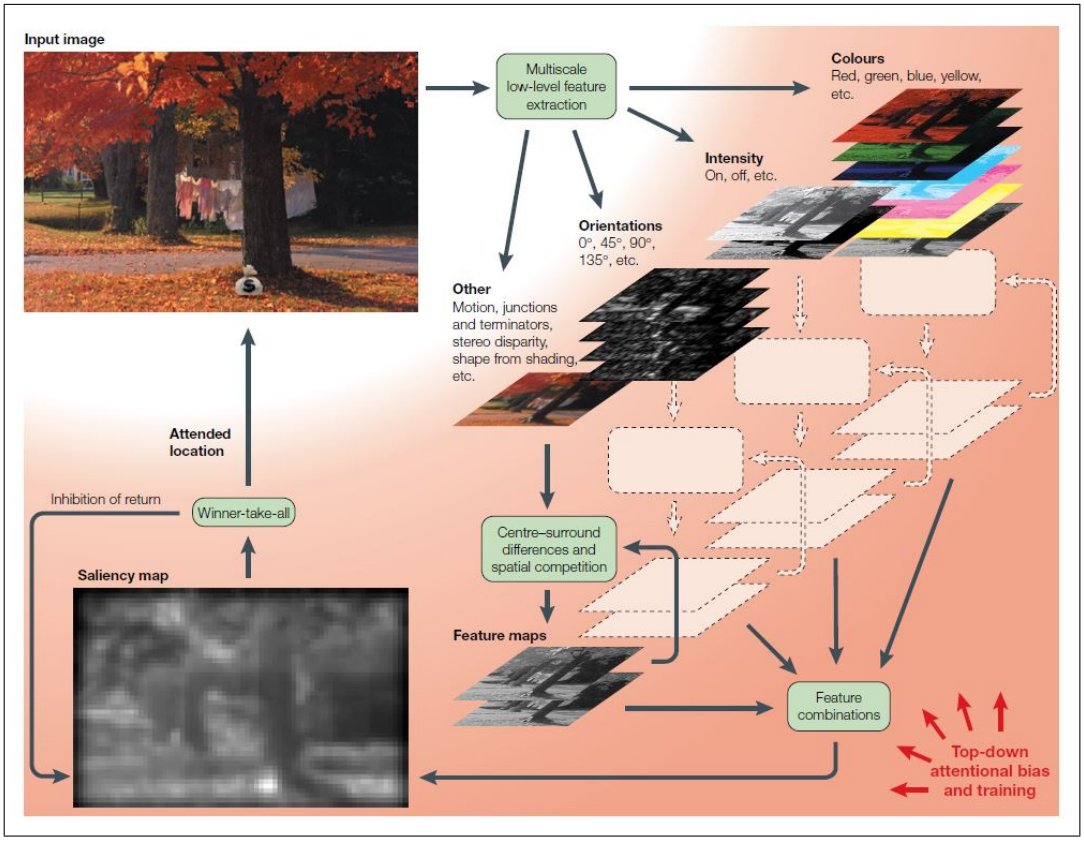


Figure 1.1: Diagram from [86] to demonstrate computational architecture proposed by Koch and Ulman in [107].

features, along with k-means, to determine bottom up saliency. These features included Region size (RS), Sum of color difference (SCD), Variance of middle frequency coefficients (VMF), Color variance (CV), Mean of distance of each pixel to the position center (MDC), Variance of distance of each pixel to the position center (VDC), and Mean of distance of each pixel of a region to the position center (MDI).

When studying the saliency importance of bottom-up saliency attributed in videos without considering cognitive bias, [79] found that warm (example: red and pink), bright colors (example: cyan and yellow), dense textures, patterns with high contrast were more prone to be salient. Vertical, high speed or low speed movements were also able to grab more attention from the users.

To help solve the problem of heavy computation and resource expenses associated with deep learning convolutional neural networks, dominant in saliency prediction, [131] tested thirty five different variations of their bottom up saliency models. Their findings suggest that models pre-trained for object classification provide the optimal implementation for saliency predictions. They also reported negative correlation between saliency prediction and depth of deep convolutional neural network.

A simple framework proposed by [35] showed that position alignment information and long term temporal information was an effective method to compute saliency in videos with dynamic backgrounds.

1.2.2 Top-Down Saliency

When discussing saliency, cognition plays an important role in how something becomes salient in a scenario. This type of saliency is known as top-down saliency. The context of the stimuli, the task at hand, goal accomplishment, prior knowledge, and the situation are examples of what could impact saliency. Researchers have even shown that in some cases top down saliency overrides bottom-up saliency.

In earlier works, [119] used a Self-Organizing Feature with Fuzzy Association neural network model and multi layer perceptron for better control at attention when recognizing patterns.

[146] showed how scene guidance can help in directing the gaze of the observer when searching for the target. If there is no scene context, then attention is directed by bottom-up saliency. However, in the presence of scene context, a bias can be introduced by creating areas that most likely contain the search target.

Contextual pooling was used in an attempt to predict goal oriented saliency in [230]. Block wise spatial pooling was reported to be effective in utilizing context information on three data sets.

To create class specific saliency maps, [106] improved on dictionary learning strategy from [100, 221] and used super pixel strategy from [5]. Considering the problem as a graph labeling problem, they were able to get higher accuracy than the individual approaches using this technique.

A multi exemplar deep convolutional neural network introduced in [77] aimed to study the relationships between exemplars and how they can effect saliency prediction. Their results showed that networks trained with greater number of exemplars could give out stronger associations. The model also was able to do this for unseen objects based on its prior knowledge.

An automatic top-down fusion model proposed in [160] used modified VGG-16 [176] to create an encoder-decoder network. The network was made for efficient and automatic flow of semantic information to bottom layers to improve prediction of salient objects. Results showed that automatic top-down fusion model had higher performance on six different data-sets than sixteen other methods.

1.2.3 Bottom-up and Top-Down Saliency Combination

In actual life situations, our senses and our cognition work together to determine if what grabbed our attention is relevant according to the situation or task we are performing. Therefore, recently there have been many saliency models that incorporate both bottom-up and top-down information for better saliency detection and prediction.

In earlier works of top-down saliency, [149] claimed that eye movements and areas observed by people in a search task show enough correlation that top-down models can be used as an efficient means to create object detection systems.

[126] proposed a model to compute perceptual quality significance map using both bottom-up features and top-down features. Bottom-up features used were color contrast, texture contrast and motion. For top-down features skin color and face were used. Integration of these features resulted in improvement of just noticeable detection and visual quality gauge.

[96] further reinforced that in nature scenes appearance based Bayesian network is a better solution for saliency modelling.

With linear regression. support vector machine and boosting, [23] inferred

saliency with the thirty different bottom-up features and three top-down features. Top accuracy was achieved by AdaBoost.

To determine saliency in news videos, a static saliency model, a motion saliency model and a top-down saliency model is used in [217]. After computing a color image matrix and suppressing background motion noise, and fusion with top-down object detection such as people, faces, flashes, vehicles etc. results in the entire saliency map.

[229] put forth a tag saliency model to predict top-down saliency. By using auto-tagging in images high level information extraction was performed and then integrated with bottom-up features to determine regions that were highly segmented.

Traffic saliency detection was explored in [44] by using a vanishing point based top-down model with combination of a more classic bottom-up model. Results were successful in producing saliency maps for complex traffic scenes.

A combination of object semantic information and pixel information improved object segmentation in [219]. An objectness map constituted with multi-scale saliency, color contrast, edge density and super pixel straddling when introduced with top-down features such as pixel saliency value, area size, object position, average saliency and regional variance gave good performance for object segmentation.

[174] used orientation feature maps for bottom-up portion of their model. The top-down module was made of edge pointers and corners as prior knowledge instead of other feature clues from images. Using three data sets, this model showed superior performance over bottom-up only or top-down only models.

DeepFeat [130], a deep feature based saliency model, exploited different pretrained deep models for saliency prediction. Using VGG-16 [176], GoogleNet [192] and ResNet [76] deep convolutional neural networks over MIT1003 [94] and VIU [108] data set showed great performance over four different metrics.

1.2.4 Visual Saliency

When using our eyes, we are either performing a search task or just free viewing. What grabs our attention based on the stimulus received by our eyes is visually salient stimuli. Visual saliency models, to predict or detect these visually salient areas or objects in the stimuli, have been around for a long time. Efforts have been made to understand and model what grabs our attention at an individual and at a global level. Out of all the other senses, our

vision contributes the most in determining what will grab the attention of the observer.

One of the earliest works in visual saliency were done on monkeys to determine what part of our brain contributes to processing attention. [139,142,164,165] suggested that extrastraite cortex to be of significance when concerning visual attention. Moving forward, primary visual cortex V1 was established to have impact on saliency using electrophysiological data [83, 136, 141]. Functional magnetic resonance imaging (fMRI) on human subjects also suggested the same results [66, 171, 212, 213].

As discussed in previous Section 1.2.1 and Section 1.2.2, different low level and high level features determine what is salient visually. Faces, text, people in general attract our attention in images [94] and motion plays an additional role when viewing videos [79].

Other than understanding visual attention, visual saliency has also been useful in computer vision for different purposes. Examples include salient object detection [91, 124], non photo realistic rendering of photographs [94], image compression [122, 181], exploration in aerial robotics [42] and retrieval of complex images [204] to name a few.

Recently the focus of saliency detection and prediction has been in developing models that utilize deep convolutional neural networks. These models will be discussed in Section 2.1.

1.2.5 Auditory Saliency

We, as humans, are capable of ignoring irrelevant sounds or categorize sounds quite effortlessly based on the surrounding and contextual clues [27]. Auditory saliency helps animals and humans into sorting audio information from a complex scenario. We are able to identify changes the loudness, pitch etc. quite easily.

Similar to visual saliency, auditory saliency can also be determined using bottom-up low level feature or high level top-down features. Bottom-up saliency in sounds would be catering the involuntary attention grabbing characteristics and top-down would be using prior knowledge to focus on specific sounds [49]. [187] explored how audio saliency can be used to filter more relevant speech in hearing aids. Similar to visual saliency, audio saliency can also be used for video segmentation purposes [43]. Audio saliency also seems to be able to predict how pleasant a sound is more than amplitude of sound alone [60].

Early works in modelling human auditory saliency computationally [49,98] were inspired by visual saliency models. General architecture of the model proposed in [49] can be seen in Figure 1.2.

In more recent works, LSTM [78] has been used to utilize auditory saliency with saliency pooling for speech intelligibility classification [65]. A two dimensional weighted model of time frequency for speech using audio saliency for noise-robust automatic speech recognition was successful in reducing relative word error rate [46]. Using Functional magnetic resonance imaging brain activity patterns and dictionary learning [227] studied the relationship between audio saliency and brain activity patterns. A model based on natural sounds divided into different categories was proposed in [197] was able to match up with human identified salient sounds. Electroencephalography has also been shown to provide extensive information about audio perception [82].

Although the number of studies on auditory saliency is not as much as visual saliency, considerable advancements are being made in this field as time passes.

1.2.6 Emotional Saliency

Just like filtering of visual and audio stimuli is done according to the scenario to drive attention to achieve the task at hand, similar filtering is also done for stimuli that trigger emotions. A stimulus that produces an emotional response and effects attention is an emotionally salient stimulus. This response can be physiological, psychological or in the form of action [29].

A number of studies have been done to determine how arousal plays a role in diverting attention and focusing attention. It has been reported that stimuli that elicit arousal can be associated with memory and attention narrowing [31,37,182]. In [7], it was further reinforced by the authors that arousal aids the memory in getting the overall information of the situation and not detailed series of information. Associations between essential memories and objects are also impacted by emotional stimuli [99, 134]. An emotional saliency map was introduced in [45] by aggregating pixel level attributes and emotion invoking color space information that performed better than visual saliency models.

Stimuli that does not elicit any negative or positive emotion is likely to be forgotten and not salient [105, 188]. Saliency has been studied in terms of how different emotions effect the motion of eyes in different situations [53]. Emotional stimuli is more likely to affect saccade behavior of observers and

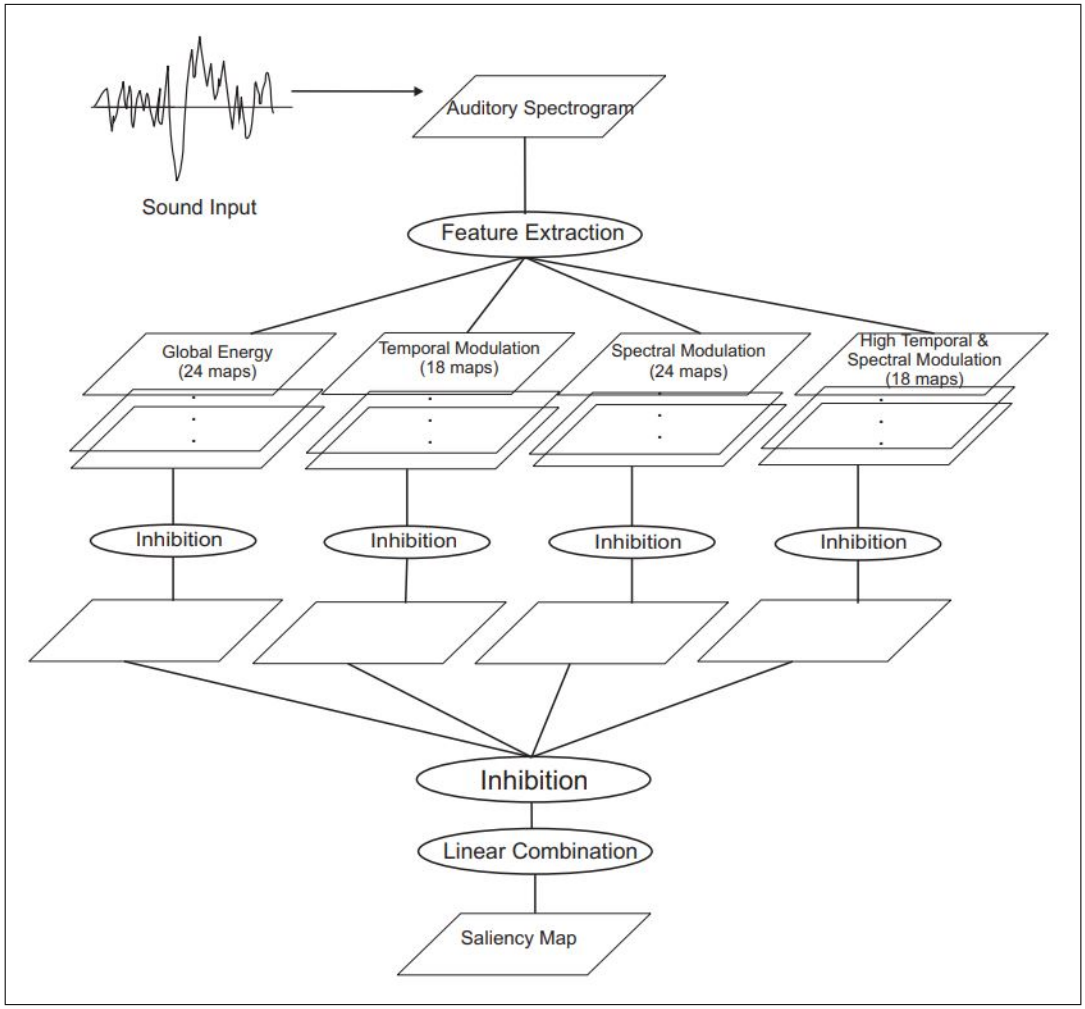


Figure 1.2: General architecture of Auditory Saliency Model proposed in [49]

create interference when trying to focus on neutral stimuli [143]. A visual saliency model that took consideration of what kind of emotions are induced by images was developed in [56] and showed effectively how emotions impact visual attention.

1.3 Types of Environments

Immersion is a feeling of being so present in an environment that the person or user looses awareness of the actual environment they are in [89]. We classify the environments for which applications, that can utilize the research on personalized saliency in this thesis, as non-immersive and immersive environments. With the advancement of technology, we can do shopping, play games, access entertainments etc. on our devices. These devices can be non-immersive in nature or immersive in nature.

Place Illusion and Plausibility Illusion were two concepts introduced in [178]. Place Illusion can be understood as having a feeling and illusion being actually in an environment even though the person or user knows that is not true. There is a feeling of being present in the environment. Plausibility Illusion. on the other hand, is having a feeling or illusion that an occurring in the environment is actually occurring even though the person or user knows that is not true.

Based on this, we consider non-immersive environments as applications on mobile phones or computers. On these devices, the user interacts with the environment with external devices such as touch screen, buttons, mouse, or keyboard. Movement of the user does not have any impact or interaction with the environment.

In non-immersive environments saliency is usually applied in terms of visual saliency. Image compression [85], image re-targeting [172], object detection [67] and website design [4, 185] are some high level applications of saliency in non-immersive environments.

Immersive environments, on the other hand, have strong place illusion and plausibility illusion factors. Virtual reality is an example of immersive environment, where the movement of the user can have impact or some kind of feedback from the environment. Currently, out of all other mediums, virtual reality is regarded to be the most immersive environment [184]. Due to this, virtual reality is heavily used in research to understand and induce emotions [205]. Virtual reality has also gaming experience more immersive than

ever before for players [178].

Detailed applications of personalized saliency in both immersive and non-immersive environments will be discussed in Section 5.1.

1.4 Research vs Practical Applications

1.4.1 Emotional Saliency

There is a lot of literature on recognizing, inducing and predicting different types of emotions. These studies use different physiological signals such as ElectroDermal Activity (EDA), Respiratory Rate (RR), or Heart Rate (HR) that are measured using specialized biofeedback sensors and gadgets [13,22,101,102,186]. The problem is that these sensors and gadgets are not something everyday consumers have readily available. Therefore, both developers and consumers cannot utilize these research studies to their full extent because the result would be an inaccessible application.

Some attempts in emotion recognition have been made using smart watches but not for prediction and induction [159, 161, 194, 211]. To exploit emotional saliency for personalization purposes in practical applications we need to use devices that are accessible to both everyday consumers and developers that will create applications.

1.5 Motivation

Even though, saliency is a well research field, in both immersive environments such as virtual reality and non immersive environments such as desktops and phones, there are limitations when it comes to applying the research to practical applications.

Currently all personalized saliency prediction models use deep learning for training and prediction. If the models are deployed on cloud, there are major network bandwidth and privacy concerns regarding data handling. if the model is deployed locally on mobile devices, the resources would not be enough.

Therefore, although saliency and personalized saliency prediction is advancing, practical applications cannot be connected to them efficiently. Since users are now more concerned with their privacy than ever, giving away personal information to ensure personalized models also becomes a problem in practical applications where users are skeptical.

Furthermore, in immersive environments where emotional saliency applications are dominant, the parameters used in such research require sensors and devices that either everyday consumer does not have or is out of reach for them. Therefore, again increasing the gap between research and practical applications. Observing such a gap between personalized saliency in both immersive and non immersive environments, this research aims to bridge the gap in both the environments and attempt to solve them.

With the above motivation in mind, in this thesis, we provide the following contributions:

1. Propose a simple machine learning algorithm that works for visual saliency using just colors spaces
2. Discussion on how colors and personality are related
3. Propose a different approach for personalized saliency prediction in immersive environment using auditory saliency and emotional saliency calculated using heart rate from the users themselves
4. Show that this research can be applied in real life environments not just laboratory setting
5. Discussion on how environment and gender effects emotional saliency.

1.5.1 Research Questions

Following are the questions we are aiming to answer in this thesis:

1. Are color spaces enough to compute personalized saliency?
2. Does individual gaze behavior effect personalized saliency?
3. Does image content effect personalized saliency?
4. Is heart rate from smart watch enough for emotion induction?
5. How does replay effect personalized saliency?
6. How does environment impact heart rate and personalized saliency?
7. Does heart rate and self reported results correlate with each other?
8. How does gender impact personalized saliency?

1.6 Thesis Organization

This dissertation is organized in the following structure. Chapter 1 introduces the topic of saliency and highlights the limitations of current research regarding personalized saliency and the motivation of this dissertation. discussion on different types of research is also included in this chapter. Chapter 2 establishes some background concepts and literature review for this research. Chapter 3

focuses on personalized saliency in non immersive environments and how visual saliency can be calculated using just machine learning and color spaces. Chapter 4 discusses personalized saliency in immersive environments such as virtual reality. Chapter 5 opens up some discussion regarding personalized saliency in non immersive and immersive environments and what considerations should be taken in both and provides some practical applications of this research. Chapter 5 also concludes this thesis.

Chapter 2

Related Works

2.1 Visual Saliency

In visual saliency, [220] used the terms of universal saliency and personalized saliency to differentiate between different types of saliency. Universal saliency maps are created using salient regions of an image. These salient regions are where people, regardless of their age, gender, culture, or personality. Personalized saliency maps, on the other hand, are created individually for each person based on the salient regions that attracted their attention. These regions may be influenced by the personal characteristics of the individual such as their age, gender, personality, or culture. This is demonstrated in Figure 2.1.

2.1.1 Universal Saliency in Non-Immersive Environments

In visual saliency, detection and prediction of universal saliency is very well established and is an active area of research. Earlier saliency models usually took bottom-up approach.

Universal saliency models that utilize bottom-up features, top-down features, or a combination of both. These models have been discussed in Section 1.2.1, 1.2.2 and 1.2.3.

Currently most of the universal saliency models use different variations of deep convolutional neural networks. According to the MIT/Tuebingen Saliency Benchmark [113], the top performing models on MIT300 [93] use some variation of convolutional neural networks. UNISAL [48] takes an interesting approach of modelling saliency of images and videos together with encoder- recurrent neural network-decoder design. Using this design, UNISAL is currently at the second place on MIT/Tuebingen Saliency Benchmark. CAS-

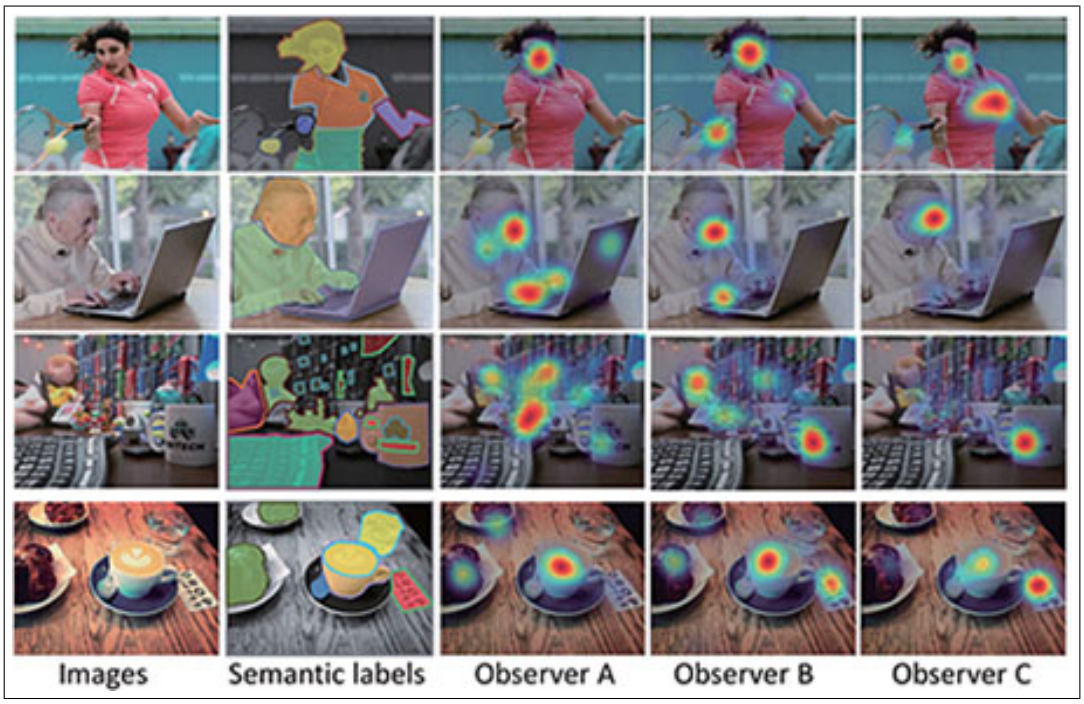


Figure 2.1: Different people focus on different things with different intensity. Image from [220] © 2018 IEEE

NET [57] studied the influence of emotion on image saliency and also created EMOtional attention dataset (EMOd) for this purpose. Their results showed intense but short lived bias towards objects that incited emotions.

The top performing models on CAT2000 [24] have variations of different types of models that take up the top of the benchmark. DeepGaze II [114] tops the benchmark for CAT2000 database and is also included in the top ten models for MIT2000 data set. Ensembles of Deep Networks (eDN) [203] used hierarchical feature learning and did not specify any low level or high level features by hand. The model trained itself on gaze patterns of the subjects and resulted in a generalized model. [54] used region covariance for saliency prediction. Subspace selection based on the eigenvalues from principal component analysis done on image patches is also another technique that performed well for fixation prediction [58].

2.1.2 Universal Saliency in Immersive Environments

As with universal saliency in non immersive environments, progress is also being made in the area of universal saliency for immersive environments. In the past decade, a few models have been proposed to understand and predict

how people explore virtual reality, 3D images, videos, and dynamic videos as well.

[177] provided important insights and understand of how people explore virtual environments. Their study showed that saliency maps were similar for people under three different viewing conditions:

- Exploring virtual environment standing with head mounted headset
- Exploring virtual environment sitting with head mounted headset
- Exploring virtual environment on desktop with computer mouse

An equator bias was also observed in [177] under all conditions. Inter-user variance was realized as a potential problem in saliency prediction for longer time periods in virtual environments.

[10] reported on how depth effected the changes in viewing behavior of people in 2D and 3d conditions. In high depth of field, fixations were recorded farther in 3D, where as in 2D it was they were reported closer. in low depth of field, the fixation behavior was similar in both 3D and 2D. However, the authors concluded with discussion that depth, while important, cannot determine saliency on its own. The importance of depth cue was also validated in [9, 16].

A learning-based visual saliency prediction model (LBVS-3D) introduced in [19] for 3D videos, used object segmentation and spatial information for feature extraction. Using random forest model for learning and saliency map generation gave this model high performance results.

SaLGAN360 [34], proposed as an extension of SalGAN (for 2D images) [153], a deep convolution neural Newark model produced local and global maps for 360 images and achieved higher performance than models live SalNet360 [138] and GBVS360 [117].

A few graph based models have also been put forth to predict saliency in 360 degree videos in [8, 231].

2.1.3 Personalized Saliency

A few personalized saliency prediction have been proposed in the last five years. The concept of personalized saliency in images was introduced by [220]. This study also created the first data set catered towards personalized saliency. They used multi-task convolutional neural networks for personalized saliency prediction and achieved 0.8588 AUC Judd [32] score when SALNet [154] was

used with convolutional neural network in combination with Person-specific Information Encoded Filters (CNN-PIEF).

Age has also been known to create heterogeneity in fixations of individuals [111]. [223] used a modified conditional generative adversarial network (GAN) to compute personalized saliency based on age. They achieved 0.74 AUC Judd score for seniors and 0.76 AUC Judd score for juniors. [123] introduced personalized attention Network (PANet) which was trained based on preferences of only two individuals and pseudo ground truths. The pseudo ground truths that were generated based on the ground truths of the two individuals.

Two few-shot Personalized saliency prediction were also proposed in [140] and [128]. In [140] collaborative Gaussian process regression (CoMOGP) was utilized for personalized saliency prediction. CoMOGP exploits the gaze similarity between a target individual and other individuals as weights, and then image similarity features as input. [128] aims to make model adaption on a new subject easier by utilizing meta-learning based model. Meta-learning can be done on any universal saliency model that uses gradient descent.

The models mentioned above were designed for images and meant for non-immersive environments. To the best of my knowledge, no personalized saliency models have been proposed for immersive environments such as virtual reality.

2.1.4 Limitations

Personalized visual saliency prediction, currently, is following the same trend of universal visual saliency and using different variations of deep neural networks such as multi-task convolutions neural networks or generative adversarial network. Theoretically, there is no problem with using deep learning for saliency detection and prediction but practically, resource utilization needs to be taken into consideration.

Considering personalized visual Saliency, most applications would be in mobile environments and when considering mobile environments the biggest issue is making sure that the limited resources available are being utilized efficiently. Network, storage, and memory limitations should all be taken into consideration when designing applications for mobile devices.

Deep neural networks training, especially for modern models, require powerful CPUs and GPUs which are way above the level that are used in mobile devices and even some computers. Therefore, usually pre-trained models are

used to infer results when utilizing deep learning in mobile devices. There are two inference approaches that can be taken when performing deep learning on mobile devices, on-device and on-cloud.

On-device inference uses pre-trained models to infer results for new data. Frameworks that provide deep learning for mobile devices such as Caffe2 [3] and TensorFlow Lite [1] provide such service by using pre-trained exported models. On-cloud inference sends the requests for required results to the cloud servers which host models. The main problem that arises here is the time required to send and receive these requests over network and the speed of the network.

Overall the following issues need to be taken into consideration when deep convolutions neural networks are being used on mobile devices [148]:

- **Network** On-Cloud inference require high speed networks for efficient and quick results. On-mobile inference, however can only be used in a practical manner on higher end and newer mobile devices.

- **Storage** Models can be compressed to reduce storage requirement but this can have impact on speed and performance.

- **Memory** Deep convolutional neural networks require much more and frequent garbage collection. They also take up extensive amounts of memory resources for model loading and computation.

- **Time** On-device inference requires more time to load the model in memory and this time increase as the model complexity increases. On-cloud inferences requires more time to send image data to the cloud. Image re-scaling time also needs to be considered.

- **Privacy** On-cloud frameworks needs to take privacy laws of different countries into considerations when sending data to servers in different countries.

Due to these factors, we propose using machine learning instead of deep learning for personalized saliency prediction. We also propose having an individual model for each subject rather than having a common model for all. This, way we reduce the data size and the resulting model would be personalized for that subject and that subject only, it will not be influenced by the preferences and behaviors of other subject. The details of this model are discussed in Chapter 3.

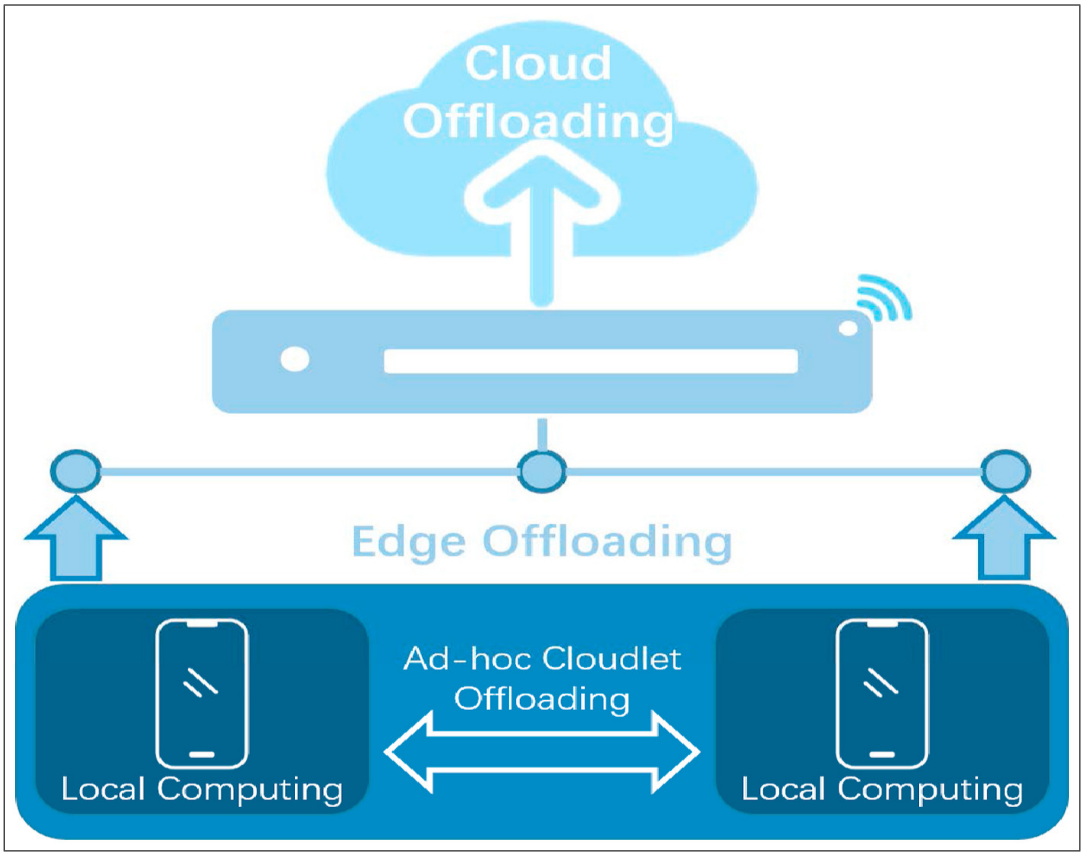


Figure 2.2: Mobile Deep Learning Applications depicting offloading computation to the cloud and device to device communication within closer distance for light computation offloading to cloud-lets [210].

2.2 Emotional Saliency

Emotions have a substantial influence on everything we do. Even when free viewing images, emotions guides our attention [57]. How emotions effect our attention has been studied in psychology [40,163]. Stroop test [189] introduced by John Ridley Stroop, is done to test the attention and information processing capabilities of an individual. In literature, stroop test how emotions delays the attention of individual when naming something even if there is no relevance to the task at hand [163, 215]. Depending on how emotionally salient the content of a post on social, also correlates with what kind of reactions it will receive from the audience [50].

As with visual saliency, emotional saliency also depends on the race, gender, age and life experience of the person. [52] found that when media is more personally relevant to the participant, there is an increase in positive affect but no increase or decrease was noticed in negative affect. Differences between young and old adults were studied in [59]. Their results showed that older adults had high subjective arousal and young adults had higher physiological arousal when experiencing negative emotions and tenderness. Fear and amusement was also experienced more intensely by young adults.

2.2.1 Emotion Induction

Emotion induction is one of the most important factor that contributes to affective computing in human-computer interaction research. The term affective computing was coined by Rosalind Picard in 1995 in [158] as computing that is influenced by emotions of the user and influences emotions of the user. Affective computing takes the considerations of human emotions and tries to make the system or application cater to individualistic behaviour of the user [195]. This can be done through input from sensors, camera or microphone.

With considerations, different emotion induction techniques discussed in [55] are mentioned below:

- **Imagination** Participants imagine scenarios, situations or memories that will elicit the desired emotion. These imaginations can be based on reality based or completely hypothetical. A potential draw of this technique is that participants need to be isolated, so that they can concentrate but that contradicts most real-life situations and emotional experiences [39].

- **Film** Films are used extensively in emotion induction as they provide both

audio and visual stimuli while simulating real world experiences. Care should be taken while preparing the environment in which the film will be watched to encourage immersion. Some people might react strongly to the films while others make not give much of a reaction. These reactions depend on the sensitivity of the participants.

- **Sound** Using sounds that are usually heard in daily life such as animal, human, objects, transportation etc. are used in studies to evoke different emotions. These studies are preformed with sounds playing on either speakers or the participant is required to wear headphones.

- **Music** Music is a another emotion induction technique where personal relevance to the participant results in stronger emotion. Due to the complex emotions experienced while listening to music, some studies regard music elicited emotion different from real emotions [127, 155].

- **Images** Images of wide variety are used for emotion induction such as families, nature images, animals playing, funerals. art etc. These images are usually shown on a screen and the participant is questioned on the emotions that they feel using self assessment tests.

- **Reading Passages** In this type of induction, participants read different types of stories or passages and then recall how it made them feel after they are finished reading.

- **Writing Passages** There are few different ways to use writing for emotion induction. Participants may be required to write a past experience that elicited different emotions such as anger or joy. The participants may also be asked to rank these experiences.

- **Embodiment** By using the muscles of the face, tone of their voice or postures, participants are asked to embody different emotions. The researcher may instruct the participants to contract or relax certain muscles. This method is helpful in evoking emotions without any stimuli to trigger the participant.

- **Virtual Reality** Using virtual reality, participants can experience situations artificially and feel real emotions from this. Virtual reality is one the most immersive medium and is used extensively in emotion related research. It has also been used to treat different phobias with clinically significant results [104].

- **Feedback** After performing a task, participants in this technique receives positive or negative feedback. The feedback might be true or made up by the experimenters to elicit different type of emotion. This type of study may effect the self-esteem of the participants and this ethical consideration should be kept in mind [73].

- **Self-Referent-Statement** In this emotion induction technique developed by Velten [200], participants are to evaluate themselves positively or negatively and asked to feel those statements. However, this method causes higher and longer depressive feeling. Music was reported as superior emotion induction method in [11].

- **Social Interaction** Extensive social situations are created in laboratories using to create high level of realism. This method is useful in inducing emotions that are otherwise difficult to induce such as anger. Personal contact was reported as a significant factor for inducing anger in [125].

- **Physiological Manipulations** This techniques requires the use of pharmaceutical drugs or psychedelic drugs for inducing or manipulating emotions. However, this method requires high levels of expertise and ethical precautions to prevent any unfortunate situation.

- **Motivated Performance Tasks** To study the impact of environment in social psychology, motivated performance tasks are useful. In this technique, participants are required to perform a task in front of other people. However, these tests are not easy to administer and arrange.

- **Combined Techniques** Different methods when used together, create combined techniques. Usually, music is paired with other techniques such as images or guided imagery tasks to achieve better results.

Emotion induction methods can be tricky because one emotion might cause another emotion to flare up. This can create discrepancy in results. Music has high success rate in inducing emotions but other techniques in combination also produced good results. However, what method should be used depends on the nature of experiment, number of participants and tools available. Considerations should also be taken regarding the differences among self-reported results, physiological and behavioral data.

2.2.2 Heart Rate for Emotion Induction

Heart rate and heart rate variability have been used extensively in research emotion recognition. Heart Rate corresponds to the number of times a heart beats per minute. Heart rate variability [173], is the fluctuations of heart rate between consecutive beats. Heart rate variability is known to give appropriate indications of emotional changes in the presence of strong stimuli [36]. Low heart rate variability corresponds to a state of calmness because there is less variations between heart beats. High heart rate variability corresponds to a state of short term stress or excitement due to high variations between heart rate [170, 218]. Heart rate asymmetry, which corresponds to the acceleration and deceleration of heart rate variability, is also effected in the same way when experiencing calm or stressful situations [95].

Heart rate activity also seems to be longer when negative emotions are experienced. The effects of positive emotions on heart rate activity seems to be short lived. These findings were reported in [30]. Heart rate variability measures are also effective in not only recognition but also prediction of emotions based on valence and arousal values [38, 144].

Heart rate is an important measure in affective computing. Efforts have also been made to utilize heart to create affective games. [15] aimed to increase the heart rate of players by guiding players according to the increase in their heart rate. [191] utilized fixation and saccade of the players during horror game play. Heart rate and pupil changes were depicted as a good measure for immersion levels of the player. An increase in immersion and longer play time resulted in decrease of fixation frequency and length. However, the claims of these two studies were not backed by results.

A biofeedback system to amplify fear included two pseudo heart rates [198]. One adapted to the real time heart rate of the subject and the other increase in a step wise manner. Emotion induction performed in [206] using virtual reality showed an increase in arousal and a decrease in valence when playing horror game. A decrease in arousal and increase in valence was noticed when a virtual nature scene was experienced. These findings were backed by self reported results and heart rate variability.

[196] studied the induction of anxiety using augmented reality and virtual reality with heart rate variability and self-reported questionnaires for analysis. The results showed that heart rate variability indicated anxiety in both environments. Height related anxiety was research in [103] with hear rate as physiological measure. The results showed that presence was adequate enough

in virtual reality to trigger anxiety and produce heart rate changes. Social situation related emotions can also be studied using self reported questionnaires, heart rate and virtual reality [74]. Distress and habituation regarding public speaking fear was also reported to increase heart rate in [193].

2.2.3 Limitations

Although emotion has been studied for so long using various techniques, it is hard to apply this knowledge in practical applications such as creating virtual reality therapy applications, training applications, games etc. There are two reasons for that:

- **Laboratory Setting** current emotion related studies are done in a controlled environment which might induce different emotions if the same situation happened in real world environment. Even with the use of virtual reality, the experimenters are present to make sure the apparatus is correctly calibrated and fitted properly on the head of the participant. Therefore, user related errors that might occur in practical applications and effect the results may be ignored in these studies.

- **Inaccessible Gadgets** Although using heart rate is prevalent in emotion research along with other physiological signals such as sweat using electrodermal activity, brain activity using electroencephalogram, heart rate and heart using sensors or monitors stuck to the chest or wrists. However, these devices are either not available to the general consumer or people would have to go out of their way to buy these specialized gadgets. Therefore, these measures are not considered by developers when trying to create an accessible application due to their inaccessibility to the general public.

To address these limitations, we perform an experiment to induce emotion in real life where people participated from their own homes, with the devices they already owned. They were not required to buy any new device to participate in this experiment. We also test the use of smartwatches as effective way to utilize heart rate in this experiment. This experiment is discussed in detail in Chapter 4.

2.3 Personalization

2.3.1 Colors and Personality

Colors are expressed by our eyes when stimulation occurs on the cone cells at different intensities by electromagnetic radiations in the visible spectrum. Depending on the absorption and reflection levels of rays on an object, different colors are interpreted by our eyes. Numerically colors spaces are used for representation of colors. In our eyes, there are three types of cone cells. These three types of cone cells are used during interpretation of red, green and blue light bands. Similar to these cone cells, the RGB color space that is made up of red, green and blue color values. Other colors are represented using other color spaces. Cyan, Yellow, Magenta and Key (Black) colors are represented in the CYMK color space.

Introversion and extroversion are personality types that can be indicated using preference of different types of environment. Introverts tend to have a preference for calmer environments whereas extroverts prefer to surround themselves in exciting environments [135]. What defines a calmer environment and what defines an exciting environment depends on the person and where they are on the spectrum of introversion or extroversion. Colors are know to play an important role in creating the atmosphere and emotion of the environment [151, 222]. There are differences among individuals for their preference of hue and saturation of colors [179].

When we individuals into categories based on factors like personality types, gender, culture or race a common traits or behaviors might be noticed [118]. [151] reported that difference on how colors had emotions associated with them between British and Chinese participants. Significant difference was not observed when participants were grouped based on gender. Follow-up of this research in [152], showed that female participants had higher accuracy regarding color-combination emotion than male participants.

All these factors contribute towards the idea that colors and their attribute preferences are a individualistic characteristic. As discussed before, traits like personality usually exist on a spectrum rather than being absolute and extreme. Therefore, a personalized saliency model will be better and superior when applied in personalized services for practical purposes. With this discussion, we can hypothesize that if we graph the RGB and CYMK values of fixations of different people, we would see different patterns. This hypothesis is tested in 3.

2.3.2 Gender and Emotions

The relationship between gender and emotions has been researched for a long time. In 1988, [11] reported that gender played an important role in emotion induction. When music was used to induce emotions, women had a higher increase in depressive and anxious mood than men. Women were also shown to be more susceptible to mood induction using music. The best predictor for an increase in anxious mood increase was also indicated to be gender.

Ten years later, using films to induce emotion, [109] studied the relationship between gender and emotion expression. The results indicated that women tend to be more emotionally expressive, even if the same emotion is felt by both genders with the same intensity. Reporting of intense emotional experience by women was also found in [70].The skin conductance reactivity of the participants also showed that men had more reaction to films that incite fear than women. Overall, men tend to be internalizers and low responders, whereas women tend to be externalizers and high responders according to this research. In another experiment, it was noted that women were more accurate than men with the experience of fear and disgust in reports.

Female preference in "soft", "relaxed", "light", or "feminine" color pairs was observed in [152]. Consistency was also observed in genders in color emotions except in the case of masculine-feminine and like-dislike. In a study involving thirty seven countries, women were known to have more experience with expressing powerless emotions and men were known to have more experience with powerful emotions [61]. Men were also reported to be more antagonist than women in this study. When using films to induce emotions, women tend to experience higher level of sadness than men [87]. Affective reactivity was also self-reported more by women than for post positive and negative emotions [72].

2.3.3 Environment and Emotions

As mention in Section 2.3.1, people have preferences for different types of environment based on their personality. Factors such as indoor or outdoor settings, sounds, and lights all contribute towards making an environment favorable or unfavorable.

Different types of environments elicit different emotional reactions. Ulrich in [199] proposed a stress reduction theory that suggests that exposure to nature can reduce stress. Attention restoration theory from [97] suggests that

nature can help in replenishment of human attention. Both theories suggest that these effects are unconscious. [81] reported, in a virtual reality experiment, that nature environment was more effective in restoration than concrete environments without vegetation. For people living alone in confined areas, virtual reality exposure to nature can improve mood and reduce stress [14]. Stress reduction after viewing nature was also confirmed in [207].

Immersion is also impacted by environment, when using virtual reality to induce emotions. Different sounds and lights can influence how a person feels in an environment [207]. Open or closed spaces, especially for people with claustrophobia, can elicit different types of emotions. Sudden noises or even silence can contribute towards making an environment stressful [132,175]. Furthermore, unprecedented events happening in a familiar environment can make a person uncomfortable and trigger unsafe feelings [115].

Chapter 3

Personalized Saliency in Non-Immersive Environments

3.1 Machine Learning in Non-Immersive Environments

As discussed in 5.2, deep learning in non immersive environments such as smart phones is not feasible in terms of saliency due to resource and time complexities. Therefore, in this study we will be focusing on using machine learning algorithms for the purpose of computing personalized saliency in non-immersive environments.

After testing out different regression models, we decided to use gradient boosted tree regression. The results of these initial testing have been included in the supplementary section.

Gradient boosting in machine learning [62] has been utilized for both classification and regression purposes. It works by building multiple models and optimizing a differential loss function. Gradient boosting builds up a single more accurate model with the help of multiple weaker models. This makes the model become more robust.

3.2 Personalized Saliency Data set

As we are moving towards a society whose demands for personalized services and experiences are increasing day by day, there are also data sets beings developed to keep the research in this area active and constantly thriving. For this purpose, to build different saliency models eye tracking data sets are

created.

A saliency data set is mainly composed of images and eye fixation data of individuals who viewed those images. Tracking devices aid in obtaining the exact x and y coordinates where the observer looks in the image. Due to our viewing behaviors, every individual has multiple fixations recorded on a single image. This process is repeated for every image in the data set with can be in hundreds or thousands for all the participants in the study. Data sets can be compiled based on categories such as single object, humans, faces, crowds, multiple objects, animals, user interfaces etc.

In the presence of highly salient features in an image such as faces, text, signs etc. human have a tendency of fixating on them immediately [94]. However, the nature of fixations become more spread out when there are multiple salient areas cluttered in an image. In traditional universal saliency data sets every individual looks at an image only once, which does properly give us the information if the fixation points were due to random search or that area definitely caught the attention of the observer.

To solve this problem, the data set introduced in [220] was made with every individual looking at every image at least four times in different sessions. This was done so a more definite ground truth can be obtained. Their results showed that saliency maps produced with traditional saliency maps where viewing was done only once, performed poorly in comparison to when viewing was done multiple times.

With the help of different available universal saliency data sets and some of their own original images, [220] created the first and only data set, to the best of our knowledge, that was collected and designed specifically for personalized saliency map computation. The data set was made up of 1600 images. 1100 images were chosen from other universal saliency data sets, 375 images were obtained form the internet and the rest were original images.

3.3 Proposed Method

As we hypothesized in section 2.3.1 that different individuals would show different color patterns when the colors values of their fixations are graphed, we test this hypothesis by plotting the fixations of different individuals from the personalized saliency data set by [220]. Using Matlab, RGB values of the image pixels for fixations can be obtained. The corresponding CYMK values can be derived from RGB values.

Figure 3.1: Original 000004 image from [220].

To calculate CYMK values, the RGB values have to be normalized from 0.255 to 0..1. To achieve this the original RGB values are divided by 255 to get R', G', and B' [147].

$$R' = R/255$$

$$G' = G/255$$

$$B' = B/255$$

Using R', G', and B' we can calculate the black color called Key (K)

$$K = 1 - max(R', G', B')$$

Cyan(C), Yellow(Y) and Magenta(M) are subsequently calculated using R', G', B' and K

$$C = (1 - R' - K)/(1 - K)$$

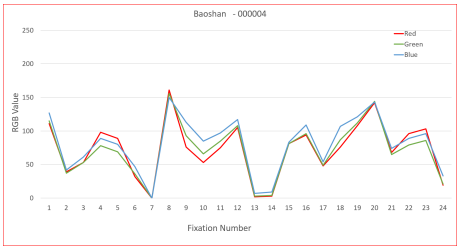
$$Y = (1 - G' - K)/(1 - K)$$

$$M = (1 - B' - K)/(1 - K)$$

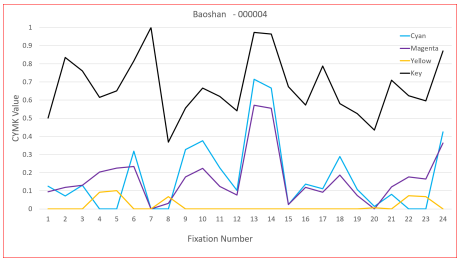
We can see different patterns of RBG and CYMK values after they were plotted and inspected. This further proves that different individuals process and explore the same image differently.

After graphing the fixations of all 30 subjects in the data set we could conclude that our hypothesis was correct and different patterns can be seen when different subjects view the same image.

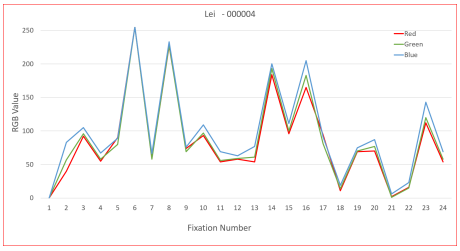
Figures 3.2, 3.3, 3.5, 3.6, 3.8, 3.9 depict the RGB and CYMK patterns of



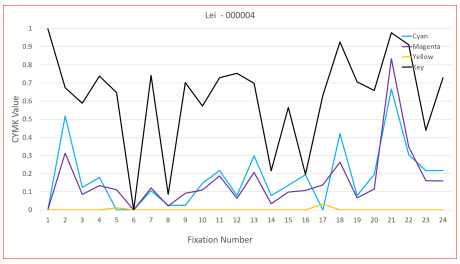
(a) Baoshan - RGB



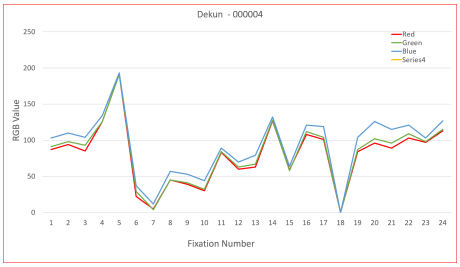
(b) Baoshan - CYMK



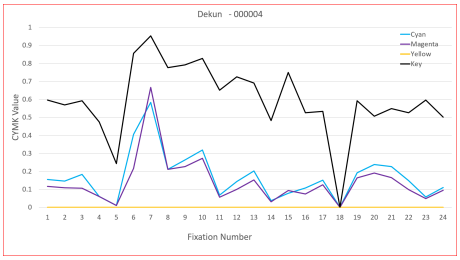
(c) Lei - RGB



(d) Lei - CYMK



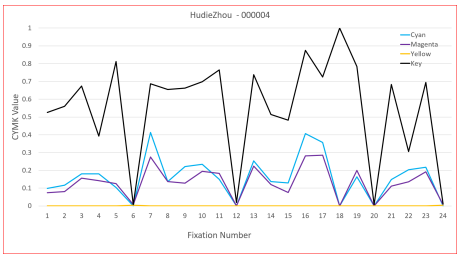
(e) Dekun - RGB



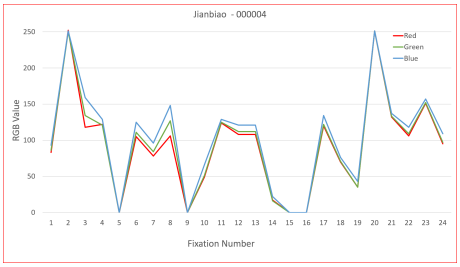
(f) Dekun - CYMK

Figure 3.2: RGB and CYMK patterns for fixations of different subjects for image 000004 from [220]. Graphs from [224].

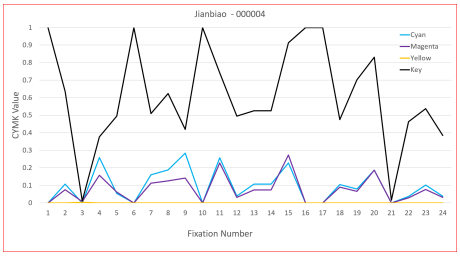
(a) HudieZhou - RGB



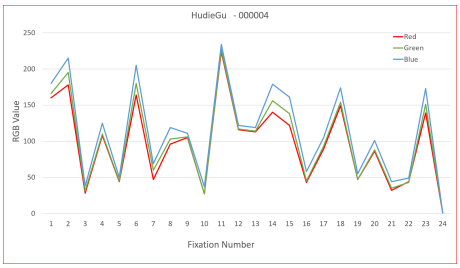
(b) HudieZhou - CYMK



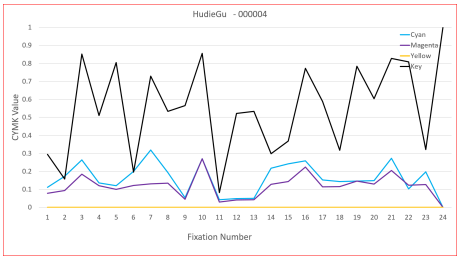
(c) Jianbiao - RGB



(d) Jianbiao - CYMK



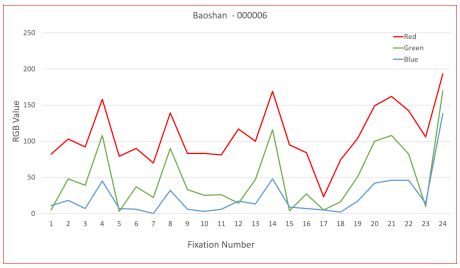
(e) HudieGu - RGB



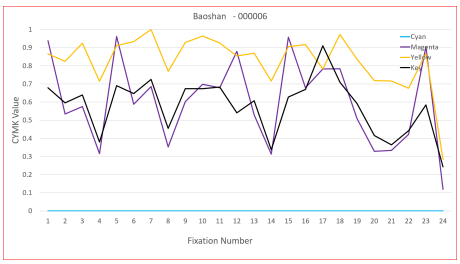
(f) HudieGu - CYMK

Figure 3.3: RGB and CYMK patterns for fixations of different subjects for image 000004 from [220]. Graphs from [224].

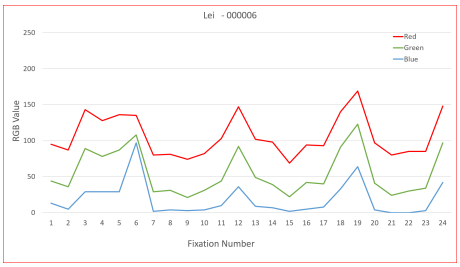
Figure 3.4: Original 000004 image from [220].



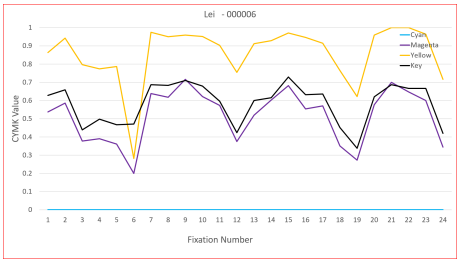
(a) Baoshan - RGB



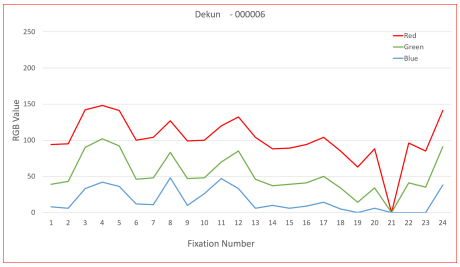
(b) Baoshan - CYMK



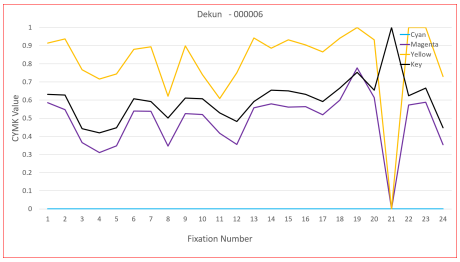
(c) Lei - RGB



(d) Lei - CYMK

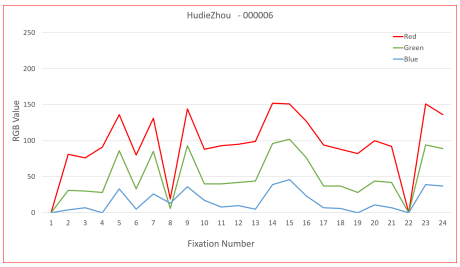


(e) Dekun - RGB

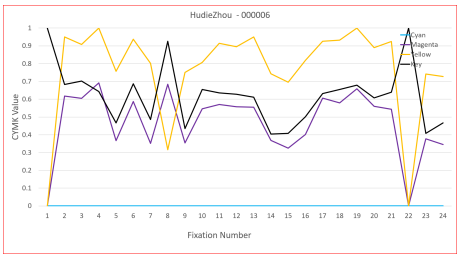


(f) Dekun - CYMK

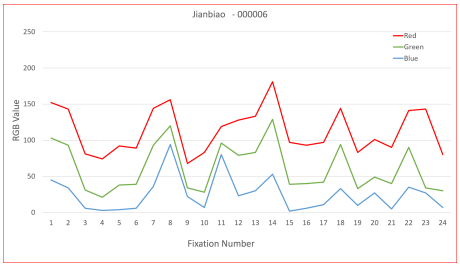
Figure 3.5: RGB and CYMK patterns for fixations of different subjects for image 000006 from [220]. Graphs from [224].



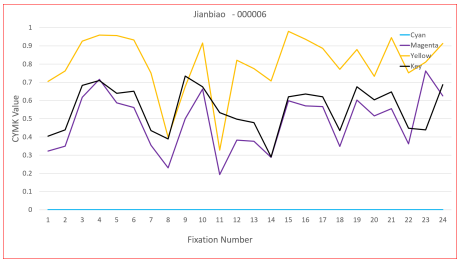
(a) HudieZhou - RGB



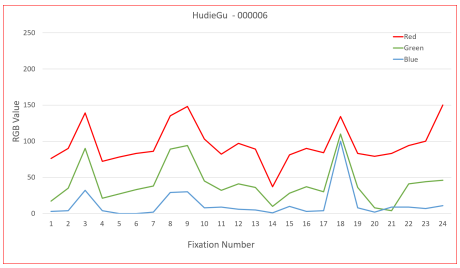
(b) HudieZhou - CYMK



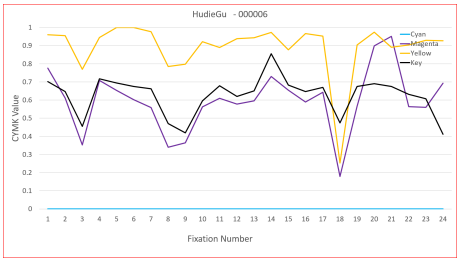
(c) Jianbiao - RGB



(d) Jianbiao - CYMK



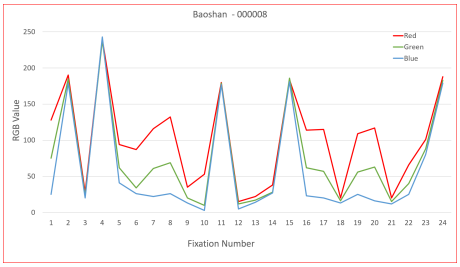
(e) HudieGu - RGB



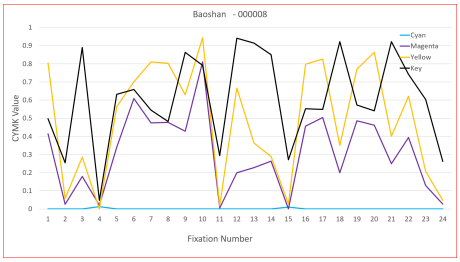
(f) HudieGu - CYMK

Figure 3.6: RGB and CYMK patterns for fixations of different subjects for image 000006 from [220]. Graphs from [224].

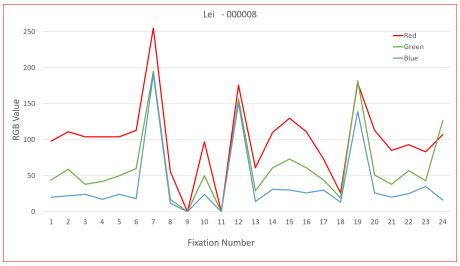
Figure 3.7: Original 000004 image from [220].



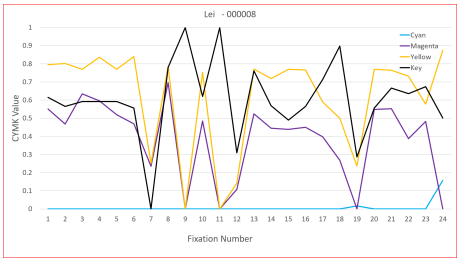
(a) Baoshan - RGB



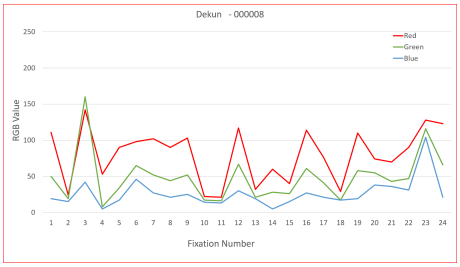
(b) Baoshan - CYMK



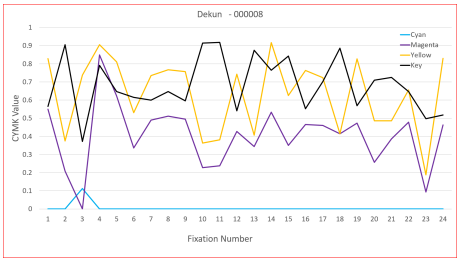
(c) Lei - RGB



(d) Lei - CYMK



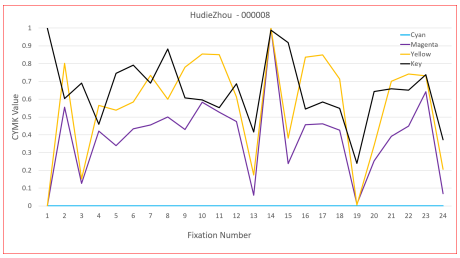
(e) Dekun - RGB



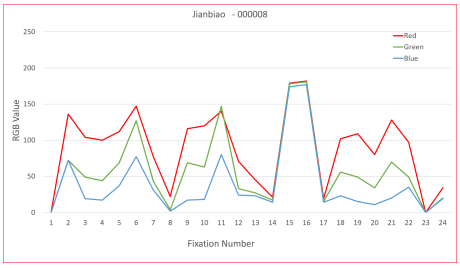
(f) Dekun - CYMK

Figure 3.8: RGB and CYMK patterns for fixations of different subjects for image 000008 from [220], Graphs from [224].

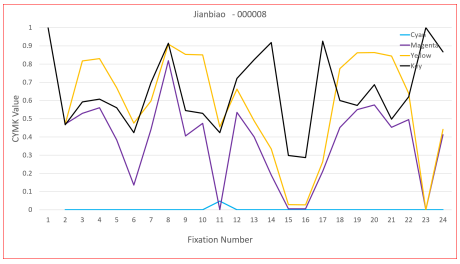
(a) HudieZhou - RGB



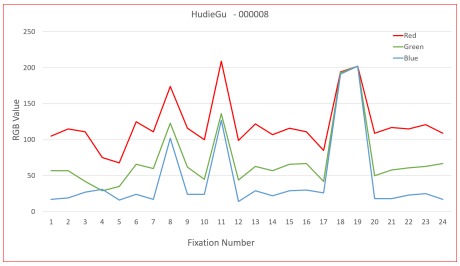
(b) HudieZhou - CYMK



(c) Jianbiao - RGB



(d) Jianbiao - CYMK



(e) HudieGu - RGB

(f) HudieGu - CYMK

Figure 3.9: RGB and CYMK patterns for fixations of different subjects for image 000008 from [220]. Graphs from [224].

six different individuals when observing the three different images.

Upon further analyses of these graphs, we observe that for image 000004 (Figure 3.1) and image 000006 (Figure 3.4) the overall RGB and CYMK graphs (Figure 3.2, Figure 3.3, Figure 3.5, Figure 3.6) have some similarities in peaks but ultimately are quite different from each other. However, for image 000008 (Figure 3.7) all the subjects show a bit of similarity in terms of the difference between the RBG and CYMK line graphs (Figure 3.8 and Figure 3.9) and the values being in the similar range.

With the confirmation that different individuals show different color patterns in their fixations, we propose using these color patterns to create personalized saliency maps. Using regression and without any invasive personal information, we can model and predict these color patterns to create a simple personalized saliency prediction model.

In [180], it was observed that different individuals have varying preferences for hues and saturation. Therefore, these low level features can also be added for our personalized saliency computation, i.e. Hue, Saturation, Value (HSV) and Hue, Saturation, Lightness (HSL).

Using R', G', and B' from before and [147]

$$C_{max} = max(R', G', B')$$

$$C_{min} = min(R', G', B')$$

$$\Delta = C_{max} - C_{min}$$

$$H = \begin{cases} 0 & \text{if } \Delta = 0 \\ 60 * (\frac{(G'-B')}{\Delta}) mod 6 & \text{if } C_{max} = R' \\ 60 * (\frac{(B'-R')}{\Delta}) + 2 & \text{if } C_{max} = G' \\ 60 * (\frac{(R'-G')}{\Delta}) + 4 & \text{if } C_{max} = B' \end{cases}$$

$$S = \begin{cases} 0 & \text{if } C_{max} = 0 \\ \frac{\Delta}{C_{max}} & \text{if } C_{max} \neq 0 \end{cases}$$

$$V = C_{max}$$

$$L = \frac{C_{max} + C_{min}}{2}$$

With the aid of universal saliency maps we can the areas that are universally saliency among individuals. Therefore, we can extract the personalized saliency map from it rather than creating it entirely from scratch on its own. A personalized saliency map is an aggregation of universal saliency map and the difference between universal saliency map and personalized saliency map [220].

$$PSM(S_n, I_i) = USM(I_i) + \Delta(S_n, I_i)$$

where $PSM(P_n, I_i)$ represents the personalized saliency map for nth Subject S corresponding to ith Image I, $USM(I_i)$ represents the universal saliency map corresponding to ith Image I and $\Delta(S_n, I_i)$ represents the difference between $PSM(S_n, I_i)$ and $USM(I_i)$.

After the computation of universal saliency maps of the images in personalized saliency data set. we can create a gradient boosted tree regression model for personalized saliency map computation. Although any universal saliency model can be used for universal saliency map generation, we chose Deep Gaze II [114] which has the highest performance on the saliency benchmark [113] for CAT2000 data set [24]. Deep gaze II is also trained on high level features such as object detection, face recognition etc. which can strengthen our otherwise low level feature based model. We have seen in Section 1.2.3 how combination models are useful in saliency prediction.

From the universal saliency density maps generated using Deep Gaze II, the pixels needed to form the personalized saliency map that fit our regression model can be extracted. The personalized saliency maps are then evaluated using AUC Judd and NSS metrics.

An overall summary of the process can be seen in Figure 3.10

3.4 Experiment Design

3.4.1 Machine Setup

The machine used for this research was Dell Latitude 7280 with Windows 10 operating system. The machine had 16 GB memory and 2.60GHz 7.70GHz Intel i5-7300U CPU. For data preparation and result evaluation MATLAB was used. Python was used for producing universal saliency maps and Knime Analytics Software was used for building the gradient boosted tree regression model.

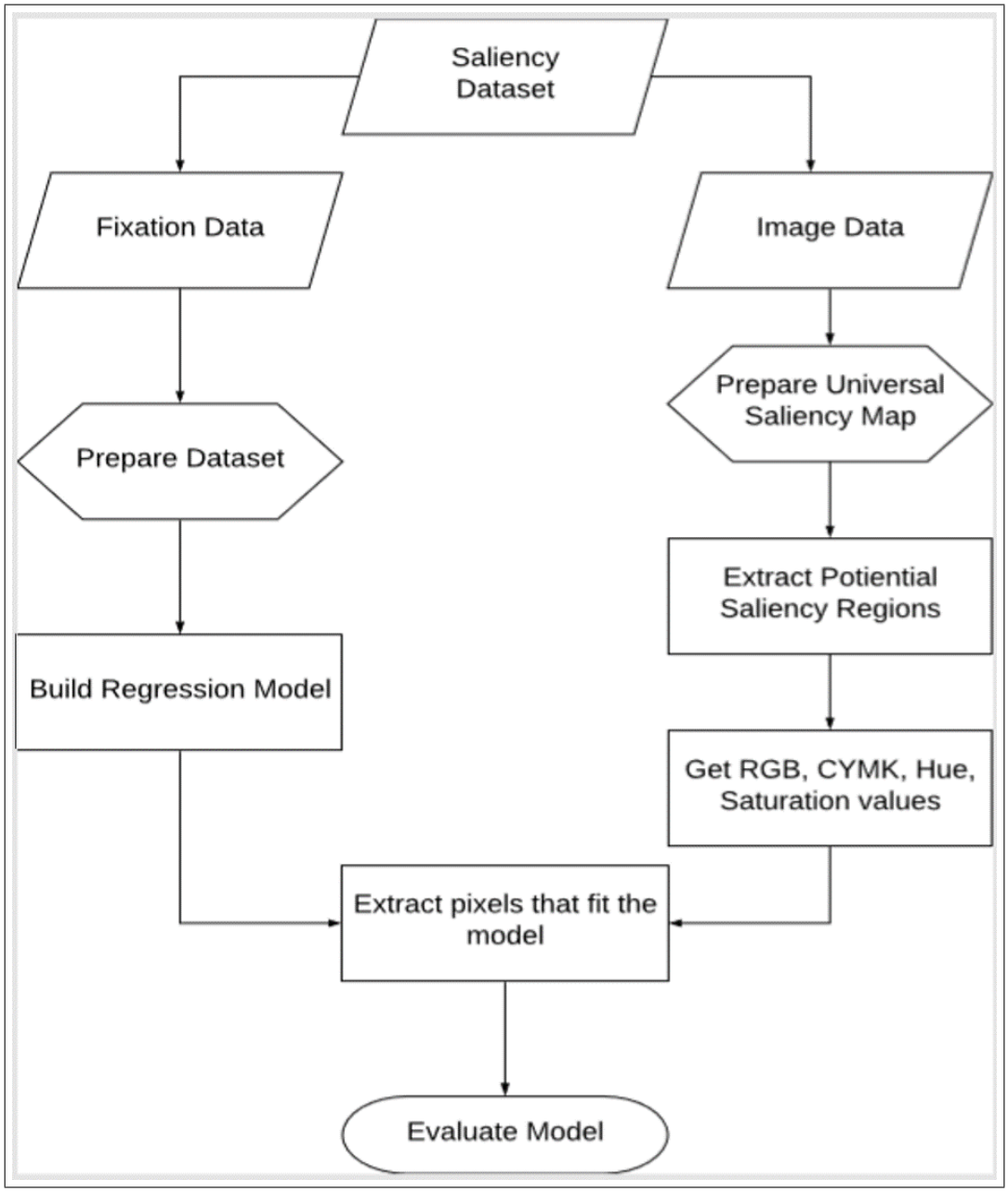


Figure 3.10: Summary of proposed model [224].

3.4.2 Universal Saliency Map

We used Deep Gaze II [114] for universal saliency map generation. Deep Gaze II is a probabilistic model which improves accuracy based on log-likelihood. Deep Gaze II is a modified VGG-19 network [176] that is pre-trained on SALICON data set [112]. To reduce computation time, all fully-connected layers are removed, unit variance feature maps are obtained after re-scaling the filter for layer consistency.

After the image is processed in the VGG-19 modified network, the feature maps then serve as an input for another neural network, called the readout network, which gives the point wise non linearity in the VGG features. The center bias is modelled after the output of the readout network is regularized by convolving it with a Gaussian. After pre-training the model on SALICON Data set, the MIT1003 data set is used for fine-tuning purposes and to stop the model from over-fitting.

Using this framework, Deep Gaze II tops the Saliency Benchmark for CAT2000 data set and is included in the top ten models for MIT300 data set. Using AUC Judd metrics for evaluation, with center bias Deep Gaze II has an accuracy of 0.88 on MIT300 data set and 0.86 on CAT2000 data set It also achieved 0.77 according to shuffled AUC without center bias.

3.4.3 Personalized Saliency Map: Data Preparation

Using the x and y coordinate fixation data of all session from personalized saliency data set for all 1600 images, for 30 subjects required some preparations. The fixation data of every subject for all sessions for each image was compiled into excel files. Therefore, for each subject, there would be 1600 files. For all 30 subjects there would be a total of 48000 files. These files contained the (x,y) coordinates and the corresponding RGB (Red, Green, Blue) values, CYMK (Cyan, Yellow, Magenta, Key) values, HSV (Hue, Saturation, Value) values and HSL (Hue, Saturation, Lightness) values. This data will serve as training data for the gradient boosted tree regression model for each individual.

Using the Deep Gaze II model [114], we can obtain the density maps from the universal saliency model output. We used density maps instead of image saliency map for our study. As with fixation data of the subject, the salient pixels and the corresponding RGB, CYMK, HSV, and HSL values were prepared for the 1600 universal saliency maps. The salient pixels chosen to be

included were those with density value greater than 0.1. This data will be fed into the prepared model and personalized saliency maps will be extracted out from this data.

3.4.4 Personalized Saliency Map: Gradient Boosted Tree Regression Model

When creating the first personalized saliency model, [220] raised two points:

- Personalized saliency map is a part of the universal saliency map. In fact, personalized saliency map is an aggregation of universal saliency map and the difference between universal saliency map and personalized saliency map. Therefore, it is more efficient to extract personalized saliency map from the universal saliency map rather than creating it from nothing.

- The nature of images and subjects contribute greatly to the difference between the personalized saliency maps and the universal saliency maps. Since for all subjects, the images would be same [220] used multi-task convolutional neural network to build one model that shares four convolutional layers and then splits into as many layers as the number of subjects. Therefore, subjects learn from each other and then split to produce individual personalized saliency maps.

However, considering that homogeneity among individuals is already included in the universal saliency maps, the heterogeneity among individuals should be kept separate. With this reasoning in mind, we trained our gradient boosted tree regression model separately for each individual. For each subject all the attributes data from 1600 images was combined and then the pre-processing was done. During this experiment Red was treated as the target variable or the dependent variable and all the other variables were treated as independent variables.

The following steps were taken for data pre-processing:

1. Missing values were replaced by 0. During Matlab excel file writing process, 0 could result in an empty cell.

2. Data is normalized using min-max normalization.

3. Correlation of variables is checked. Although strong correlation was not observed between variables, removing any of them resulted in poor accuracy. Hence, no feature elimination was performed and all the variables were used in model building.

4. Data splitting was done using linear sampling. 80% data was used for training and 20% data was used for testing the model during the training phase.

Since testing and evaluating results for 1600 images corresponding to one subject took around 32 hours, we decided to test out different regression models initially with 100 images rather than all 1600. From all the different regression models, three models gave the highest and almost similar results. The results of this initial testing have been included in the appendix.

After the initial testing, the gradient boosted regression model with the following setting was chosen as the most appropriate model for our study.

- Tree depth = 4
- Number of models = 100
- Alpha (percentage of data not treated as outlier) = 0.96
- Learning Rte = 0.1

Using the these parameters gradient boosted tree regression model was tested for all 1600 images. The models for all subjects had R^2 value of 1, the Mean Absolute Error was either 0.002 or 0.003 and the Root Mean Squared Error was either 0.004 or 0.005 for all the subjects.

3.4.5 Personalized Saliency Map: Map Extraction

On completion of model training for every subject, the data prepared using universal saliency maps in 3.4.3 is fed to the gradient boosted tree regression model predictor. This results in 1600 data files for every individual. These files are then used for personalized saliency map extraction and then evaluation of the extracted map.

The personalized saliency map is made up pf all the pixels that has error an error value of less than 0.5 and density value greater than 0.1. These pixels and the corresponding x and y coordinated make up the personalized saliency map. AUC Judd [32] and NSS functions are then use this map and the ground truth for evaluation with jitter value set to true,

3.5 Results

There are many evaluation metrics that are used for testing the accuracy of saliency maps. These metrics can be location based or distribution based [32].

Discrete fixation make up the saliency map values for location based metrics, whereas continuous fixation make up the saliency map values for distribution based metrics.

- **Location Based Metrics** Area Under the curve (AUC), Normalized Scanpath Saliency (NSS), Information Gain (IG).
- **Distribution Based Metrics** Pearson's Correlation Coefficient (CC), Earth Mover's Distance (EMD), Similarity (SIM), Kullback-Leibler divergence (KL)

In our research AUC Judd and NSS were used for evaluation purposes. The reasoning for this is as follows:

Deep Gaze II universal saliency model is a probabilistic model and hence produces probabilistic saliency maps. In case of probabilistic saliency maps zero indicates that fixation cannot be at that pixel. Other saliency models could have another definition and meaning. Non-probabilistic saliency maps that have abundantly more zero value pixels benefit more from metrics such has Information Gain, Pearson's Correlation Coefficient, and Similarity [32].

Shuffled AUC (sAUC) lowers the rank of models with center bias compared to models that have more sporadic predictions. Since, we take center bias into consideration we did not use shuffled AUC.

Location based metrics, AUC Judd and Normalized Scanpath Saliency were used in our evaluation purpose. This was done in order to maintain consistency with Deep Gaze II universal saliency model and the reasons mentions above.

AUC Judd is one of the variants of AUC that has built in center bias. AUC Judd true positive rate (TP Rate) is the ratio of true positive fixations to total number of fixations that are more than fixated pixels. AUC Judd false positive rate (FP Rate) is the ratio of false positives to total number of fixations that are more than unfixated pixels [32].

Normalized Scanpath Saliency [88] is a hybrid approach metric that compares the ground truth saliency map and the scanpath. It takes the average distance between normalized saliency values and fixation. Close similarity between the fixations and the predicted points is depicted when the NSS values is greater than zero.

Since every subject had their own model, evaluating the results with AUC judd and NSS, following are the average results:

- **AUC Judd** Lowest score was 0.75 and highest score was 0.86. On average, the score for 30 subjects was 0.80.

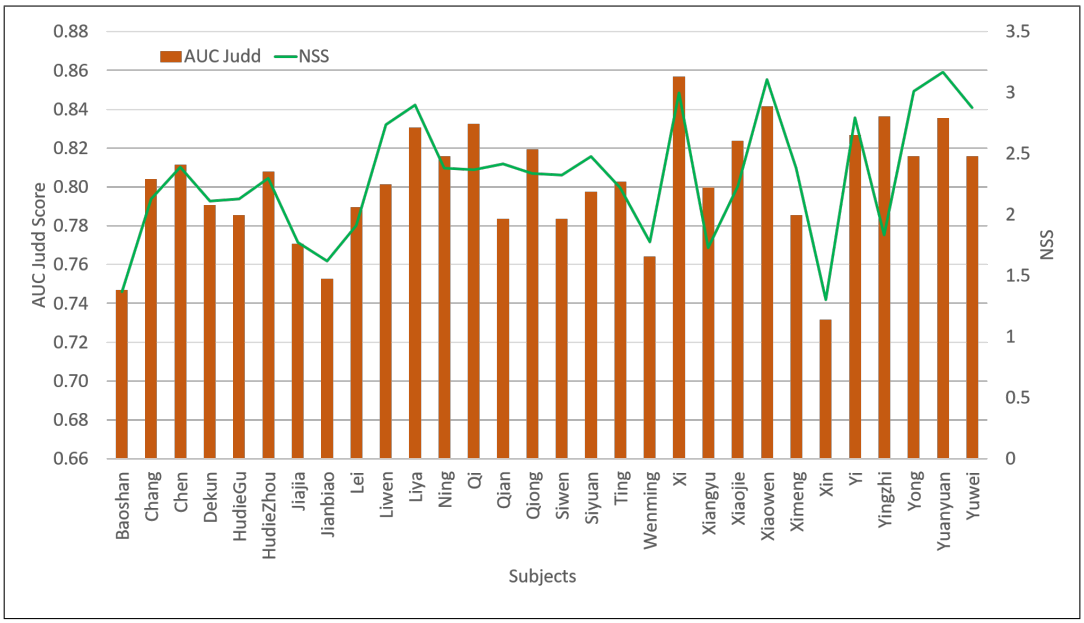


Figure 3.11: Average AUC Judd score and NSS for all subjects [224].

- **NSS** Lowest score was 1.30 and highest score was 3.17. On average, the score for 30 subjects was 2.30.

Therefore, we can get quite good results from this technique with the aid of complicated deep neural networks. However, we notice that two subjects perform much worse when compared to others. Furthermore, we also observed in our analysis that for some images the model performance was low for all subjects and for some images it was excellent for all subjects. These observations have been discussed in the next section.

3.6 Discussion

During evaluation and result aggregation two prominent observations stood out in the results.

First, there were some vast differences between the results of some subjects as compared to others. The second observation was similar but instead of subjects, it was regarding the images. Some images had great accuracy for all the subjects whereas some images have poor accuracy for all the subjects.

In the subsequent sections, we will try to understand why this was so. In section 3.6.1 we discuss how contents of an images can effect the performance of the model and in section 3.6.2 we discuss how the gaze behavior of an individual effects the performance of their model.

3.6.1 Image Analysis

We started out with the hypothesis that bright, contrasting, colorful images would perform good. In comparison, we hypothesized that darker images would perform worse in comparison. However, the best performing images (Figure 3.13) and the worst performing images (Figure 3.12) prove our initial hypothesis wrong.

People unconsciously or consciously, look towards the center of the image. This is called center bias. Faces, texts and sign boards are also highly salient object, as reported in [94]. In [150], four scenarios are put forth that impact the detection probability.

- High contrast object in an image has high detection probability.
- Low contrast object in an image has low detection probability.
- Modified image to have higher target contrast with lower background contrast has high detection probability
- Modified image to have lower target contrast with higher background contrast little to no impact on detection probability

Based on these reasons we can infer why the two images in figure 3.12 (Figure 3.12(a) and Figure 3.12(b)) are the best performing images despite not being very colorful and bright. In both the images, the main salient regions stand out very distinctly from the background and are towards the center of the image.

In the first image, the face and text stand out more than their surroundings and the rest of the images. The face being in the center also helps in increasing its detection probability. In the second image, the color of the faces stand out than the surrounding and the four faces are spread out relatively evenly through the picture along the center horizontally. On average this image had an AUC Judd score of 0.93 for all subjects. Other images, similar in nature, also performed the same way.

In the presence of clutter or multiple objects together in a scene, our natural categorization ability starts to deteriorate and we take longer to process the information to determine salient regions of an image or scene [209]. When there is a search task involved, the detection probability of the target object would increase but no such task was given in the collection of personalized saliency data set [220] that we used. The participants freely viewed the images.

(a) fiftyshadesofgrey-officialtraileruniversalpictureshd.mp4-00.01.10.640 [220]

(b) cloudatlasextendedtrailer12012-tomhankshalleberrywachowskimoviehd.mp4-00.03.46.184 [220]

Figure 3.12: Best images for all subjects with an average AUC Judd score of 0.93 [220].

Therefore, in this case, in the presence of clutter or multiple potentially salient object the observer looks around to gather as much information as possible and as a result we get fixations that are scattered around.

Along with categorization, in the presence of clutter or multiple objects our parallelism also starts to deteriorate. Humans depend on their skill to perform and analyze information about objects in parallel to properly identify and process objects quickly [92]. Therefore, our information processing becomes longer with the increase of objects in an image.

Based on these reasons we can infer why the two images in figure 3.13 (Figure 3.13(a) and Figure 3.13(b)) are the worst performing images despite being very colorful and bright. Due to presence of clutter, the gaze of the spreads around to gather as much information as possible, rather than being fixating on any specific area or areas.

In Figure 3.13(a), the Deep Gaze II universal saliency model only two objects out of all are considered salient and the rest are not represented in the universal saliency map with similar intensity as salient. Due to the fact that personalized saliency map is extracted from the universal saliency map, it is not possible for a region to be absent from universal saliency map and be present in personalized saliency map. This could be the reason behind the poor performance of this this image. The fixations of the individuals might be outside the region deemed not salient by the universal saliency model.

In Figure 3.13(b), for our brains it is easy to recognize that the potentially salient area of this image are the three girls playing and ball. We could also include the fourth girl from the bottom of the image. However, when the universal saliency map of this image was observed it included majority of the background in the saliency map as salient. This could be due to the presence of multiple objects and in this image considering these objects are faces which are considered highly salient. Therefore, inclusion of background which was not salient as salient in the universal saliency map also effected the performance of our model. Other images, that were similar in nature, also performed in the same way.

3.6.2 Subject Analysis

In visual saliency, eye movements dictate where our attention lands and stays [120]. Due to this high quality eye trackers are important when creating data sets for research in visual saliency. The information rec vied by our eyes, help our brain in determining whether this information is important, if it should be

(a) idx$_1$68$_i$*lsvrc*2014$_t$*rain*$_0$00334011001[220]

(b) idx$_2$52$_i$*lsvrc*2014$_t$*rain*$_0$00147431001[220]

Figure 3.13: Worst images for all subjects with an average AUC Judd score of 0.64 [220].

remembered or if does any past memory can associated with this information needs to be brought up.

Therefore, eye movements contribute a lot towards understanding visual saliency and consequently the behavior of the gaze of the individual contributes toward their personalized visual saliency. When studying visual saliency, eye movements can be divided into two types:

- **Fixation** indicates if the subject focuses and spends longer time looking at certain areas of the image.
- **Saccade** is when an individual moves their eyes from one area to another in a short period of time to change the area of attention.

The information from fixation and saccade of an individual tells us whether the individual was focusing on something definite or just looking over the image to get the overall information from the image.

With the exclusion of two subjects, the average AUC Judd score for the rest of the subjects in 3.11 is about 0.80 to 0.84. The personalized saliency model of the two subjects identified as Baoshan and Xin in the personalized saliency data set performed exceptionally worse than others. Conversely, the personalized saliency model of an the subject identified as Xi performed exceptionally better than others by having an average AUC Judd score of 0.86. To understand these exceptions, we mapped the fixations of these subjects against all other subjects (Figure 3.14 and Figure 3.15) and looked for distinguishing behaviors.

The better performing subject i.e. Xi, had two features that stood out in their fixations. The first feature was the fixations of Xi were very focused and grouped together. The second feature was that the fixations of Xi were mostly in the areas that were deemed salient by the Deep Gaze II universal saliency model. This increased the detection probability of the fixations and as a result the personalized saliency model of Xi was more accurate at predicting fixations that would be salient for Xi. Figure 3.14 shows an example saliency map with fixations of Xi and other subjects.

The worse performing subjects i.e. Baoshan and Xin also had two features that stood out in their fixations that were entirely different from the features noticed the fixations of Xi. Firstly, The fixations of Baoshan and Xin were more sporadic in nature, They were spread out more either all over the image or in areas that were not considered salient by the Deep Gaze II universal saliency model. This meant their personalized saliency model would have a

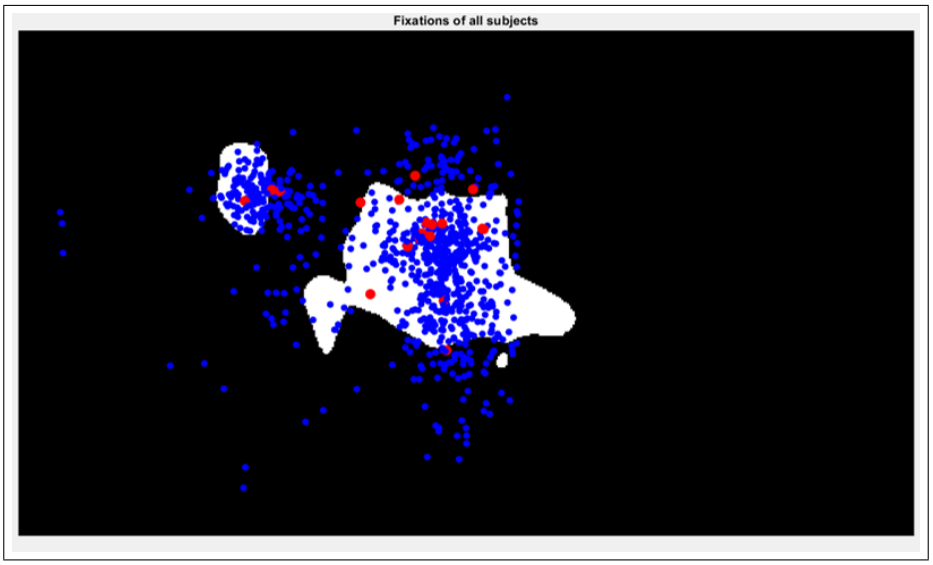


Figure 3.14: Saliency Map with fixations of all subjects (denoted by blue dots) and the best performing subject (denoted by red dots). The white area represents the salient regions predicted by Deep Gaze II [224].

lower probability of predicting fixations withing the salient regions. This could be the reason behind the poor performance of these two subjects. Figure 3.15 shows an example saliency map with fixations of one of the poor performing subject and other subjects.

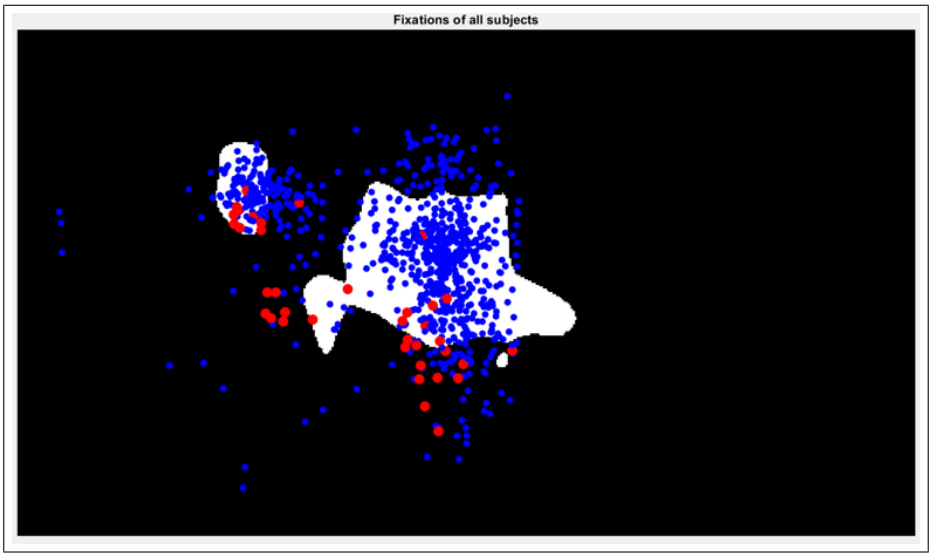


Figure 3.15: Saliency Map with fixations of all subjects (denoted by blue dots) and the worst performing subject (denoted by red dots). The white area represents the salient regions predicted by Deep Gaze II [224].

Chapter 4

Personalized Saliency in Immersive Environments

4.1 Limitations of Visual Saliency

Visual saliency is important to study and understand in both immersive and non-immersive environments. However, in an immersive environment where we look and what we think about our view changes dramatically. When concerning personalized saliency in immersive environments for practical applications, we should look beyond visual saliency.

When trying to understand what grabs the attention of a person in an immersive environment such as real world or virtual reality, context and the situation going on plays an important role. Top down saliency models do incorporate some semantic information by introducing object labels and task information (Section 1.2.2 and Section 1.2.3) but the labels do not consider the social situation or background information that might be relevant to why attention is being paid to that particular area.

For example, in an image of person, face would be considered highly salient and visual saliency model would recognize the face as salient but visual saliency models would not be able to tell us why. In [168], it was reported that faces are preferentially processed by the brain without any regards of spatial placement and task. This information cannot be explained by models that consider only low level features. It was also suggested that this pre-processing was done in participants before saccade occurred. Therefore, this pre-processing was done before the stimulus had been reacted upon the attention was not paid to it.

Research has also shown that in dynamic situation or when viewing video content, eye movements are more consistent than when viewing static im-

age content [47]. Therefore, in immersive situation exploration based saliency might be a better avenue for attention prediction than gaze.

In a study done using eye trackers where participants were plain visitors exploring museum to understand what people actually viewed and the scanning patterns of the visitors, four main observations were reported [51]:

- Our eyes move rapidly in real-life situations and conventional eye tracking methods are bound to miss out some information due to this.
- In cases where instead of eye movements, head and body movements are tracked, the consideration of observing without moving the body i.e. moving eyes without moving the body or head is not taken into consideration.
- Distractions by obstructions due to the presentation of the salient stimuli or due to other people cannot be accounted by in visual saliency models. This can change focus of the observer based on the proximity of the stimuli or the number of obstructions in between.
- We cannot determine the goal, emotion or the thought process behind why eyes of the observer focused on the stimuli

Other than these reasons, personalized saliency cannot be computed for immersive environments simply because of the inaccessibility of eye tracking devices. Currently, there are very limited headsets available to general public that can capture gaze data effectively. Making application that utilize eye gaze data very limited and inaccessible to the general public and the developers of such applications.

There are also other disadvantages of using mobile eye tracking which include detailed in [162]:

- **Covert attention and mental spotlight** The accuracy of eye fixations does not give us information about why the fixation occurred or if it was even relevant to what the person was thinking about.
- **Limited conclusions about cognitive processing** The subjective meaning behind the eye movements is based on assumptions about the cognitive processes that take place alongside [75]. Although top-down saliency can be determined when a task is set but the same knowledge cannot be applied in the same manner in a task-free setting.

- **Obtrusiveness of measurement** Wearing an eye tracking device makes a person aware of their gaze and this interferes with how the gaze would be in a more natural environment.

- **Selective sampling** Not every type of eye can be tracked using an eye tracker. People with impaired vision, glasses or corneal irregularity increase the inaccessibility of eye trackers due to difficultly in calibration.

- **Limited temporal and spatial accuracy** Our eyes move faster than the eye trackers leading to missing fixation information due to device technical limitation. Eye tracking also works better when fixed to a certain point, this again is not what happens in real or immersive situations.

- **Laborious data analysis** Due to constant changes in the background along with behaviour and movements, eye tracking has to be analyzed by professionals manually rather than automatically. Developments of software to aid this process can be helpful in utilizing mobile eye tracking.

- **Price** Mobile eye trackers are expensive and limited. This however, could change in the future as technology advances.

- **Ethical concerns** Unconscious eye movements might reveal information about the participants that they would be embarrassed about. This could happen even if the process was explained before hand and consent was acquired.

In real life situations, where we look might not be the exact indicator of our cognitive attention. Reading a book or watching a movie while seated are perceived, due to context, very differently even if both are being done while seated. Moving around, more regions of the brain are involved with the task to ensure knowledge from memory is being associated with the visual stimulus being received [17].

Hence, a need beyond visual saliency arises when trying to understand attention in an immersive environment that can be consumed and utilized for the benefit of everyday people in practical applications.

4.2 Emotional Saliency

As discussed in Section 2.2, emotional saliency is a field of research rich with literature aimed towards understanding how our emotions play a role in deter-

mining what grabs our attention and why. Emotions also play a huge role in determining whether what we paid attention to will be remembered or not.

However, in section 2.2, we saw how extensive research in emotions is not being utilized in practical applications due to it being done in controlled environments with inaccessible gadgets and sensors.

Therefore, if we are able to predict how a stimulus might make a person feel, we can create a personalized saliency model catered to that person based on their emotion. For this purpose, we need to first experiment if we can create a model that can induce the emotion we want in the participant using devices that will be available to the participant in real life.

4.2.1 Audio Saliency

In Section 2.2.1, among different techniques for emotion induction, audio stimulus when combined with other induction methods work better for emotion induction.

Although music and sounds both induce emotions, music related emotion induction has been studied in more detail than sound related emotion induction [208]. In [21] the relationship between sounds and emotion was deemed important as it effects categorization and perception of the sound. The features of sounds, such as loudness, also effected the arousal values of the participants but this observation was not noticed for valence.

Due to the complex interconnection of audio stimuli and emotion, in our research we aim to utilize different kind of audio stimuli to induce emotions.

4.3 Proposed Method

As discussion in the previous section, visual saliency is not enough or feasible for modelling and predicting personalized saliency in immersive environments with the current technology for practical applications. Therefore, in this study, we propose using emotional saliency instead in combination with audio stimuli and appropriate environment.

How people feel different emotions is a very personal and individual experience [216]. It differs significantly from person to person based on their gender, culture, life experiences and personality [61].Research has shown that men and women have different coping mechanism when dealing with stress and other negative emotions [110]. Emotions, such as fear and anxiety, that are associated with horror games are even more personal.

There are different types of emotions, negative and positive. Emotional saliency can be researched for both kinds of emotions. However, since this can be too overwhelming for one research, we will be focusing on one emotion only.

4.3.1 Emotion Selection

Keeping personalized saliency in mind, we want to focus on an emotion that is the most personal for human beings and that would have most applications. For the focus of this research, we chose the emotion fear. Considering how mental health therapy is a huge application for immersive environments, negative emotions like fear, which includes anxiety and panic [157], would be good candidate for this research. [99]

4.3.2 Understanding

Before we induce fear, we need to first understand fear. Fear is an emotion that has been evolved with time for the sake of survival. Based on different life experiences, different objects, people, or even concepts make people anxious and afraid. When a person senses threat or danger, fear incites the individual to change behavior to protect themselves. This change also involves memory and attention. Therefore, this need for behavior change in response to stimulus that might be endangering situation [6].

Fear can be triggered both consciously and unconsciously. Research shows that even with hidden or indirect stimuli fear can be triggered. Fear is expressed when people notice physical changes in their body and changes in action to get themselves out of the situation that caused fear. People may also become aware of change on their thoughts and background information [20].

4.3.3 Induction

Although fear is a negative emotion, the popularity of media genre horror, that incites fear, cannot be denied. People are fascinated with fear because it is different from daily life [2] and can be experienced safely using horror movies and games. Extensive research has been done in understanding fear and inducing fear to improve horror media.

In section 2.2.1, different methods for inducing emotions were discussed. In our experiment, we used a combination of virtual reality and sounds to create an environment of fear. For this purpose, we create a virtual reality based

horror game. we analyze changes in heart rate of the participants and their self reported results to determine if fear was successfully induced.

Different factors need to considered when creating a horror game

- **Audio Visual Stimuli** In order to create a personal experience in horror games audio visual stimuli should be adapted according to the player. Some people may be more affected by visual stimuli and some more with audio stimuli. Some people might need a combination of both to experience fear [69]. Graja and Lopez [69] built a game using finite state machines that make use of audio and visual effects to study the impact of Horror Games on Galvanic skin Response. Sounds were reported to be the most important factor to induce greater stress and anxiety in this research.

- **Environment** Environment plays an important role in creating a realistic and immersive horror game experience. Different lighting and sounds play a role in making the player feel vulnerable. Open spaces or closed spaces trigger different kinds of emotions for people, especially if the player has claustrophobia [133, 175].

  Maintaining an element of obscurity by limiting the vision of the player also helps in making the player feel vulnerable and unsafe from potential dangers. Obscurity can be created not only visually but also through sounds. Sudden noises or complete silence will trigger negative emotions in the player and make them uncomfortable [69, 196].

  It can be imagined that having a familiar environment would likely not trigger fear and vulnerability. However, if unusual and unexpected things happen in a familiar setting it can make the player uncomfortable and question their safety in the game [116].

- **Immersion** Immersion is a feeling described as being present in the game while loosing awareness of real time and place [90]. Currently out of all game formats, Virtual Reality (VR) provides the most immersive experience to the player [183]. It creates an illusion of actually being present in the game. Real time movement in the game make the experience even more close to reality [178]. Due to this VR is used extensively in emotion research. Inducing different kinds of emotions creates real life reactions and helps the researchers in understanding emotions better [1]. Extensive research is also being done using VR to help people suffering from different kinds of phobia [156].

Madsen [129] studied immersion in horror games by doing a comparative study of horror game players and watchers using electrodermal activity (EDA), respiratory rate (RR), heart rate (HR) and self reported fear data. Results showed players had greater variance in physiological data than watchers but no significant difference was noticed in self reported fear data.

4.3.4 Evaluation

To evaluate if fear has been induced, we use Self-Assessment Manikin test and heart rate of the participants.

Self-Assessment Manikin (SAM) [26] is a pictorial questionnaire to determine emotion using valence, arousal and domination. Different emotions have different valence and arousal values.

This is shown in Figure 4.1 Low valence and high arousal corresponds to horror and fear [205]. In our study, a nine point SAM questionnaire was filled by the participants after every session to collect valence and arousal values.

After each game play session, the participants were required to fill a 9 point SAM questionnaire about their pleasantness feeling (valence) and state of calmness (arousal) during the game. 9 represented most pleasant and least calm. 1 represented most unpleasant and most calm.

As discussed in section 2.2.2, heart rate is a good indicator for emotion detection. In order to mimic real life situations, we did not use any specialized sensors or gadgets to capture hear rate fort in this experiment. The participants were required to use their own smart watch that could sync data to google fit. This way we could check if consumer smart watches are viable to be used in practical applications. This would help developers in incorporating heart rate and other physiological features from these smart watches into application. This would help in applying emotion research in actual practical applications.

4.4 Experimental Design

In order to investigate the use machine learning algorithms and heart rate for the induction and personalization the experience of fear in participants, we developed a Virtual Reality Horror Game using Unity Engine.

Every participant was required to play the game at least three times and each session could last 10 minutes long maximum. The session could end early

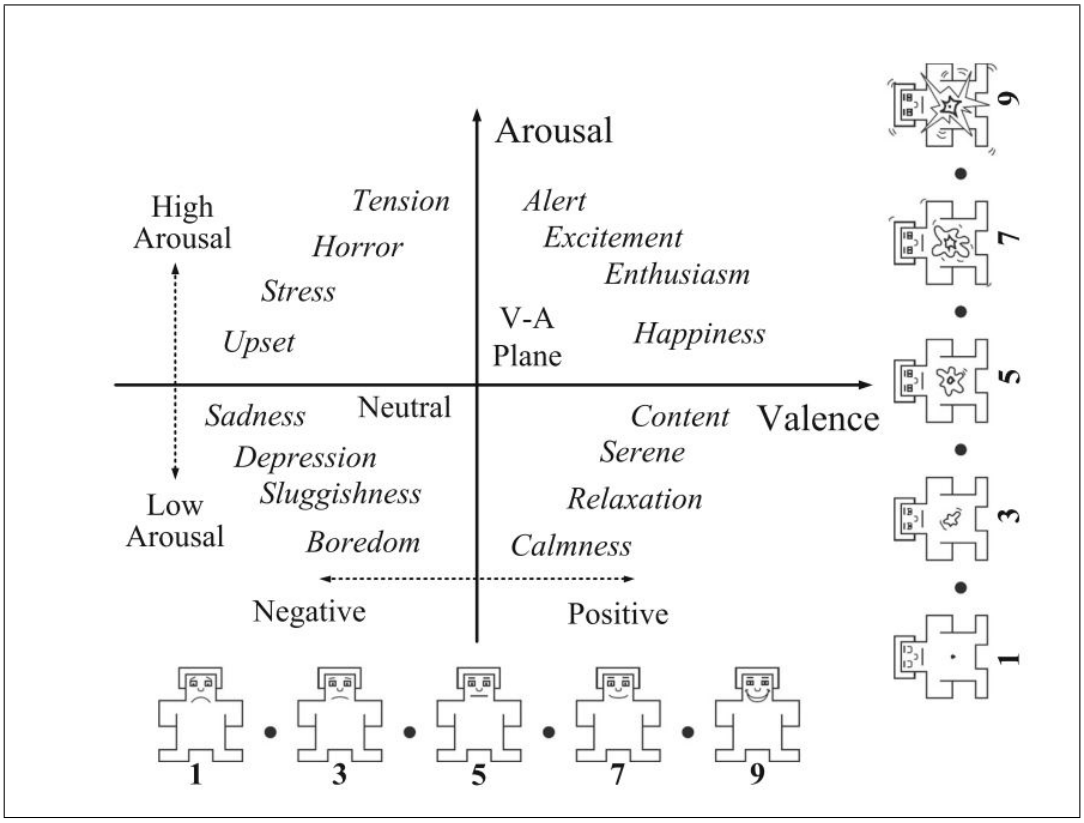


Figure 4.1: 2D Map representation of emotions when using Self-Assessment Manikin questionnaire. The image is from [41].

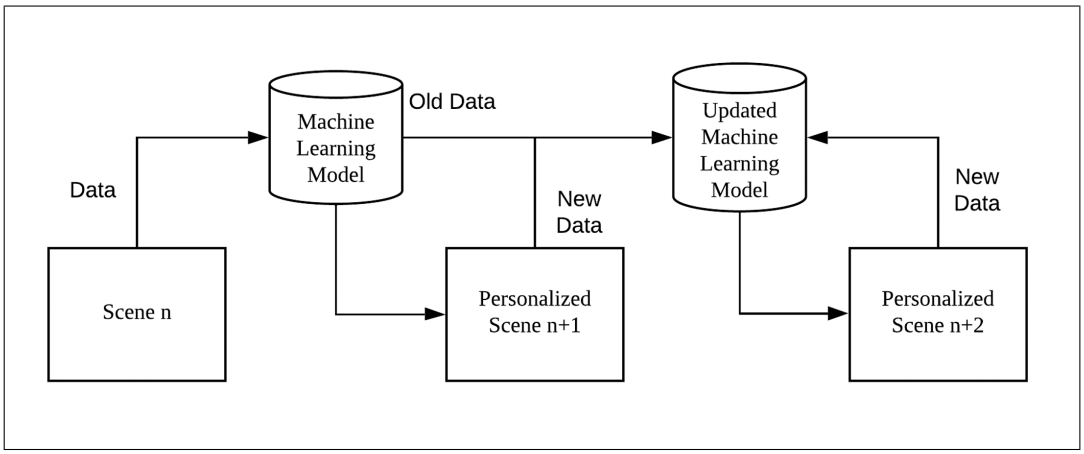


Figure 4.2: Summary of data flow for personalization [225]. © 2022 IEEE

if the goal of the scene was reached. The light and sounds setting in the first session were generated randomly. For subsequent sessions, the setting values were retrieved from machine learning models. Figure 4.2 shows a simple flow of this model.

There were no jump scares or enemies like zombies in the game. This was done to ensure a realistic experience of exploring an abandoned hospital and forest.

To complete the entire experiment the following steps were followed by the participants in the given order:

- Install virtual reality game on headset
- Install mobile application on android phone
- Wear smart watch and ensure Google Fit synchronization is working

After successful installation of virtual reality game on headset and mobile application on android phone, the following steps were required to be repeated at least three times while wearing smart watch for successful completion of participation:

- Start virtual reality game by entering provided user name
- Select Scene (Forest or Hospital)
- Play the 10 minute scene
- Fill out Self-Assessment Manikin questionnaire provided in the mobile application

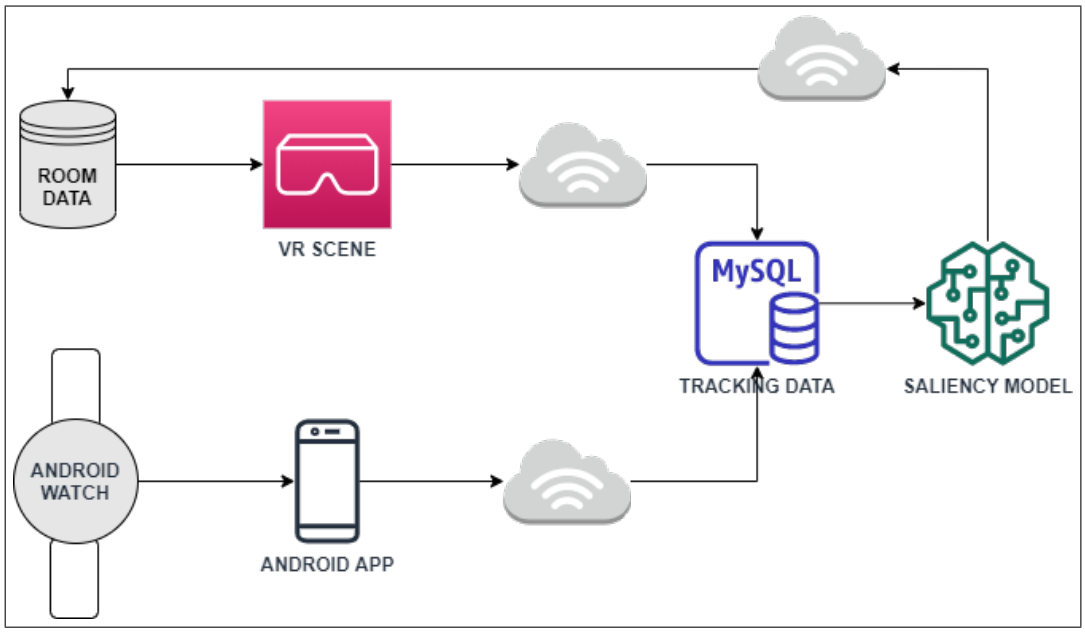


Figure 4.3: Data flow between virtual reality environment, smart watch, android application and the saliency model.

- Upload heart rate data using mobile application

4.4.1 Devices

Participants were required to own all the devices to participate in this study. The devices were as follows:

- Oculus Quest or Oculus Quest 2
- Android Smartphone (Minimum Android version: Oreo)
- Smart watch to detect heart rate that can sync with Google Fit

The participants were also required to enable Continuous Heart Rate monitoring setting on their smart watch that allows heart rate capture without any activity specification.

During the development of this game, the following devices were used:

- Oculus Quest
- OnePlus 7 Pro (Android version 11)
- Mi Band 5 smart watch

Unity XR was used for the development of virtual environment and MySQL database was used for data storage.

4.4.2 Participants

When doing emotion or induction research it is necessary to take considerations from ethical point of view. Especially, when inducing negative emotions including sadness, fear or anger. The participants should be participating in these studies with their free will after the understanding of the research purpose and possible effects [55].

This study was approved by the ethics committee of Tokyo Institute of Technology. The experiment conducted for this research was not done in laboratory setting. Two important points were kept in mid while making this decision.

First, with the rampant Coronavirus pandemic, which is still ongoing, online recruitment of participants and having them play our designed horror game in their own environment was safer for both the participants and the researchers. Due to this, we were able to recruit diverse people from all over the world. This removed the assumption that our results could be due to people belonging to the same ethnicity or living in the same kind of culture. Secondly, participants at home are more engaged and experience more immersion than in laboratory [120].

Generally people do not use these kind of applications and games in designated locations with special sensors and wearing different gadgets to detect different physiological responses. People use them whenever and however they want. They might play standing, sitting, with headphones, without headphones and so on. Therefore, it is difficult for application designers and developers to practically translate emotion research into their applications. Considering how virtual reality requires proper set up and play area, it would always be preferable to play at home. Therefore, for this research, participants were recruited online on Reddit and Facebook virtual reality communities. They were required to already own all the devices needed for this research.

30 people (15 Males, 15 Females) participated in this research with ages from 22 years old to 38 years old. The participants were used to playing virtual reality games. The game was shared with the participants after completing the consent form and confirming the ownership of the required devices. No incentive was given for participating in this research. Participants were free to quit the research at any moment.

Usually in saliency related experiments control group is not used unless a specific trait or factor is being studied. Similarly in emotion induction studies, control group is usually only used when a factor's impact on emotion induction

is being studied.

In our study, we want to know if our research can be used in practical real-life environments without having to make any adjustments. We want to study if in real life environments our method is still successful in emotion induction. For this purpose, we created a control group that was required to prepare their environment according to the following conditions when taking part in the experiment.

- Prepare a quiet room with no sounds
- Be alone in the room during participation
- Headphones were required
- VR headset correctly calibrated to ensure no external light is visible
- Required to stand during participation

The experimental group were not required to any adjustment to their environment. Using this control group and experimental group we were able to find out that the following factors do not impact emotion induction in our experiment

- Environment sounds such as people talking, media playing etc.
- Incorrect VR headset usage disrupting the immersive experience
- Presence of other people during usage

4.4.3 Scenes

To study how different kinds of environment impact heart rate and fear induction in participants, two different types of scenes were created in our virtual reality application. The scenes had sound and lights in them that after the initial session, were adapted according to the heart rate of the participant.

- **In-Outdoor Scene** An outdoor forest scene was designed using Flooded Grounds asset from the unity asset store. The scene was modified so that the player could explore the forest with sounds and lights adapted according to the player's heart rate. The scene consisted of an outdoor forest area with abandoned houses and buildings that the player could go inside.

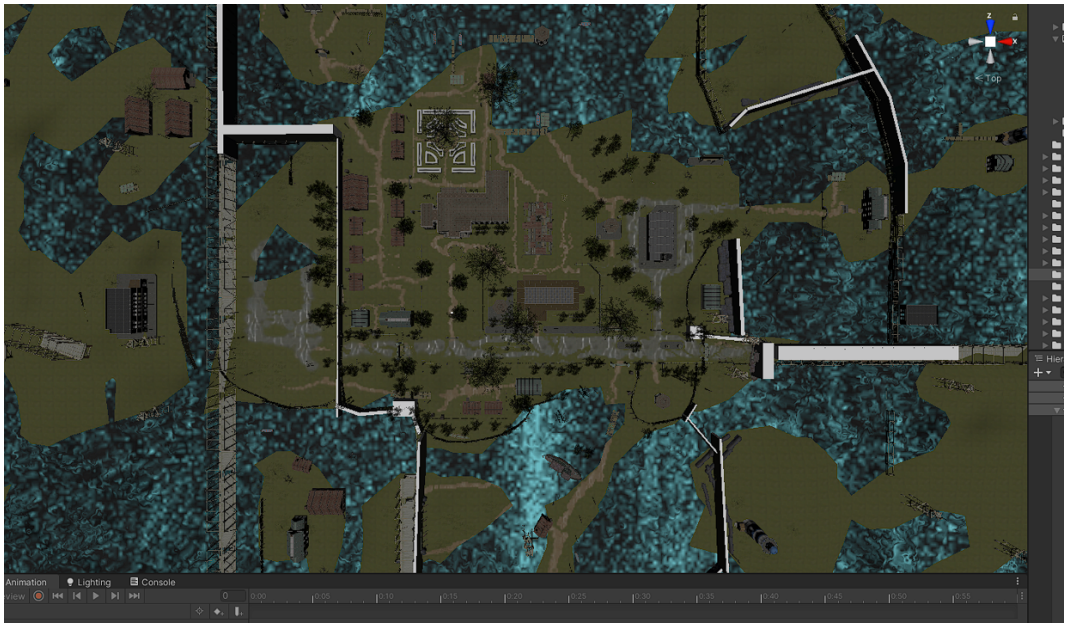


Figure 4.4: Map of the forest used in this experiment (In-Outdoor scene) [225]. © 2022 IEEE

A feature was added such that every time the scene was initiated a random door was assigned to be the glowing exit door. The goal of the forest scene was to find a glowing door. Forest allowed the players to have an indoor and outdoor experience in one scene. A zoomed-out map of the forest has been shown in Figure 4.4.

- **Indoor Scene** An indoor abandoned hospital scene was designed using HE - Abandoned Hospital v.1 asset from the unity asset store. Similar to the forest, this hospital was modified to adapt the sounds and lights of the hospital according to the player's heart rate. The hospital consisted of three floors: main floor, operating floor and the morgue.

  Upon initialization, a key was placed randomly in the hospital and the goal of this scene was to find the key and exit the main door. The key could be on the ground, inside drawers or on top of any object. This was done to encourage exploration and so that the player would be more attentive of their surroundings. A model of this hospital has been shown in Figure 4.5

4.4.4 Sounds

In the game, there were four sounds playing at all times. One background sound and three sounds that surrounded the player. The background sounds

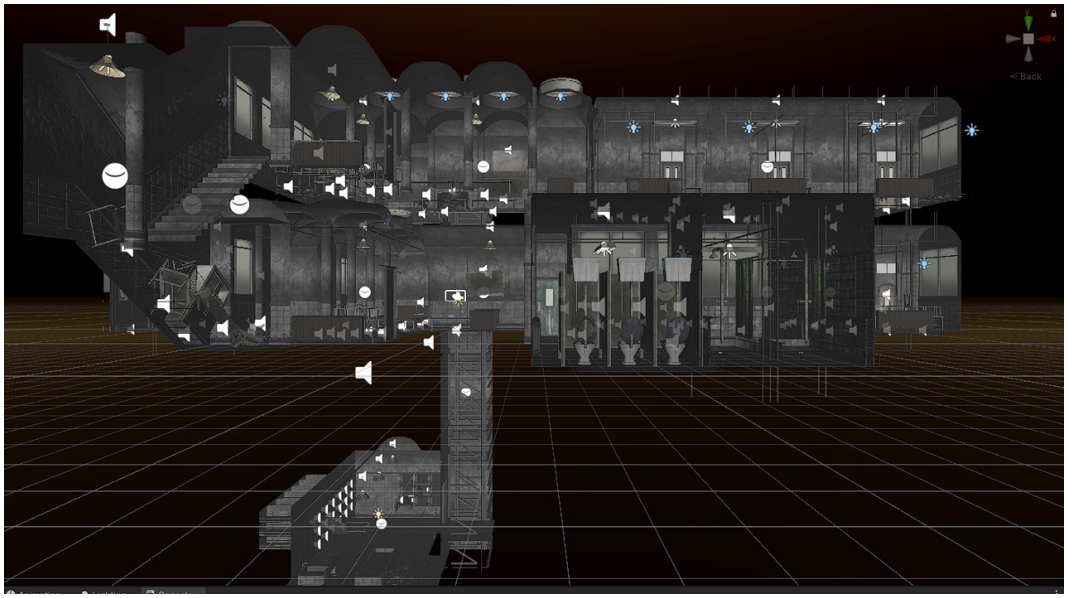


Figure 4.5: Model of a three floor hospital used in this experiment (Indoor scene) [225]. © 2022 IEEE

were stationary and could be heard depending on the distance of the player from the sounds. The other three sounds were attached to the player, The three sounds were placed at different distances from the player and overlapped each other.

Background sound could be from four categories: Wind, Rain, Insects and Others. The other three sounds were selected from 19 different categories namely:

Crow, Dog, Frog, Others, Man_Crying, Woman_Crying, Baby_Crying, Bell, Ground, Footsteps_indoor, Footsteps_outdoor, Knocking_iron (knocking sounds on iron door), Knocking_wooden (knocking sounds on wooden door), Breathing_monster, Eating_monster, Heartbeat_monster, Laughing_monster, Ear_monster and Mouth_monster.

In total 187 sounds from 23 different categories were used in this game.

4.4.5 Lights

Both indoor and in-outdoor scenes had building structures with lamp fixtures. Different types of lamps were used in both the scenes. However, the intensity of lamp light could be on, off or flickering in both the scenes.

Additional feature was added so that the lamps turn on or off automatically when the player was around the lamp.

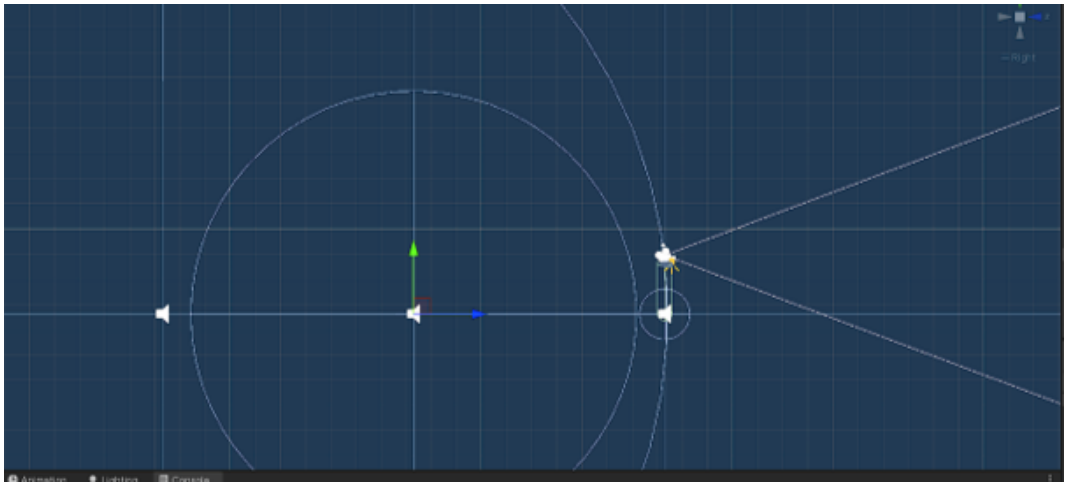


Figure 4.6: Player Composition.

4.4.6 Controls

The player was made up of four components:

- Camera
- Left hand
- Right hand
- 3 sound sources

The three sound sources were placed at different distances from the camera and overlapped with each other. Player composition can be seen in Figure 4.6.

This game for this experiment was designed to be played on Oculus Quest or Oculus Quest 2. Navigation of the player was controlled using the joysticks of the controllers. One joystick was used for movement and the other was used for camera control.

The player was equipped with a torch in left hand. The intensity of torch light could be controlled using the left grip button. The right grip button was used to interact with doors, drawers, cupboards and hold the key in the game.

4.4.7 Mobile Application

An Android mobile application was designed that could collect heart rate data in beats per minute (bpm) from Google Fit. This was done so the players with any kind of smart watch that could sync with google fit could play the game.

The participants were required to play the game at least three times and send the data three times to our servers for successful participation.

After every session they were required to make sure Google fit data is synced and then use the provided mobile application to upload their heart rate data to our servers. This was also designed after recommendation from the ethics committee so that the participants could take a little break after playing one session and prevent possible sickness induced due to virtual reality use [33]. The user interface can be seen in Figure 4.7

4.4.8 Machine Learning Model

Ensemble learning is used to achieve high predictive performance using multiple learner models and combining their predictions. Ensemble learning also reduces the risk of over-fitting in small data sets [167], which is perfect for our small data for every subject. We use three different ensemble machine learning algorithms namely Gradient Boosted Tree Regression, Random Forest Regression, and Tree Ensemble Regression.

Gradient boosting [63] works by building multiple models and optimizing a differential loss function. Gradient boosting combines multiple weak models into a single more accurate and robust model.

Random Forest Regression [28] is a supervised machine learning algorithm that creates a single model by using predictions from ensembles of different machine learning algorithms.

Lastly, Tree Ensemble Regression [169] uses multiple weighted regression trees to make a more accurate prediction. These algorithms can be used for both classification and regression.

To study the impact of environment on the participants, 14 participants were required to play the in-outdoor (forest) scene first and 16 participants were required to play the indoor (hospital) scene first. Every participant was required to play both the scenes and could choose the third scene on their own.

A remote server was created to collect data from the game. After every minute in the game the sounds and lights would change and the updated values would be added to the database. The communication of different servers can be seen in Figure 4.8.

The heart data was collected from Google Fit provided heart rate every minute. Therefore, a 10 minute session resulted in 10 heart rate data points. Heart Rate data was then matched with the game data using timestamps to prepare training data for prediction models.

Figure 4.7: User interface of android mobile application.

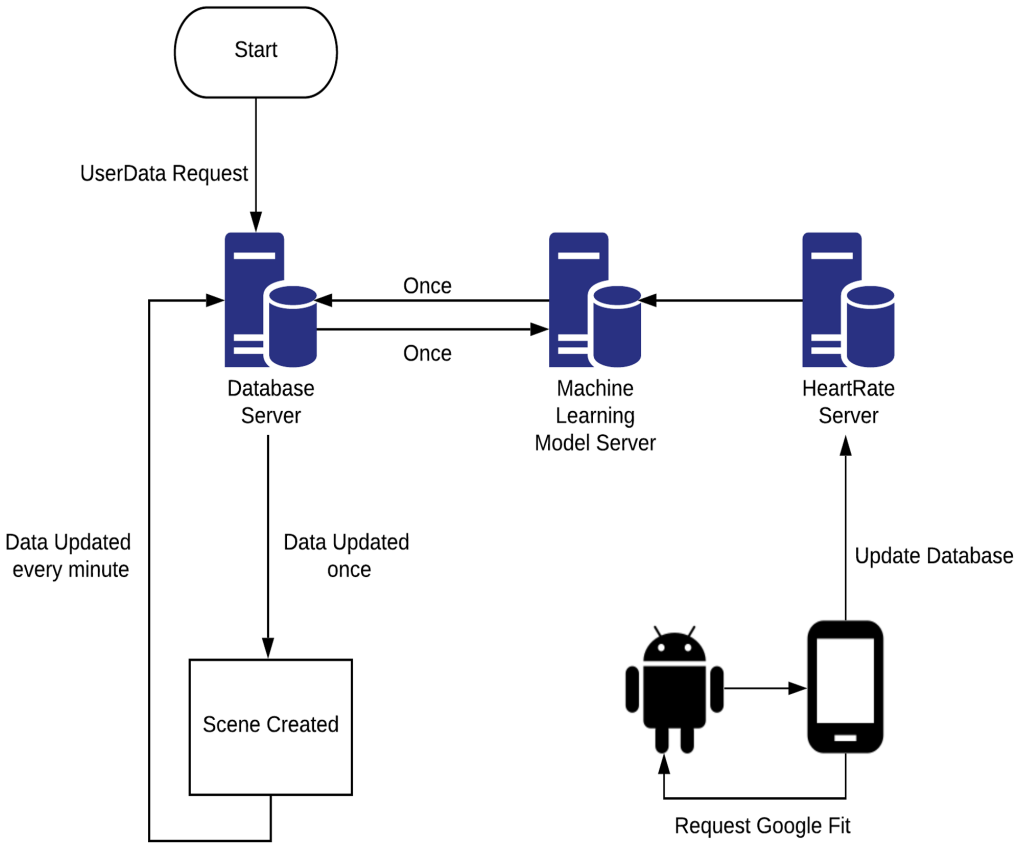


Figure 4.8: Communications of different servers [225]. © 2022 IEEE

To personalize the game for increased horror aspect, the heart rate of the player needed to be increased. For this purpose, regression was performed using three different algorithms with heart rate as target variable. Three models were used because if the model is being re-trained after every session, the accuracy of the new model could be higher or lower than the previous model.

Therefore, to improve chances of good accuracy three different regression algorithms were used. The model with the highest accuracy was then selected for personalization settings. Considering that the model is personalized and the training data size is not huge (less than 1000 rows for each participant), it is feasible to do model training at run time at the start of game session.

Based on the data collected, the following variables were prepared for the models

- Volume (Value from 0 to 1)
- Pitch (Value from -3 to 3)
- StereoPan (Value from -1 to 1)
- Sound (Value from 1 to 187)
- Sound_Category (Value from 1 to 23)
- Light_Value (Value from 0,1,2)
- Heart_Rate (Value from 50 and above)

Light value 0 corresponds to off, 1 corresponds to on and 2 corresponds to flickering light.

At the start of every session,

- Random Forest Regression (Models: 100, Static Random Seed)
- Gradient Boosted Tree Regression (Models: 100, Learning Rate: 0.1, Alpha: 0.95, XGBoost missing value, Static Random Seed)
- Tree Ensemble Regression (Models: 100, Fraction of Data for single model: 1, Static Random Seed)

models were trained with 80 percent data and tested with 20 percent data using linear sampling.

To initialize new session of the game with different sound and light setting, 500 rows of random data was generated without the heart rate variable. The

model with the highest accuracy was chosen at run-time and was used to predict the heart rate of the randomly generated test data. The data was then sorted from highest to lowest heart rate. The 40 rows with highest heart rate were fed to the scene to initialize the game and change the settings of 4 sounds and lights every minute.

4.5 Results

To check if there was any difference between the control group and experimental group, Wilcoxon rank-sum test was done with the null hypothesis that control group and the experimental group are not different. Wilcoxon rank-sum test [214], also known as Mann–Whitney U test, is a non parametric test to check if two samples are different. The results of this test can be seen in Table 4.1. With alpha 0.05, the null hypothesis was accepted.

Table 4.1: P-values for Wilcoxon Rank Sum test done with control group and experimental group.

	Session 1	**Session 2**	**Session 3**
Average Heart Rate	0.495296	0.179654	0.082115
Valence	1	0.837722	0.357424
Arousal	0.449278	0.517888	0.704816

To check replay-ability and to detect differences in repeated measures Friedman test [64] was done. Friedman test is a non parametric test to determine if repeated test had any differences. Although average heart rate increased from session one to session three, this change was not statistically significant different over the three sessions. Average valence decreased and average arousal increased from session one to session three. this change was statistically significant.

Average heart rate can be observed in Figure 4.14. Average valence and average arousal can be observed in Figure 4.10. Friedman test results can be seen in Table 4.2.

4.5.1 Environment Analysis

Another Wilcoxon rank-sum test was done to determine if the same setting was successful in inducing fear when it was machine learning boosted compared to random audio and video settings. The results can be seen in Table 4.3.

Table 4.2: Friedman Test Results (* indicate significant results.)

	Control		**Experiment**		**All**	
	Statistic	**p**	**Statistic**	**p**	**Statistic**	**p**
Heart Rate	0.2	0.904837	3	0.22313	2.762712	0.251238
Valence	4.764706	0.092333	12.09375	0.002365*	16.71429	0.000235*
Arousal	1.354839	0.507926	8.342857	0.01543*	7.980198	0.018498*

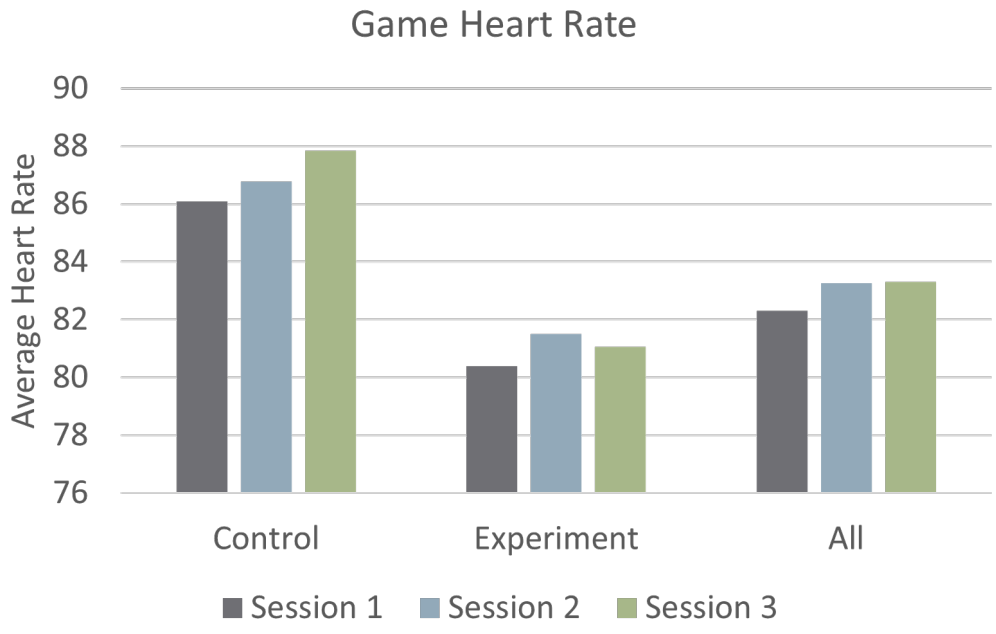


Figure 4.9: Average heart rate.

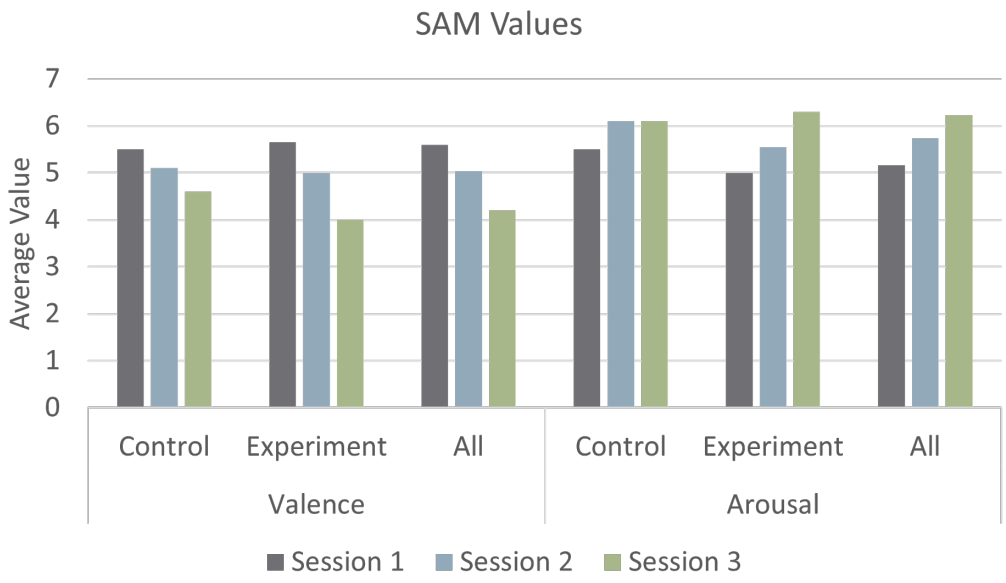


Figure 4.10: Average SAM Valence and Arousal.

Graphical representations can be observed in Figure 4.11 and Figure 4.12 for in-outdoor scene. Only valence for in-outdoor scene was statistically significant. Graphical representations can be observed in Figure 4.13 and Figure 4.14 for indoor scene.

Table 4.3: Wilcoxon rank-sum test Results (* indicate significant results.)

	Average HR		**HR Variability**		**Valence**		**Arousal**	
	Statistic	**p**	**Statistic**	**p**	**Statistic**	**p**	**Statistic**	**p**
Indoor ML boosted	22	0.574	24.5	0.949	2	0.188	4.5	0.438
In-outdoor ML boosted	46	0.549	50	0.970	0	0.039*	4	0.089

4.5.2 Gender Analysis

Another Friedman test was done to determine if gender impacted the heart rate, valence and arousal of the participants. There was no statistically significant differences in heart rate for both male and female participants. However, only for female participants valence and arousal differences were statically significant.

The results can be observed in Table 4.4. Graphical representations of

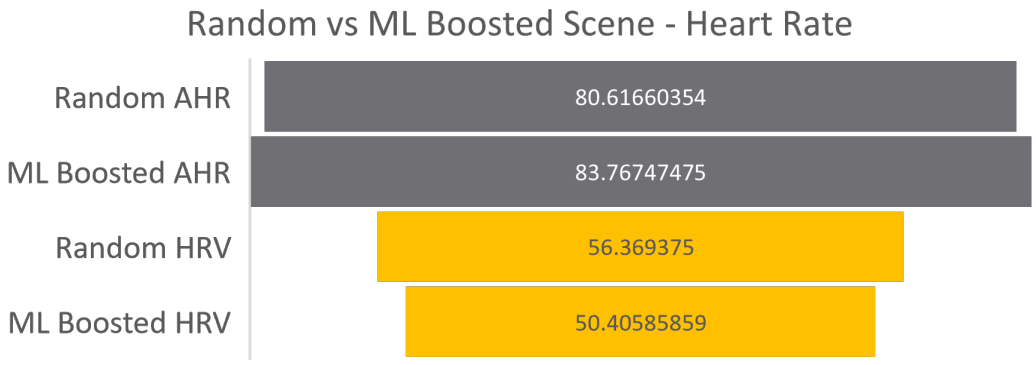


Figure 4.11: Wilcoxon rank-sum test for differences between average heart rate (AHR) and heart rate variability (HRV) for in-outdoor environment when ML-boosted.

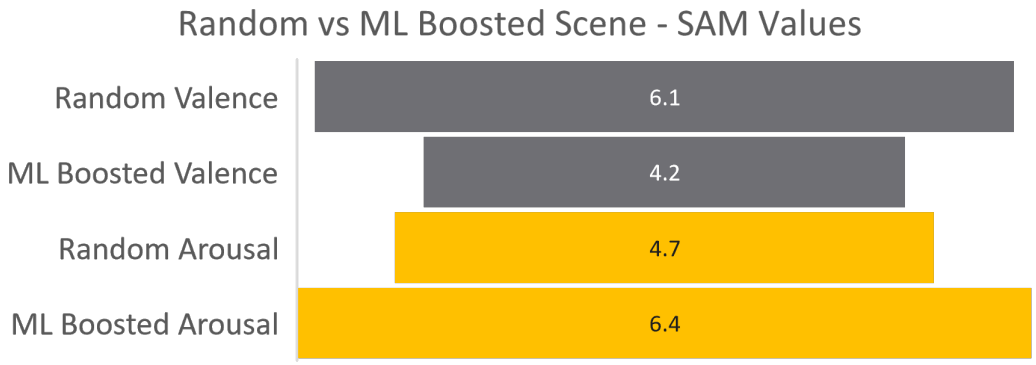


Figure 4.12: Wilcoxon rank-sum test for differences between SAM Valence and SAM Arousal for in-outdoor environment when ML-boosted.

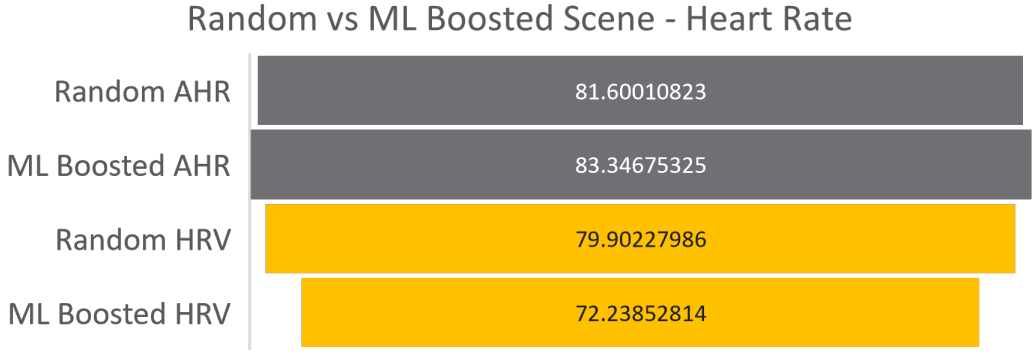


Figure 4.13: Wilcoxon rank-sum test for differences between average heart rate (AHR) and heart rate variability (HRV) for indoor environment when ML-boosted.

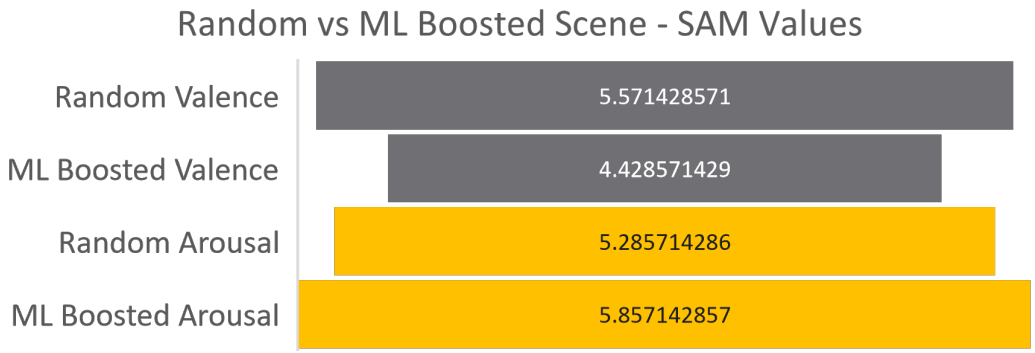


Figure 4.14: Wilcoxon rank-sum test for differences between SAM Valence and SAM Arousal for indoor environment when ML-boosted.

average heart rate and self-assessment manikin test values can be observed in Figure 4.15 and Figure 4.16 respectively. Correlation of average heart rate and heart rate variability for male and females participants with valence and arousal can be observed in Table 4.5.

Table 4.4: Friedman Test Results (* indicate significant results.)

	Average HR		**HR Variability**		**Valence**		**Arousal**	
	Statistic	**p**	**Statistic**	**p**	**Statistic**	**p**	**Statistic**	**p**
Male	1.2	0.55	0.4	0.82	4.72	0.09	1.12	0.57
Female	0.65	0.72	0.4	0.82	13.18	0.00*	9.09434	0.01*

Table 4.5: Gender Correlation

	Valence		**Arousal**	
	Average	**Variance**	**Average**	**Variance**
Male AHR	0.334618283	-0.03904	-0.98341	0.952802
Male HRV	0.928318523	1	-0.14218	0.265507
Female AHR	-0.642422138	-0.63114	0.739257	-0.50262
Female HRV	-0.08343127	0.994324	-0.05076	-0.25182
All AHR	-0.831111418	-0.47887	0.903001	-0.99398
All HRV	0.364873609	0.894349	-0.4961	0.878626

4.5.3 Individual Analysis

For individual subject analysis, Mann-Kendall test [190] was done to check whether there were trends in individual subject heart rate over three sessions.

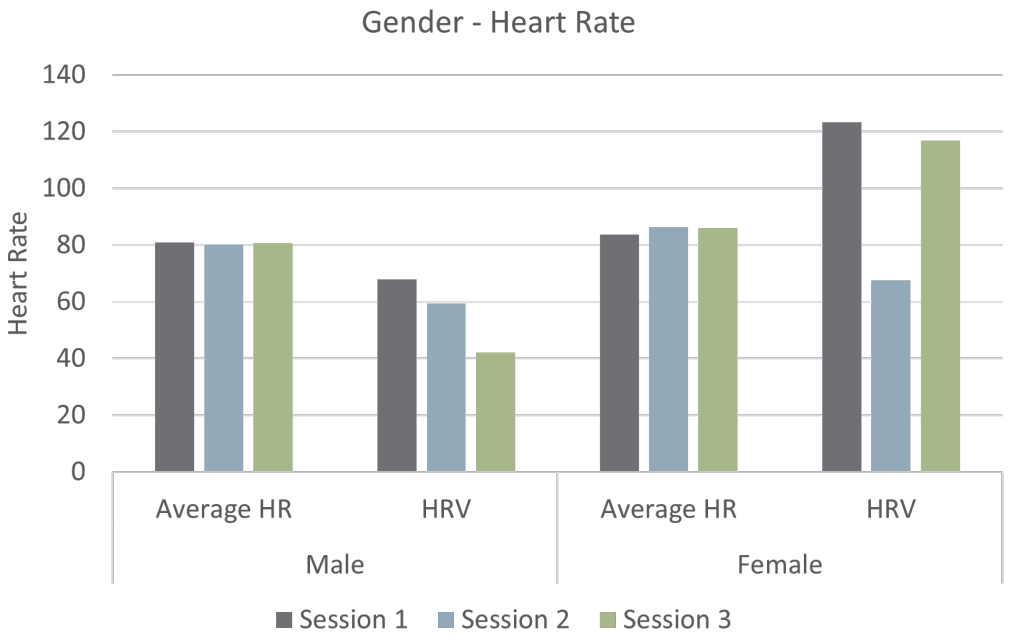


Figure 4.15: Average heart rate based on Gender.

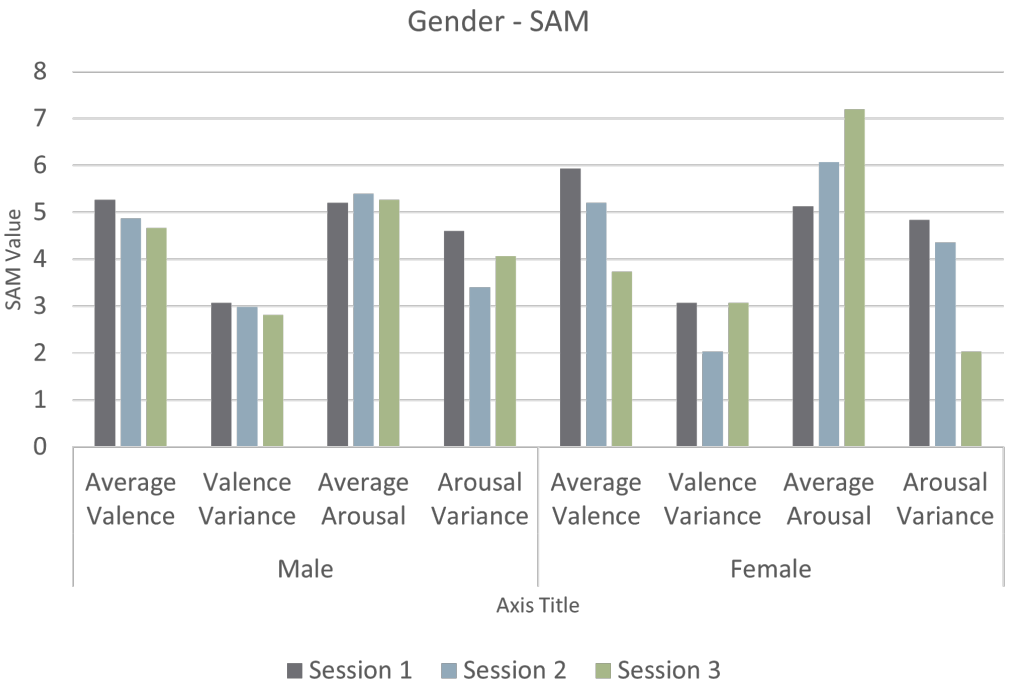


Figure 4.16: Average SAM Valence and Arousal based on Gender.

A high positive S value shows an increasing trend. A low negative S values shows a decreasing trend. The test results showed both increasing and decreasing trends for subjects. The results can be seen in Table 4.6.

4.5.4 Audio

To determine which feature of audio or category of audio impact personalized saliency of participants correlation of these features with heart rate was studied. However, no strong correlation was observed. These correlations can be observed in Table 4.7.

4.6 Discussion

In this research, we aimed to utilize heart rate from common consumer smart watches and machine learning to model personalized emotional saliency of the participants.

Individual heart rate trends show how for some participants in case of experiencing fear their heart rate increased whereas for others it decreased. We also saw differences based on gender. In case of audio, even though no strong correlation could be made between heart rate and sound features, no pattern could be devised either. Every participant was effected differently by different features of sounds. Therefore, we see more evidence that what grabs our attention, even based on emotions, is different for everyone.

Going back to the questions raised in subsection 1.5.1, we can analyze the results in section 4.5 to answer these questions.

Is heart rate from smart watch enough for emotion induction? Changing the lights and sounds setting of the virtual environment based on our saliency model, designed predict settings that would increase heart rate, we can say that heart rate was enough to induce fear in the participants. We saw a steady increase in arousal and a steady decline in valence over the three sessions. This indicates that fear was induced.

This experiment was only conducted to personalize the induction of fear, there have been other researches done that show the induction of positive emotions such as calmness [206].

How does replay effect personalized saliency? Average heart rate does not differ much for both male and female participants from session 1 to session 3 (Figure 4.15). For interactive task based games replay value increases due to better flow of the game with exposure and enjoyment remains stable [66].

Heart rate variance decreased for male subjects and fluctuated for female participants. However, this change was not statistically significant.

Self-assessment Manikin results show that even after replays, there was consistent decrease in valence and increase in arousal. This shows that replays did not impact the personalized emotional saliency negatively, participants experienced fear even after being exposed to the same environment more than once (Figure 4.16).

Does heart rate and self reported results correlate with each other? Strong correlation between the following pairs was observed for all participants

- Average heart rate and average valence (negative)
- Average heart rate and average arousal (positive)
- Heart rate variability and valence variance (positive)
- Heart rate variability and arousal variance (positive)
- Average heart rate and valence variance (negative)

Based on this, both average heart rate and heart rate variability correlate strongly with average and variance of self-assessment manikin test value respectively.

Based on these results, we can see that horror experience is very different for different individuals and heart rate obtained from player's own smart watch can be used by developers to utilize emotion research in their application. However, more research is required to get more reliable results for positive emotions. Combinations of indoor and outdoor experience can also be used to control the intensity of emotion induced.

4.6.1 Impact of Environment

How does environment impact heart rate and personalized saliency? In Table 4.3, we observe that for combination in-outdoor environment, the self reported results were statistically significant. In further analysis of the data from scene we had more observations. For all participants going from an indoor building structure to an open outdoor environment decreased their heart rate.

For participants going from outdoor to indoor there were differing behaviors. 20 participants experienced an increase in heart rate (13 participants

had increasing heart rate trend and 7 had decreasing heart rate trend), 5 participants experienced a decrease in heart rate (2 participants had increasing heart rate trend and 3 had decreasing heart rate trend). The remaining 5 participants did not go inside any indoor structure.

These results are in line with the current literature that suggests that nature settings help people to calm down [206].

4.6.2 Impact of Gender

How does gender impact personalized saliency? In subsection 2.3.2, we discussed how that although men and women experience emotions equally, women tend to be more expressive about them. In our results, average heart rate did not differ much for both male and female participants from session 1 to session 3 but heart rate variability did. Male participants showed decreased in heart rate variability but female participants had more fluctuating heart rate variability.

Based on self-assessment manikin results, female participants had statistically significant results indicating fear but male participants did not. Therefore, we can say that for at least female participants, our method of personalized saliency was successful but for male participants the results were not as significant.

In individual analysis it was observed that men had both equal number of increasing trend and decreasing trend in heart rate (7 male participants had increasing trend and 8 male participants had decreasing trend). However, for female participants increasing trend was more common (11 out 15 female participants had increasing trend, rest had decreasing trend).

Both genders had positive correlation between heart rate variability and valence variance but only male participants showed strong correlation between average heart rate and average arousal. These differences in results further show that just one factor of gender impacts how we feel emotions, especially fear.

Table 4.6: Mann-Kendall Test for Trend Detection (Experimental group: 1 to 20, Control Group: 21 to 30). * indicates significant results. M represents Male and F represents female.

Subject	Gender	Minimum HR	Maximum HR	Mean HR	Standard Deviation	Kendall's tau	S	Var(S)	p
1	M	58	97	72.41935	13.43571	0.579958	265	3442.333	6.28164E-06*
2	M	59	92	78.67857	7.911799	-0.3185	-117	2531.667	0.020054892*
3	M	51	82	71	7.709734	0.283544	90	2040	0.046301595*
4	M	50	93	75.3125	9.508697	-0.40664	-198	3778	0.001276005*
5	M	64	90	79.18182	4.798792	-0.08777	-44	4062.667	0.489996346
6	M	50	94	83.77778	10.55146	0.049576	17	2279.667	0.721801735
7	M	62	93	79.18182	7.903538	-0.18258	-94	4131.333	0.143616361
8	F	65	94	79.28125	7.292216	0.269966	132	3787.333	0.031960797*
9	F	72	98	84.90909	7.5472	0.032987	17	4127	0.791297976
10	F	51	92	74.96875	8.185291	-0.12979	-63	3775	0.305187007
11	F	51	96	79.875	10.95371	0.155435	76	3786.667	0.216811352
12	F	54	95	82.33333	7.824907	0.296887	153	4131	0.017290281*
13	M	53	91	79.48387	9.003225	0.087638	40	3440.667	0.495284312
14	M	71	98	86.67742	6.300196	-0.14022	-64	3440.667	0.275235384
15	M	62	94	82.46154	7.527182	-0.09361	-30	2048	0.507386527
16	F	77	99	89.13333	5.981197	-0.2652	-112	3106	0.04446978*
17	F	78	109	88	7.10152	0.041667	22	4142	0.744198995
18	F	52	98	76	9.619029	0.312605	186	4938	0.008471698*
19	F	78	111	91	7.055857	0.117424	62	4142	0.343222298
20	F	62	114	87	15.85595	0.655914	305	3455.667	0.000000232*
21	M	66	90	78.96552	6.94099	-0.51637	-206	2826.667	0.000106788*
22	M	75	109	95.28571	9.071147	0.050265	19	2547	0.721344124
23	M	68	111	79.6	9.107026	0.108046	-47	3120.333	0.410230472
24	F	65	148	100.6875	21.86976	-0.15726	-78	3794	0.211265851
25	F	72	113	87.28571	11.77838	0.261905	99	2555	0.052526995
26	F	62	105	81.51515	19.34412	-0.0625	-33	4155.667	0.619613956
27	M	67	115	86.75	17.39855	0.28629	142	3789.333	0.021990085*
28	M	67	110	82.85185	11.57817	0.133903	47	2292.333	0.336668368
29	F	60	107	84.36364	13.79497	0.486742	257	4153	7.11E-05*
30	F	71	113	93.16129	12.94887	0.490323	228	3452.667	0.000111906*

Table 4.7: Audio feature Correlation with heart rate. M represents male and F represents female.

Subject	Gender	Volume	Pitch	Stereopan	Sound	Sound Category
1	M	-0.30	-0.43	0.27	0.63	0.27
2	M	0.02	0.09	-0.12	-0.18	-0.14
3	M	0.42	0.06	0.10	0.12	-0.22
4	M	0.23	0.35	-0.04	-0.48	0.46
5	M	0.05	-0.41	-0.44	-0.10	-0.20
6	M	-0.19	-0.26	-0.13	-0.33	-0.31
7	M	-0.04	-0.02	-0.06	-0.13	-0.16
8	M	0.01	0.22	0.37	-0.01	-0.02
9	M	-0.30	-0.08	-0.29	0.02	0.14
10	M	-0.01	-0.07	-0.06	-0.24	-0.25
11	M	-0.07	0.16	-0.06	-0.14	-0.15
12	M	-0.51	-0.12	-0.09	-0.28	-0.01
13	M	-0.12	-0.36	0.02	0.18	0.11
14	M	0.03	0.00	-0.16	0.21	0.16
15	M	0.13	-0.15	-0.24	0.42	0.32
16	F	0.11	-0.11	-0.04	0.14	0.12
17	F	-0.12	-0.06	0.09	0.31	0.33
18	F	-0.32	-0.28	-0.01	0.39	0.47
19	F	0.09	-0.13	0.18	0.56	0.55
20	F	-0.28	-0.36	0.01	0.00	-0.05
21	F	-0.02	-0.11	-0.04	-0.41	-0.30
22	F	0.04	0.27	-0.07	-0.02	0.17
23	F	-0.14	0.03	-0.23	-0.12	-0.13
24	F	-0.32	-0.02	0.31	0.21	-0.28
25	F	-0.43	0.16	-0.21	0.63	0.54
26	F	0.16	-0.07	0.05	-0.16	0.00
27	F	0.46	0.18	0.32	0.43	0.42
28	F	0.04	0.00	0.07	-0.21	-0.20
29	F	-0.24	0.54	0.03	0.36	0.42
30	F	-0.18	0.22	-0.05	0.33	0.50

Chapter 5

Discussion and Conclusion

5.1 Applications of Personalized Saliency

Personalized Saliency in both non-immersive and immersive environments have a number of applications. A few of these practical applications have been discussed in the upcoming sections.

5.1.1 Non-Immersive Environments

- **Personalized user experience** With companies moving forward to provide as much personalized experience as possible, personalized saliency can aid in this quest. Based on the color preferences determined by personalized saliency model of individuals their phone or computer theme settings can be customized.

- **Image compression** Universal saliency has already been studied for image compression [85]. This can also be extended for personalized saliency by determining what area of the image would be salient to the user and compressing them less compared to non salient regions.

- **User Interface Design** User interface of websites and application can be designed to be dynamic in order to change according to the preferences and gaze behavior of the users. A user interface in tune with user exploration behavior can lead to better user experience. Visual saliency for mobile interfaces has been studied in [121].

- **Video Quality Compression** Similar to image compression, components of the video not salient to the user can be compressed more. This way

quality of what matters to the user can be of higher quality. A saliency guided video compression algorithm was introduced in [71].

- **Search Optimization** To make it easier for users to find certain buttons or text, personalized saliency help cater according to own unique gaze behavior [12].

5.1.2 Immersive Environments

- **Educational Therapy** For students that have trouble with social settings and concentrating in class, tools can be developed using virtual reality for combination of exposure therapy and emotional regulation. [202] verified that different emotions trigger learning and effect task performance in student using virtual reality. However, proper guidelines would be need to be set in place [80].

- **Exposure Therapy** Exposure therapy is a form of treatment for different kinds of phobia that works by slowly exposing the phobic person to their phobia and increasing their tolerance. Virtual reality with emotional personalized saliency can help monitor and use triggers in exposure therapy safely. [80] used virtual reality to help participants overcome their phobias with promising results.

- **Emotion Regulation** If certain social situations or other trigger points incite strong emotional reaction, this can be made better using an immersive environment such as virtual reality with virtual people to ensure safety of the other people and the one needing emotional regulation. Virtual reality based treatments for mental issues is an upcoming and rapidly flourishing field [68, 145].

- **Immersive Gaming Experience** In the past decade, there has been alot of interest in making games as immersive and as personal as possible. currently virtual reality provides the most immersive form of gaming available. with the utilization of emotional saliency, gaming experience can become even more affective [69].

- **Training** For training, that are otherwise dangerous, a personalized saliency equipped application can detect and induce right amount of emotions to train for different situations and how to deal with them before training in the real world [25].

- **Empathy Invocation** In social application, virtual reality has been used to make people more empathetic towards people, towards whom they would otherwise be prejudiced [201]. An extension of this application can help understand people in understanding the struggles of other people by living out and feeling their experience virtually.

- **Tourism and Architecture** Tour guides and museums can be designed with emotional saliency in mind to evoke different kinds of emotional impact in the visitors [226].

5.2 Limitations

- **Small personalized saliency data set** The personalized saliency prediction data set used for non-immersive environments includes only 1600 images. From saliency prediction perspective, this is not a very huge data set. A bigger data set would be better for research. however, if the suggested personalized saliency model is applied in actual applications it would be retraining itself periodically on new data increasing the data size.

- **Positive emotions not studied** Only negative emotion fear was studied in personalized saliency for immersive environments. a better understanding of emotional saliency would require studying of different kinds of emotions and determining if the same techniques can be applied to them.

- **Environmental study** The heart rate data frequency used in personalized saliency for immersive environments was in minutes. Since the participant would be moving constantly in virtual reality, the relationship between heart rate and the presence of environmental factors could not be studied. In the future, if heart rate can be obtained at smaller intervals, the study of environmental factors could be helpful in practical applications.

- **Real Life applications** This purpose of this thesis was to determine if personalized saliency could be predicted using our techniques in both immersive and non-immersive environments. Therefore, this research was fundamental in nature and the first step. The next step would be

to apply this knowledge in real life applications and determine if this knowledge when applied is successful and useful.

- **General Conclusions** To predict personalized saliency in non-immersive environments in the least invasive way possible, we only used color spaces and fixations of individuals. Therefore, our conclusions for this portion of the research are more general and individualistic. In the future, with more personal information, more detailed conclusions and observations could be drawn.

5.3 Conclusion and Future Work

In this thesis, we provide a comprehensive study on personalized saliency in both non-immersive and immersive environments with keeping practical applications in mind. We propose using a visual saliency based personalized saliency model for non-immersive and emotional saliency based personalized saliency model for immersive environments. For both environments, we try to achieve personalized saliency in a way that least invasive in terms of gathering personal information. We also present our outlook on how there is a gap between saliency related research and its applications in the real world.

With our personalized saliency model in non-immersive environments we show that using just color spaces of pixels where a person looks we can predict where the user will pay attention. This method also achieve good accuracy without the use of computation and storage heavy deep learning models that are not feasible for mobile applications.

Furthermore, we present our discussion on how image content and individual gaze behavior effects personalized saliency. Cluttered images show poor performance and images with more target contrast have better performance. Subjects with higher fixation and saccade also showed poor comparison.

With our personalized saliency model in immersive environments, we utilize emotional saliency to understand user behavior. We show how with normal consumer smartwatch and machine learning algorithms, emotions can be induced to create applications that will be of benefit to the society. We show how every individual has different effects on their heart rate when they experience emotions such as fear based on their environment exploration behavior and gender. These findings were supported by results. We also observed strong correlation between heart rate and self-reported results.

In conclusion, the idea that saliency of any kind, in any kind of environment is heterogeneous is further reinforced from visual, emotional and auditory saliency perspective. For applications that want to utilize saliency need to take note of this consideration. This research was fundamental in nature and was done to determine if our proposed techniques can actually determine personalized saliency. The next step would be to apply this knowledge and techniques in actual applications with greater sample size for more definite conclusions.

In the future, personalized saliency model for non-immersive environments model can be extended for use in real-time videos, animations, virtual reality and augmented reality. Similarly, personalized saliency model for non-immersive environments model can be extended for study and use regarding positive emotions, mental health services, training's and gaming. Since personalized saliency is such a new and developing topic, the possibilities could be endless.

Appendix A

Model selection and performance evaluation for Personalized Saliency in Non-immersive Environments

Since testing and evaluating results for 1600 images from personalized saliency data set corresponding to one subject took around 32 hours, we decided to test out different regression models initially with 100 images rather than all 1600. From all the different regression models, three models gave the highest and almost similar results.

First a simple neural network with four hidden layers, 100 neurons in each hidden layer and 100 iterations. Second, a deep learning neural network model with two fully connected layers containing 18 and 20 outputs respectively. Both layers had the learning rate of 0.5 and used XAVIER as weight initialization strategy with LeakyReLU as activation function along with Stochastic Gradient Descent as the optimization algorithm. For the output layer the learning rate was set to 0.1 with RELU as weight initialization strategy and Mean Squared Error for the loss function. Finally, for the gradient boosted regression model the tree depth was 4, the number of models was 100, the alpha (percentage of data not treated as outlier) was 0.96 and the learning rate was 0.1.

Since all three of these regression models produced quite similar results shown in Figure A.1 and Table A.2 , we chose gradient boosted tree regression because it gave highest average AUC Judd score. Also, 15 out 30 models with highest score were gradient boosted models with two models having the same

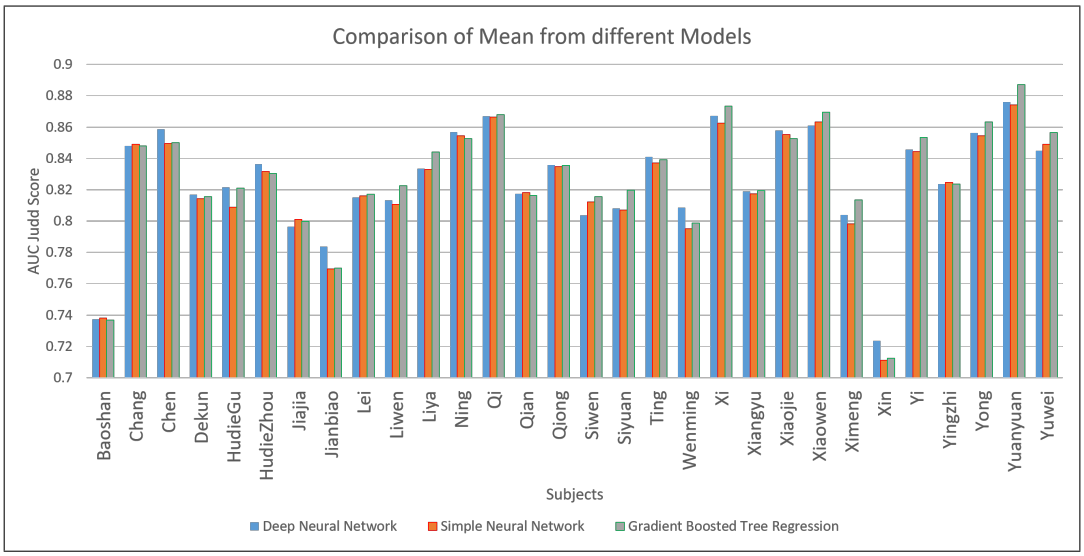


Figure A.1: The average AUC Judd score obtained using Deep Neural Network, Simple Neural Network and Gradient Boosted Tree Regression [224].

Method	Average AUC Judd
Deep NN	0.826
Simple NN	0.823
Gradient Boosted	0.828

Table A.1: Average AUC Judd scores for all subjects [224].

score as the deep neural network models. Therefore, using the same parameters as before gradient boosted tree regression model was then tested for all 1600 images. The models for all subjects had R^2 value of 1, the Mean Absolute Error was either 0.002 or 0.003 and the Root Mean Squared Error was either 0.004 or 0.005 for all the subjects.

Method	Average AUC Judd	Time (ms)
Deep NN	0.826	24105.8
Simple NN	0.823	453482.7
Gradient Boosted	0.828	6452.2

Table A.2: Average AUC Judd scores and average training time (milliseconds) for all subjects [224].

References

[1] Tensorflow lite: Ml for mobile and edge devices.

[2] *The philosophy of horror, or, Paradoxes of the heart*. Routledge, 1990.

[3] Delivering real-time ai in the palm of your hand, Jun 2018.

[4] Examining the effects of clutter and target salience in an e-commerce visual search task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1):1761–1765, 2019.

[5] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34, 05 2012.

[6] Ralph Adolphs. The biology of fear. *Current biology*, 23(2):R79–R93, 2013.

[7] Ralph Adolphs, Daniel Tranel, and Tony Buchanan. Amygdala damage impairs memory for gist but not details of complex stimuli. *Nature neuroscience*, 8:512–8, 05 2005.

[8] A. Deniz Aladagli, Erhan Ekmekcioglu, Dmitri Jarnikov, and Ahmet Kondoz. Predicting head trajectories in 360° virtual reality videos. In *2017 International Conference on 3D Immersion (IC3D)*, pages 1–6, 2017.

[9] Dilara Albayrak, Mehmet Bahadir Askin, Tolga K. Capin, and Ufuk Celikcan. Visual saliency prediction in dynamic virtual reality environments experienced with head-mounted displays: An exploratory study. In *2019 International Conference on Cyberworlds (CW)*, pages 61–68, 2019.

[10] Dilara Albayrak, Tolga K Çapın, Ufuk Çelikcan, and Mehmet Bahadır Askın. Deep into visual saliency for immersive vr environments rendered in real-time. 2020.

[11] Frans A Albersnagel. Velten and musical mood induction procedures: A comparison with accessibility of thought associations. *Behaviour research and therapy*, 26(1):79–95, 1988.

[12] Eshaa Alkhalifa, Khulood Gaid, Jeremiah Still, and Christopher Masciocchi. *Considering the Influence of Visual Saliency during Interface Searches*, pages 84–97. 01 2012.

[13] Allison P Anderson, Michael D Mayer, Abigail M Fellows, Devin R Cowan, Mark T Hegel, and Jay C Buckey. Relaxation with immersive natural scenes presented using virtual reality. *Aerospace medicine and human performance*, 88(6):520–526, 2017.

[14] Allison P Anderson, Michael D Mayer, Abigail M Fellows, Devin R Cowan, Mark T Hegel, and Jay C Buckey. Relaxation with immersive natural scenes presented using virtual reality. *Aerospace medicine and human performance*, 88(6):520—526, June 2017.

[15] Hayato Araki, Taichi Ikeda, Takumi Ozawa, Kenta Kawahara, and Yasuo Kawai. Development of a horror game that route branches by the player's pulse rate. In *Proc. Intelligent User Interfaces Companion*, IUI '18 Companion, New York, NY, USA, 2018. Association for Computing Machinery.

[16] Mehmet Bahadir Askin and Ufuk Celikcan. Learning based versus heuristic based: A comparative analysis of visual saliency prediction in immersive virtual reality. *Computer Animation and Virtual Worlds*, n/a(n/a):e2106.

[17] Alan Baddeley. *Working memory, thought, and action*, volume 45. OuP Oxford, 2007.

[18] Shumeet Baluja and Dean A. Pomerleau. Expectation-based selective attention for visual monitoring and control of a robot vehicle. *Robotics Auton. Syst.*, 22:329–344, 1997.

[19] Amin Banitalebi Dehkordi, Mahsa Pourazad, and Panos Nasiopoulos. A learning-based visual saliency prediction model for stereoscopic 3d video (lbvs-3d). *Multimedia Tools and Applications*, 76, 11 2017.

[20] Lisa Feldman Barrett, Batja Mesquita, Kevin N Ochsner, and James J Gross. The experience of emotion. *Annual review of psychology*, 58:373, 2007.

[21] Penny Bergman, Daniel Västfjäll, Ana Tajadura-Jiménez, and Erkin Asutay. Auditory-induced emotion mediates perceptual categorization of everyday sounds. *Frontiers in Psychology*, 7:1565, 09 2016.

[22] Sarah V Biedermann, Daniel G Biedermann, Frederike Wenzlaff, Tim Kurjak, Sawis Nouri, Matthias K Auer, Klaus Wiedemann, Peer Briken, Jan Haaker, Tina B Lonsdorf, et al. An elevated plus-maze in mixed reality for studying human anxiety-related behavior. *BMC biology*, 15(1):1–13, 2017.

[23] Ali Borji. Boosting bottom-up and top-down visual features for saliency estimation. pages 438–445, 06 2012.

[24] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research, 2015.

[25] Tibor Bosse, Charlotte Gerritsen, Jeroen de Man, and Jan Treur. Towards virtual training of emotion regulation. *Brain Informatics*, 1:27–37, 12 2014.

[26] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[27] Albert Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*, volume 95. 01 1990.

[28] Leo Breiman. Machine learning. *Machine Learning*, 45:5–32, 10 2001.

[29] Tobias Brosch, Gilles Pourtois, and David Sander. The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion*, 24(3):377–400, 2010.

[30] Jos Brosschot and Julian Thayer. Heart rate response is longer after negative emotions than after positive emotions. *International journal of*

psychophysiology : official journal of the International Organization of Psychophysiology, 50:181–7, 12 2003.

[31] Alafair S. Burke, Friderike Heuer, and Daniel Reisberg. Remembering emotional events. *Memory & Cognition*, 20:277–290, 1992.

[32] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[33] Eunhee Chang, Hyun-Taek Kim, and Byounghyun Yoo. Virtual reality sickness: A review of causes and measurements. *International Journal of Human-Computer Interaction*, 36:1–25, 07 2020.

[34] Fang-Yi Chao, Lu Zhang, Wassim Hamidouche, and Olivier Déforges. Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks. 07 2018.

[35] Chenglizhao Chen, Yunxiao Li, Shuai Li, Hong Qin, and Aimin Hao. A novel bottom-up saliency detection method for video with dynamic background. *IEEE Signal Processing Letters*, 25(2):154–158, 2018.

[36] Kwang-Ho Choi, Junbeom Kim, O. Sang Kwon, Min Ji Kim, Yeon Hee Ryu, and Ji-Eun Park. Is heart rate variability (hrv) an adequate tool for evaluating human emotions? – a focus on the use of the international affective picture system (iaps). *Psychiatry Research*, 251:192–196, 2017.

[37] Sven Christianson, Elizabeth Loftus, Hunter Hoffman, and Geoffrey Loftus. Eye fixations and memory for emotional events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17:693–701, 07 1991.

[38] Joanne Chung, Henry So, Marcy Choi, Chun Man Vincent Yan, and Thomas Wong. Artificial intelligence in education: Using heart rate variability (hrv) as a biomarker to assess emotions objectively. *Computers and Education: Artificial Intelligence*, 2:100011, 01 2021.

[39] James A Coan, John JB Allen, et al. *Handbook of emotion elicitation and assessment*. Oxford university press, 2007.

[40] Rebecca Compton. The interface between emotion and attention: A review of evidence from psychology and neuroscience. *Behavioral and cognitive neuroscience reviews*, 2:115–29, 07 2003.

[41] Yixiang Dai, Xue Wang, Pengbo Zhang, Weihang Zhang, and Chen Junfeng. Sparsity constrained differential evolution enabled feature-channel-sample hybrid selection for daily-life eeg emotion recognition. *Multimedia Tools and Applications*, 77, 09 2018.

[42] Tung Dang, Christos Papachristos, and Kostas Alexis. Visual saliency-aware receding horizon autonomous exploration with application to aerial robotics. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2526–2533, 2018.

[43] Delmotte and Varinthira Duangudom. *Computational auditory saliency*. PhD thesis, Georgia Institute of Technology, 2012.

[44] Tao Deng, Kai-Fu Yang, Yongjie Li, and Hongmei Yan. Where does the driver look? top-down-based saliency detection in a traffic driving environment. *IEEE Transactions on Intelligent Transportation Systems*, 17:1–12, 03 2016.

[45] Xinmiao Ding, Lulu Huang, Bing Li, Congyan Lang, Zhen Hua, and Yuling Wang. A novel emotional saliency map to model emotional attention mechanism. In Qi Tian, Nicu Sebe, Guo-Jun Qi, Benoit Huet, Richang Hong, and Xueliang Liu, editors, *MultiMedia Modeling*, pages 197–206, Cham, 2016. Springer International Publishing.

[46] Cong-Thanh Do and Yannis Stylianou. Weighting time-frequency representation of speech using auditory saliency for automatic speech recognition. 09 2018.

[47] Michael Dorr, Thomas Martinetz, Karl R Gegenfurtner, and Erhardt Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, 10(10):28–28, 2010.

[48] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified image and video saliency modeling. In *Computer Vision – ECCV 2020*, pages 419–435. Springer International Publishing, 2020.

[49] Varinthira Duangudom and David V. Anderson. Using auditory saliency to understand complex auditory scenes. In *2007 15th European Signal Processing Conference*, pages 1206–1210, 2007.

[50] Jakob-Moritz Eberl, Petro Tolochko, Pablo Jost, Tobias Heidenreich, and Hajo Boomgaarden. What's in a post? how sentiment and issue salience affect users' emotional reactions on facebook. *Journal of Information Technology Politics*, 17, 01 2020.

[51] Kira Eghbal-Azar and Thomas Widlok. Potentials and limitations of mobile eye tracking in visitor studies evidence from field research at two museum exhibitions in germany. *Social Science Computer Review*, 31:103–118, 02 2013.

[52] Kristen Ellard, Todd Farchione, and Barlow David. Relative effectiveness of emotion induction procedures and the role of personal relevance in a clinical sample: A comparison of film, images, and music. *Journal of Psychopathology and Behavioral Assessment*, 34, 06 2011.

[53] Léa Entzmann, Nathalie Guyader, Louise Kauffmann, Juliette Lenouvel, Clémence Charles, Carole Peyrin, Roman Vuillaume, and Martial Mermillod. The role of emotional content and perceptual saliency during the programming of saccades toward faces. *Cognitive Science*, 45(10):e13042, 2021.

[54] Erkut Erdem and Aykut Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, 13(4):11–11, 2013.

[55] Shabnam Fakhrhosseini and Myounghoon Jeon. *Affect/Emotion Induction Methods*, pages 235–253. 12 2017.

[56] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[57] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[58] Shu Fang, Jia Li, Yonghong Tian, Tiejun Huang, and Xiaowu Chen. Learning discriminative subspaces on random contrasts for image saliency analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 28(5):1095–1108, 2017.

[59] Luz Fernández-Aguilar, José Latorre, Arturo Martinez Rodrigo, Jose Moncho-Bogani, Laura Ros, Pablo Latorre Doménech, Jorge Ricarte, and Antonio Fernández-Caballero. Differences between young and older adults in physiological and subjective responses to emotion induction using films. *Scientific Reports*, 10:14548, 09 2020.

[60] Karlo Filipan, Bert De Coensel, Pierre Aumond, Arnaud Can, Catherine Lavandier, and Dick Botteldooren. Auditory sensory saliency as a better predictor of change than sound amplitude in pleasantness assessment of reproduced urban soundscapes. *Building and Environment*, 148:730–741, 2019.

[61] Agneta Fischer, P. Rodriguez Mosquera, Annelies van Vianen, and Antony Manstead. Gender and culture differences in emotion. *Emotion (Washington, D.C.)*, 4:87–94, 04 2004.

[62] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[63] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[64] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[65] Ascensión Gallardo-Antolín and Juan M. Montero. An auditory saliency pooling-based lstm model for speech intelligibility classification. *Symmetry*, 13(9), 2021.

[66] Sunil Gandhi, David Heeger, and Geoffrey Boynton. Spatial attention affects brain activity in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 96:3314–9, 04 1999.

[67] Zhifan Gao, Heye Zhang, Shizhou Dong, Shanhui Sun, Xin Wang, Guang Yang, Wanqing Wu, Shuo Li, and Victor Hugo C de Albuquerque. Salient

object detection in the distributed cloud-edge intelligent network. *IEEE Network*, 34(2):216–224, 2020.

[68] Azucena Garcia-Palacios, Cristina Botella, Rosa Baños, Verónica Botella, and Marivi Navarro Haro. Inclusion of virtual reality: A rationale for the use of vr in the treatment of ptsd. *Future Directions in Post-Traumatic Stress Disorder: Prevention, Diagnosis, and Treatment*, pages 275–287, 01 2015.

[69] Sarra Graja, Phil Lopes, and Guillaume Chanel. Impact of visual and sound orchestration on physiological arousal and tension in a horror game. *IEEE Transactions on Games*, 13(3):287–299, 2021.

[70] James J Gross and Robert W Levenson. Emotion elicitation using films. *Cognition & emotion*, 9(1):87–108, 1995.

[71] Rupesh Gupta, Meera Khanna, and Santanu Chaudhury. Visual saliency guided video compression algorithm. *Signal Processing: Image Communication*, 28:1006–1022, 10 2013.

[72] Dirk Hagemann, Ewald Naumann, Stefanie Maier, Gabriele Becker, Alexander Lürken, and Dieter Bartussek. The assessment of affective reactivity using films: Validity, reliability and sex differences. *Personality and Individual differences*, 26(4):627–639, 1999.

[73] Eddie Harmon-Jones, David Amodio, and L.R. Zinner. *Social psychological methods in emotion elicitation*, pages 91–105. Oxford University Press, 2007.

[74] Dwi Hartanto, Isabel Kampmann, Nexhmedin Morina, Paul Emmelkamp, Mark Neerincx, and Willem-Paul Brinkman. Controlling social stress in virtual reality environments. *PloS one*, 9:e92804, 03 2014.

[75] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9:188–94, 05 2005.

[76] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 06 2016.

[77] Shengfeng He, Rynson W.H. Lau, and Qingxiong Yang. Exemplar-driven top-down saliency detection via deep association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[78] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[79] Jila Hosseinkhani and Chris Joslin. Significance of bottom-up attributes in video saliency detection without cognitive bias. In *2018 IEEE 17th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, pages 606–613, 2018.

[80] Hsiu-Mei Huang, Ulrich Rauch, and Shu-Sheng Liaw. Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach. *Computers Education*, 55:1171–1182, 11 2010.

[81] Qiuyun Huang, Mingyan Yang, Hao-ann Jane, Shuhua Li, and Nicole Bauer. Trees, grass, or concrete?the effects of different types of environments on stress reduction. *Landscape and Urban Planning*, 193, 10 2019.

[82] Xun-Yi Huang, Fu-Yin Cherng, Jung-Tai King, and Wen-Chieh Lin. Eeg-based measures of auditory saliency in a complex context. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA, 2019. Association for Computing Machinery.

[83] Minami Ito and Charles D. Gilbert. Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron*, 22(3):593–604, 1999.

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

[85] Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE transactions on image processing*, 13(10):1304–1318, 2004.

[86] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001.

[87] Kinga Izsóf Jurásová and Marián Špajdel. Development and assessment of film excerpts used for emotion elicitation. *Activitas Nervosa Superior Rediviva*, 55:135–140, 01 2013.

[88] Robert J, Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005.

[89] Charlene Jennett, Anna Cox, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. Measuring and defining the experience of the immersion in games. *International Journal of Human-Computer Studies*, 66:641–661, 09 2008.

[90] Charlene Jennett, Anna L. Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, 66(9):641–661, 2008.

[91] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[92] Guillaume A.RousseletSimon J.ThorpeMichèleFabre-Thorpe. Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Research*, 44:877 – 894, 2004.

[93] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 01 2012.

[94] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. pages 2106–2113, 09 2009.

[95] Lukasz D. Kaczmarek, Maciej Behnke, Jolanta Enko, Michał Kosakowski, Brian M. Hughes, Jaroslaw Piskorski, and Przemysław Guzik. Effects of emotions on heart rate asymmetry. *Psychophysiology*, 56(4):e13318, 2019.

[96] Christopher Kanan, Mathew H. Tong, Lingyun Zhang, and Garrison W. Cottrell. Sun: Top-down saliency using natural statistics. *Visual Cognition*, 17(6-7):979–1003, 2009. PMID: 21052485.

[97] Rachel Kaplan and Stephen Kaplan. *The experience of nature: A psychological perspective*. Cambridge university press, 1989.

[98] Christoph Kayser, Christopher Petkov, Michael Lippert, and Nikos Logothetis. Mechanisms for allocating auditory attention: An auditory saliency map. *Current biology : CB*, 15:1943–7, 12 2005.

[99] Elizabeth A Kensinger. Remembering the details: Effects of emotion. *Emotion review*, 1(2):99–113, 2009.

[100] Nazar Khan and Marshall Tappen. Discriminative dictionary learning with spatial priors. pages 166–170, 09 2013.

[101] Young Kim, JunYoung Moon, Nak-Jun Sung, and Min Hong. Correlation between selected gait variables and emotion using virtual reality. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–8, 2019.

[102] Joanna Kisker, Thomas Gruber, and Benjamin Schöne. Behavioral realism and lifelike psychophysiological responses in virtual reality by the example of a height exposure. *Psychological research*, 85(1):68–81, 2021.

[103] Joanna Kisker, Thomas Gruber, and Benjamin Schöne. Behavioral realism and lifelike psychophysiological responses in virtual reality by the example of a height exposure. *Psychological Research*, 85, 02 2021.

[104] E. Klinger, S. Bouchard, P. Légeron, S. Roy, F. Lauer, I. Chemin, and P. Nugues. Virtual reality therapy versus cognitive behavior therapy for social phobia: A preliminary controlled study. *CyberPsychology & Behavior*, 8(1):76–88, 2005. PMID: 15738695.

[105] Marisa Knight and Mara Mather. Reconciling findings of emotion-induced memory enhancement and impairment of preceding items. *Emotion*, 9(6):763, 2009.

[106] Aysun Kocak, Kemal Cizmeciler, Aykut Erdem, and Erkut Erdem. Top down saliency estimation via superpixel-based discriminative dictionaries. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[107] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4 4:219–27, 1985.

[108] Kathryn Koehler, Fei Guo, Sheng Zhang, and Miguel Eckstein. What do saliency models predict? *Journal of vision*, 14, 03 2014.

[109] Ann Kring and Albert Gordon. Sex differences in emotion: Expression, experience, and physiology. *Journal of personality and social psychology*, 74:686–703, 03 1998.

[110] Ann Kring and Albert Gordon. Sex differences in emotion: Expression, experience, and physiology. *Journal of personality and social psychology*, 74:686–703, 03 1998.

[111] Onkar Krishna, Andrea Helo, Rämä P, and Kiyoharu Aizawa. Gaze distribution analysis and saliency prediction across age groups. *PlosOne*, PLOS ONE, 13(2): e019., 02 2018.

[112] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[113] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/.

[114] Matthias Kummerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[115] Fanny Lantz. Exploring the impact of familiarity on the emotional response to acousmatic sound effects in horror film, 2021.

[116] Fanny Lantz. Exploring the impact of familiarity on the emotional response to acousmatic sound effects in horror film, 2021.

[117] Pierre Lebreton and Alexander Raake. Gbvs360, bms360, prosal: Extending existing saliency prediction models from 2d to omnidirectional images. *Signal Processing: Image Communication*, 69, 03 2018.

[118] Jerry Lee, Patricia Jones, Yoshimitsu Mineyama, and Esther Zhang. Cultural differences in responses to a likert scale. *Research in nursing health*, 25:295–306, 08 2002.

[119] Su-In Lee and Soo-Young Lee. Top-down attention control at feature space for robust pattern recognition. In Seong-Whan Lee, Heinrich H. Bülthoff, and Tomaso Poggio, editors, *Biologically Motivated Computer Vision*, pages 129–138, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[120] Tai Lee and Stella X. Yu. An information-theoretic framework for understanding saccadic eye movements. *Advances in Neural Information Processing Systems*, 09 2002.

[121] Luis A. Leiva, Yunfei Xue, Avya Bansal, Hamed R. Tavakoli, Tuçe Körolu, Jingzhou Du, Niraj R. Dayama, and Antti Oulasvirta. Understanding visual saliency in mobile user interfaces. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '20, New York, NY, USA, 2020. Association for Computing Machinery.

[122] Shenda Li, Jin Wang, and Qing Zhu. High dynamic range image compression based on visual saliency. In *2018 Picture Coding Symposium (PCS)*, pages 21–25, 2018.

[123] Sikun Lin and Pan Hui. Where's YOUR focus: Personalized attention. *CoRR*, abs/1802.07931, 2018.

[124] Yi Liu, Jungong Han, Qiang Zhang, and Long Wang. Salient object detection via two-stage graphs. *IEEE Transactions on Circuits and Systems for Video Technology*, PP:1–1, 04 2018.

[125] Jill Lobbestael, Arnoud Arntz, and Reinout W. Wiers. How to push someone's buttons: A comparison of four anger-induction methods. *Cognition and Emotion*, 22(2):353–373, 2008.

[126] Zhongkang Lu, W. Lin, X. Yang, EePing Ong, and Susu Yao. Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. *IEEE Transactions on Image Processing*, 14(11):1928–1942, 2005.

[127] Lars-Olov Lundqvist, Fredrik Carlsson, Per Hilmersson, and Patrik Juslin. Emotional responses to music: Experience, expression, and physiology. *Psychology of Music*, 37:61–90, 01 2009.

[128] Xinhui Luo, Zhi Liu, Weijie Wei, Linwei Ye, Tianhong Zhang, Lihua Xu, and Jijun Wang. Few-shot personalized saliency prediction using meta-learning. *Image and Vision Computing*, page 104491, 2022.

[129] Kyle E. Madsen. The differential effects of agency on fear induction using a horror-themed video game. *Computers in Human Behavior*, 56:142–146, 2016.

[130] Ali Mahdi and Jun Qin. Deepfeat: A bottom up and top down saliency model based on deep features of convolutional neural nets. *IEEE Transactions on Cognitive and Developmental Systems*, PP, 09 2017.

[131] Ali Mahdi and Jun Qin. Bottom up saliency evaluation via deep features of state-of-the-art convolutional neural networks. In *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 247–250, 2018.

[132] Gantz Matan. *Analysis of the Horror Genre and its Implementation in Video Games*. PhD thesis, TH Köln University of Applied Sciences, 06 2019.

[133] Gantz Matan. *Analysis of the Horror Genre and its Implementation in Video Games*. PhD thesis, TH Köln University of Applied Sciences, 06 2019.

[134] Mara Mather. Emotional arousal and memory binding: An object-based framework. *Perspectives on Psychological Science*, 2(1):33–52, 2007.

[135] Gerald Matthews and Kirby Gilliland. The personality theories of h.j. eysenck and j.a. gray: a comparative review. *Personality and Individual Differences*, 26(4):583 – 626, 1999.

[136] Carrie J. McAdams and John H. R. Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. *Journal of Neuroscience*, 19(1):431–441, 1999.

[137] Ruggero Milanese, Sylvia Gil, and Thierry Pun. Attentive mechanisms for dynamic and static scene analysis. *Optical Engineering*, 34:2428–2434, 1995.

[138] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljosa Smolic. Salnet360: Saliency maps for omni-directional images with cnn. *Signal Processing: Image Communication*, 69, 09 2017.

[139] Jeffrey Moran and Robert Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–784, 1985.

[140] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Few-shot personalized saliency prediction using person similarity based on collaborative multi-output gaussian process regression. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1469–1473, 2021.

[141] B. C. Motter. Focal attention produces spatially selective processing in visual cortical areas v1, v2, and v4 in the presence of competing stimuli. *Journal of Neurophysiology*, 70(3):909–919, 1993. PMID: 8229178.

[142] Vb Mountcastle, Ra Andersen, and Brad Motter. The inuence of attentive fixation upon the excitability of the light-sensitive neurons of the posterior parietal cortex. 09 2013.

[143] Manon Mulckhuyse. The influence of emotional stimuli on the oculomotor system: A review of the literature. *Cognitive, Affective, & Behavioral Neuroscience*, 18(3):411–425, 2018.

[144] Mimma Nardelli, Gaetano Valenza, Alberto Greco, Antonio Lanata, and Enzo Pasquale Scilingo. Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Transactions on Affective Computing*, 6(4):385–394, 2015.

[145] Marivi Navarro Haro, Marta Modrego-Alarcón, Hunter Hoffman, Alba López-Montoyo, Mayte Navarro-Gil, Jesus Montero-Marin, Azucena Garcia-Palacios, Luis Borao-Zabala, and Javier Garcia-Campayo. Evaluation of a mindfulness-based intervention with and without virtual reality dialectical behavior therapy® mindfulness skills training for the treatment of generalized anxiety disorder in primary care: A pilot study. *Frontiers in Psychology*, 10:55, 01 2019.

[146] Mark B. Neider and Gregory J. Zelinsky. Scene context guides eye movements during visual search. *Vision Research*, 46(5):614–621, 2006.

[147] Pm Nishad and R Chezian. Various colour spaces and colour space conversion algorithms. *Journal of Global Research in Computer Science*, 4:44–48, 01 2013.

[148] Samuel S. Ogden and Tian Guo. Characterizing the deep neural networks inference performance of mobile applications, 2019.

[149] Aude Oliva, Antonio Torralba, Monica Castelhano, and John Henderson. Top-down control of visual attention in object detection. volume 1, pages 253–256, 01 2003.

[150] Wilfried Osberger and Ann M. Rohaly. Automatic detection of regions of interest in complex video sequences. In *Human Vision and Electronic Imaging*, 2001.

[151] Li-Chen Ou, Ming Luo, Andrée Woodcock, and Angela Wright. A study of colour emotion and colour preference. part i: Colour emotions for single colours. *Color Research Application*, 29:232 – 240, 06 2004.

[152] Li-Chen Ou, Ming Luo, Andrée Woodcock, and Angela Wright. A study of colour emotion and colour preference. part ii: Colour emotions for two-colour combinations. *Color Research Application*, 29:292 – 298, 05 2004.

[153] Junting Pan, Cristian Canton, Kevin McGuinness, Noel O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giró-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. 01 2017.

[154] Junting Pan, Elisa Sayrol, Xavier Giró-i Nieto, Kevin McGuinness, and Noel OConnor. Shallow and deep convolutional networks for saliency prediction. pages 598–606, 06 2016.

[155] Jaak Panksepp and Günther Bernatzky. 'emotional sounds and the brain: the neuro-affective foundations of musical appreciation'. *Behavioural processes*, 60:133–155, 12 2002.

[156] Thomas Parsons and Albert Rizzo. Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of behavior therapy and experimental psychiatry*, 39:250–61, 10 2008.

[157] Jennifer N Perusini and Michael S Fanselow. Neurobehavioral perspectives on the distinction between fear and anxiety. *Learning & Memory*, 22(9):417–425, 2015.

[158] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[159] David Pollreisz and Nima TaheriNejad. A simple algorithm for emotion recognition, using physiological signals of a smart watch. In *2017 39th annual international conference of the ieee engineering in medicine and biology society (EMBC)*, pages 2353–2356. IEEE, 2017.

[160] Yu Qiu, Yun Liu, Hui Yang, and Jing Xu. A simple saliency detection approach via automatic top-down feature fusion. *Neurocomputing*, 388:124–134, 2020.

[161] Juan C. Quiroz, Min Hooi Yong, and Elena Geangu. Emotion-recognition using smart watch accelerometer data: Preliminary findings. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, page 805–812, New York, NY, USA, 2017. Association for Computing Machinery.

[162] Mattia Rainoldi, Barbara Neuhofer, and Mario Jooss. Mobile eyetracking of museum learning experiences. pages 473–485, 01 2018.

[163] Anne Richards and Isabelle Blanchette. Independent manipulation of emotion in an emotional stroop task using classical conditioning. *Emotion (Washington, D.C.)*, 4:275–81, 10 2004.

[164] B. J. Richmond, R. H. Wurtz, and T. Sato. Visual responses of inferior temporal neurons in awake rhesus monkey. *Journal of Neurophysiology*, 50(6):1415–1432, 1983. PMID: 6663335.

[165] D. L. Robinson, M. E. Goldberg, and G. B. Stanton. Parietal association cortex in the primate: sensory mechanisms and behavioral modulations. *Journal of Neurophysiology*, 41(4):910–932, 1978. PMID: 98614.

[166] Michael H. Robinson. Predator-prey interactions, informational complexity, and the origins of intelligence. *Journal of the Washington Academy of Sciences*, 75(4), 1985.

[167] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[168] Elisa Scheller, Christian Büchel, and Matthias Gamer. Diagnostic features of emotional expressions are processed preferentially. *PloS one*, 7(7):e41792, 2012.

[169] Fritz Schiltz, Chiara Masci, Tommaso Agasisti, and Daniel Horn. Using regression tree ensembles to model interaction effects: a graphical approach. *Applied Economics*, 50(58):6341–6354, 2018.

[170] Christian Schubert, M Lambertz, R.A. Nelesen, Wayne Bardwell, J.-B Choi, and J.E. Dimsdale. Effects of stress on heart rate complexity—a comparison between short-term and chronic stress. *Biological psychology*, 80:325–32, 12 2008.

[171] Adriane Seiffert, David Somers, Anders Dale, and Roger Tootell. Functional mri studies of human visual motion perception: texture, luminance, attention and after-effects. *Cerebral cortex (New York, N.Y. : 1991)*, 13:340–9, 05 2003.

[172] Vidya Setlur, Saeko Takagi, Ramesh Raskar, Michael Gleicher, and Bruce Gooch. Automatic image retargeting. In *Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia*, MUM '05, page 59–68, New York, NY, USA, 2005. Association for Computing Machinery.

[173] Fred Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:258, 2017.

[174] Zahra Sadat Shariatmadar and Karim Faez. Visual saliency detection via integrating bottom-up and top-down information. *Optik*, 178:1195–1207, 2019.

[175] Niedenthal Simon. *Patterns of Obscurity : Gothic Setting and Light in Resident Evil 4 and Silent Hill 2*. McFarland, 2009.

[176] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[177] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018.

[178] Mel Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557, 2009.

[179] Don Spiegel and Patricia Keith-Spiegel. Manifest anxiety, color preferences and sensory minimizing in college men and women. *Journal of Clinical Psychology*, 27(3):318–321, 1971.

[180] Don Spiegel and Patricia Keith-Spiegel. Manifest anxiety, color preferences and sensory minimizing in college men and women. *Journal of Clinical Psychology*, 27(3):318–321, 1971.

[181] Paramveer Kaur Sran, Savita Gupta, and Sukhwinder Singh. Segmentation based image compression of brain magnetic resonance images using visual saliency. *Biomedical Signal Processing and Control*, 62:102089, 2020.

[182] Nancy Steblay. A meta-analysis review of the weapon effect. *Law Hum. Behav.*, 16:413–424, 08 1992.

[183] Jonathan Steuer. Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42(4):73–93, 1992.

[184] Jonathan Steuer. Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42, 07 2000.

[185] Jeremiah Still and Mary Still. Influence of visual salience on webpage product searches. 16(1), feb 2019.

[186] Christopher Stolz, Dominik Endres, and Erik M Mueller. Threat-conditioned contexts modulate the late positive potential to faces—a mobile eeg/virtual reality study. *Psychophysiology*, 56(4):e13308, 2019.

[187] L Straetmans, B Holtze, S Debener, M Jaeger, and B Mirkovic. Neural tracking to go: auditory attention decoding and saliency detection with mobile EEG. *Journal of Neural Engineering*, 18(6):066054, dec 2021.

[188] Bryan A Strange, R Hurlemann, and Raymond J Dolan. An emotion-induced retrograde amnesia in humans is amygdala-and β-adrenergic-dependent. *Proceedings of the National Academy of Sciences*, 100(23):13626–13631, 2003.

[189] J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

[190] Alan Stuart. Rank correlation methods. by m. g. kendall, 2nd edition. *British Journal of Statistical Psychology*, 9(1):68–68, 1956.

[191] Chuen-Tsai Sun, Holin Lin, and Hsueh-Yu Lu. Using physiological response data to examine horror video game enjoyment. In *DIGRA Game, Play and the Emerging Ludo-Mix*, 2019.

[192] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. pages 1–9, 06 2015.

[193] Marcel Takac, James Collett, Kristopher Blom, Russell Conduit, Imogen Rehm, and Alexander Foe. Public speaking anxiety decreases within repeated virtual reality training sessions. *PLOS ONE*, 14:e0216288, 05 2019.

[194] Reika Takeshita, Aya Shoji, Tahera Hossain, Anna Yokokubo, and Guillaume Lopez. Emotion recognition from heart rate variability data of smartwatch while watching a video. In *2021 Thirteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*, pages 1–6, 2021.

[195] Jianhua Tao and Tieniu Tan. Affective computing: A review. pages 981–995, 10 2005.

[196] Chai-Fen Tsai, Shih-Ching Yeh, Yanyan Huang, Zhengyu Wu, Jianjun Cui, and Lirong Zheng. The effect of augmented reality and virtual reality on inducing anxiety for exposure therapy: A comparison using heart rate variability. *Journal of Healthcare Engineering*, 2018:1–8, 11 2018.

[197] Tomoki Tsuchida and G. Cottrell. Auditory saliency using natural statistics. *Cognitive Science*, 34, 2012.

[198] Ryoko Ueoka and Kouya Ishigaki. Development of the horror emotion amplification system by means of biofeedback method. In Sakae Yamamoto, editor, *Human Interface and the Management of Information. Information and Knowledge in Context*, pages 657–665, Cham, 2015. Springer International Publishing.

[199] Roger S Ulrich. Biophilia, biophobia, and natural landscapes. *The biophilia hypothesis*, 7:73–137, 1993.

[200] Emmett Velten. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 6(4):473–482, 1968.

[201] Sara Ventura, Laura Badenes-Ribera, Rocio Herrero, Ausiàs Cebolla, Laura Galiana, and Rosa Baños. Virtual reality as a medium to elicit empathy: A meta-analysis. *Cyberpsychology, Behavior, and Social Networking*, 23, 07 2020.

[202] Mikko Vesisenaho, Merja Juntunen, Paivi Häkkinen, Johanna Pöysä-Tarhonen, Janne Fagerlund, Iryna Miakush, and Tiina Parviainen. Virtual reality in education: Focus on the role of emotions and physiological reactivity. *Journal For Virtual Worlds Research*, 12, 02 2019.

[203] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. pages 2798–2805, 06 2014.

[204] T. Vijayakumar and Vinothkanna R. Retrieval of complex images using visual saliency guided cognitive classification. *Journal of Innovative Image Processing*, 2:102–109, 06 2020.

[205] Jan-Niklas Voigt-Antons, Robert Spang, Tanja Kojić, Luis Meier, Maurizio Vergari, and Sebastian Möller. Don't worry be happy - using virtual environments to induce emotional states measured by subjective scales and heart rate parameters. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 679–686, 2021.

[206] Jan-Niklas Voigt-Antons, Robert Spang, Tanja Kojić, Luis Meier, Maurizio Vergari, and Sebastian Möller. Don't worry be happy - using virtual environments to induce emotional states measured by subjective scales and heart rate parameters. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 679–686, 2021.

[207] Jan-Niklas Voigt-Antons, Robert Spang, Tanja Kojić, Luis Meier, Maurizio Vergari, and Sebastian Möller. Don't worry be happy - using virtual environments to induce emotional states measured by subjective scales and heart rate parameters. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 679–686, 2021.

[208] Daniel Västfjäll. Emotion induction through music: A review of the musical mood induction procedure. *Musicae Scientiae*, 5(1_suppl):173–211, 2001.

[209] Sarah Walker, Paul Stafford, and Greg Davis. Ultra-rapid categorization requires visual attention: Scenes with multiple foreground objects. *Journal of Vision*, 8(4):21, 2008.

[210] Yingchun Wang, Jingyi Wang, Weizhan Zhang, Yufeng Zhan, Song Guo, Qinghua Zheng, and Xuanyu Wang. A survey on deploying mobile deep learning applications: A systemic and technical perspective. *Digital Communications and Networks*, 8(1):1–17, 2022.

[211] Zhu Wang, Zhiwen Yu, Bobo Zhao, Bin Guo, Chao Chen, and Zhiyong Yu. Emotionsense: An adaptive emotion recognition system based on wearable smart devices. *ACM Trans. Comput. Healthcare*, 1(4), sep 2020.

[212] T. Watanabe, Alexander M. Harner, Satoru Miyauchi, Yuka Sasaki, Matthew Nielsen, D. Palomo, and Ikuko Mukai. Task-dependent influences of attention on the activation of human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 95 19:11489–92, 1998.

[213] Takeo Watanabe, Yuka Sasaki, Satoru Miyauchi, Benno Putz, Norio Fujimaki, Matthew Nielsen, Ryosuke Takino, and Satoshi Miyakawa. Attention-regulated activity in human primary visual cortex. *Journal of Neurophysiology*, 79(4):2218–2221, 1998. PMID: 9535981.

[214] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[215] J Mark G Williams, Andrew Mathews, and Colin MacLeod. The emotional stroop task and psychopathology. *Psychological bulletin*, 120(1):3, 1996.

[216] Kathy A. Winter and Nicholas A. Kuiper. Individual differences in the experience of emotions. *Clinical Psychology Review*, 17(7):791–821, 1997.

[217] bo wu and Linfeng Xu. Integrating bottom-up and top-down visual stimulus for saliency detection in news video. *Multimedia Tools and Applications*, 73, 12 2013.

[218] Yan Wu, Ruolei Gu, Qiwei Yang, and Yue-jia Luo. How do amusement, anger and fear influence heart rate and heart rate variability? *Frontiers in Neuroscience*, 13:1131, 2019.

[219] Yang Xu, Jun Li, Jianbin Chen, Guangtian Shen, and Yangjian Gao. A novel approach for visual saliency detection and segmentation based on objectness and top-down attention. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, pages 361–365, 2017.

[220] Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu. Personalized saliency and its prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2975–2989, 2019.

[221] Jimei Yang and Ming-Hsuan Yang. Top-down visual saliency via joint crf and dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 39, 03 2016.

[222] KEMAL Yildirim, A Akalin-Baskaya, and ML Hidayetoglu. Effects of indoor color on mood and cognitive performance. *Building and Environment*, 42(9):3233–3240, 2007.

[223] Bingqing Yu and James J. Clark. Personalization of saliency estimation. *CoRR*, abs/1711.08000, 2017.

[224] Sumaira Zaib and Masayuki Yamamura. Personalized saliency prediction using color spaces. *Multimedia Tools and Applications*, 81, 05 2022.

[225] Sumaira Erum Zaib and Masayuki Yamamura. Using heart rate and machine learning for vr horror game personalization. In *2022 IEEE Conference on Games (CoG)*, pages 213–220, 2022.

[226] Yanfang Zeng, Lihua Liu, and Rui Xu. The effects of a virtual reality tourism experience on touristrsquo;s cultural dissemination behavior. *Tourism and Hospitality*, 3(1):314–329, 2022.

[227] Shijie Zhao, Junwei Han, Xi Jiang, Heng Huang, Huan Liu, Jinglei Lv, Kaiming Li, and Tianming Liu. Decoding auditory saliency from brain activity patterns during free listening to naturalistic audio excerpts. *Neuroinformatics*, 16, 10 2018.

[228] Xiaosong Zhou and Ping Zeng. A bottom-up saliency detection method. In *2021 International Conference on Intelligent Computing, Automation and Applications (ICAA)*, pages 496–504, 2021.

[229] Guokang Zhu, Qi Wang, and Yuan Yuan. Tag-saliency: Combining bottom-up and top-down information for saliency detection. *Computer Vision and Image Understanding*, 118:40–49, 2014.

[230] Jun Zhu, Yuanyuan Qiu, Rui Zhang, Jun Huang, and Wenjun Zhang. Top-down saliency detection via contextual pooling. *Journal of Signal Processing Systems*, 74, 01 2014.

[231] Yucheng Zhu, Guangtao Zhai, Yiwei Yang, Huiyu Duan, Xiongkuo Min, and Xiaokang Yang. Viewing behavior supported visual saliency predictor for 360 degree videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4188–4201, 2022.