Computational Modeling of Free-viewing Attention on Multimodal Webpages - A Machine Learning Approach

Thesis submitted in partial fulfilment of the requirements for the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering

by

Vidyapu Sandeep

Under the supervision of

Dr. V Vijaya Saradhi
Dr. Samit Bhattacharya

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

August, 2020

Dedicated to the inspirational soul of Dr. Avul Pakir Jainulabdeen Abdul Kalam (A P J Abdul Kalam) who selflessly contributed to the technology and humanity, and received the highest of accolades such as Bharat Ratna, Padma Bhushan, and Padma Vibhushan.

Acknowledgements

Wow! What an amazing journey it has been, from a teacup cleaner to the doctorate holder (possibly!). There are many great human beings whose kind help put me in this position. I take this opportunity to acknowledge their precious contribution. However, in case if I fail to acknowledge anyone, please forgive me, I am no way intended to discredit your contribution but was only limited in recollecting by the time of this thesis submission.

The acknowledgments must begin with the school days. I would like to wholeheartedly thank L Jeevan Prakash Sir (Maths Teacher in 9th and 10th classes) who picked this son of a watchman to give high-quality education by investing the most of his valuable time. His contribution extends from the school days to the doctoral days. Thoughtfully, he advertised my academic performance in the newspaper, took pain in visiting the door of every generous donor, and streamlined every donation to systematically spend for my notebooks, college fees, and other educational requirements. I wonder if anyone does so much to even their own son. I would also like to thank each of those generous donors who extended their helping hand, ranging from donating their kids' textbooks to the senior citizens who donated from their pension and savings. I am grateful to each one of you; without your help, this journey would not have been possible.

I would like to wholeheartedly thank Late Shri Y S Rajasekhara Reddy (ex-Chief Minister of Andhra Pradesh) for providing free Engineering education, which greatly helped me to complete the B.Tech. I would also like to thank New Arihant Opticians (Himayat Nagar, Hyderabad) for providing free textbooks throughout my under-graduation and spectacles as long as I wore them. I want to extend my gratitude to Ernst Young (EY) foundation, especially Bharathi Madam (the coordinator), for their voluntary financial help from my +2 to the completion of under-graduation. Again, I obediently extend my gratitude to Late Shri Pandu Ranga Rao, who used to donate a major proportion of his limited pension amount towards my educational expenses. I am grateful to his sons, who extended their helping hand after the sad demise of their father. With the collective contribution of all these great human beings and many other generous people, I comfortably joined the Dual-degree course at IIT Guwahati.

Now, pertaining to the prominent doctoral work, I am wholeheartedly thankful to my supervisors, Dr. V Vijaya Saradhi and Dr. Samit Bhattacharya, for not only their esteemed guidance and financial support but also empowering me with two different yet powerful approaches of guiding a student. I believe both the techniques would come to my rescue when I had to guide any student in the future. I convey my special gratitude to the ever

helpful Dr. V Vijaya Saradhi, whose timely feedbacks, emotional, and financial support significantly contributed to the successful conclusion of this thesis work. I would also like to thank my doctoral committee for their regular inputs to improve my understanding and clarity in presentation.

I want to extend my wholehearted thanks to the seniors and colleagues for their kind support, timely guidance, and voluntary participation in the experiments. I would offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

Lastly, I am indebted to the family for their warm support. Even after reaching adulthood, they let me pursue the doctorate while keeping the financial struggles to themselves. I am highly thankful to the three sisters (Swarajyam, Swapna, and Swaroopa) who are always ready to even give away their own food for me. A special thanks (truly, it can never be 'special' nor be 'thanks') to my mother for everything she has done for me. I still remember, if I request her to wake me up at 3 AM, she will remain awake till 3 AM (as she thinks, what if the alarm doesn't work), wakes me up, and then only goes to sleep. Their dedicated and selfless efforts helped me to cross all the barriers until now. Always indebted to all of you! Thanking God for such a beautiful life...

August 24, 2020 Vidyapu Sandeep

Declaration

I certify that

- The work contained in this thesis is original and has been done by myself and under the general supervision of my supervisors.
- The work reported herein has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (concepts, ideas, text, expressions, data, graphs, diagrams, theoretical analysis, results, etc.) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- I also affirm that no part of this thesis can be considered plagiarism to the best of my knowledge and understanding and take complete responsibility if any complaint arises. The similarity (if exists any) would only be with the works published from this thesis work.
- I am fully aware that my thesis supervisors are not in a position to check for any possible instance of plagiarism within this submitted work.

August 24, 2020 Vidyapu Sandeep

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

Dr. V Vijaya Saradhi
Associate Professor
+91-361-2582367
saradhi@iitg.ac.in

Dr. Samit Bhattacharya
Associate Professor
+91-361-2582362
samit@iitg.ac.in

Certificate

This is to certify that this thesis entitled "**Computational Modeling of Free-viewing Attention on Multimodal Webpages - A Machine Learning Approach**" submitted by **Vidyapu Sandeep**, in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy, to the Indian Institute of Technology Guwahati, Assam, India, is a record of the bonafide research work carried out by him under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam, India. To the best of my knowledge, no part of the work reported in this thesis has been presented for the award of any degree at any other institution.

Date: August 24, 2020
Place: Guwahati

Dr. V Vijaya Saradhi
(Main-supervisor)

Dr. Samit Bhattacharya
(Co-supervisor)

Abstract

Webpages are the visual gateways into the information from the World Wide Web. The information retrieved from the Web gets rendered on webpages, predominantly through textual and image elements. The visual characteristics of the rendered elements determine the user's visual attention allocation. With the progressive expansion of competitive e-commerce and Web resources, attention modeling is essential for Web authors, information creators, advertisers, and Web-designers to understand and predict the user attention on webpages.

The existing attention models evolved from `position`-based analyses to standalone image-processing techniques as applied on webpage screenshots. The state-of-the-art computational models are centered on binary predictions to determine whether a given webpage *region* (pixel or group of pixels) is salient or not. Consequently, these approaches are limited in predicting the saliency (attention-drawing ability) of a whole web element and to further segregate among salient elements, which would be beneficial for preferential webpage rendering. The element-driven attention models are desired as the webpage is an ensemble of web elements (unlike an image), and designers provide visual features for individual web elements to achieve the desired visual representation. However, the *bottleneck* was to incorporate the elements' heterogeneous features into the model as texts are represented using features such as 'text-size' and 'text-color' whereas images are represented using 'brightness', 'intensity' and 'color histograms'. This thesis work is predominantly centered around overcoming the *heterogeneity bottleneck* to predict the users' element-driven free-viewing attention on multi-modal webpages, precisely consisting of text and images.

Owing to the prominence of `position`, primarily, the position-based free-viewing attention allocation is investigated at the higher-granularity for text and image elements. Subsequently, the designer-oriented informative text and image visual features are identified in explaining the free-viewing attention. Additionally, the association between visual features and the allocated user attention is computationally modeled, separately for text and images. The analyses revealed: (i) the elements positioned in the Right and Bottom regions of a webpage are not always ignored; (ii) *Space*-related (`column-gap, line-height, padding`) and font *Size*-related (`font-size, font-weight`) intrinsic text features, and *Mid-level Color Histogram* intrinsic image features are informative, while `position` and `size` are informative for both the types; (iii) the informative visual features predict the ordinal visual attention on an element with 90% average accuracy and 70% micro-F1 score; (iv) For the prominent images, the visual features also help in predicting the *weighted-voting-based* attention and *multiple discrete* levels of attention.

The aforementioned segregated analysis established the prominence of text and image visual features in explaining the users' free-viewing attention. However, both the modalities, in unison, determine the user attention on bi-modal webpages. The feature differences between both the modalities act as a bottleneck to unify them and subsequently model the attention. In the second contribution, a computational approach is presented and empirically analyzed to overcome the heterogeneity bottleneck, and to unify the text and images. For this, Canonical Correlation Analysis (CCA) is employed. The eye-tracking analyses on real-world webpages reveal: (i) text and images are unifiable if the interface idiosyncrasies alone or along with user idiosyncrasies are constrained; (ii) The *font-families* of text are influential and are comparable to *color histogram* visual features of images in achieving the unification between both modalities.

The achieved unification is utilized to overcome the limitations of existing approaches. That is, element-granular attention is predicted as opposed to region-based prediction, and multi-level attention is predicted as opposed to binary (salient or not) prediction. For this, the Support Vector Machine (SVM) based multi-class classification is employed. The analyses of the user attention collected on real-world webpages reveal: (i) the achieved unification outperforms the random baseline in predicting the element-granular user attention; (ii) the presented model demonstrates the efficacy in predicting the attention on all the elements unified into text as well as unified into image visual spaces; (iii) the model also outperforms the baseline for predicting the attention on unseen webpage data.

In the fourth contribution, the element-granular attention prediction is extended to predict a prominent user scanpath on webpages for *free-viewing*. For this, users' positional preference accumulation and *adherence to zero-sum-rule* are utilized besides the element-granular attention prediction. The analyses on the eye-tracking experiment conducted on two real-world webpage datasets reveal: (i) the considered text and image visual features outperform (average accuracy of more than 87% and micro-F1 score of more than 63%) the random baseline in predicting the users' scanpath on webpages; (ii) the presented approach demonstrates the prediction efficacy for scanpath prediction on unseen webpage data (average accuracy of more than 82% and micro-F1 score of more than 57%).

Contents

List of Figures

List of Tables

List of Abbreviations

AOI Area-Of-Interest. xvii, 19, 20, 85

CCA Canonical Correlation Analysis. 5, 7, 103

CDF Cumulative Distribution Function. 43

CMIR Cross-Modal Information Retrieval. 5, 103

CSS Cascading Style Sheets. 12, 17, 21, 37, 42, 133, 135

DOI Data-Of-Interest. xv, 14, 42–44, 85, 87–90, 92

DOM Document Object Model. 9, 11, 12, 22, 35, 39

FI Fixation-Index. xix, 15, 18, 23, 27, 28, 39, 40, 42, 44, 46, 48, 50–52, 54–59, 65–68, 71–75, 77–79, 84, 103, 124

GUI Graphical User Interface. 1

HTML Hypertext Markup Language. xvii, 11, 12, 22, 37, 41, 54, 135

MUMLA Multi User Multi Level Attention. xix, 76–80, 83, 84

PCCR pupil center/corneal reflection. xvii, 13–15

SVM Support Vector Machine. xvii, 23, 24, 27, 55, 56, 71

UAB User Attention-based. 2, 3, 20–22, 131

UG User-oriented Grouping. 43, 89, 92, 96, 97

VFB Visual Feature-based. 2, 3, 22, 131

WG Webpage-oriented Grouping. xix, 8, 9, 43, 89, 92–94, 96–98, 101, 103, 110, 116, 117, 125, 130, 132

WUG Webpage-and-User-oriented Grouping. xix, 8, 9, 43, 89, 92, 93, 95–98, 101, 103, 110, 116, 117, 126, 130, 132

WWW World Wide Web. 1

List of Definitions

Attention	As per William James (1890), "Attention is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought".
Classification	The procedure to determine a class (or category) of a given entity.
Document object model	A model that provides logical structure (also called, tree) of the webpage so that the underlying web elements can be traversed to access and manipulate the associated features.
Eye-tracking	A technique to record and store the eye-movement data for further analysis.
Free-viewing	The gazing to a webpage in the absence of a specific task.
Fixation	Considered as proxy for overt visual attention. Typically, at least a duration of 80–100 milliseconds at a location is considered as a fixation for free-viewing attention analysis on webpages.
Fixation-index	The ordinal number (starting with 1) assigned to each fixation of a user on a webpage. For free-viewing, it is also considered as an *attention-level*, and an indicator of *ordinal visual attention*.
Modality	Type of the data entities, such as text and images.
Multi-class classification	Classification technique to determine a class from a set of all possible classes.
Multi-label classification	Classification technique to determine a (non-empty) subset of classes from a set of all possible classes.
Scanpath	The sequence of fixation allocations on a webpage.
Visual attention	Attention achieved through visual perception.

List of Symbols

Symbol		Descirption
m	—	number of visual text features
n	—	number of visual image features
h	—	dimension of homogeneous space; $h \leq \min(m, n)$
$\mathcal{T}$	—	text visual space $\subseteq \mathbb{R}^m$
$\mathcal{I}$	—	image visual space $\subseteq \mathbb{R}^n$
$\mathcal{H}$	—	homogeneous visual space $\subseteq \mathbb{R}^h$
T	—	all fixated text elements (in $\mathcal{T}$) obtained from an eye-tracking experiment
I	—	all fixated image elements (in $\mathcal{I}$) obtained from an eye-tracking experiment
$\mathbf{w}_t$	—	linear transformation direction for text elements
$\mathbf{w}_i$	—	linear transformation direction for image elements
$T^{'}$	—	all fixated text elements in $\mathcal{H}$
$I^{'}$	—	all fixated image elements in $\mathcal{H}$
E	—	union of text and image fixated elements in $\mathcal{H}$
$\mathcal{L}$	—	List of possible Fixation-Indices
θ	—	scanpath threshold, or saliency threshold

1

Introduction

Webpages are the visual interfaces between users and the World Wide Web (WWW), the largest source of transformable information [1]. The users across the globe visually perceive the information from WWW as Web-browsers render on Graphical User Interface (GUI)s called webpages. The progressive expansion of the Web [62] and the number of Internet users [130] drew the research community's interest towards web-designing and web-interface based user attentional analyses. Especially, the user attentional understanding and modeling obtained prominence due to the ever-increasing commercial value of Web-based businesses such as business-to-consumer model (valued at 2.3 trillion USD in 2017 and projected to be 4.48 trillion USD by year 2021[1]). Moreover, understanding of user's attentional behavior helps in the improvement of user perception and satisfaction which may further help in user retainment towards the website [81].

As information on webpages is represented through elements of different *modalities*, such as text and images, researchers explored their prominence in drawing the user attention [51, 61, 143]. For example, Hsieh and Chen [69] found that video-based and picture-based webpages are better than text-based webpages in drawing the user attention towards advertisement; and, the associated information modality can be manipulated to prolong the attention [70].

The foundational and theoretical work by Faraday [51] suggested a decreasing preferential order (in drawing the user attention) for six salient visual elements as motion, size, color, text-style, images and position. That is, motion element is attended over a large-sized element which in turn is preferred over a colored element and so on. However, the subsequent empirical analyses [61, 143] invalidated the suggested preferential ordering. Both these analyses highlighted the prominence of element's position in drawing the attention for which Faraday assigned the least priority. Highlighting the prominence of position-bias, pattern-based models were proposed such as F-pattern [105], Z-pattern, Zig-Zag pattern, etc. [16]. However, these generic pattern-based models are limited in the applicability [16, 137] as they

[1] https://www.statista.com/topics/871/online-shopping/

do not incorporate the underlying visual features. The subsequent attentional analyses were centered on overcoming this limitation and predicting the attention at higher granularity. The existing attention prediction approaches may broadly be classified into *User Attention-based* approaches and *Visual Feature-based* approaches. The former is centered on eye-gaze data, and the latter is based on stand-alone image processing techniques.

User Attention-based (UAB) approaches: Leveraging the *eye-mind hypothesis* [80] that established the connection between eye-gaze and visual attention, UAB approaches elicit a prominent attention pattern from an ensemble of gaze allocations. The approaches include: (i) *fixation*—proxy for visual attention [14]—heatmap based F-pattern [105], Z-pattern, and Zig-Zag pattern [16]; (ii) *scanpath*—sequence of fixation allocations on a webpage—based *trending* patterns [46, 47, 49]. The former approaches help in the determination of prominent regions while the latter approaches further arrange them into a prioritized sequence concerning the user attention. As the UAB approaches are based on direct measurement (through non-intrusive eye-tracking [165]) of visual attention, the findings are easily interpretable and more reliable. However, these approaches are limited in as follows.

1. *Generalizability*: the underlying webpage visual features are not incorporated into these approaches, consequently limiting the applicability. For example, F-pattern applicability is restricted to the text-rich webpages [16, 137].

2. *Feasibility*: the UAB approaches are pattern eliciting than attention predictive; thus, users' attention data is required for every test webpage. However, collecting this data on every webpage and especially after every small modification enforced on a webpage may not be practical due to the time-constraints, money-constraints, and user-availability constraints.

Visual Feature-based (VFB) approaches: Computational VFB approaches help to overcome the limitations of UAB approaches. The typical procedure is to extract multiple feature maps, all are of same dimension (as that of the considered webpages), and integrate them to determine the *salient* or attention-drawing locations [133, 134, 135, 142, 144]. These approaches are the adaptations of saliency prediction techniques proposed for standalone images [14, 76]. However, these approaches are limited in as follows.

1. *Interpretability*: due to non-direct incorporation of human attention. For example, scanpath cannot be evident, unlike in the UAB approaches.

2. *Granularity*: the dichotomous prediction (salient or non-salient) conceals the *true* saliency of underlying location. That is, a location with very high saliency and the location with saliency score just above the threshold are considered equivalent;

3. *Extensibility*: the prominence of webpage constituting elements is overlooked due to the processing of the whole webpage as an image. Consequently, these approaches

cannot be extended to other web element-driven applications such as user preferential rendering [88, 115], and incorporation of dynamic (motion-related) characteristics.

The Heterogeneity Bottleneck: The limitations of UAB and VFB approaches can be overcome through the incorporation of (i) human visual attention, (ii) idiosyncrasies of individual elements, and (iii) higher-granularity in attention (say, scanpath). However, the predominant bottleneck in their simultaneous incorporation is the **heterogeneity of data modalities constituting a webpage**. The text and image elements are represented using a distinct set of visual features. Moreover, the number of text features may not equal the number of image features, unlike the VFB approaches, where all feature maps are of the same dimension. Even in case same feature count is achieved for both modalities, the existing feature-integration approaches cannot be applied due to the absence of one-to-one correspondence between text and image features.
This thesis work is predominantly centered on overcoming the heterogeneity bottleneck to predict the element-driven attention on webpages at the higher granularity.

1.1 Scope of the Thesis

There are five key areas of study associated with the thesis work including, (i) attention modeling, (ii) eye-tracking, (iii) interface, (iv) users, and (v) webpages; as shown in Figure 1.1.

Attention Modeling: Based on the attention guiding features, the attention models may broadly be classified into following three categories.

1. *Bottom-Up*— the underlying visual features guide the attention. For the current analysis, we limit ourselves to Bottom-Up attention modeling. The Bottom-Up attention models may further be dichotomized into following two categories.

   (i) *Computational*—learning the association between visual features and attention. The *Computational* modeling may further be classified into *Uni-modal* approaches and *Multi-modal* approaches. Concerning the attention modeling on webpages, the image-processing techniques cater to Uni-modal approaches and the element-driven approaches (on multi-modal webpages) cater to Multi-modal approaches. As the current thesis work presents the element-driven approaches for attention modeling on bi-modal webpages, the work registers in Multi-modal category.

   (ii) *Non-Computational*— observation-based approaches such as F-pattern [105].

2. *Top-Down*— user preferences guide the attention.

3. *Mixed*— a combination of the former two. Among the three, the Mixed attention modeling is complicated as it is difficult to attribute the allocated attention to either of Bottom-Up features or the Top-Down features.

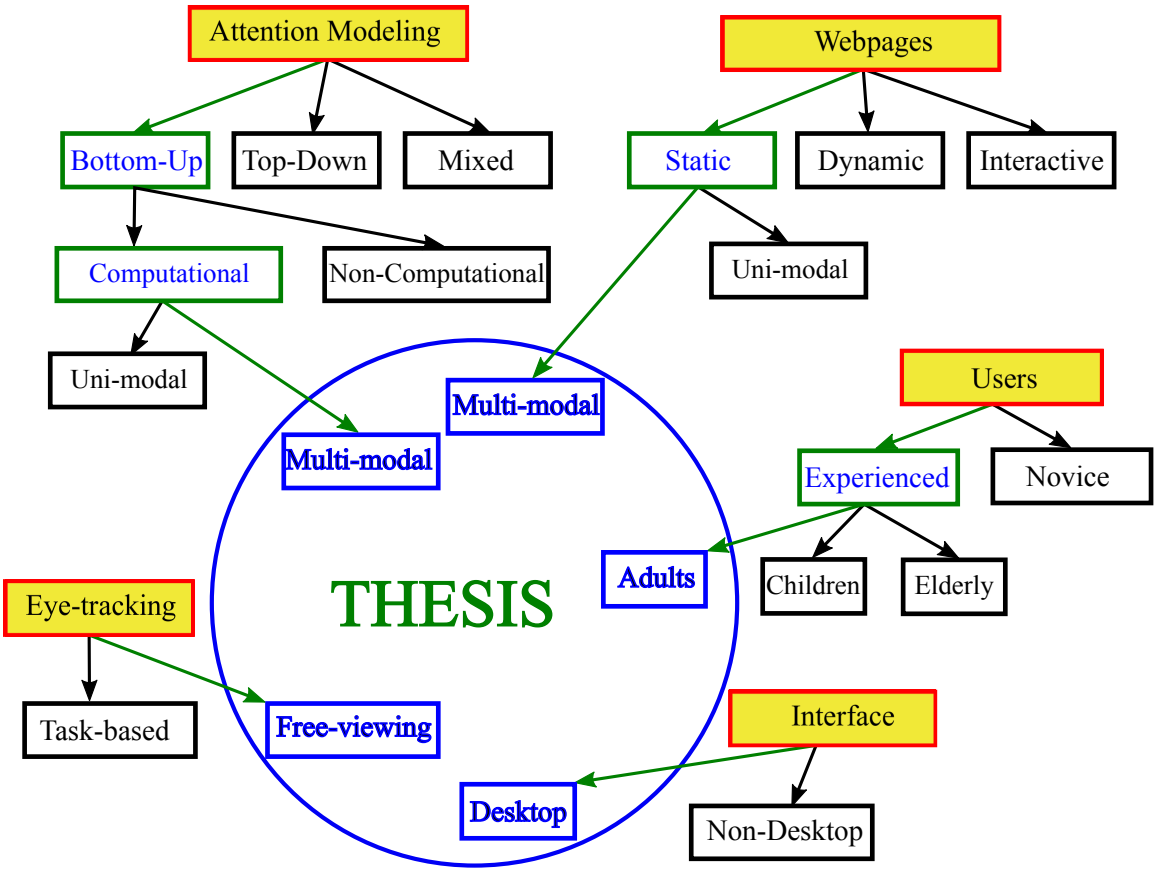


Figure 1.1: Scope of the thesis concerning the attention modeling, eye-tracking, webpages, users, and interface components.

Eye-tracking: It is a technique to capture, record, and store the eye-movements for further analysis. Based on the activity for which the eye-movements are captured, the eye-tracking approaches are classified into following two categories.

1. *Task-based*— capture the eye-movements when the user is gazing voluntarily so as to perform a specific task.

2. *Free-viewing*— capture the eye-movements when the user is gazing involuntarily. As the thesis work is centered on Bottom-Up attention modeling, that is, the webpage visual feature influence on the user attention is modeled, the eye-tracking experiments are limited to free-viewing.

Interface: With the advances in hardware technologies, wide variety of interfaces are available for the users to visualize the webpages. The interfaces may broadly be classified into *Desktop*s, and *Non-Desktop*s which include the displays such as mobile phone interfaces and foldable interfaces. Owing to their prevalent usage, existing literature, and easier experimental setups the current thesis work limits to attention modeling on regular Desktop displays (precisely, with the dimension of 22").

Users: Concerning the webpage access, the users are either *novice* with limited or no web usage experience, or *experienced*. Owing to the progressive expansion of the Internet and the users, the attention models for experienced web users are desired. The experienced web users may either be *Children*, *Adults*, or *Elderly*. The thesis limits to adult and experienced web users towards the attention modeling.

Webpages: The existing webpages may broadly be classified into following three categories.

1. *Static*— rendered content does not change through-out the web session. Owing to the simplicity of static webpages, the thesis work limits to static webpages. Though static webpages may further be classified into *Uni-modal* and *Multi-modal*, the latter is considered due to their prevalence than their counter-parts (say, text-only webpages).
2. *Dynamic*— presence of motion elements.
3. *Interactive*— web content changes based on user input such as through pop-ups.

To summarize, the thesis work aims to model the experienced adult users' free-viewing attention on multi-modal static webpages as displayed on desktop monitors. Towards the attention modeling, the bottom-up computational approaches incorporating multiple modalities are considered.

1.2 Motivation

The thesis work derives the motivation from the information-retrieval domain towards overcoming the prominent *heterogeneity bottleneck* to predict the attention on bi-modal webpages. Rasiwasia et al. proposed an approach to bridge the semantic gap, the bottleneck between text and images through correlation maximization between respective semantic features [32, 118, 119]. The authors considered Wikipedia articles and paired each text paragraph with an image element from the corresponding section. The pairing presumes that the paired elements are associated with the same semantic concepts. For example, an image of 'Statue of Liberty' is paired with a text paragraph describing the same. In other words, the text element and the image element are considered as two different views of the same semantic concept. Subsequently, a *space* that achieves the homogeneous representation for both text and images was learned using the paired elements' semantic features. Towards the construction of homogeneous space, *Canonical Correlation Analysis (CCA)*— a multivariate analysis technique to maximize the correlation between two sets of features [64, 68], was employed. Consequently, the learned space enabled the comparison of text and images which was utilized for *Semantic Cross-Modal Information Retrieval (CMIR)*— given a text, retrieve the semantically equivalent image, and vice-versa [118, 119].
The overcoming of heterogeneity bottleneck enables the attention prediction at the granularity of web elements. That is, the prominence of the whole element on a given webpage

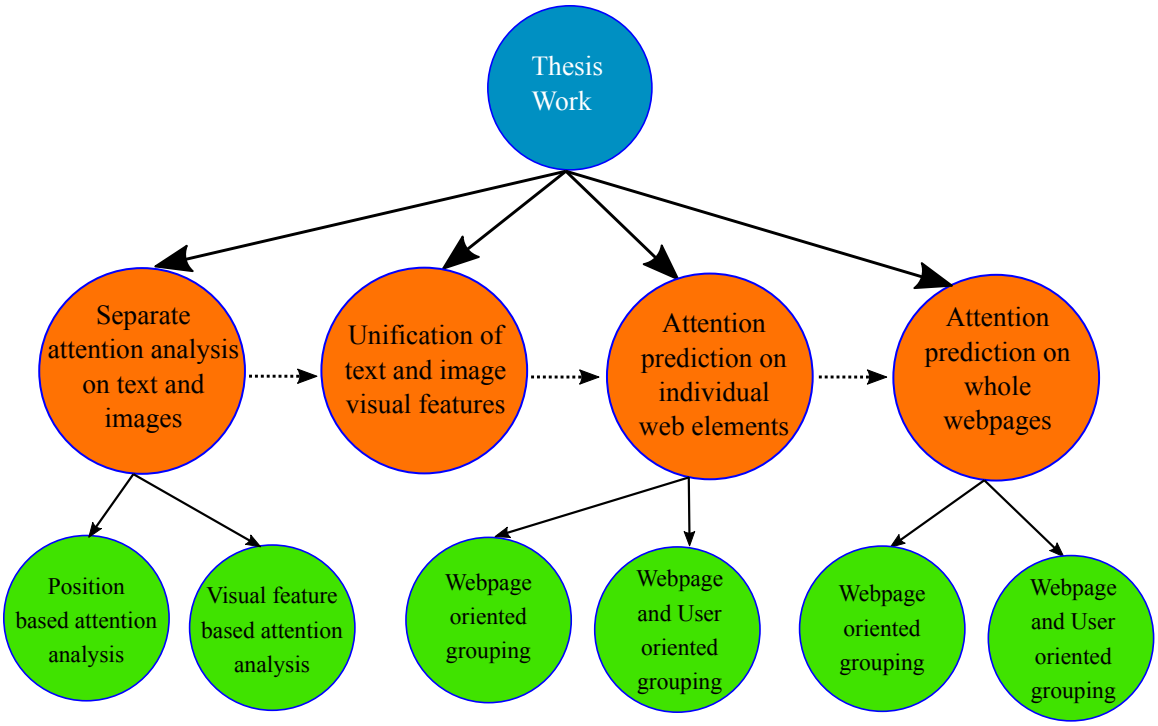


Figure 1.2: Organization of the contributions

can be estimated as opposed to the *region*-based predictions in the state-of-the-art. To elaborate, the application of existing approaches may determine some regions of an element (say, text) as salient while the remaining as non-salient whereas web-designers assign the visual characteristics to the element as a whole. Additionally, the element-based attention predictions are useful for applications such as user preferential webpage rendering [88, 115], and incorporation of dynamic (motion-related) characteristics.

1.3 Objectives and Contributions

The overall objective of the thesis is to analyze, establish, and predict the association between web elements' visual features and the free-viewing user attention. Among the multiple possible modalities, only text and images are perceived through visual-alone inspection on webpages. Accordingly, the formulated objectives and the corresponding contributions are as follows. See Figure 1.2 for the organization of the contributions.

I. *Position-based and Modality-specific Attention Analyses:* As the majority of the earlier attentional analyses are position-based, more fine-grained analyses are performed to answer the following research questions.

   R1: How users allocate *free-viewing* ordinal attention on text and image elements positioned in 3×3 webpage regions?

R2: Which *intrinsic visual features* are informative in explaining the free-viewing ordinal attention on web elements?

R3: How the informative intrinsic visual features perform in *predicting the free-viewing ordinal visual attention?*

The attention analyses on real-world webpages (described in Chapter 4) resulted in following key findings.

1. Though users predominantly allocate the initial attention to Middle and Top regions, the elements in Right and Bottom regions are not completely ignored.
2. The textual elements' *Space* and *Font-Size* determining intrinsic visual features, and image elements' *Mid-level Color Histogram* intrinsic visual features are informative, while `position` and `size` are informative for both the modalities.
3. The informative visual features predict the ordinal visual attention on an element with 90% average accuracy and 70% micro-F1 score.
4. The analyses concerning the image elements revealed that the image visual features outperform the random baseline in predicting the free-viewing user attention.

II. *Unification of Text and Images:* As text and images influence the user attention allocation on bi-modal webpages, a Canonical Correlation Analysis (CCA) based computational approach is presented to unify the cross-modal visual features of text and images. Through this analysis, we tried to answer the following research questions.

R1: Are the text and image visual features correlated based on the free-viewing user attention allocation on bi-modal webpages? Do the user idiosyncrasies and the interface idiosyncrasies affect such correlations?

R2: Which cross-modal visual features are *comparable* with each other based on the free-viewing user attention allocation on bi-modal webpages? Do the user idiosyncrasies and the interface idiosyncrasies affect such comparisons?

R3: Can the text visual features *delineate* the free-viewing user attention on image visual features and vice-versa for the bi-modal webpages? Do the user idiosyncrasies and interface idiosyncrasies affect such delineations?

The CCA based computational approach (described in Chapter 5) resulted in following findings.

1. Cross-modal text and image visual features are correlated when the interface idiosyncrasies, alone or along with user idiosyncrasies are constrained.
2. The *font-families* of text are *comparable* to *color histogram* visual features of images in drawing the users' attention.

3. Text visual features and image visual features can delineate each other's free-viewing attention drawing ability.

III. *Element Attention Prediction:* The unification achieved through the *Webpage-oriented Grouping (WG)*— considers all users' attention on each webpage separately, and *Webpage-and-User-oriented Grouping (WUG)*— considers each user's attention on each webpage separately towards the unification; are utilized for element-granular attention prediction. Through this analysis, we tried to answer the following research questions concerning both the groupings.

R1: Can attention on elements be predicted if all the elements are unified into a text modality?

R2: Can attention on elements be predicted if all the elements are unified into an image modality?

R3: How well the achieved unification perform to predict the attention on unseen data for both the above research questions?

The multi-class classification based computational approach (described in Chapter 6) resulted in following findings.

1. The element attention prediction outperforms the random baseline when all the elements are unified into the text modality.
2. The element attention prediction outperforms the random baseline when all the elements are unified into the image modality.
3. For both, WG and WUG, the element attention prediction outperforms the random baseline for unseen webpage data, while both achieved comparable predictive performance.

IV. *Scanpath Prediction* The unification achieved through the WG, and WUG, are utilized to predict the prominent scanpath on webpages. For this, the element attention prediction model is extended to incorporate the users' positional-bias. The following research questions are investigated concerning both the groupings.

R1: Can scanpath be predicted if all the elements are unified using WG?

R2: Can scanpath be predicted if all the elements are unified using WUG?

R3: How well the achieved unification perform to predict the scanpath on unseen data for both the above research questions?

Following are the scanpath prediction findings (described in Chapter 7).

1. The scanpath attention prediction outperforms the random baseline for WG
2. The scanpath attention prediction outperforms the random baseline for WUG

3. For both, WG and WUG, the scanpath prediction outperforms the random baseline for unseen webpage data, while both achieved comparable predictive performance.

1.4 Thesis Organization

This thesis comprises eight chapters. The chapter wise organization of the thesis is given as follows:

Chapter 1: This chapter introduces the prominence of the research problem besides the summarization of thesis contributions and organization.

Chapter 2: This chapter introduces the relevant literature including Document Object Model (DOM) of webpages, eye-tracking, visual attention terminology, and the prominent attention models on webpages.

Chapter 3: This chapter describes the free-viewing eye-tracking experiments conducted on real-world webpages whose data was utilized throughout the analyses presented in the thesis. Among the three experiments, two were conducted by us while the remaining experiment data was curated from the open-source.

Chapter 4: This chapter illustrates the prominent position-based user attention analyses and highlights the prominence of elements' visual features. Further, the computational modeling of association between modality-specific features and the modality-specific allocated attention are presented.

Chapter 5: This chapter presents the computational approach to unify the text and image modalities concerning the user attention. Further, three groupings are presented to account for the user and webpage idiosyncrasies.

Chapter 6: This chapter presents a computational approach to predict the quantitative visual attention on web elements, utilizing the unified representation achieved in Chapter 5. The attention prediction is performed in text visual space as well as the image visual space.

Chapter 7: This chapter elucidates an approach to determine a visual feature-based scanpath on webpages. Accordingly, this chapter introduces an approach to quantify the users' positional preference which in association with element attention prediction is leveraged for scanpath prediction.

Chapter 8: This chapter discusses the applications, limitations, concluding remarks, and the future research directions.

2

Background & Related Work

Our thesis incorporates the knowledge of key concepts including, webpage, eye-tracking, visual attention, multi-modalities[1], and computational modeling. Accordingly, this chapter introduces the relevant background of these related concepts, and subsequently summarizes the previous works to posit the current thesis work.

The background includes the (i) introduction of webpage as a Document Object Model (DOM) in Section 2.1, (ii) eye-tracking as a technique to capture the eye movements in Section 2.2, and (iii) the association between eye movements and visual attention in the terminology of eye movements in Section 2.3. The related works include the (i) uni-modal (text-only or image-only) analyses that motivated the feature selection for our analysis (Section 2.4), (ii) related webpage-based analyses to describe the limitations in state-of-the-art (Section 2.5), and (iii) classification-based prediction approaches from machine learning literature that we utilized in the current thesis work (Section 2.6).

Our work considers the webpage as an ensemble of segregable elements. These elements are traversable through a structure called DOM which is described in the following section.

2.1 Webpage as Document Object Model (DOM)

DOM defines a logical structure for webpages (broadly, valid Hypertext Markup Language (HTML)[2] documents), and means of accessing and manipulating the constituent elements [27]. That is DOM assists in building the webpages as well as traversing through the associated structure and content. An example HTML webpage snippet along with the corresponding graphical DOM representation is shown in Figure 2.1. For example, to access or modify the attributes of tag <H1>, the arrow paths shown in Figure 2.1(right) are traversed: DOCUMENT → <HTML> → <BODY> → <H1>. Web browsers read the HTML page and render the content as per the *tags* (see HTML, HEAD, TITLE, BODY, H1 and IMG in

[1] through-out, *modalities* means data modalities such as text and images

[2] https://www.w3.org/html/

```
<!DOCTYPE html>
<HTML>
<BODY>
<H1>Hello World!</H1>
<P style="font size:2em;margin top:0.67em;
margin bottom:0.67em;margin left:0em;
margin right:0em;font weight:bold">
Hello World!</P>
<IMG src="sun.png" height="22" width="22">
</BODY>
</HTML>
```

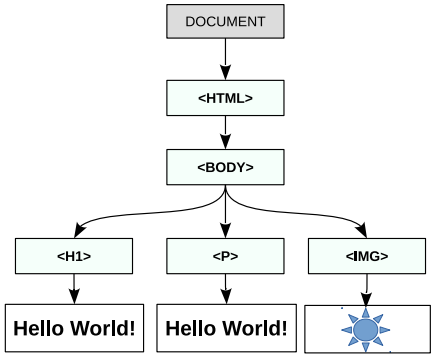


Figure 2.1: (left) Sample HTML webpage; and (right) its equivalent DOM

Figure 2.1(left)) whose visual representation is further attuned by the associated attributes (see **height** and **width** attributes of IMG tag).

Cascading Style Sheets (CSS)[3] are widely utilized to enhance the visual representation of elements through extended visual attribute setting functionality. That is, CSS attach the styling to web elements, such as fonts, coloring, and spacing. The CSS helps in overcoming the inherent tag-based representational limitations. For example, by default, the majority of web-browsers assign HTML tag <H1> with the CSS attributes **font-size**: 2em, **margin-top**: 0.67em, **margin-bottom**: 0.67em, **margin-left**: 0em, **margin-right**: 0em, **font-weight**: bold, to the enclosed text. However, assignment of same attributes to an HTML text paragraph tag <P> (see **style** attribute) results in same visual representation (as shown in Figure 2.1(right)) which may be further extended by adding **color, border**, etc. In summary, the HTML tags represent the logical separation of the content whereas their visual rendering is guided by CSS attributes. Thus, investigating the CSS attributes' influence on user attention helps the web designers in setting the attribute values accordingly.

Though the DOM helps in webpage exploration, not many studies leveraged it for attentional analyses on webpage. The analyses evolved from theoretical approaches to adapting the image processing techniques and towards the DOM based approaches as described in the related works (Section 2.4 and Section 2.5).

To understand and analyze the user's visual interaction with webpages, the user's eye movement behavior needs to be captured. For this, a technique called *eye-tracking* has been widely employed which is described in the following section.

2.2 Eye-tracking

Eye-movements are important body movements that reveal valuable information about inner and outer world. So, studying eye-movements has huge applications in diverse fields [42]. Consequently, *Eye-tracking*—a technique to record and store the eye-movement

[3] https://www.w3.org/Style/CSS/Overview.en.html

data for further analysis— obtained the prominence from technological as well as application perspective.

The eye-tracking technologies evolved over the years in capturing the human physiological eye-movements. The four broad categories of eye-movement capturing techniques [41, 116] are:

- *electro-oculography* techniques utilize the electric potential generated from the skin surrounding ocular activity for eye-movement estimation.
- *scleral contact* techniques physically mount the contact lens or search coil onto the eye for direct eye-movement measurement.
- *photo-oculography* or *video-oculography* techniques utilize the eye-features such as pupil size, limbus position, and corneal reflections to measure the translatory or rotatory eye-movements.
- *pupil center/corneal reflection (PCCR)* techniques utilize the orientation of corneal reflections with respect to the pupil center to measure the eye-movements. Among the four, PCCR techniques are widely employed for non-intrusive and remote eye-tracking [41].

The genesis of PCCR eye-tracking involves eye illumination with an infrared light source to form a bright eye, and a corneal reflection called *glint* or *Purkinje image* as shown in Figure 2.2. The reflections are captured by the specialized camera(s), subsequently measuring the orientation of Purkinje image with respect to the pupil center to estimate the gaze direction (see the sample illustrations in Figure 2.2). On combining with the appropriate calibration, the gaze-direction is mapped to a point-of-regard in the visual scene. A sample PCCR eye-tracker, Tobii X2-60 which employed for the experimentation (described in Chapter 3) is shown in Figure 2.3. With the advancement in hardware and computational architectures, the eye-tracking was achieved with a lesser number of light sources [136], easier calibration [24, 92], statistical modeling [5], and in tandem with other measurements such as brainwave measurements [7].

The progressive robust eye-tracking techniques invited numerous applications in multiple domains [42]. In *Industrial engineering*, eye-tracking was used to identify the pilot's primary source of information during flight, driver's search strategy during driving, visual inspection of medical X-ray, etc [42]. In *Marketing/Advertising* domain, eye-tracking helped to identify the consumer preferences for advertisements [93, 126, 158, 163]. In *Psychology*, human reading behavior [120], scene perception, art perception, film perception, visual search task and more are studied [42]. In *Neuroscience*, eye-tracking was studied to establish the relationship between eye-movements and the visual attention. In *Computer science*, the eye-gaze was explored for interactive mechanisms such as pointing modality similar to a keyboard or mouse [8, 77, 138], and for understanding student's web-based learning [79].

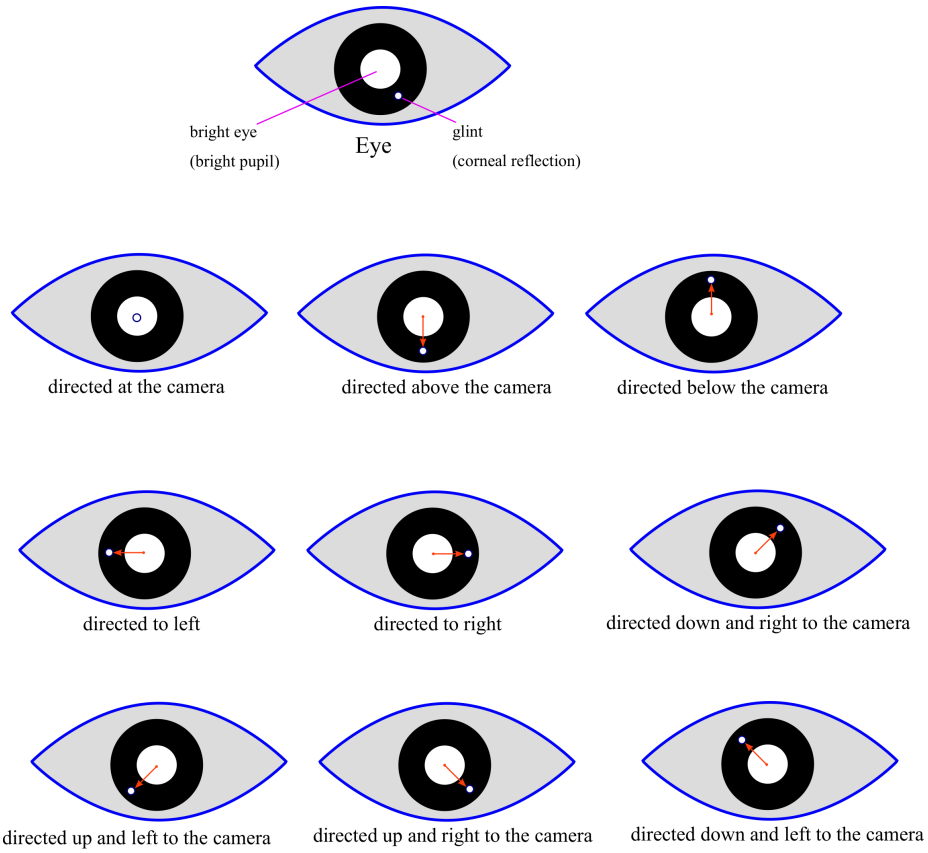


Figure 2.2: pupil center/corneal reflection (PCCR) based gaze directions

Especially in interface usability studies, eye-tracking was employed to understand the effectiveness of interface in presenting the information and user attention allocation strategies on interfaces [42].

2.3 Visual Attention and Terminology

During the visual inspection, only a fraction of visual-scene (webpage in the current setting) is perceived in higher detail [41]. This contributes a visual angle of 1–2°, called *foveal visual angle*, out of possible 220° visual angle whereas the remaining visual scene is perceived blurry as shown in Figure 2.4. The eye-movements are often stationed to process the visual information under gazing location. These stable eye-movements which indicate the user perception are called *fixations* based on which following useful terminology is developed.

- **Fixations** are standstill in eye-movements (typically for a duration of 80–100 ms [19]) which act as *proxy* for visual attention [63, 80]. An element underlying a fixated location is called an *attended element* or *fixated element*, or a Data-Of-Interest (DOI) according to terminology of Jianu and Alam [79].

- **Saccades** are the rapid eye-movements in-between the fixations during which no

Figure 2.3: A prototype PCCR eye-tracker, Tobii X2-60

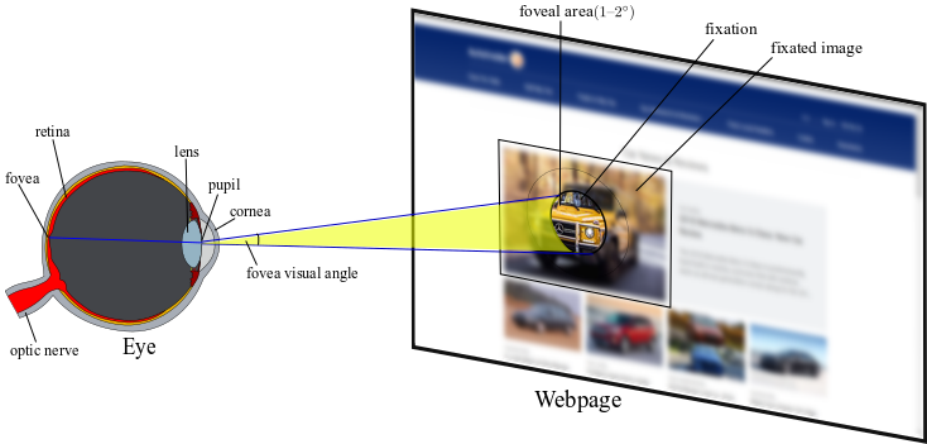


Figure 2.4: Illustration of eye-gazing towards a web image element. The stable allocation of foveal visual angle (1-2°) results in *fixation*, an indicator of underlying element's visual perception.

significant information is processed.

- **Scanpath** is an alternate sequence of fixations and saccades. The ordinal sequence of fixations by a user on a webpage constitutes a scanpath. Each user's gazing session on a webpage results in a scanpath. Example scanpaths are shown in Figure 2.5.

- **Fixation-indices (FIs)** are ordinal numbers associated with the fixations to indicate the sequential allocation of fixations. FIs are numbered as 1, 2, and so on. For *free-viewing*, the lower the FI value the quicker is the attention-drawing ability of the underlying element. Thus, FI is an indicative of *ordinal visual attention*. Adapting the machine learning terminology, all fixated elements associated with same FI form a *class*. For example, all the fixated elements with FI=1 are the class of elements with quickest attention-drawing ability.

Three prominent factors influence the user's fixation allocation [19]: (i) elements' saliency on a webpage, (ii) user's expectations and memory bias, and (iii) given task or information

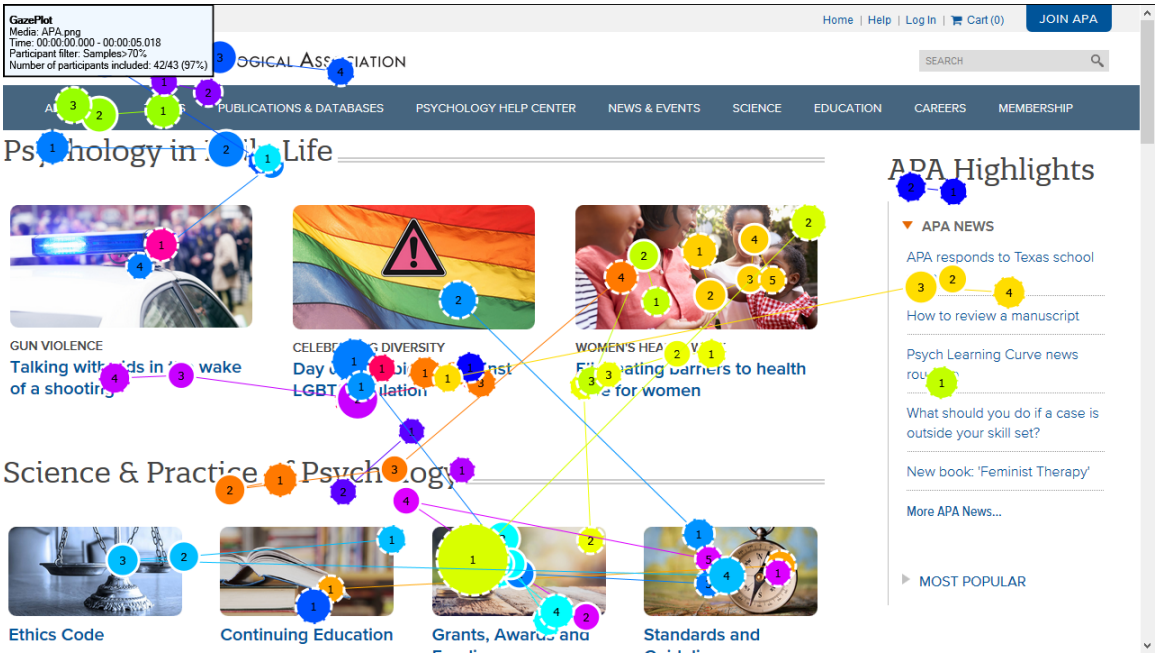


Figure 2.5: Example users' scanpaths on a webpage: Each bubble is a fixation with the inscribed number (fixation index) indicating the ordinal number of the fixation in a scanpath. Each user's scanpath is indicated with a different color.

at the hand. The attention driven by the first factor is called *bottom-up visual attention* and the latter two is called *top-down visual attention*. Though eye-tracking helps in measuring the attention through fixations, attributing it to any of the three factors is an intricate task in their simultaneous presence. To overcome this difficulty, often, any of the two factors are restricted to understand the other factor's influence on attention. For the current thesis work, we limit the latter two factors to analyze and model the intrinsic visual features' influence on attention.

2.4 Modality-specific Visual Feature based Analyses

Literature explored the restricted set of visual features from text and image modalities for their influence on psychology, cognition, and perception. This section describes the explored visual features from either modality to motivate the feature selection for our analysis.

2.4.1 Text Modality

The webpages during 1990s and early 2000s [1] were predominantly unimodal consisting of textual elements. Consequently, wide-variety of literature focused on text feature based analyses. The important results from the literature are as highlighted.

1. **Text and Background colors:** Users find desaturated color combinations more satisfactory for text and background and their preference is influenced by the intensity

of the color (called, *saturation*) [108]. Additionally, achromatic background color helps in reduced reading time [72]. Text and background color combinations influence the understanding of the text [153]. These works motivated us to consider text and background related `color` features for the attention analysis.

2. **Font families:** For a specific font-size, 'arial' was easier to read than 'times new roman' while text with larger font size is more readable [9]. Motivated by this work, we included `font-family` and `font-size` features for the analysis of our proposed model.

3. **Spaces and Font sizes:** Text length and character density influence the reading rate with the longer text being read faster than shorter length text [40, 45]. Medium size lines (with 55 characters per line) support the medium and fast reading of the text effectively [43, 44]. This motivated us to consider the `position` of the text from the left, right, top, and bottom edges of the screen. As we conducted the experiments on fixed size screens, these positional parameters account for text length. We also utilized `font-size` to account for the number of characters per line besides its own influence.

For searching, bigger line spacing reduces the reaction time in completing the search task [96] while users prefer justified text for searching. Better legibility was reported for colors with greater luminance contrast having dark backgrounds on Cathode Ray Tubes (CRTs) [73]. As spacing, text justification influence the user's attention, we considered `indentation`, `padding` and `margin` features of the text in analyzing our model.
For the webpages, all the aforementioned feature values are manipulated through Cascading Style Sheets (CSS) features by the web designers. Accordingly, CSS based visual features (described in detail in Section 3.5) were considered for the attentional analyses in our thesis work.

2.4.2 Image Modality

Multiple works were presented in computer vision and cognitive science domains to predict the attention on images [14]. Majority of these works used features such as `color` and `intensity` [75, 76]. Lin et al. [95] reported that composition of icons (which are typically pictorial information) and background significantly affects the attention. These works motivated us to consider the image elements' intrinsic visual characteristics including `color histograms`, color spread (mean and variance of R, G, B color components) and other features such as `brightness, intensity, luminance, hue, saturation, value` for the feature-based analyses.
Ma and Zhang [99] reported that contrast significantly influences the attention besides aspects such as color and intensity. Whether an element is perceived or not depends on its ability to *stand out* from the surroundings. Contrast helps to quantify this standing-out

nature. So, we considered `contrast` features of the image elements for our computational approaches.

Recently, Jana et al. [78] proposed a regression-based model to predict the attention (representative FI in a scanpath) on web images (i.e. image elements on a webpage). The authors divided the screen into three vertical regions and every image is constrained to entirely present in either of the regions. The total number of image elements on the screen are restricted to 6. The linear combination of size, position, and chromatic contrast features were utilized to learn an *attention-factor* which is further exploited for assigning the ranking order (FI) to images. However, in real-world webpages, the image position and size cannot be restricted to any specific region as they are prominent to draw user attention [61, 143]. Especially, different webpages intended for different set of users require the varied image position and size. Additionally, all the features may not affect the attention linearly. That is, a linear change in feature value may not correspond to a linear change in attention value. Most importantly, the Fixation-Index (FI) is a categorical value, but not continuous. Consequently, mapping the linear-regression output (attention-factor), a continuous value, to a categorical value (FI) results in inconsistent predictions based on the used mapping scheme. For example, in [78], attention-factor value of all six images were sorted to accordingly assign the fixation-indices from 1 to 6. Consequently, even if all the users made their first three fixations on three images and fourth fixation on some other element (say, text) and subsequent three fixations on remaining three images, their model still assigns the fixation-indices from 1 to 6 instead of $\{1, 2, 3, 5, 6, 7\}$. In contrast, if the magnitude of attention-factor value (instead of sorted index) is utilized for ranking, different fixation-indices (beyond 6) may be assigned for the images. On the contrary, the classification model's output variable itself is categorical which do not require external mapping, unlike the regression. Thus, classification-based modeling may be more appropriate than regression-based modeling.

2.5 Attention Models for Webpages

Web 2.0 [28] made the webpages multimodal. The webpages became rich of content consisting text, image, and video modalities. Among these modalities, only the former two modalities (text and images) are perceived dedicatedly through visual-alone sensory. This also motivated the HCI research community to propose attentional models for multimodal webpages.

Faraday proposed the first visual attention model of webpage users [51]. As per this theoretical model, a user selects a salient region on webpage followed by its scanning in a left-to-right pattern. A salient region is selected based on the presence of elements, called *Salient Visual Elements* (SVEs), in the decreasing order of preference: motion, size, image, color, text style and position. However, empirical studies demonstrated otherwise and invalidated the theoretical preferential ordering suggested by [51]. Especially, the element's position (center and top-left regions of a webpage) [61] and size [143] found to be predominant

in drawing the user attention. The subsequent analyses, focused dedicatedly on the user attention (we call them as, *User Attention-based approaches*) or dedicatedly on the webpage visual features (we call them as, *Visual Feature-based approaches*).

2.5.1 User Attention-based (UAB) approaches

Heatmap-based Inferences

The typical approach is to accumulate the users' attention allocations (fixations) on a webpage, and indicate the accumulated value with a color. Warmer colors (such as, red and orange) are used to designate a larger value, and cooler colors (such as, blue and green) are used to designate a lower accumulated value to result in a *heatmap*. A standing-out pattern consisting of warmer colors is assigned as users' attention allocation pattern. The prominent one of these patterns is the *F-pattern* of web content reading [105], according to which the users first scan in a long horizontal movement at the top of a webpage followed by a small downward movement on the left. Then, a shorter horizontal movement is observed followed by a complete downward movement on the left to result in an 'F' shape resembling heatmap pattern. Analogously, Alistair and Abdallah [4] investigated the user attention allocation on different designs of webpages in the context of browsing and searching tasks. The authors found that higher amount of fixation densities were allocated in the upper parts of the webpages, and users focused on information-scent (likely to contain the information corresponding to the goal) and salient objects during the search task. Subsequently, based on the observations, a heatmap prediction model was suggested for browsing and search tasks.

However, as the observed patterns are not associated with the underlying visual features, a multitude of other patterns emanated based on the webpage content and user attention allocations, such as the Z-pattern, the Zig-Zag pattern, the Layer-cake pattern, the Spotted pattern, the Commitment pattern, and more [16, 110]. Moreover, the heatmap-based approaches limit the webpage-specific interpretation. That is, even if the users' fixation data is available on a webpage, the designer need to apply one of the existing patterns (such as F-pattern) than eliciting a prominent pattern from the available data.

Scanpaths and AOI-based Inferences

The scanpaths help to overcome the limitations of heatmap-based inferences. Typically, the visible webpage region is divided into an ensemble of Areas-Of-Interest (AOIs), and the scanpath traversals across them is visualized. An example AOI selections are shown in Figure 2.6. As demonstrated by Eraslan et al. [48, 49, 50], such approaches help in identifying a trending scanpath on a webpage from the collection of scanpaths. However, as the visual attributes of AOIs are overlooked in such analyses, the identified trends are not generalizable to other webpages with varied visual features, or the same webpage with varied AOI selection.

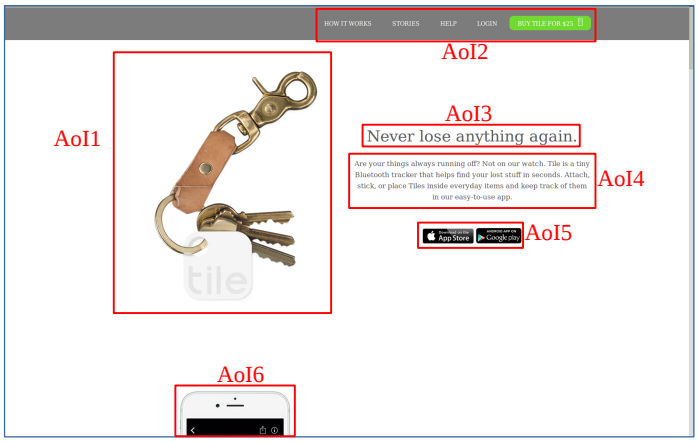


Figure 2.6: Example Areas-Of-Interest (AOIs) on a webpage

Eraslan et al. [48] presented the survey of eye-tracking based scanpath analysis techniques for webpages. All those techniques categorize under UAB approaches. In contrast to our terminology of *web elements* (motivated by Buscher et al. [19]), the authors utilized *visual elements* as each visual element may consist of multiple web elements based on the webpage segmentation level provided to the Vision based Page Segmentation (VIPS) algorithm [2]. That is, each visual element was designated as an AOI and four categories of scanpath analysis techniques were presented, as summarized.

- **Similarity/Dissimilarity Calculation:** these techniques helps to compare and determine if two scanpaths are similar or dissimilar. The corresponding approaches include *String-edit algorithm* [67], its improvized version *String-edit algorithm with a substitution matrix* [147], and *Scan-Match* technique [33]. These techniques are not centered on predicting a trending (or common) scanpath.

- **Transition Probability Calculation:** these techniques helps to determine the transition probabilities across the visual elements. These techniques are also not centered on predicting a trending (or common) scanpath. The approaches include *Markov Models* [25, 82, 159] and *eSeeTrack* [151] techniques.

- **Pattern Detection:** these techniques helps to either search for a given pattern or to discover all the patterns matching with the given scanpath pattern. The approaches include, (i) *eyePatterns* analysis tool— searches/ discovers the specific pattern among all available scanpths [159]; (ii) *Sequential PAttern Mining (SPAM)* algorithm— similar to eyePatterns tool but with a tolerance to additional visual elements present in the patterns; [65] (iii) *T(emporal)-pattern* detection— detects the patterns within scanpaths [18, 39, 101]. Here, T-pattern indicates two events (say, fixation allocation) either occurred more than once or appeared invariantly over time in a behaviour (attention allocation, in current settings) sequence.

- **Common Scanpath Identification:** in contrast to the aforementioned approaches, these techniques helps to achieve a representative scanpath (called *common scanpath*) from the given group of scanpaths. The approaches include, Multi Sequence Alignment [66], Hierarchical Clustering with Dotplots algorithm [57], Shortest Common Supersequence [117], Position-based Weighting model [4], and *eMine* scanpath algorithm [47].

Besides *generalizability*, the UAB approaches are limited by *feasibility* constraint. The UAB approaches are pattern eliciting than attention predictive; thus, users' attention data is required for every test webpage. Collection of this data on every webpage and especially after every small modification enforced on a webpage may not be practical due to the time-constraints, money-constraints, and user-availability constraints.

2.5.2 Visual Feature-based (VFB) approaches

To overcome the aforementioned limitations, data-driven computational approaches were proposed for attention modeling on webpages [14, 94, 100, 134, 144, 156]. The typical approach is to consider the whole webpage as an image (through screenshot capture), and apply the image processing-based saliency prediction techniques to identify the salient (potentially attention drawing) locations (a group of pixels). These approaches are adaptations of seminal work by Itti and Koch [74, 75, 76], which were directly applied on webpages (screenshots) to identify the web-viewing entry points [144].

In Masciocchi and Still [100], Itti's image saliency prediction model [76] was applied on webpages to correlate the saliency points with fixations (indicative of visual attention). In [156], the low-level (based on pixels) and high-level (based on semantics) feature maps were integrated to predict the saliency on webpages. Shen and Zhao [134] further extracted low spatial frequency (for images) and high spatial frequency (for text) feature maps, both are of the same dimension and integrated them to estimate the possible fixation on a given webpage position. Similarly, Li et al. [94] achieved saliency prediction through an integration of high frequency representation, sub-band features, position and face feature maps. In summary, the **saliency based models process the whole webpage as an image to predict the salient regions**.

The saliency-based computational approaches do not directly associate the webpage visual features with the user attention. Additionally, unlike an image, webpage is a segregable ensemble of elements from multiple modalities. The processing of whole webpage as an image (through screenshot) overlooks the intrinsic visual features of the underlying web elements (text and images). Moreover, these saliency-based techniques may determine a part of an element (say, text) as salient and the other part of the same element as non-salient, while, the designers may like to associate the visual characteristics for the whole element, such as through the Cascading Style Sheets (CSS) features.

2.5.3 Element Visual Features' and User Attention based Analysis

Buscher et al. [19] made efforts to overcome the limitations of UAB and VFB approaches by preserving the visual characteristics of individual web elements. They considered a restricted set of visual features to predict the attention on web elements for **information-foraging and page-recognition tasks**. The considered features include (i) 44 binary features corresponding to the HTML tags such as <A>, <H1>, <DIV>, <IMG>, etc. (assign '1' to the corresponding tag and '0' for all other tags); (ii) ten binary features to indicate the element's position in either of 3×3 grid regions or below the fold; (iii) one binary feature to indicate whether link element <A> points to top page of a website; (iv) one binary feature to indicate the presence of a logo; (v) one binary feature to indicate the depth in DOM tree; (vi) the remaining features represent the element's size and aspect ratio. However, there are several issues with the feature set as listed below.

- Predominantly, using a binary feature to represent an element discards all its intrinsic visual features. For example, a red color link and a blue color link are considered equivalent as both are identified using <A> tag.

- The tag (alone) consideration results in data inconsistencies as different tag nesting results in different feature values whereas user perceives the same data. For example, between '<DIV><H1>Hello World!</H1></DIV>' and '<H1><DIV>Hello World!</DIV></H1>', though same visual content 'Hello World!' is perceived, the former snippet considers <DIV> tag while the latter considers <H1> tag as the feature value resulting in data inconsistency.

- Additionally, '<DIV><DIV><DIV><H1>Hello World!</H1></DIV></DIV></DIV>' increases the depth in DOM by 3 levels when compared with <H1>Hello World!</H1> even though user notices no disparity in visual perception.

Broadly, the majority of HTML tags correspond to two prevalent modalities, text and image. For example, <H1>, <H2>, <H3>, <H4>, <H5> and <H6> all represent the text on webpage with varying visual representation. Similarly, <IMG> and <PICTURE> represent the images on a webpage.

Thus, studying the user attentional behavior with respect to the visual representation determining features (intrinsic visual features) would be more beneficial and helps the web-designers in estimating the prominence of a given web element. Further, these analyses help in incorporating the feature values directly into the web-designing. For example, user attentional analysis on intrinsic features `font-size, margin-top, margin-left, margin-right, margin-bottom`, and `font-weight` would help the designers in setting these feature values for any text element, beyond <H1> to <H6>.

2.6 Prediction Approaches

In this thesis work, classification-based prediction approaches are utilized to predict the attention on web elements at the higher granularity (Fixation-Indices). This section describes the corresponding literature that we employ for our analysis in Chapter 4, Chapter 6, and Chapter 7.

2.6.1 Multi-class Classification using SVM

This section describes the multi-class classification based on Support Vector Machine (SVM). The objective is to predict **an expected class** (a representative FI in the current setting) from the set of multiple possible classes for a given data-point (visual features of a web element in the current setting). Towards achieving this objective, classification techniques learn a *separating boundary* among the multiple classes based on the training data. Then, for a test data-point, the separating boundary is utilized to determine the expected class.

For multi-class classification, learning a single separating boundary is a complex task, whereas learning a separating boundary for binary classification is simpler. Thus, often, an ensemble of binary classifiers are constructed through which multi-class classification is achieved.

The SVM is the most popular classification technique primarily introduced for binary classification [15, 30]. Subsequently, it has been employed for multi-class classification through multiple reformulations [71]. Among them, "one-Vs-one" based formulation was shown to perform better for practical applications [71]. In this reformulation, for θ classes, $\binom{\theta}{2} = \theta(\theta-1)/2$ binary classifiers are trained to classify between any pair of classes. For a test data-point, the prediction results are obtained from all the trained binary classifiers and the class with the highest votes is assigned as the predicted class label. That is, a *majority-voting-scheme* is employed to assign the class label. In remainder of this subsection 2.6.1, the SVM based binary classification is described which we utilize for multi-class predictions.

Binary Linear SVM formulation

Consider d fixated web elements (text or images) represent points in $n-$dimensional feature space $\mathbb{R}^n$, where n is the number of visual features. That is, $\mathbf{x} = [\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_d]^\top \in \mathbb{R}^{n\times d}$. The respective class-labels of the d points are denoted as $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_d]^\top$ with $\mathbf{y}_k \in \{+1, -1\}\ \forall k$. Then, SVM achieves the classification by maximizing the margin (separation) between the data-points of two classes (fixation-indices). Without loss of generality, assuming the binary classifier need to be built for classes i (positive class) and j (negative class), the

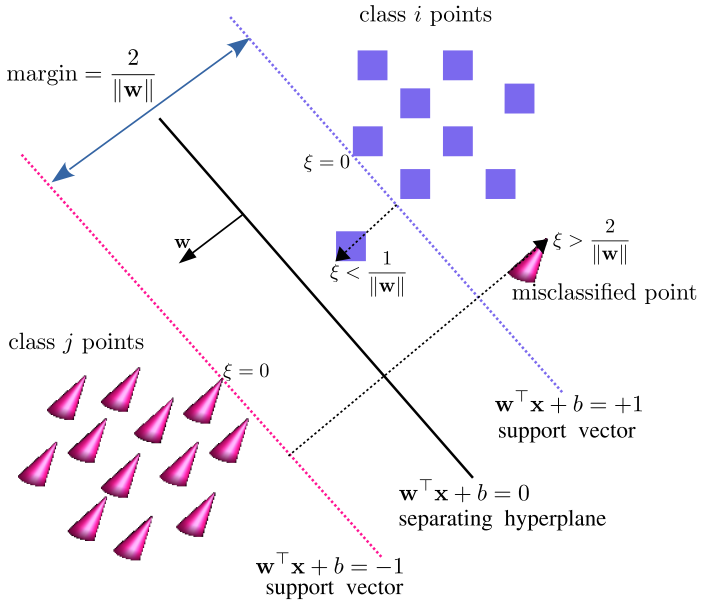


Figure 2.7: Linear SVM illustration for binary classification of classes i and j

SVM is formulated as the following constrained optimization problem [30].

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^\top \mathbf{w} + \lambda \sum_{k=1}^{d} \xi_k \tag{2.1}$$

$$\begin{aligned} \text{such that} \quad & \mathbf{w}^\top \mathbf{x}_k + b \geq 1 - \xi_k, \text{ if } f_k = i; \\ & \mathbf{w}^\top \mathbf{x}_k + b \leq -1 + \xi_k, \text{ if } f_k = j; \\ \text{and} \quad & \xi_k \geq 0; \ \forall \ k = 1, \ldots, d \end{aligned} \tag{2.2}$$

That is, the objective is to obtain a best separating boundary called *hyperplane*, $\mathbf{w}^\top \mathbf{x} + b = 0$, between the two classes as shown in Figure 2.7. The optimal hyperplane maximizes the margin $2/\|\mathbf{w}\|$ between the two classes to separate the two classes which is achieved by minimizing the $\mathbf{w}^\top \mathbf{w}$. In Equation 2.1, the parameters $\mathbf{w}$ and b determine the position of hyperplane. The slack variable ξ_k is useful for accommodating the data outliers as shown in Figure 2.7. The $\lambda > 0$ is a regularization parameter.

The class i points are bounded by $\mathbf{w}^\top \mathbf{x}_k + b \geq 1 - \xi_k$ and class j points by $\mathbf{w}^\top \mathbf{x}_k + b \leq -1 + \xi_k$, where $k = 1, \ldots, d$. The vectors passing through the boundary points (with respect to the other class) are the *support vectors* (shown in Figure 2.7) that help in determining the hyperplane.

The optimization problem in Equation 2.1 is solved for $\mathbf{w}$ and b using the Lagrangian

formulation [22, 30]. The class of a test web element $\mathbf{x}_t$ is determined based on the sign of

$$\text{sign}\left(\sum_{k=1}^{d} \mathbf{y}_k \beta_k \mathbf{x}_k^\top \mathbf{x}_t + b\right) = \begin{cases} i, & \text{if } \textit{positive} \\ j, & \text{if } \textit{negative} \end{cases} \tag{2.3}$$

where β_k, $k = 1, 2, \ldots, d$ are the Lagrangian multipliers [30].

Binary Kernel SVM formulation

The aforementioned formulation is useful if the data-points are linearly separable. However, the separability may not hold always. In those non-linear settings, the data-points are projected into a high-dimensional space using non-linear transformation function $\phi(\mathbf{x})$. In the high-dimensional space, the separating hyperplane $(\mathbf{w})^\top \phi(\mathbf{x}) + b = 0$ is obtained, where parameters $\mathbf{w}$ and b decide the orientation of hyperplane. Without loss of generality, considering ith class as positive class and the jth class as negative class, the separating hyperplane is supported by class-specific linear boundaries called support vectors in the classification as given by $(\mathbf{w})^\top \phi(\mathbf{x}) + b = +1$ and $(\mathbf{w})^\top \phi(\mathbf{x}) + b = -1$ respectively. That is, class i datapoints are bounded by $(\mathbf{w})^\top \phi(\mathbf{x}) + b \geq +1$ and class j datapoints by $(\mathbf{w})^\top \phi(\mathbf{x}) + b \leq -1$. The optimal hyperplane achieves a maximum margin— magnitude separation between class-specific support vectors— of $2/\|\mathbf{w}\|$ [30] as shown in Figure 2.7. The objective is to maximize the margin $2/\|\mathbf{w}\|$ which is equivalent to the minimization of the denominator $\|\mathbf{w}\|$, to result in following constrained optimization problem analogous to Equation 2.1.

$$\min_{\mathbf{w},b,\xi} \frac{1}{2} (\mathbf{w})^\top \mathbf{w} + \lambda \sum_{k=1}^{d} \xi_k \tag{2.4}$$

$$\begin{aligned} \text{such that} \quad & (\mathbf{w})^\top \phi(\mathbf{x}_k) + b \geq 1 - \xi_k, \text{ if } f_k = i \\ & (\mathbf{w})^\top \phi(\mathbf{x}_k) + b \leq -1 + \xi_k, \text{ if } f_k = j \\ & \text{and} \quad \xi_k \geq 0 \; \forall \; k = 1, \ldots, d. \end{aligned} \tag{2.5}$$

where, $\lambda > 0$ is a regularization parameter, and d is the cumulative number of datapoints in class i and class j. The slack variable ξ controls the classification threshold and accounts for noise (outliers) in the data as shown in Figure 2.7. For $\xi < 1/\|\mathbf{w}\|$, the separating hyperplane acts as final decision boundary with the mis-classifications occurring at $\xi > 2/\|\mathbf{w}\|$ as shown in Figure 2.7. Analogous to Equation 2.1, the optimization problem Equation 2.4 is solved using Lagrangian formulation and primal-dual reformulations [22, 30]. Accordingly, class i or class j is determined for a test webpage image $\mathbf{x}_t$ based on the sign of

$$\text{sign}\left(\sum_{k=1}^{d} f_k \beta_k K(\mathbf{x}_k, \mathbf{x}_t) + b\right) \tag{2.6}$$

where, β_k is the weight of kth datapoint such that $0 \leq \beta_k \leq \lambda$; $K(\mathbf{x}_k, \mathbf{x}_t) \equiv \phi(\mathbf{x}_k)^\top \phi(\mathbf{x}_t)$ is the kernel function that elicits the non-linear association between visual features and the quantitative attention.

Multiclass SVM

The multi-class classification is achieved through the "one-Vs-one" reformulation [71]. That is, for θ classes, $\binom{\theta}{2} = \theta(\theta-1)/2$ binary classifiers are trained and the test predicted class is obtained from each of them. A majority-voting-scheme is employed to assign the class with maximum votes as a class label to the test element.

Performance Metrics of Classification

In classification and prediction tasks, the correctly classified positive and negative instances are respectively called *true positives (TP)* and *true negatives (TN)*. Accordingly, the negative instances that are incorrectly predicted as positives are *false positives (FP)* and the positive instances that are incorrectly predicted as negatives are *false negatives (FN)* [162]. Based on these notations, the following metrics are defined to analyze the performance of the classification model.

Precision is the proportion of correctly predicted instances among all the positively predicted instances.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.7}$$

Recall is the proportion of correctly predicted instances among the given positive instances.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.8}$$

Kappa κ-statistic [26] helps to check the agreement between predicted and observed labels while accounting for the agreement occurring by chance [162].

$$\kappa = \frac{\text{overall accuracy} - \text{expected chance accuracy}}{1 - \text{expected chance accuracy}} \tag{2.9}$$

Average Accuracy is the average of accuracies achieved from all the classifiers.

$$\text{Average Accuracy} = \frac{\sum_{i=1}^{\theta} (TP_i + TN_i)}{\sum_{i=1}^{\theta} (TP_i + TN_i + FP_i + FN_i)} \tag{2.10}$$

Micro Precision, Micro Recall and Micro F1-score metrics are utilized for analyzing the performance of classifier on non-uniformly distributed classes.

$$\text{Micro Precision} = \frac{\sum_{i=1}^{\theta} TP_i}{\sum_{i=1}^{\theta} (TP_i + FP_i)} \tag{2.11}$$

$$\text{Micro Recall} = \frac{\sum_{i=1}^{\theta} TP_i}{\sum_{i=1}^{\theta} (TP_i + FN_i)} \tag{2.12}$$

$$\text{Micro F1-score} = \frac{2 \times \text{Micro Precision} \times \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}} \tag{2.13}$$

2.6.2 Multi-label Classification

A data-point may associate with multiple labels simultaneously (motivated in Section 4.6 for image elements). In such settings, predicting as many *relevant* labels as possible would be the objective of a prediction model. For this settings as well, SVM is utilized. As we limit this approach to predict multiple Fixation-Indices (FIs) on image elements, the remainder of this subsection 2.6.2 is explained in the context of image elements.

Consider d unique web image elements, each represented using n visual features. Let the set of possible fixation-indices be $\mathcal{L} = \{z_1, \ldots, z_f\}$. Representing each image as a point in n-dimensional space $\mathcal{I}$, the associated labels represent points in f-dimensional space, $\mathcal{Y} = \{+1, -1\}^f$. Thus, the ith image is represented using the feature vector $\mathbf{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\right)^\top$ and the associated multilabels as $\mathbf{y}^{(i)} = \left(y_1^{(i)}, y_2^{(i)}, \ldots, y_f^{(i)}\right)^\top$. Note that $\mathbf{y}^{(i)}$ is a binary-valued vector with $+1$ indicating the presence of a fixation-index and -1 for its absence. For example, $y_2^{(i)} = +1$ if the image is associated with the label z_2 which refers to second fixation-index.

However, the latter fixation-indices may be guided by other factors such as absence of (or exhausted) salient elements, influence of semantic features, etc. To mitigate their influence on free-viewing attention prediction and to segregate the initial fixation-indices from latter fixation-indices, we introduce a fixation-index threshold θ such that $\theta \leq f$ resulting in $z = \{z_1, \ldots, z_\theta\}$. Accordingly, $\mathbf{y}^{(i)}$ is modified as $\mathbf{y}^{(i)} = \left(y_1^{(i)}, y_2^{(i)}, \ldots, y_\theta^{(i)}\right)^\top$ Then, the objective is to predict the set of fixation-indices for a test image element $\mathbf{y}^{(t)}$ using its associated visual features $\mathbf{x}^{(t)}$. The current setting is typically considered as a multilabel classification problem where we need to predict the binary values for each of $\mathbf{y}^{(t)} = \left(y_1^{(t)}, y_2^{(t)}, \ldots, y_\theta^{(t)}\right)^\top$.

Multiple approaches exists in the literature to solve this multilabel prediction problem [114, 152]. Due to the complexity of multilabel classification, often, the problem is transformed into multiple binary classification problems which is called a *problem transformation method*. The simpler binary classification is then solved through state-of-the-art approaches such as

support vector machine (SVM) [30].

As the fixation-indices obtained from one user are independent of another user, we considered the *Binary Relevance (BR)* method based problem transformation method. In this approach, θ binary classifiers are learned each corresponding to a fixation-index. For a test image element, the binary value +1 or -1 is assigned as a *relevancy* score based on the decision from corresponding binary classifier. The resulting set of Fixation-Indices (FIs) forms a label-set $\widehat{L_t}$ while the ground-truth label-set is denoted as L_t. Thus, the kth binary classifier is a mapping function

$$C_k : \mathcal{I} \longrightarrow \{+1, -1\} \qquad k = 1, \ldots, \theta \tag{2.14}$$

Accordingly, the predicted multilabels of test image $\mathbf{x}^{(t)}$ are

$$\widehat{\mathbf{y}}^{(t)} = \widehat{L_t} = \Big(C_1(\mathbf{x}^{(t)}), C_2(\mathbf{x}^{(t)}), \ldots, C_\theta(\mathbf{x}^{(t)})\Big)^\top \tag{2.15}$$

Performance Metrics

Table 2.1: Description of Multi-user Multi-level Attention Prediction Performance Metrics

Metric	Description	Computation	Desired
Subset 0/1 loss	Proportion of mis-matches where even a single mis-classification for an image instance is considered as complete mis-match	$\frac{1}{d}\sum_{i=1}^{d} \mathbb{1}_{\left(L_i \neq \widehat{L_i}\right)}$	↓ (low)
Hamming loss	Proportion of mis-matches accounting only for the mismatched label instances	$\frac{1}{d}\sum_{i=1}^{d}\frac{1}{\theta}\sum_{j=1}^{\theta} \mathbb{1}_{\left(L_i^{(j)} \neq \widehat{L_i}^{(j)}\right)}$	↓ (low)
Accuracy	proportion of correctly classified labels (analogous to average accuracy in single-label prediction)	$\frac{1}{d}\sum_{i=1}^{d}\frac{\sum_{j=1}^{\theta} \mathbb{1}_{\left(L_i^{(j)}=+1 \;\&\&\; \widehat{L_i}^{(j)}=+1\right)}}{\sum_{j=1}^{\theta} \mathbb{1}_{\left(L_i^{(j)}=+1 \;\|\; \widehat{L_i}^{(j)}=+1\right)}}$	↑ (high)
Precision (PPV)	(analogous to single-label prediction) proportion of correctly predicted labels among all the positive labels	$\frac{1}{d}\sum_{i=1}^{d}\frac{\sum_{j=1}^{\theta} \mathbb{1}_{\left(L_i^{(j)}=+1 \;\&\&\; \widehat{L_i}^{(j)}=+1\right)}}{\sum_{j=1}^{\theta} \mathbb{1}_{\left(\widehat{L_i}^{(j)}=+1\right)}}$	↑ (high)
Recall (TPR)	(analogous to single-label prediction) proportion of correctly predicted positive labels among all the actual labels	$\frac{1}{d}\sum_{i=1}^{d}\frac{\sum_{j=1}^{\theta} \mathbb{1}_{\left(L_i^{(j)}=+1 \;\&\&\; \widehat{L_i}^{(j)}=+1\right)}}{\sum_{j=1}^{\theta} \mathbb{1}_{\left(L_i^{(j)}=+1\right)}}$	↑ (high)
F1-score	(analogous to single-label prediction) harmonic mean of precision and recall	$\frac{1}{d}\sum_{i=1}^{d}\frac{2\sum_{j=1}^{\theta} \mathbb{1}_{\left(L_i^{(j)}=+1 \;\&\&\; \widehat{L_i}^{(j)}=+1\right)}}{\sum_{j=1}^{\theta}\left(\mathbb{1}_{\left(\widehat{L_i}^{(j)}=+1\right)} + \mathbb{1}_{\left(L_i^{(j)}=+1\right)}\right)}$	↑ (high)

Unlike the binary classification or multiclass classification (which take only one output label), the predicted output of multilabel classification takes multiple labels and accordingly may

partially match the ground-truth output (or set of labels). Thus, the distinct performance metrics are introduced in the literature to estimate the performance of a multilabel classifier. In Table 2.1, the performance metrics computation is summarized using the following indicator function for simplicity, where C denotes the any of $C_1, \ldots, C_k$ binary classifiers.

$$\mathbb{1}_{(\mathbf{y}^{(i)} \neq C(\mathbf{x}^{(i)}))} = \begin{cases} 1 & \text{if} \quad \mathbf{y}^{(i)} \neq C(\mathbf{x}^{(i)}) \\ 0 & \text{if} \quad \mathbf{y}^{(i)} = C(\mathbf{x}^{(i)}) \end{cases} \tag{2.16}$$

In Table 2.1, subset 0/1 loss and Hamming loss indicate the poorer prediction performance, and consequently, lower values are preferred. Between the two, Hamming loss allows for partial match whereas the subset 0/1 loss considers the complete match (that is, correct prediction of all multilabels of a test image). On contrary, other metrics are analogous to the typical performance metrics of binary and multiclass classification and higher values are preferred for these metrics.

2.7 List of Visual Features

We considered the following web-designing oriented visual features that characterize the visual representation of text and image elements on webpages.

Visual Text Features

Font-size Size of the font that corresponds to `em` square in typography [12, 29]. Larger the font-size bigger is the text character's visual appearance on the webpage. Font-size may be indicated in either of the four ways: *absolute-size, relative-size, length*, and *percentage*. The *absolute-size* values are 'xx-small', 'x-small', 'small', 'medium', 'large', 'x-large' and 'xx-large' [29]. The 'medium' value is reference medium value being the user-preferred default font-size (typically, 16 `px`). Other absolute-sizes are scaled with respect to their appearance in the aforementioned sequence. For example, 'x-large' font-size is $16 \times 6/4 = 24$ `px` for a 16 `px` medium font. The *relative-size* values are interpreted in relation to the parent element's (obtained from traversing the DOM tree) font-size and the absolute-sizes. The two possible relative-sizes are 'larger' and 'smaller'. For example, a text element's font-size results in 'small' when its relative-size is 'smaller' and its parent's absolute-size is 'medium'. The *length*-based font-size reports the absolute, fixed size of the text for rendering on the screen and to be measured in the units of `cm, mm, in, pt, pc`, and `px`. To overcome the adaptability difficulties of the fixed-size fonts, relative length units are adopted for denoting the font-sizes, such as `em, ex, ch, rem` and more. The font-size indicated in *percentage* scales the parent element's font-size to the mentioned percentage and assigns it to the considered text element. For the sake of unified representation, we transformed all the rendered text's font-sizes into `px` units for our analysis.

Font-weight indicates the weight to be applied to the font, that is, thickness or thinness to the existing font. The possible values are $100, 200, \ldots, 900$. The larger the value, larger is the thickness with value 400 indicating the "normal" font-weight and 700 indicating the "bold" font [29]. Additionally, "bolder" and "lighter" values are used to set the font thickness with respect to the font-weight of the parent element.

Line-height determines the height of text line. That is the amount of space allocated for each line of text. Increasing the line-height for a specific font-size increase the space around the corresponding line. The line-height is either specified as a number that indicates a scaling with respect to the font-size or "normal" which is a default line-height, or in *length, percentage* units as described in **Font-size**. All units are converted to the equivalent `px` units after rendering the text on webpage.

Color The color feature is associated with multiple text element's characteristics including **background, outline, text emphasis, text decoration, column rule, border** and the **text** foreground itself. The color value is reported either as a color name (such as "red", "yellow" and "blue") or using the six-digit hexadecimal code (for example "#00FF00" denotes the green color) corresponding to RGB color components or through a decimal number $\in [0, 255]$ for each of the RGB color components (for example, "rgb(0,0,255)" indicates a blue color). It is not uncommon to denote the transparency of the color through the *alpha* channel. An alpha = 0.0 indicates complete transparency while alpha = 1.0 indicates a fully opaque color. That is, rgba(255,0,0,1.0) indicates an opaque red color. For the sake of unified representation, we extracted the color features and converted them into an equivalent red, green, blue color component and alpha feature values.

Padding The area occupied by the rendered text is called *content area*. The content area is optionally surrounded by a border. The padding sets the space between the rendered content and the border. The larger the padding value larger is the space between content and the border. The padding is described using four features corresponding to the four rectangular edges of the border, padding-left, padding-right, padding-top and padding-bottom. The possible values are *length* and *percentage* based metrics, as earlier described for **Font-size**.

Border-size Border is the rectangular edge outside of the content area and padding. The width of the border along the four rectangular edges is denoted by border-size or border-width. In case of zero width border, the border edge matches with padding edge [29]. Analogous to "padding", border-size contributes four visual features, border-left, border-right, border-top and border-bottom. The possible values are "thin", "medium" and "thick" besides the earlier defined *length* based metrics. The former three values are in the increasing sequence of magnitudes whereas their absolute value is user agent-dependent but consistent throughout the document [29].

Margin Margin sets the space outside the text element, that is, outside any of the defined borders. The larger is the margin larger will be the outer space around the text element. This feature also contributes four visual features describing the space along the four rectangular edges, margin-top, margin-bottom, margin-left and margin-right. The possible values are *length* and *percentage* based metrics.

Column-gap When the text is represented as columns, the desired space between the columns is specified by "column-gap". The larger is the column-gap, larger is the spatial separation between the columns. The in-between column separation lines, if any, will be placed in the center of the column-gap. The possible values are *length*-based metrics or "normal" which is suggested to be 1 `em` by the World Wide Web Consortium (W3C).

Text-indentation indents the first line of the text block [29]. The larger the text-indentation value larger is space before starting the first line in a text block. The text-indentation value can be negative or more than the size of the text block in which case the first line overflows the text block. The possible values are *length* and *percentage* based metrics; the former indicates the absolute indentation and the latter indicates the percentage of indentation with respect to the width of the text block.

Font-family specifies the decreasing prioritized list of fonts to be applied to the corresponding text element. Each value is an alternative to their preceding value in the list and all values are separated by commas. That is, the browser tries to apply the first font-family to the text, in case of failure, tries to apply the second font-family and so on from the prioritized list. The feature is assigned with specific font-family names such as "Gill" and "Helvetica", or generic font-family names such as "serif", "sans-serif", "cursive", "fantasy" and "monospace". Typically, generic font-families are used for fallback mechanism and are placed at the last position in a prioritized list. More details about the font-family feature are available in [29]. Unlike the above text features, font-family values are not ordinal, that is, we can not order the "Gill" and "Helvetica" to indicate which is bigger than the other. Thus, font-family feature contributes a categorical feature.

Position The distance of the element's border box edges from the top and left edges of the displayed webpage. That is, top and bottom positional values indicate the element's top border edge and bottom border edge distance from top edge of the displayed webpage. Similarly, left and right positional values indicate the element's left border edge and right border edge distance from left edge of the displayed webpage. These computations ensure that "right distance" $\geq$ "left distance" and "bottom distance" $\geq$ "top distance".

Area visible text element's area occupied on the screen. The four positional features are

utilized in computing the area as (right distance - left distance)×(bottom distance - top distance).

Note that setting **Font-size, Font-weight, Line-height, Color, Padding, Border-size, Margin, Text-indentation**, and **Font-family** features to "inherit" assigns corresponding parent element's feature value to the respective element's feature. The parent element is determined from the DOM tree.

Visual Image Features

Consider the image is in RGB color format where R_i, G_i, B_i respectively denote the red (R), green (G) and blue (B) color component values at the ith pixel. Without loss of generality, we employed the following formulations to compute the visual features of each image with N pixels.

Color histograms: Each color component value of a pixel is indicated on a scale of 0 to 255. As there are 256 levels, it is typical to distribute the color frequencies into eight uniformly spaced bins ($\log_2 256 = 8$). To account for the gray images, we computed the gray color histograms. For the colored images, *decolorization*— color image to gray-scale transformation [98], was performed to determine the gray component values using $0.2989 \times R + 0.5870 \times G + 0.1140 \times B$ (rgb2gray() function in Matlab). Accordingly, red, green, blue and gray components contribute a total of 32 features.

Luminance: Perceived luminance is computed as a linear combination of R, G, B color components, Luminance $= 0.299 \times R + 0.587 \times G + 0.114 \times B$ [11, 53]. The weights used in Luminance computation corresponds to the weights used in RGB to gray-scale conversion after rounding the digits to three decimal places. We compute the mean value of luminance to assign a uni-value to the image. That is, Luminance $= \frac{1}{N} \sum_{i=1}^{N} 0.299R_i + 0.587G_i + 0.114B_i$.

Color spread To quantify the spread of the color, mean and variance are computed for each color component of an image. Accordingly, color spread provides six image visual features.

Brightness: Though 'luminance' is used synonymously with the brightness [11], the latter is typically computed as an arithmetic mean of R, G, B color components. That is, Brightness $= \frac{1}{N} \sum_{i=1}^{N} \frac{R_i+G_i+B_i}{3}$.

Hue, H: Smith [140] introduced the HSV color model to approximate the perceptual properties of 'Hue', 'Saturation' and 'Value' [132]. 'Hue' is a traversal on the color circle from red color to yellow, green, cyan, blue, magenta and back to the red color, indicated with a value in $[0, 1]$. In simple words, 'Hue' is something that left after

removing the desaturation or gray component from a given color. That is, 'Hue' is a mixture of at most two-color primaries whereas 'color' is a mixture of at most three color primaries. The normalized R, G, B color components are utilized in computing the 'Hue' using Hexcone model [140].

Saturation, S: Saturation measures the hue's departure from white or gray which are *achromatic*. That is, increasing S decreases the whiteness in an image. The saturation is zero for gray color, $R = G = B$. Hue is not defined at zero saturation points, and in fact, its value is immaterial as these are gray points [140].

Value, V: Introduced by Munsell [104], "value" corresponds to "lightness" of a color [132]. That is, "value" measures the departure of hue from the zero energy color, black [140]. The smaller the value of V the higher is the blackness in the image. Assuming the R, G, B color components are normalized to $[0, 1]$, V corresponds to the largest color component value. That is, $V = \frac{1}{N} \sum_{i=1}^{N} \max(R_i, G_i, B_i)$.

Contrast: To quantify the contrast of an image on a given webpage, all of the above visual features are computed for webpage screenshot. The difference between respective visual features of a webpage and the image results in contrast features. That is, mean luminance, mean hue, *etc.* of an image are respectively subtracted from mean luminance, mean hue, *etc.* of the respective webpage.

Position: Image element's position on the corresponding webpage. It is an ensemble of four distances, 'top distance', 'bottom distance', 'left distance' and 'right distance' where former two distances are computed from the screen's top edge and the latter two from screen's left edge.

Area: visible image area occupied on the screen. Computed using the 'Position' feature as (right distance - left distance)×(bottom distance - top distance).

2.8 Summary

In this chapter, the preliminaries and the existing works are presented to introduce the perspective for our thesis work in the context of literature. The prominent point concerning the webpages is, they are logically segregable unlike the standalone images. We also described that eye movements indicate the users' cognitive and perceptual behaviour. The fixations (standstill in eye movements for a certain threshold of time) act as proxy for visual attention. Among the modality-independent features, the element's `position` and `size` (or area) are predominant in drawing the user attention. The *tag*-based consideration discards the intrinsic visual features of the respective elements. We also described the utilized prediction approaches and the list of visual features considered for our analysis for remainder of the thesis.

3

Eye-tracking Experiments and Visual Features

Eye-tracking experiments help to collect and prepare the ground-truth user attention towards the analysis of computational attention models. In total, we utilized the fixation data from three free-viewing eye-tracking experiments. As very few works are centered on DOM based attention modeling (in contrast to image-based attention modeling), we ourselves conducted two of those free-viewing experiments, Experiment-I (Section 3.2) and Experiment-II (Section 3.3). However, to demonstrate the efficacy of proposed approach on unseen data, we utilized the free-viewing experiment dataset from [134] which we denote as Experiment-III (Section 3.4). Further, this chapter summarizes the visual features of text and image modalities (Section 3.5) that are utilized throughout the thesis work.

3.1 Common Experimental Setup

3.1.1 Apparatus

A host computer decorated with the eye-tracking experimental setup is shown in Figure 3.1. The involved apparatus components are as follows.

1. *Tobii X2-60* eye-tracker[1]— a binocular eye-tracker that samples the gaze at 60 Hz (60 times in a second) and called a "fixation picker" [86] was utilized for our fixation-based research. The eye-tracker's accuracy (the average difference between real stimuli position and measured gaze position) is 0.4° and the precision (variation of the recorded data measured via root mean square of successive samples) is 0.34°.

2. *EPU*— external processing unit to perform the gaze related calculations which reduces the computational load on the host computer.

[1] https://www.tobiipro.com/product-listing/tobii-pro-x2-60/

3. *Ethernet cable* to transfer the processed gaze-data from EPU to host CPU; *magnetic strip* to attach the eye-tracker onto the display monitor.

4. A power connector to supply the power to EPU.

5. A magnetic strip to mount the eye-tracker.

6. A host computer with Windows 7, 64-bit operating system and 8 GB RAM was used for rendering the stimuli on 22" display monitor.

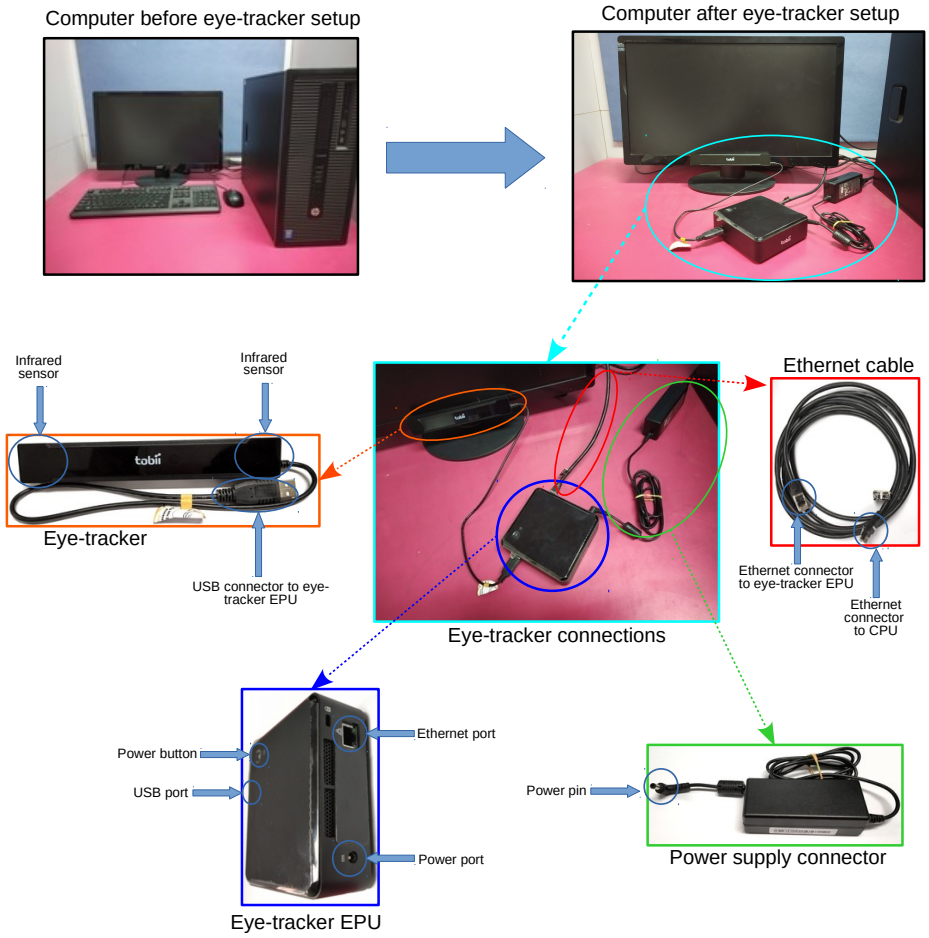


Figure 3.1: Eye-tracker setup used for collecting the gaze-data

The magnetic strip was attached at the bottom edge of display monitor onto which the eye-tracker was mounted for non-intrusive eye-movement tracking. The eye-tracker was connected to the EPU to send the sampled gaze-data for processing. The I-VT filter (Velocity-Threshold identification fixation filter [90, 128]) is used to detect the fixations. The I-VT filter assigns an angular velocity (measured in degrees/second) to each gaze data point, and all the points with the velocity less than a threshold (default 30 degrees/second is utilized as per [90, 106, 122]) are classified as *fixation*; The fixation location is an arithmetic

mean of all the points corresponding to a fixation. The EPU was connected to the host CPU using Ethernet cable for storing the processed data for further analysis. The necessary apparatus connections are shown in Figure 3.1.

3.1.2 Stimuli Preparation

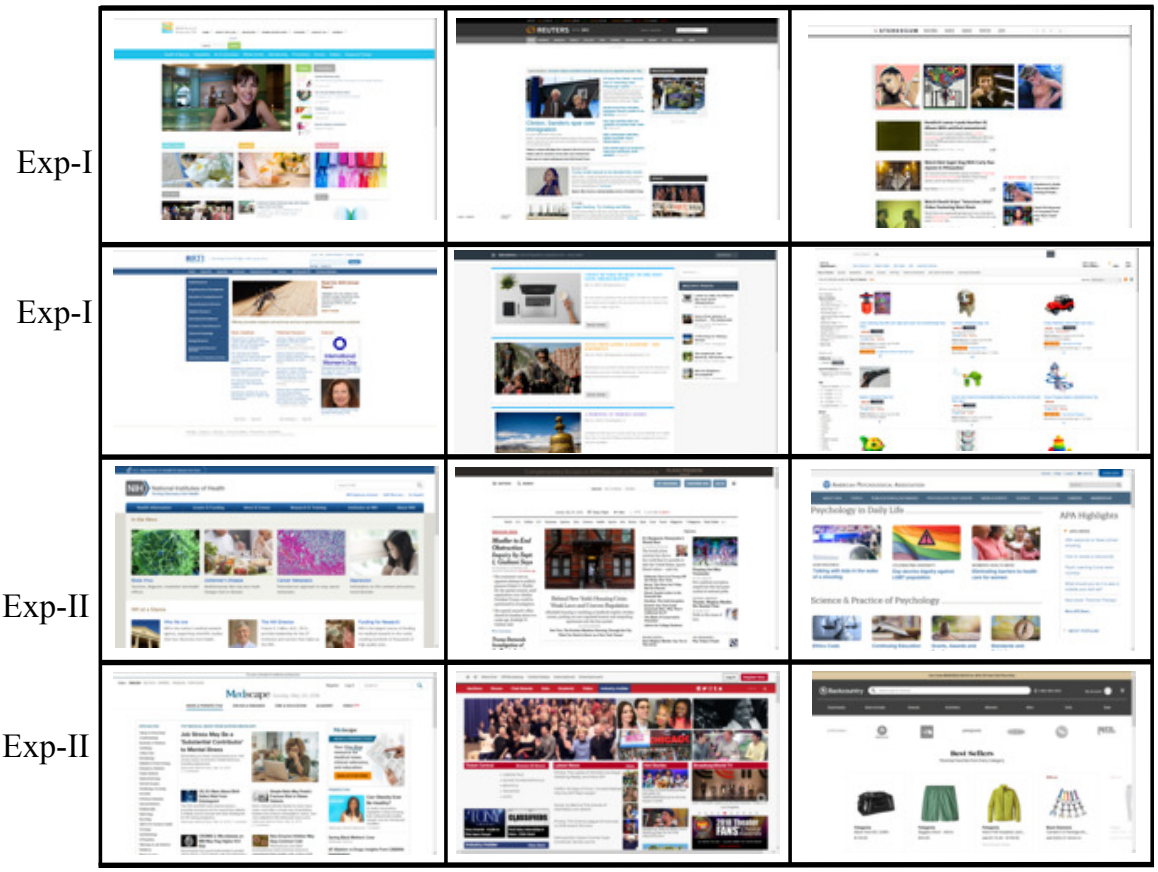


Figure 3.2: Sample stimuli webpages of Experiment-I (top two rows) and Experiment-II (bottom two rows).

In order to account for the user attentional behavior on webpages with varied visual features, we considered the webpages from multiple categories including SHOPPING, HEALTH, ARTS, HOME, NEWS, and SPORTS from Alexa[2]. The sample stimuli of Experiment-I[3] and Experiment-II[4] are shown in Figure 3.2. Each webpage was displayed in full-screen mode on a 22" display monitor and screenshot was captured (at a screen resolution shown in Table 3.1). The screen-shot helps in rendering the webpage as a whole avoiding the rendering latencies of individual web elements which may influence the attention allocation. Additionally, the stimuli webpages (including HTML, CSS, and media) were stored locally

[2] https://www.alexa.com/topsites/category
[3] Stimuli screenshots & webpages are available at http://bit.ly/2DaUoup & https://bit.ly/3l8avjv.
[4] Stimuli screenshots & webpages are available at https://bit.ly/2C8JmKP & https://bit.ly/3l8avjv.

Table 3.1: Summary Experimental Details

Characteristic	Experiment-I	Experiment-II	Experiment-III
Number of webpages	51	36	149
Number of participants	65 (33M; 32F)	42 (21M, 21F)	11 (4M, 7F)
Min. age of participants	21 years	20 years	21
Max. age of participants	34 years	32 years	25
Avg. age of participants	24.32 years	26.93 years	NA
Standard deviation of age	3.24 years	2.69 years	NA
Screen resolution	1680×1050	1280×720	1360×768
Fixation-threshold	100 ms	100 ms	100 ms
Each webpage display duration	5 seconds	5 seconds	5 seconds
Number of fixations on Text	11722	5726	NA
Number of fixations on Images	3013	2223	5762
Total number of fixations	14735	7949	5762

for the purpose of replication and reproducibility.

3.1.3 Experimental Setup

A slide-show presentation was prepared with a blank slide inserted after every webpage to reset the participants' attention after every webpage. Counter-balance mode of stimuli presentation was considered to mitigate the ordering effects [161]. A 22" desktop monitor was used for displaying the stimuli. Each slide (including the blank webpage) was presented for five seconds duration which is typical for free-viewing analyses [134].

3.1.4 Procedure

Participants (see the details in Table 3.1) were individually invited to the experimental setup. Each participant was briefed about the experimental setup. Instructions were provided that all the webpages are independent (to avoid the *expectational bias* about presentation sequence). In pre-experimentation step, eye-tracker calibration was performed using a 5-point method (participants gazed at the center of a red-color ball when it was moving along the four corners and the center of the screen as shown in Figure 3.3, second from left). Quality of the calibration was assessed and recalibration was performed, in case of errors. After successful calibration, an automated slideshow was presented using *counter-balanced* mode (to mitigate the stimuli ordering effects on attention) which was *free-viewed* by the participant while their eye-movements were captured by the eye-tracker. In the post-experimental survey, participants were asked *how many of the webpages shown in the experimentation was previously seen by you?* for which each participant reported a number from 0 to the number of stimuli webpages. The reported median of zero known webpages mitigated the possible *memory-bias* effect.

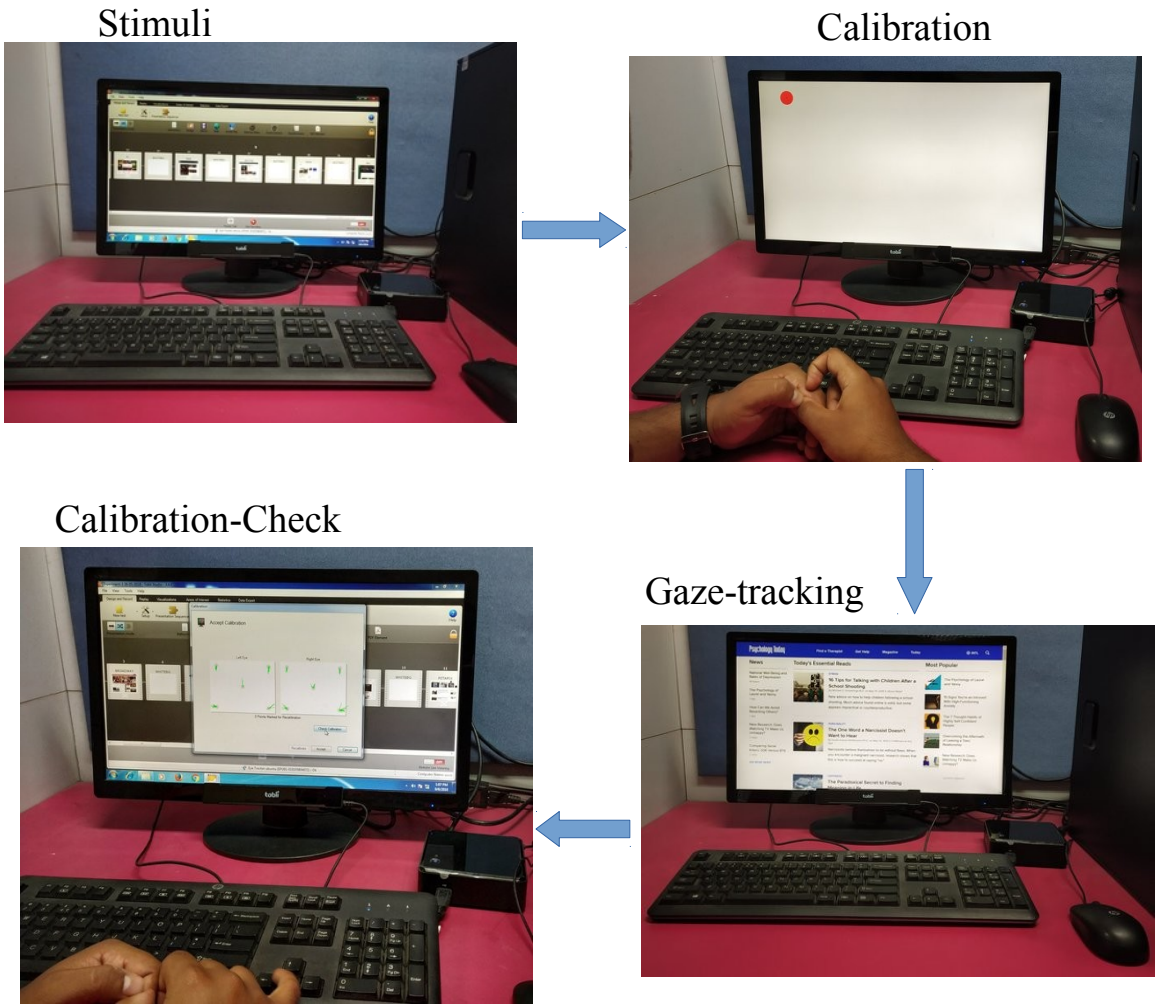


Figure 3.3: Experimental procedure in four steps: (i) stimuli loading (ii) eye-tracker calibration for a participant (iii) checking for successful calibration (iv) gaze-tracking during stimuli presentation

3.1.5 Data Extraction

From users' gaze data, fixations were extracted with the fixation duration of at least 100 milliseconds which is a typical threshold [55]. Accordingly, Fixation-Indices (FIs) were assigned. Now, the locally stored webpages were launched in a headless browser to replicate the screenshots of stimuli webpages, as seen by the participants. Based on the fixation-position, the DOM tree of the corresponding webpage is traversed (using a JavaScript program) and the leaf node element is identified, analogous to [19]. The corresponding fixated elements were identified from the fixation location. Among the extracted fixations, the repetitive fixations (multiple fixations from the same participant on a web element) are pruned. The repetitions might be due to the lack of further salient elements on a webpage, where, as per *inhibition-of-return*, participants are less likely to attend the already attended element [89]. Consequently, we considered the FI corresponding to the first attention

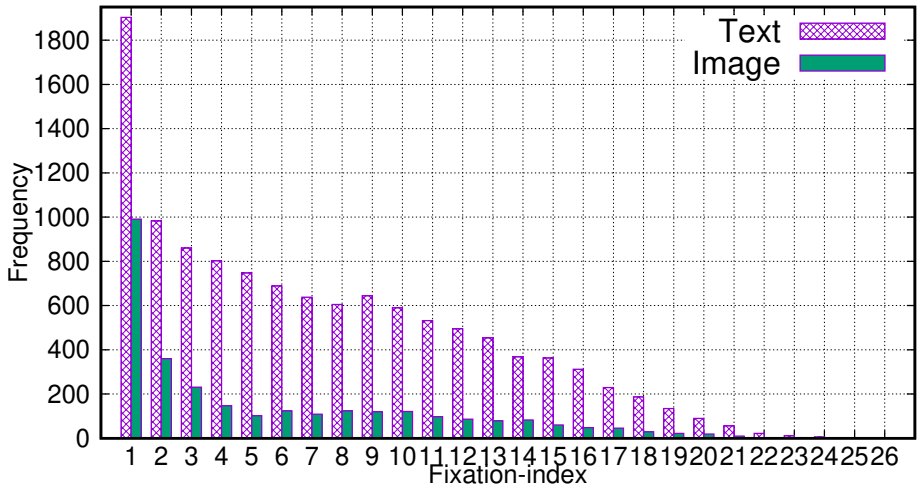


Figure 3.4: Experiment-I: Frequency distribution of fixation-indices

allocation (on the element) for the analysis.

3.2 Experiment-I Data

The frequency distribution of FIs of both the modalities obtained from Experiment-I is shown in Figure 3.4. The number of fixations on text is 11722 and the number of fixations on images is 3013.

The preliminary observation shows that more number of text elements were fixated than the image elements for respective FIs. This is attributed to the presence of relatively more number of text elements than image elements on bi-modal webpages. However, for both modalities, the frequency of an FI (attention-levels) decreases with increasing FI value. Consequently, the latter FIs are relatively sparser being associated with lesser number of elements.

This dataset is widely utilized for demonstrating the efficacy of proposed element-driven computational approaches on multimodal webpages (in Chapter 5, Chapter 6 and Chapter 7).

3.3 Experiment-II Data

The frequency distribution of FIs for both the modalities obtained from Experiment-II is shown in Figure 3.5. The number of fixations on text is 5726 and the number of fixations on images is 2223.

Analogous to Experiment-I, more number of text elements are associated with an FI than the image elements. Comparing both the experiments, the dataset size (number of fixations) of Experiment-I is larger than the Experiment-II as more number of elements are fixated from both the modalities. This is attributed to the more number of stimuli webpages and

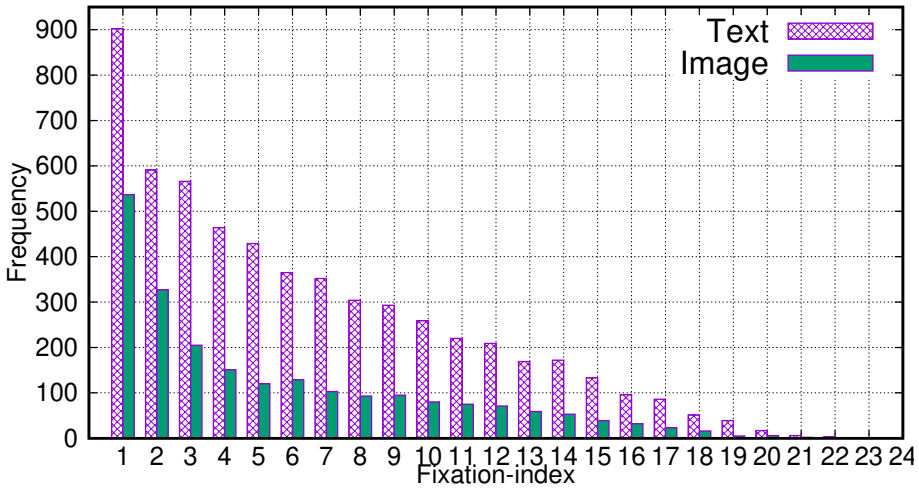


Figure 3.5: Experiment-II: Frequency distribution of fixation-indices

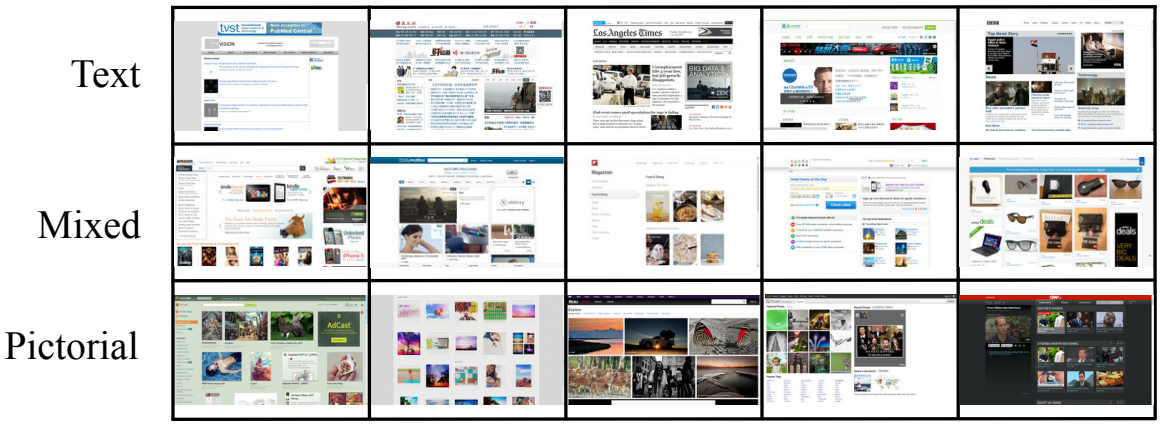


Figure 3.6: Sample stimuli webpages from each category of Experiment-III

participants in Experiment-I. See Table 3.1 for summarization of both the experiments. This dataset is utilized for preliminary analyses towards the computational approaches (in Chapter 4), and for demonstrating the efficacy of element-driven attention prediction approaches on unseen data (in Chapter 6 and Chapter 7).

3.4 Experiment-III Data

We have utilized the open-source dataset as presented in [134, 135]. The stimuli consists of three categories of webpages (Text, Mixed, and Pictorial) which were free-viewed by 11 participants where each webpage was displayed for five seconds. The sample stimuli from each category is shown in Figure 3.6. The typical threshold of 100 milliseconds [19] is used to identify the fixations and assign the fixation-indices for each user on each webpage. As, only webpage screenshots are available (unlike the HTML webpages), we manually extracted the image elements corresponding to the fixation-locations. In case only a part of the image

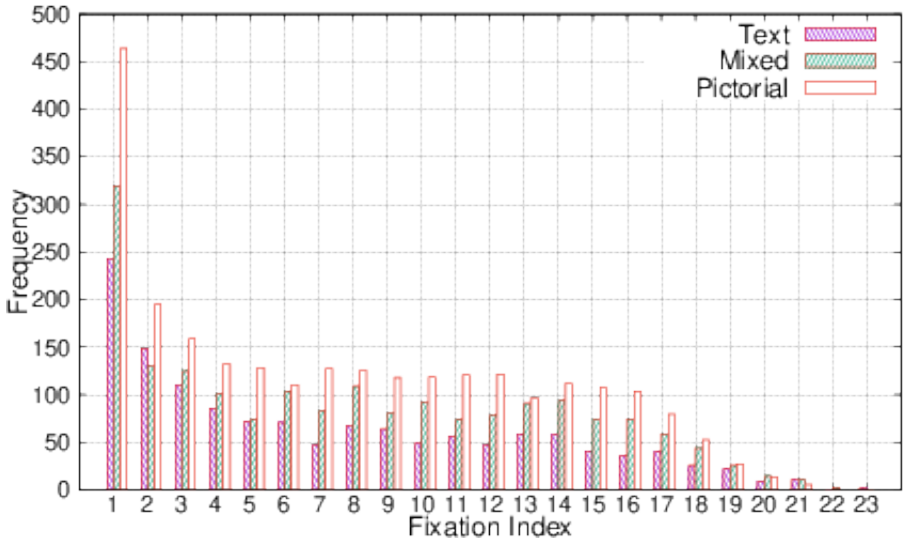


Figure 3.7: Frequency distribution of fixation-indices (attention-levels)

element is above the visual fold, the respective visual portion is considered for feature extraction and analysis. We made this curated data (consisting of extracted images, features along with attention allocations) publicly available[5] for research community's interest.

The frequency distribution of the FIs from Experiment-III are shown in Figure 3.7. In total, image elements from text category obtained 1369 fixations, mixed category obtained 1866 fixations, and pictorial category obtained 2527 fixations. Evidently, web images on Pictorial-category webpages consistently received more fixations followed by Mixed-category and Text-categories. This is attributed to the presence of more number of image elements on Pictorial-category than its counterparts. However, across the categories, the frequency of attention-levels consistently reduced with increasing attention-level indicating the reduction in salient image elements.

This dataset is utilized for image elements' based computational analysis (in Chapter 4), and for demonstrating the efficacy of element-granular attention prediction on cross-experimental data (in Chapter 6).

3.5 Considered Visual Features

Towards the application of the proposed approach, the following visual features are extracted for Data-Of-Interests (DOIs) (fixated elements) of both modalities.

3.5.1 Visual Text Features

For each text DOI, the visual representation determining Cascading Style Sheets (CSS) [29] features were extracted, analogous to [19]. The extracted features are segregated into groups

[5] at https://bit.ly/2lL1ORW

Table 3.2: Visual features of text modality

Feature Group	**# of feat.**	**Visual Text Features**
COLOR	40	Text color, Background color, Outline color, Text decoration color, Text emphasis color, Border color (left, right, top, center), Column rule color
SPACE	11	Line height, Margin, Padding, Column gap, Text indentation
SIZE	6	Font size, Font weight, Border size
FAMILY	22, 9, 15	Selected 85% of the font-families contributing to the cumulative distribution. (22 without grouping; 9 for WG & UG; 15 for WUG)
POSITION	5	left, right, top, and bottom distance from edges and area
Total		**84 without grouping; 71 for WG and UG; 77 for WUG**

Table 3.3: Visual features of the image modality

Feature Group	**# of feat.**	**Visual Image Features**
HISTOGRAM	32	Eight uniform histogram bins for each of red, green, blue, and gray color components
COMPREHENSIVE	11	Luminance, hue, saturation, brightness, value, mean, and variance of the R,G,B color component values
CONTRAST HISTOGRAM	32	Difference between the respective HISTOGRAM of the stimuli webpage and the image
CONTRAST COMPREHENSIVE	11	Difference between the respective COMPREHENSIVE features of the webpage and the image
POSITION	5	left, right, top, and bottom distance of the image from screen, and area
Total		**91**

for easy comprehension as shown in Table 3.2. Note that the WG, UG, and WUG are introduced in Chapter 5. From each COLOR attribute, four features including red, green, blue, and color opacity features were extracted. The SPACE that describes the spacing around and within the text DOI contributed eleven features. The *margin* and *padding* represent the rectangular spacing around left, right, top, and bottom sides with four features each. The *font-size* and its thickness determining *font-weight* contributed six features towards SIZE group along with a rectangular border size. The POSITION group describing the rectangular position of the text DOI and its size contributes five features.

Among the considered features, only *font-family* is consisting of categories such as 'Arial', 'Times New Roman', 'Sans', etc. We performed a *dummy variable coding*— considering each unique category as a separate dichotomous feature where the respective category occurrence is indicated with one and absence with zero [146]. However, such an approach may result in sparse dichotomous features for the rarely occurring font-family values. Consequently, we computed the Cumulative Distribution Function (CDF) of font-family frequencies and merged the sparse font-families which do not contribute to 85% of the total data. The resultant FAMILY feature counts for each criterion is shown in Table 3.2.

3.5.2 Visual Image Features

Color histograms were extracted for each image DOI. Eight uniformly spaced histograms are computed for red, green, blue, and gray color components. Besides color histograms, we considered average and variance values for each color component, brightness, luminance, hue, saturation, and intensity. For contrast features, we subtracted the feature values of images from the corresponding values associated with the respective webpage. This resulted in consideration of 91 image features as reported in Table 3.3. In summary, HISTOGRAM and COMPREHENSIVE describe the INTRINSIC visual characteristics of image DOIs while CONTRAST HISTOGRAM and CONTRAST COMPREHENSIVE features describe the contrast with respect to the underlying webpage.
The detailed description of text and image visual features is provided in Section 2.7.

3.6 Summary

In this chapter, three free-viewing eye-tracking experiments are described. Two of those were conducted by us while the remaining experiment was available in the open-source. Further, the text and image visual features considered for the rest of the analyses are summarized and segregated into cohort groups for easier comprehension. The preliminary observations of the users' attention data revealed:

- Frequency of Fixation-Indices (FIs) decreases with the corresponding increase in FI value. That is, latter FIs are relatively sparse when compared with initial FIs.
- Overall, on bi-modal webpages, more number of unique text elements are associated with a given FI than the respective unique image elements.

4

Text, Images, and Position based Attention Analysis

The attention modeling evolved from position-based analysis to computational models. Accordingly, this chapter analyzes the position-based attention allocation but, quantitatively at the higher granularity (Section 4.1). Subsequently, the prominence of the text and image visual features in explaining the free-viewing attention is analyzed (Section 4.2) and computationally modeled (Section 4.3). Owing to the prominence of image elements in drawing the user attention, ground-truth attention allocation approaches were proposed and analyzed for the same (Section 4.5 and Section 4.6).

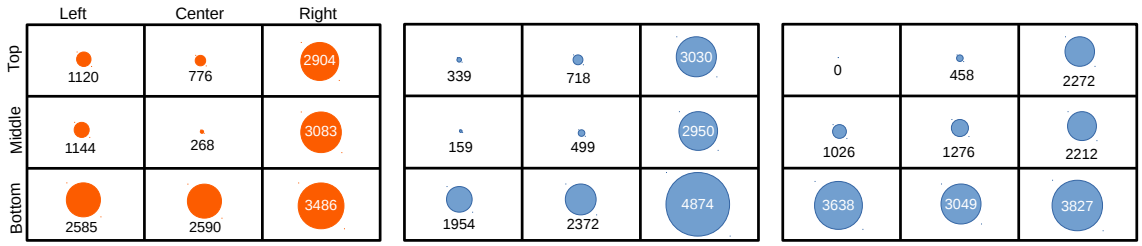
For the analysis reported in this chapter (Section 4.1, Section 4.2, Section 4.3), the data from eye-tracking Experiment-II (described in Section 3.3) was utilized. In the following section, the prominence of position in drawing the user attention for text and image modalities is analyzed.

4.1 Analysis-I— User's Positional Visual Attention Distribution

In this section, user's positional attention distribution is presented to answer the following first research question.

R1 → How users allocate *free-viewing* ordinal attention on text and image elements positioned in 3×3 webpage regions?

Towards the analysis, each webpage was divided into 3×3 uniform regions (tiles), analogous to [19]. We refer these regions as TopLeft, TopCenter, ..., BottomCenter, BottomRight where Top, Middle, Bottom indicate the row of a region while Left, Center, Right indicate the column of a region (as shown in Figure 4.1a). In each region, following attention allocation based metrics were computed.



(a) Free viewing (b) Information foraging (c) Page recognition

Figure 4.1: Median time (in milliseconds) for first fixation on 3x3 webpage regions during free-viewing, information foraging and page recognition tasks. Note that the values corresponding to information foraging and page recognition tasks were directly utilized from [19] for sanity comparison.

Median time for first fixation: The median time (in milliseconds) a participant takes to allocate their first fixation in a region. The metric considers the time taken for the first fixation among all webpages.

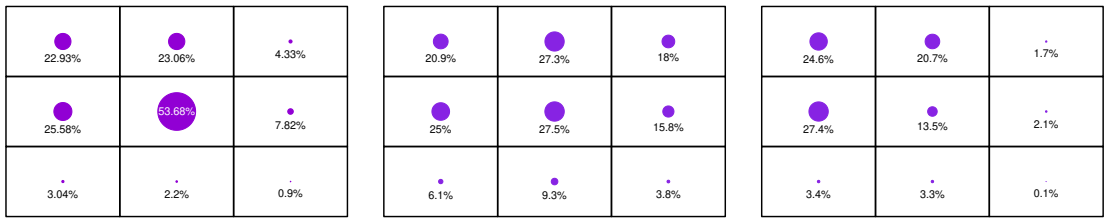
Viewing frequency: The percentage of participants fixated in a region among all the participants that fixated on the corresponding webpage at all. The metric considers the fixations of all participants on all webpages.

Fixation-Index (FI) frequency: The frequency of Fixation-Indices (FIs) obtained from all the participants.

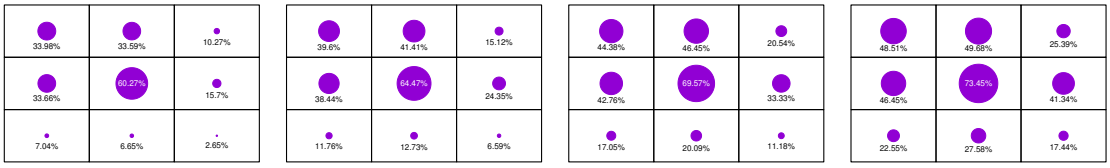
The first two metrics help to understand the user's region-based attention allocation whereas the *FI frequency* indicates the user's attention allocation at the granularity of text and image elements. The MiddleCenter region (see Figure 4.1a) is the quickest to draw user attention as opposed to MiddleLeft region in information-foraging and TopLeft region in page-recognition tasks. The MiddleCenter also obtained the highest viewing frequency up to first one second of user's gazing. It achieved 53.68% of viewing frequency as opposed to MiddleCenter region in information-foraging with 27.5% and MiddleLeft region in page-recognition tasks with 27.4%. Overall, TopLeft, TopCenter, MiddleLeft, MiddleCenter regions are more prominent in drawing the user attention and Bottom, Right regions are least prominent in drawing the user attention for all the three tasks. The detailed description of *Median time for first fixation* and *Viewing frequency* are as follows.

Figure 4.1 shows the median time for first fixation over all the webpages during free-viewing (Figure 4.1a), and its comparison with information foraging (Figure 4.1b) and page recognition (Figure 4.1c) tasks reported in [19]. The smaller the balloon size inside the tile the lower is the time taken for first fixation in that region. That is, a smaller balloon indicates the region's higher prominence in drawing the user attention [20, 113].

Clearly, MiddleCenter region (see Figure 4.1a) is the quickest to draw user attention during free-viewing with 268 milliseconds followed by TopCenter, TopLeft and MiddleLeft



(a) Free viewing— upto 1s (b) Information foraging— upto 1s (c) Page recognition— upto 1s



(d) Free viewing— 2s (e) Free viewing— 3s (f) Free viewing— 4s (g) Free viewing— 5s

Figure 4.2: Viewing frequency upto each of five seconds during free-viewing ((a), (d), (e), (f), (g)) and for the first second during information foraging (b) and page recognition (c) tasks

regions with 776, 1120 and 1144 milliseconds respectively. Although TopLeft, TopCenter, MiddleLeft, MiddleCenter regions are prominent for free-viewing, information-foraging and page-recognition tasks, their preferential dominance varied across the tasks. That is, the MiddleCenter, MiddleLeft and TopLeft regions respectively drew user's earliest attention during free-viewing, information foraging and page recognition tasks. However, Bottom and Right regions were least prominent in drawing the first fixation for free-viewing, analogous to information-foraging and page-recognition tasks. The users preferring MiddleCenter region over either of the MiddleLeft or TopLeft is attributed to the absence of a task. On the contrary, during the latter two tasks, users tried to explore the most relevant information containing regions which emanated from the user experiences.

The *median time for first fixation* helps in understanding how prominent a region is in drawing the user's first attention. However, it does not emphasize the proportional contribution of all the participants. For example, say, out of six participants five participants made their first fixation in TopRight region after $500, 550, 570, 600, 620$ milliseconds whereas only one participant fixated in BottomRight region after 400 milliseconds (ms). Accordingly, 570 ms and 400 ms are their respective median time for first fixations giving an impression that BottomRight draws quicker attention than TopRight among all participants, whereas it could only draw the attention of one participant. To overcome this limitation of the aforementioned metric *viewing frequency* was computed.

Viewing Frequency: We computed the viewing frequency— the percentage of participants fixated in a region among all the participants that made at least one fixation on the respective webpage [19], to understand overall participants' viewing behavior across all

webpages. Figure 4.2 shows the viewing frequency during the first second of free-viewing (Figure 4.2a) in comparison with information foraging (Figure 4.2b) and page recognition (Figure 4.2c) tasks reported in [19]. Additionally, the evolution of viewing frequency during the gazing session are shown in the intervals of one second, Figure 4.2a, Figure 4.2d, Figure 4.2e, Figure 4.2f, and Figure 4.2g. During the first second of gazing, MiddleCenter draws the dominant users' attention with a viewing frequency of 53.68% as compared to dominant 27.5% in MiddleCenter region for information foraging and 27.4% in MiddleLeft region for page recognition tasks. Though the four regions, TopLeft, TopCenter, MiddleLeft, MiddleCenter, are dominant than rest of the Bottom and Right regions in all the three cases, the viewing frequency is almost evenly distributed across these four regions for information foraging and page recognition than free-viewing. Among the dominant four, TopLeft, TopCenter, MiddleLeft regions' viewing frequency of free-viewing is comparable to information foraging and page-recognition. It indicates the users' bias towards MiddleCenter region for free-viewing, in agreement with the findings of [61].

To understand the participants' viewing behavior with the elapse of time, we computed the viewing frequency for the incremental duration of one second. The viewing frequency is shown in Figure 4.2a, Figure 4.2d, Figure 4.2e, Figure 4.2f, and Figure 4.2g indicates that users drift their attention to other regions of the webpage with the passage of time. The lowest metric value (17.44%) associated with the BottomRight region as shown in Figure 4.2g indicates the participants' lower interest in this region. The relatively lower viewing frequency associated with the Bottom region over the gaze-session may also attribute to the earlier generic, visual feature discarded web design guidelines such as F-pattern [105], following which designers are less likely to position the visually appealing content in the Bottom region of webpages.

Viewing Behavior on Text and Image Elements

The *Fixation index frequency (FI frequency)* is computed for text and image elements to understand their prominence in drawing the user attention based on their position in each region. Figure 4.3 and Figure 4.4 show the Top-5 most frequent FIs in each region for text and image elements respectively. Overall, text elements received a higher number of fixations than images for each region. Even the frequency of ordinal attention (fixation-index) is more for text elements than image elements. This may be attributed to the presence of more number of text elements than image elements on webpages. In congruence with the median time for first fixation and viewing frequency, the MiddleCenter, TopCenter, TopLeft, MiddleLeft regions drew a relatively higher number of initial fixations than Bottom, Right regions. Especially, for the TopLeft region, FI=1 is most frequent for images whereas FI=3 is most frequent for texts. This may be essentially due to the presence of dominant logo images in this region. In TopCenter region, FI=1 is most frequent for both text and images. However, the FI frequency of text is significantly higher than the images. This may be due to the presence of

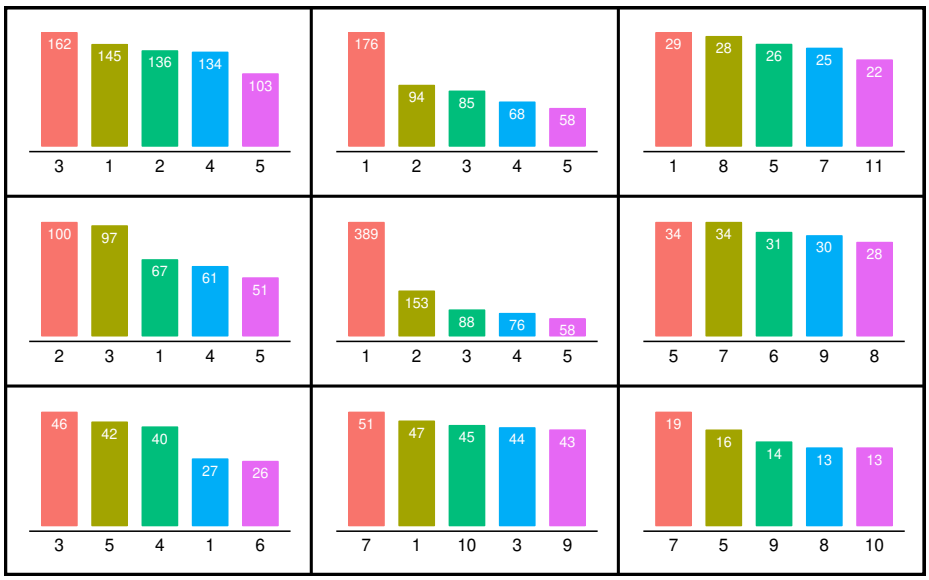


Figure 4.3: Fixation Indices distribution for Text elements in 3x3 plot. In each subplot, horizontal-axis indicates the fixation indices and vertical-axis indicates the corresponding frequency. Note that the horizontal-axis is not sorted, and limits along the vertical axis varied across subplots for better visualization.

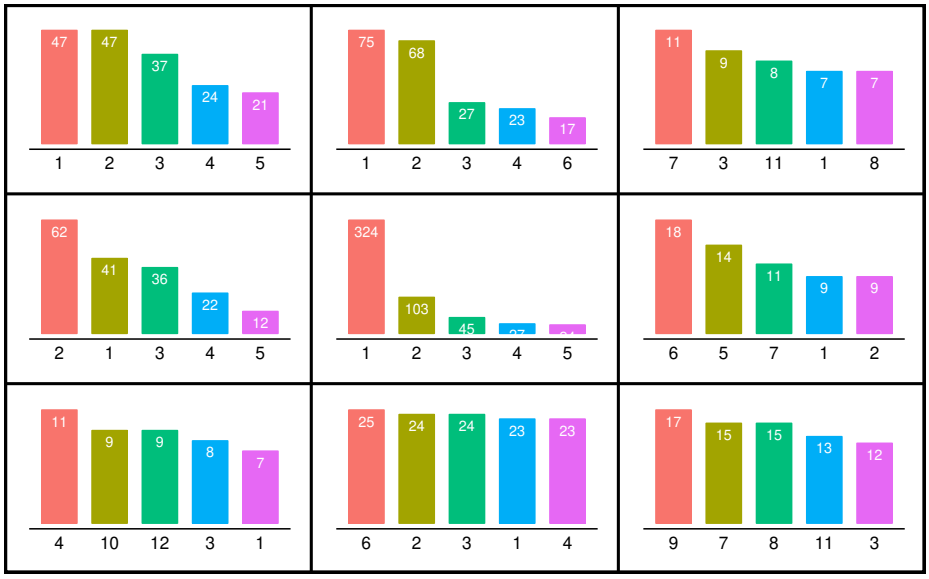


Figure 4.4: Fixation Indices distribution for Image elements in 3x3 plot. In each subplot, horizontal-axis indicates the fixation indices and vertical-axis indicates the corresponding frequency. Note that the horizontal-axis is not sorted, and limits along the vertical axis varied across subplots for better visualization.

majority webpage titles in this region.

For the text and images positioned in MiddleCenter region of webpages, the percentage of difference (difference between frequencies×100/mean of frequencies) between most frequent FI (FI=1) and the second-most frequent FI (FI=2) as shown in Figure 4.3 and Figure 4.4 is very high (103.51% for images and 87.08% for text) in comparison to TopCenter, TopLeft, MiddleLeft regions' (9.79%, 0%, 40.78% for images and 60.74%, 11.07%, 3.05% for text). The very high percentage of difference indicates that the MiddleCenter region is predominant in drawing the first fixation than any other fixation-index with respect to the other regions. Subsequently, for text, TopCenter region achieves dominant FI=1, TopLeft region achieves either of the third or first, MiddleLeft region achieves either of second or third fixation. On the contrary, TopCenter, TopLeft, MiddleLeft regions obtain either of first or second fixation for images.

Though the Bottom, Right regions are less prominent in drawing the user attention, the elements in these regions are not always discarded. The FI=1 is among the Top-5 frequent FIs in the TopRight, BottomLeft, BottomCenter regions for both the text and images (see Figure 4.3 and Figure 4.4). For images, FI=1 and 2 are among the Top-5 in MiddleRight region. It indicates that visually salient elements succeed in drawing the user attention even if they are positioned in Bottom or Right regions on a webpage.

Key Findings of Analysis-I

- MiddleCenter region is the quickest to draw user's first fixation during free-viewing in contrast to MiddleLeft and TopLeft regions respectively for the information foraging and page recognition tasks.
- In free-viewing, MiddleCenter region dominates in drawing higher proportion of users' attention followed by MiddleLeft, TopCenter, TopLeft regions.
- Proportion of users fixating in MiddleLeft, TopCenter, TopLeft regions are comparable for the three cases, free-viewing, information foraging and page-recognition.
- Overall, text elements receive more number of fixations as well as more frequent number of fixation-indices when compared with the image elements.
- For both text and images, MiddleCenter, MiddleLeft, TopCenter, TopLeft regions are prominent to draw the user's initial fixations.
- However, the text and image elements in Bottom and Right regions are not always discarded and salient elements in this region may draw the user's first fixation.

The Analysis-I highlights the prominence of position in drawing user attention. However, it is not clear whether the positional bias is because of the idiosyncratic user behavior or web designers preference to place the visually appealing content in the upper and left part of the

webpages, following the earlier design guidelines. Additionally, to investigate the prominence of elements' intrinsic visual features (besides the position) in drawing the user attention, we performed a quantitative relational analysis between visual features and the ordinal visual attention (fixation indices) as described in Section 4.2.

4.2 Analysis-II— Identification of Informative Visual Features

In this section, we perform the information-gain based analysis to identify the informative visual features in explaining the achieved attention. The visual features with higher information gain are considered as highly influential as they better explain the attention. That is, the analysis investigates the answer to the following research question.

R2 → Which *intrinsic visual features* are informative in explaining the free-viewing ordinal attention on web elements?

Information Gain: indicates the information associated with a feature apropos to the achieved attention, that is, fixation indices. That is, it indicates the relevance (attention explaining ability) of the feature with respect to the ordinal visual attention. The relative information gain scores are of interest than their absolute values in identifying the influential features [52, 103]. We computed the entropy-based information gain for text and image features using

$$\text{Information Gain}(Feature) = H(Class) + H(Feature) - H(Class, Feature) \tag{4.1}$$

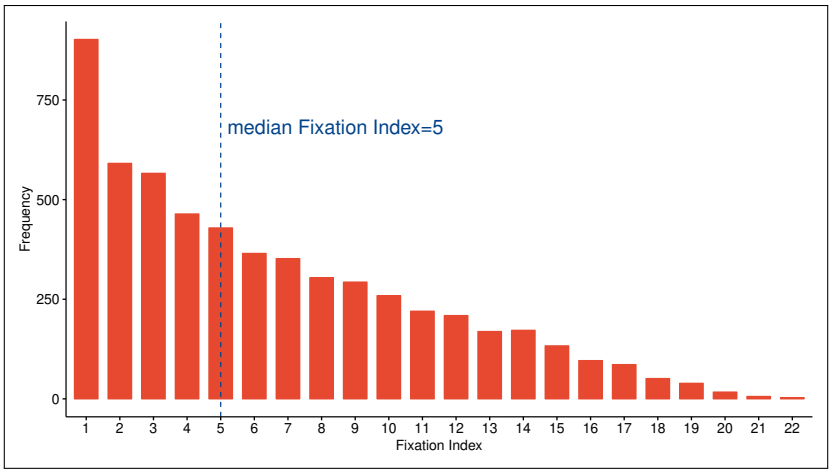
where $H(.)$ indicates the entropy and $H(.,.)$ indicates the joint entropy.

The frequency distribution of Fixation-Indices (FIs) from Experiment-II are visualized in Figure 4.5. The latter FIs are sparser when compared with the initial FIs. To segregate the initial fixations from the latter, we define the following thresholding (called *saliency thresholding*, θ) on the ordinal visual attention (fixation-indices). This is essential as the latter fixations might occur on visually less salient elements as all the salient elements are already explored.

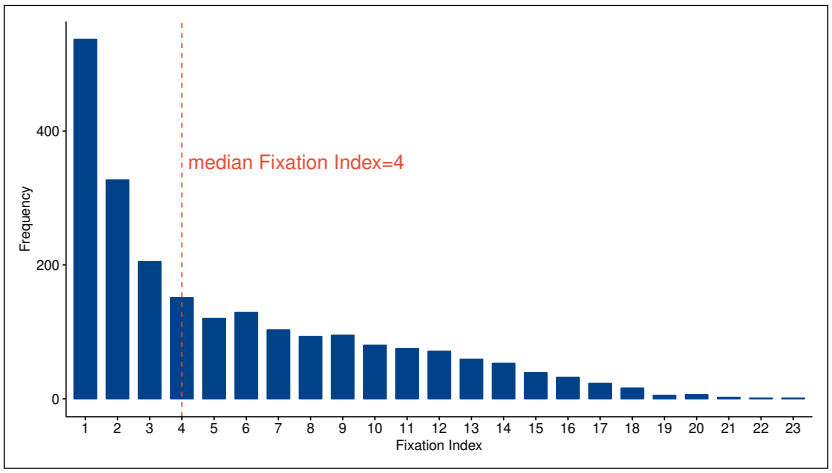
$$\theta \in \{1, 2, \cdots, f_{\max}\} \tag{4.2}$$

$$f_e = \min(f_e, \theta) \tag{4.3}$$

The fixation-indices less than the thresholding variable θ persist with the same fixation-index, highlighting its prominence in drawing the user attention. The fixation-indices from and above the θ are denoted with θ to indicate the less salient visual attention associated with the element.



(a) Text



(b) Images

Figure 4.5: Frequency distribution of fixation indices over text and image elements

θ selection: To select the θ value, we computed the median FI on both text and image elements. The frequency distribution of fixation indices along with the median FI is shown in Figure 4.5. The median FI on the text is five and the median FI on the image is four. That is, on an average, a user makes five fixations on text elements and four fixations on image elements during their gazing session on a webpage. For the sake of uniformity and analytical comparison, we consider the thresholding as $\theta = 5$. To further investigate the influence of other FIs on attention, we performed the analysis in the multiple of θ. That is, thresholding values are $\theta = 5, 10, 15, f_{\max}$. $\theta = f_{\max}$ indicates the performance analysis over all the FIs, that is, without thresholding.

4.2.1 Informative Text Visual Features

The informative gain (IG) scores for Top ten text features are shown in Table 4.1. The text element's `size` is the most informative visual feature with respect to ordinal attention (see

Table 4.1: Information gain of text visual features. Color intensity represents relative importance of a feature in comparison to others.

Feature	**Info. gain** $\theta = 5$	**Info. gain** $\theta = 10$	**Info. gain** $\theta = 15$	**Info. gain Overall**
size	0.0426	0.0443	0.0356	0.0362
rect.left	0.0342	0.0285	0.0295	0.0303
rect.top	0.0268	0.0246	0.0248	0.0256
rect.right	0.0221	0.0148	0.0154	0.0166
padding-bottom	0.0053	0.0056	0.0064	
rect.bottom	0.0172	0.0076		
font-size	0.0069	0.0088		
column-gap	0.0069	0.0088		
line-height	0.0070			
font-weight	0.0038			

Table 4.1). It confirms the intuition that larger sized text elements draw better attention than their smaller counterparts. Additionally, highlighting the prominence of position, the four position-related visual features were found to be among the Top-10 informative visual features. These are in agreement with [19] findings where `size` and `position` were found to be informative for information foraging and page recognition task settings.
The notable intrinsic informative visual features are `padding-bottom`, `font-size`, `column-gap`, `line-height` and `font-weight`. These five text-related intrinsic visual features were discarded in [19] due to tag-only consideration. Regarding the feature groups (shown in Table 3.2), SPACE (with features `padding-bottom, column-gap, line-height`) and SIZE (with features `font-size, font-weight`) groups are informative where especially the former group was little explored for user attentional analyses on the text. Surprisingly, none of the COLOR related intrinsic visual features were found to be informative which are widely explored in the literature [17, 108, 153]. This may be attributed to the wide usage of black color for text and white color for the background with little variation in both of them. To summarize, besides element's `size` and `position`, text's font and its thickness, and the height of text line, the spacing between the columns of text, the spacing between text content and the border are informative in explaining the user's free-viewing ordinal attention on text elements.

4.2.2 Informative Image Visual Features

The informative gain (IG) scores for Top-15 image features are shown in Table 4.2. The image element's position from the top of the webpage is the most informative visual feature followed by its size. Among the color histograms, the MID-LEVEL HISTOGRAM (third, fourth, fifth, and sixth histogram bins) are more informative than lower (first two bins corresponding to lower color component values) and higher histograms (last two bins corresponding to higher color component values), indicating pure colors and absence of colors are least informative in explaining the user attention. Surprisingly, neither of COMPREHENSIVE and CONTRAST

Table 4.2: Information gain of image visual features. Color intensity represents relative importance of a feature in comparison to others.

Feature	**Info. gain** $\theta = 5$	**Info. gain** $\theta = 10$	**Info. gain** $\theta = 15$	**Info. gain Overall**
rect.top	0.1364	0.1258	0.1273	0.1056
size	0.1130	0.1044	0.1064	0.1091
hist_gray_3	0.1111	0.1174	0.1004	0.0772
hist_R_5	0.1130	0.0975	0.0988	0.0996
hist_gray_5	0.1117	0.0909	0.0922	0.0931
hist_gray_6	0.1059	0.1077	0.0892	0.0895
hist_G_5	0.0977	0.0882	0.0890	0.0898
hist_G_3	0.0875	0.0893	0.0899	0.0912
hist_B_5	0.1083	0.0878	0.0881	0.0893
hist_G_6	0.1007	0.0870	0.0885	0.0888
hist_B_3	0.0993	0.0862	0.0867	0.0878
hist_B_4	0.1019	0.0813	0.0823	0.0835
hist_R_4	0.0971	0.1005	0.0712	0.0717
hist_gray_4	0.0879	0.0909	0.0920	0.0669
rect.left	0.0790	0.0813	0.0852	0.0884

visual features found among the Top-15 most informative intrinsic visual features.
In [19], intrinsic image features were discarded by utilizing the binary-valued "IMG" feature to indicate the image presence. However, the "IMG" HTML tag was not found to be informative among Top-10 features for information foraging and page recognition tasks. On the contrary, the IG scores of image intrinsic visual features were higher as well as densely packed than text features across all the FI thresholding as shown in Table 4.1. We attribute this observation to the inherent representation of the images, as multiple pixels constitute an image with each pixel associating three simultaneous values corresponding to the red, green and blue color components. Typically, in case of a modification in pixel value, all the associated color component values altered resulting in tighter IG scores for images.

Key Findings of Analysis-II

- The text intrinsic visual features `padding-bottom, column-gap, line-height` (SPACE group) and `font-size, font-weight` (SIZE group) are informative besides `size` and `position` in explaining the free-viewing ordinal attention on text elements.
- For images, MID-LEVEL HISTOGRAM are highly informative besides the `size` and `position`.
- Surprisingly, the widely explored COLOR features of text and COMPREHENSIVE, CONTRAST features of images were not found to be among the highly informative features.
- Relatively, image intrinsic visual features are more informative (with higher IG scores) than the text's intrinsic visual features.

- The modality-independent features, `size` and `position`, are informative for both text and images, analogous to the information foraging and page recognition task settings [19].

4.3 Analysis-III — Modeling the Attention on Web Elements

The Analysis-II helped in identifying the informative visual features with respect to the attention. Further extending the investigation, we explored how well the informative visual features perform in *predicting the ordinal visual attention* on a given web element. That is, we investigate the following research question which is solved through the multi-class classification procedure described in subsection 2.6.1.

R3 → How the informative intrinsic visual features perform in *predicting the free-viewing ordinal visual attention?*

It is essential to understand the influence of informative visual features (obtained from Section 4.2) in segregating and predicting the attention on individual web elements. Towards the segregation, we performed decision tree based analysis to predict whether an element obtains a fixation-index less than or equal to the given threshold θ or not, in subsection 4.3.1. Towards the attention prediction, we built a multi-class classification model using linear SVM to predict the actual fixation-index of a given element in subsection 4.3.2.

4.3.1 Salient Element Identification through Decision Tree

Table 4.3: Element saliency prediction metrics using Decision Trees

Metric	**Text**		**Image**		**Task-dependent**
	$\theta = 5$	$\theta = 10$	$\theta = 5$	$\theta = 10$	[19]
Precision	77.29%	55.93%	**95.83%**	63.64%	75%
Recall	**93.72%**	34.38%	82.14%	50.00%	53%
Kappa (κ)	0.2505	0.2278	**0.6739**	0.3642	0.59

Procedure: All the fixated text elements along with their associated fixation indices (FIs) are considered for the analysis. An element whose FI is less than θ is considered as a salient element, that is, the positive class is assigned. Accordingly, the negative class is assigned for the elements with FI equal to θ. For the elements with multiple class assignments (based on FIs from multiple participants), *majority-voting-scheme* is pursued to assign the true class label with the random resolving of conflicts (that is, an equal number of positive and negative class labels). Subsequently, elements are divided into 80:20 train:test sets through random sampling. Using the informative features obtained from Section 4.2, a decision tree is constructed using the train set whose performance is evaluated using the test set of elements. Additionally, 10-fold cross-validation with 10 iterations is performed and averaged

performance is considered. The same procedure was followed for image elements and the analysis was conducted using 'rpart' (Recursive partitioning for classification) and 'caret' (Classification and Regression Training) packages from R language [54, 149].

Results: The decision tree based metrics for $\theta = 5$ and 10 are shown in Table 4.3. The metrics indicate a very good performance of the informative features in identifying the salient elements at the fixation-index threshold ($\theta = 5$) closer to the median FI. Both the precision and recall performances are higher than the respective task-dependent performance reported in [19] where a decision tree is constructed to investigate whether an element obtains a fixation or not. The higher performance in the free-viewing case is attributed to prominence of visual features in guiding the initial fixations whereas user idiosyncratic expectational bias is an addendum in directing the attention in task-dependent settings. The relative higher precision of the images (95.83%) than the text (77.29%) indicates that image visual features are more influential in deciding whether an element is salient or not than their counterpart text features. In contrast, the recall of text (93.72%) is higher than the images (82.14%) indicating the visual features of the salient text elements better segregates them from non-salient element features than their counterpart image features. The overall performance in identifying the salient elements is *substantial* [102] for images ($\kappa = 0.6739$) than text ($\kappa = 0.2505$). However, the model performance reduces with an increase in θ (see performance for $\theta = 10$ in Table 4.3) as informative visual features reduced and the later fixations are less likely driven by the visual features.

The decision tree based analysis helped in segregating the salient elements from non-salient elements. To further quantify the saliency of the identified elements and to predict the ordinal visual attention, we performed multiclass classification using Linear Support Vector Machine (SVM) as described in subsection 2.6.1.

4.3.2 Ordinal Visual Attention Prediction on Web Elements

Ordinal visual attention or Fixation-Index (FI) prediction is a multiclass classification problem. Each possible value of FI represents a class. Unlike the binary classification (salient or non-salient determination), multiclass (number of classes > 2) classification (prediction of FI) is a difficult problem to solve. One can not obtain a single boundary to separate all the classes from each other. For θ classes, $\binom{\theta}{2} = \theta(\theta-1)/2$ separating boundaries need to be determined to separate any class from every other class. That is, $\theta(\theta-1)/2$ binary classifiers are required. Accordingly, multi-class classification requires to learn `number of parameters` $\times\binom{\theta}{2}$ parameters. For this ordinal visual attention prediction, we utilized the approach as described in subsection 2.6.1. The analysis was carried out using 'rminer' (Data Mining Classification and Regression Methods) [31] and 'kernlab' (Kernel-Based Machine Learning Lab) packages [83] in R.

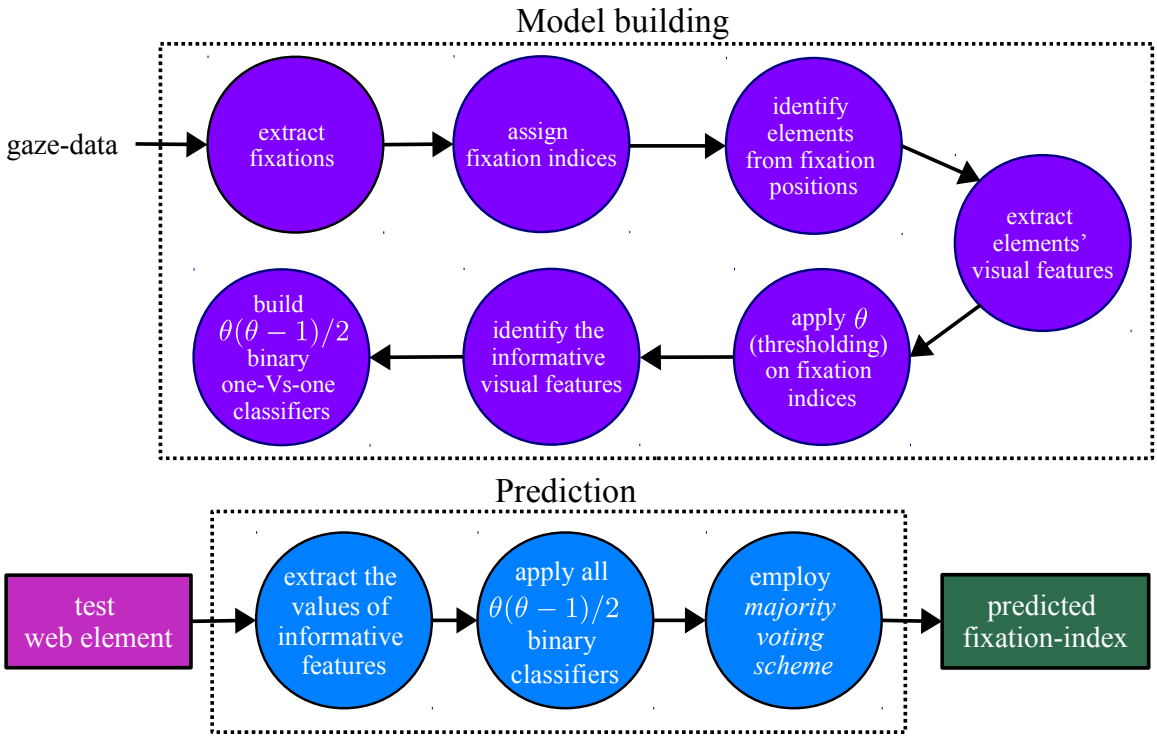


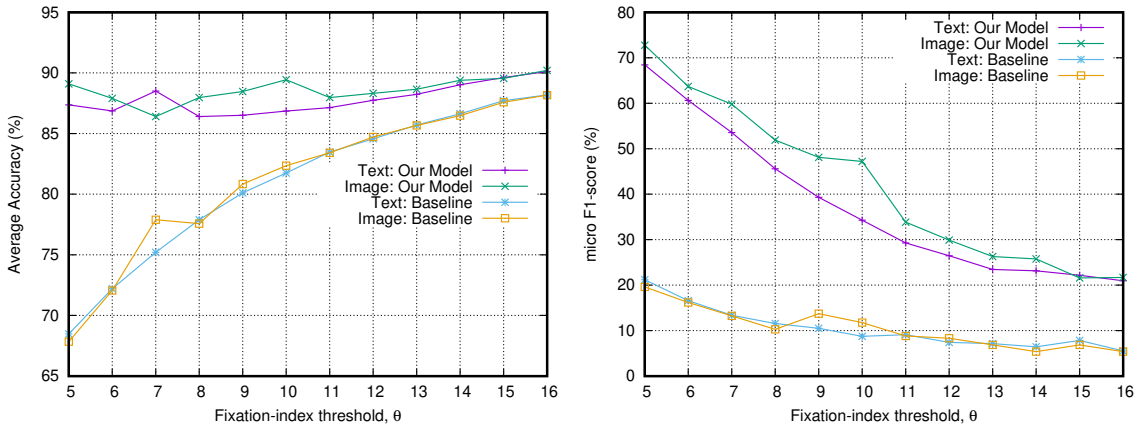
Figure 4.6: Ordinal visual attention prediction procedure

Table 4.4: Ordinal visual attention prediction performance

Type	Model	Average Accuracy(%)	micro F1-score(%)
Text	Our model	**87.37**	**68.44**
	Baseline	68.48	21.20
Image	Our model	**89.10**	**72.75**
	Baseline	67.84	19.61

Data Preparation: The frequency distribution of the FIs for text and images are shown in Figure 4.5. The FIs are reassigned using Equation 4.3 to persist with the original FI in case it is less than the threshold θ or else replace with the threshold FI. For the elements with multiple FIs, *majority-voting-scheme* is followed to assign the ground-truth FI with the random resolving of conflicts.

Results: The ordinal visual attention prediction results at median FI thresholding are shown in Table 4.4. Clearly, the informative intrinsic visual feature-based multiclass model outperformed the baseline model. The text-based model outperformed the baseline micro F1-score by 222.83% with a value of 68.44% and Average Accuracy by 27.58% with a value of 87.37%. Similarly, the image-based model outperformed the baseline micro F1-score by 270.98% with a value of 72.75% and Average Accuracy by 31.34% with a value of 89.10%.



(a) Average Accuracy

(b) micro F1-score

Figure 4.7: Prediction performance of informative intrinsic visual features with variation in θ

The relatively superior performance of the images over text is attributed to the higher information gain scores of image intrinsic visual features. The seemingly higher baseline average accuracies (68.48% for text and 67.84% for images) are attributed to the class imbalance (see Figure 4.5), where accuracy tend to bias towards the prediction of most frequent class.

Variation in θ: To further understand the influence of θ on prediction performance, the prediction procedure is repeated for each increment in θ. Average Accuracy and micro F1-scores are shown in Figure 4.7. The Average prediction Accuracy of text and images remained higher than 86% (and closer to 90%) throughout the variation in θ and remained outperformed the baseline models. However, the Average Accuracy of baseline models (for both text and images) progressively increased with each increment in FI threshold θ. This corresponds to the increase in class-imbalance with an increase in FI threshold. In contrast, the class-imbalance (or bias) overcoming micro F1-score remained lower for baseline models throughout the variation in θ. The prediction performance reduced with an increase in FI threshold. This is attributed to the reduction in remaining salient elements of a webpage on which user attention allocation varied diversely, resulting in relatively inadequate performance.

Key Findings of Analysis-III

- Informative visual features help in segregating the salient elements (fixation-index up to the median FI) from non-salient.
- Image features are better than text features in deciding whether an element is salient or not.

- Text features are better than image features in correctly identifying the salient elements as salient.
- The informative visual features help in predicting the salient fixation-index of an element with nearly 90% average accuracy.
- The best fixation-index prediction is achieved at the thresholding near the median FI of the elements with 68.44% and 72.75% micro F1-score respectively for text and images.
- The average accuracy remains around 90% and micro F1-score reduces with increased FI threshold θ.

Motivation for Remainder of the Chapter: On prevalent bi-modal webpages consisting of text and images[1], the latter stand-out from rest-of-the text [21, 165] and draw relatively more attention [112]. Images are highly probable (60–70%) in drawing human's first attention [121] and overall attention of more than 50% as compared to text [123]. Consequently, Jana et al. [78] proposed a regression-based model to computationally predict the attention on image elements. However, the constrained model suffers from the limitations as discussed in subsection 2.4.2. In remainder of this chapter, we present the computational approaches towards overcoming those limitations. In Section 4.4 possible non-linear association between image visual features and the user attention is investigated. In Section 4.5, multiple ground-truth attention assignment techniques are investigated. In Section 4.6 possibility of assigning multiple ground-truth attention to the same image element are explored.

4.4 Kernel-based Attention Prediction

We investigated for the non-linear association between image visual features and the user attention. That is, instead of assuming linear separation (boundary) between the classes, multiple non-linear separations were considered during the classification. We utilized the following popular kernel functions [141] in Equation 2.6 for our analysis.

Gaussian Radial Basis Function (Gaussian RBF) kernel:

$$K(I_k, I_t) = \exp\left(-\sigma \|I_k - I_t\|^2\right) \tag{4.4}$$

It is a general-purpose kernel, especially used in the absence of prior-knowledge about the data. $\phi()$ of Gaussian RBF kernel projects data from $\mathbb{R}^n$ to $\mathbb{R}^\infty$. The hyper-parameter $\sigma > 0$ is the inverse kernel width, that is, larger σ narrows the width of bell-shaped curve of Gaussian RBF and vice-versa.

[1] any non-textual information such as pictures, icons, graphics and photographs are collectively referred as images

Polynomial kernel:

$$K(I_k, I_t) = (scale\langle I_k, I_t\rangle + offset)^{degree} \tag{4.5}$$

This kernel is especially popular in natural language processing domain. The hyper-parameters *scale* helps to normalize the pattern without modifying the data, $offset$ provides the bias, and $degree > 0$ is the degree of polynomial.

Linear kernel:

$$K(I_k, I_t) = \langle I_k, I_t\rangle \tag{4.6}$$

It is useful for linearly separable data and is fastest to compute. Especially popular in text classification.

Hyperbolic tangent kernel:

$$K(I_k, I_t) = \tanh\left(scale\langle I_k, I_t\rangle + offset\right) \tag{4.7}$$

Hyperbolic tangent kernel or sigmoid kernel or multilayer perceptron kernel is well-known in neural networks field. Usage of sigmoid kernel with SVM is equivalent to an application of two layer perceptron. The hyper-parameters are similar to polynomial kernel's.

Laplacian kernel:

$$K(I_k, I_t) = \exp\left(-\sigma\|I_k - I_t\|\right) \tag{4.8}$$

Similar to Gaussian RBF kernel but less sensitive to variations in σ.

Bessel kernel:

$$K(I_k, I_t) = -\text{Bessel}_{order}^{degree}\left(\sigma\|I_k - I_t\|^2\right) \tag{4.9}$$

The hyper-parameters *order* and *degree* are the parameters of Bessel function. Typically, *degree* is set to 1 and the kernel is popular in theory of function spaces of fractional smoothness [141].

ANOVA RBF kernel:

$$K(I_k, I_t) = \sum_{j=1}^{degree}\prod_{l=1}^{n} K^{'}(I_{k_l}^j, I_{t_l}^j) \tag{4.10}$$

where $K^{'}(I_{k_l}^j, I_{t_l}^j)$ is a Gaussian RBF kernel with I_{k_l} and I_{t_l} indicating the lth feature of I_k and I_t respectively. Accordingly, the kernel is similar to Gaussian RBF and

Laplacian kernels.

In the aforementioned kernels, I_k and I_t are the webpage image data-points where former is from training set and latter from test set. The $\langle I_k, I_t \rangle = I_k^\top I_t$ is the inner product and $\|I_k - I_t\|$ is the Euclidean distance between I_k and I_t. For the analysis, the *hyper-parameters* are set to the default values, that is, $scale = 1$, $offset = 1$, $degree = 1$, and $order = 1$.
To analyze the performance of the proposed approach we utilized the gaze data from Experiment-II.

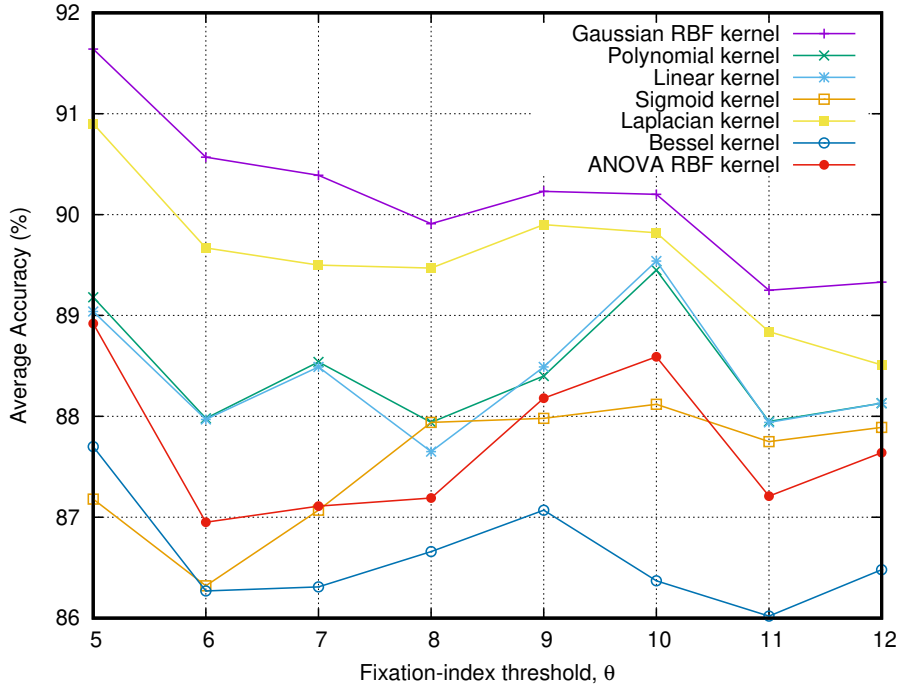
Prediction Performance

The webpage image and associated fixation-index data was split into 80:20 train test proportions for the analysis. In total $\theta(\theta - 1)/2 = 10$ binary classifiers were constructed with the train-data and majority-voting-scheme was applied to predict the fixation-index for test-data. The 5-fold cross-validation was employed to measure the performance. Prediction procedure was repeated for 10 times after randomly permuting the data. The random permutation overcomes the possible special structures occurring in splitting the data. The performance over all the iterations were averaged to obtain the overall model performance.

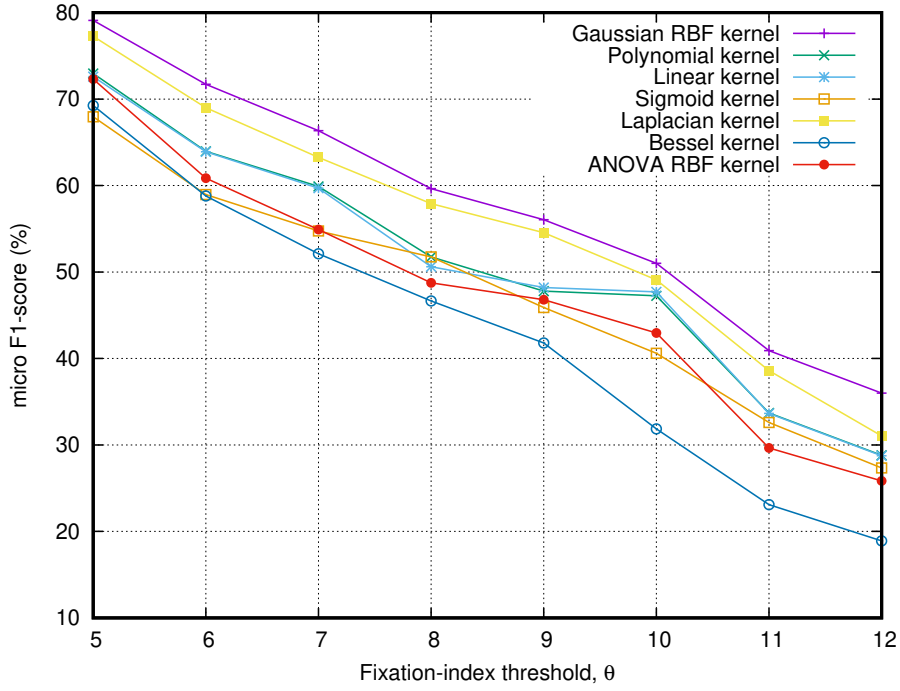
Table 4.5: Prediction performance at median fixation-index

Method	Average Accuracy	micro-F1
Gaussian RBF kernel	**91.64**	**79.10**
Laplacian kernel	90.90	77.25
Polynomial kernel	89.18	72.95
Linear kernel	89.04	72.60
ANOVA RBF kernel	88.92	72.30
Bessel kernel	87.70	69.25
Sigmoid kernel	87.18	67.95
Jana and Bhattacharya model [78]	73.92	34.80
Random prediction (RP) model	67.63	19.07

The standard multiclass classification performance metrics, *Average Accuracy* and *micro F1-score* are computed. Especially, for the class imbalanced data (such as shown in Figure 4.5b), micro F1-score is best indicator of prediction performance as accuracy biases towards the prediction performance on most frequent class (FI=1 in Figure 4.5b).
The averaged performance metrics are shown in Table 4.5. Among the kernels, Gaussian RBF achieved the best performance (average accuracy=91.64% and micro-F1=79.10%) followed by Laplacian (average accuracy=90.90% and micro-F1=77.25%) and polynomial kernels. Sigmoid kernel achieved the relatively lowest performance with an average accuracy of 87.18% and micro F1-score of 67.95%. As shown in Table 4.5, kernels' performance is consistent across the metrics. That is, the kernel with the highest average accuracy



(a) Average accuracy



(b) micro F1-score = micro Precision = micro Recall

Figure 4.8: Quantitative visual attention prediction performance of each kernel with variation in θ.

has shown the highest micro F1-score as well, and so on. However, micro F1-score better highlighted the varied prediction performance of the kernels (ranging from 67.95% to 79.10%) than the average accuracy (ranging from 87.18% to 91.64%). The linear kernel which linearly models the association between image visual features and fixation-indices shown median performance (average accuracy=89.04% and micro-F1=72.60%) among the kernels. This performance indicates the significant linear association between considered features and the quantitative visual attention.

Comparison with State-of-the-art: To compare our model performance with state-of-the-art model, the constrained regression model [78] was applied on our dataset. As their model restricts the image elements' properties (a total of 6 images distributed into 3 vertical columns with each column containing 2 images completely inside them), webpage satisfying the constraints was considered for learning the model parameters. Among the five features, 'Intensity-contrast', 'Chromatic-contrast', 'Size', 'X-position', and 'Y-position', former two features were discarded due to low correlation magnitude (<0.09 [78]) with the fixation-index (-0.03 and 0.05 respectively). The significant correlations -0.30, -0.19, and 0.25 of 'Size', 'X-position', and 'Y-position' features resulted in regression model coefficients as 0.41, 0.26, and 0.33 respectively. As shown in Table 4.5, their model achieved an average accuracy of 73.92% and micro F1-score of 34.80%. Clearly all our kernel based models outperformed the state-of-the-art performance. The rationale for poor performance of the regression-based model is, our dataset consists of real-world webpages with no constraints on images' properties whereas their model expects the presence of only 6 images on webpage but no other element.

Performance Vs. θ: To further understand the influence of fixation-index threshold θ on performance, metrics were computed with variation in θ from 5 to 12 as shown in Figure 4.8. The Gaussian RBF and Laplacian kernels showed persistently better performance than other kernels. However, the prediction performance of the kernels reduced with increasing θ. Especially, micro F1-score reduction was substantial whereas the average accuracy reduced moderately. The performance reduction (beyond median fixation-index) is attributed to the reduced number of salient webpage images, as on an average only four webpage images drew user attention on each webpage.

In summary, the performance metrics demonstrated the efficacy of the proposed approach in predicting the user's attention on webpage images. The sample user attention predictions are shown in Figure 4.9.

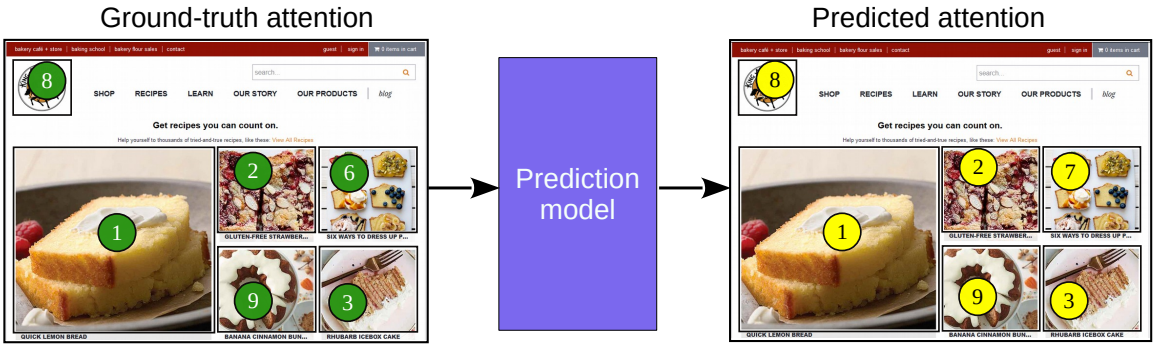
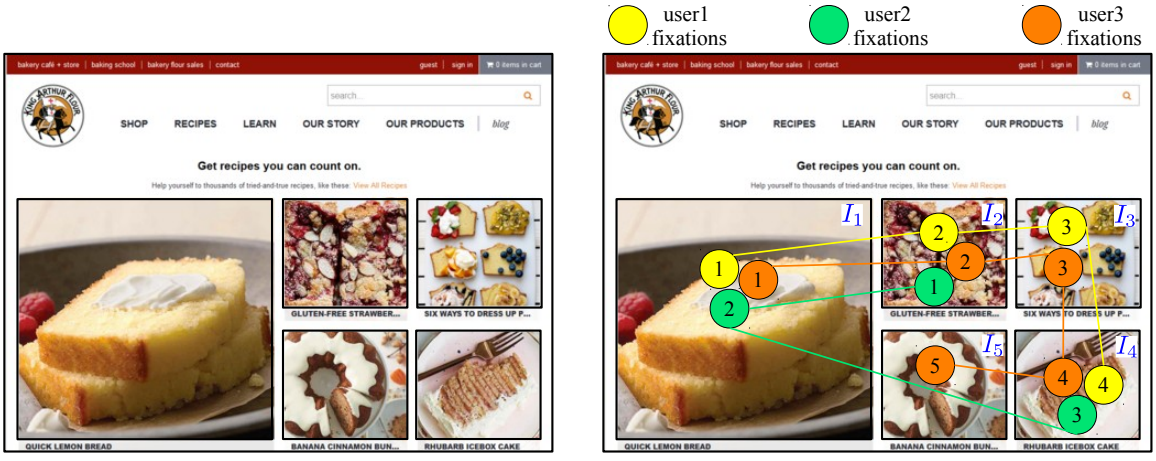


Figure 4.9: Predicted quantitative visual attention (on *right*) for the images of a webpage (on *left*, along with Ground-truth attention).

4.5 Weighted Voting based Attention Prediction on Images

Motivation

Towards the representative attention assignment (prior to prediction), subsection 4.3.2 considered all user attention allocations as equivalent. That is, users' initial attention and the latter attention are uniformly weighted. In contrast, typically, the user's initial attention is of more interest and important than the latter attention as the latter attention may be influenced by other factors such as the reduction in attention-drawing elements or semantic features. Thus, it is essential to incorporate the prominence of each attention allocation through adequate *weighting*.



(a) sample webpage (b) user attention

Figure 4.10: Example users' attention on a sample webpage. The bubbles in (b) are the *fixations* with the inscribed numbers indicating *fixation-indices*. The connected sequence of fixations constitute a *scanpath* for each user.

For example, in Figure 4.10, image I_1 able to draw the first attention (FI=1) from user1 and user3, and the second attention (FI=2) from user2. Similarly, image I_2 able to draw the second

attention (FI=2) from user1 and user3 and the first attention (FI=1) from user2. Overall, images I_1, I_2, I_3, I_4 and I_5 able to draw the unique fixation-indices of $(1,2), (1,2), (3), (3,4), (5)$ with the respective frequencies $(2,1), (1,2), (2), (1,2), (1)$. Thus, the frequencies indicate the supporting *votes* for the corresponding fixation-indices. Accordingly, the process of obtaining an effective visual attention (an *effective fixation-index*) for the attended image element should utilize both the components, fixation-index and the associated frequency. Accordingly, following research problem is formulated.

R4 → How the image visual features perform in predicting the weighted-voting-based effective *free-viewing* visual attention?

In this section, we present a computational approach to assign and subsequently predict an *effective free-viewing visual attention* on image elements.

4.5.1 Effective Visual Attention Assignment

Consider, an image I is associated with a bag of fixation-indices (FIs) from the possible set of fixation-indices $\{1, 2, \dots\}$. Though the initial fixations (lower fixation-indices) indicate the element's attention drawing ability, the latter fixations (higher fixation-indices) may be guided by other factors such as scarcity of salient elements or influence of semantic features which are non-prominent. To segregate the prominent FIs from the non-prominent FIs, we introduce a *saliency threshold* (θ) as,

$$f = \min(f, \theta) \tag{4.11}$$

where, f is the fixation-index of I. The θ preserves the individual representation of prominent FIs ($f < \theta$) and assigns a composite representation for all the non-salient FIs ($f \geq \theta$) limiting the number of possible FIs to θ.

Now, the objective is to assign a representative fixation-index, called *effective fixation-index* for the image I from $\{1, 2, \dots, \theta\}$. Towards the assignment, we utilize the implicit prominence associated with the FIs (the lower the FI the higher is the prominence) and the explicit prominence as designated by the users. The former prominence is indicated with a *weight* for each FI while the latter prominence is determined from the frequency (*votes*) of respective FIs. Owing to the visual feature-based free-viewing analysis, we introduce following intuitive weighting strategies to assign the implicit prominency for each FI.

Uniform Weighting

Assigns an equal weight to all the fixation-indices. This is an adequate weighting if all the fixation-indices are equally prominent. To note, the work in [155] corresponds to the uniform weighting.

$$w_f = \frac{1}{\theta} \quad \forall\, 1 \leq f \leq \theta \tag{4.12}$$

Algorithm 1: Effective Fixation-Index Assignment

```
for k ∈ {1, ..., d} do                                   // iterate over all the images
    for f ∈ {1, ..., θ} do                               // iterate over all the associated fixation-indices of I
        v_f ← Frequency(I_k, f)                          // computes the frequency of FI f as the vote count
        wv_f ← v_f × w_f                                 // weights w_f s are obtained from either of Eq. Equation 4.12 or
                                                            Equation 4.13 or Equation 4.14 or Equation 4.15
    end
    e_k ← argmax_f (wv_f)                                // assign FI with maximum weighted votes as effective fixation-index
end
```

Linear Weighting

Assigns the weight as a decreasing linear function of fixation-index value with the highest weight assigned to the first FI and the lowest weight assigned to the last FI, i.e., θ. This is an adequate weighting if the FI's prominence is consistently reducing with the increase in its value.

$$w_f = \frac{\theta - f + 1}{\theta + 1} \quad \forall\, 1 \leq f \leq \theta \tag{4.13}$$

Proportional Weighting

The previous two strategies assign explicit weight and do not account for the inherent strength (probability of occurrence) of fixation-index in the overall attention data. To incorporate the inherent strength of a fixation-index, the frequency-based weighting strategy is introduced with the highest weight assigned to the most frequent fixation-index (due to the higher likelihood of its occurrence).

$$w_f = \frac{\text{frequency}(f)}{\sum_{i=1}^{\theta} \text{frequency}(f)} \quad \forall\, 1 \leq f \leq \theta \tag{4.14}$$

Inverse Proportional Weighting

Analogous to *Proportional Weighting*, Inverse Proportional Weighting accounts for the frequency-based prominence of the fixation-index. However, the least frequent (in the overall attention data) fixation-index is assigned with the highest weight.

$$w_f = 1 - \frac{\text{frequency}(f)}{\sum_{i=1}^{\theta} \text{frequency}(f)} \quad \forall\, 1 \leq f \leq \theta \tag{4.15}$$

For each of the aforementioned weighting-strategies, the effective fixation-index is achieved using the algorithm 1. That is, the product of implicit prominence (w_f) and the explicit

prominence (v_f) is computed to assign the FI with the highest product as the effective fixation-index of I. The assigned effective fixation-index acts as the ground-truth attention for I.

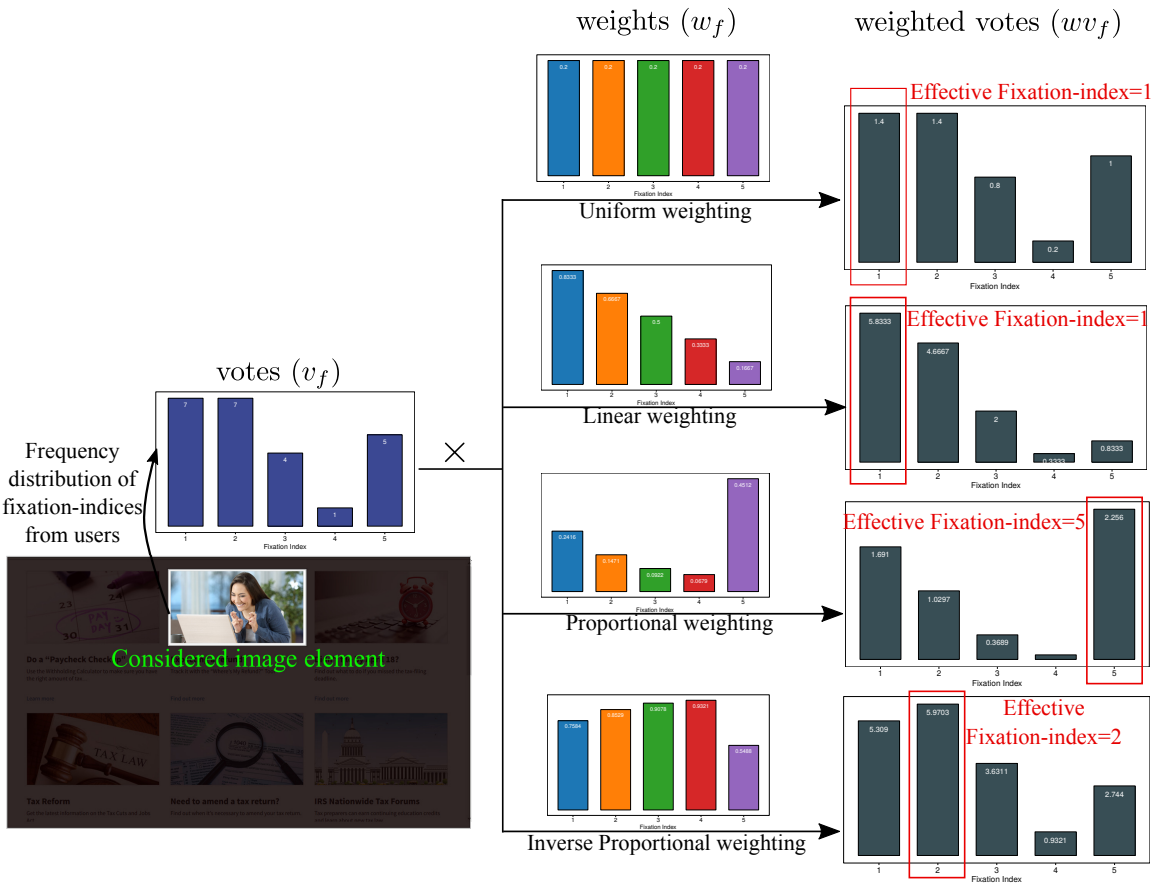


Figure 4.11: Illustration of weighting methods on a sample web image. The fixation-index with the highest weighted votes (wv_f) is assigned as the effective fixation-index in each weighting strategy.

Figure 4.11 shows a sample effective fixation-index assignment using the four weighting strategies. The considered image element obtained five unique fixation-indices 1,2,3,4,5 with the respective votes of 7,7,4,1,5. The application of each weighted-voting strategies and the selection of FI with maximum weighted votes resulted in the effective fixation-indices of 1,1,5,2 respectively for uniform weighting, linear weighting, proportional weighting, and inverse proportional weighting. Note, in case of a tie, the lower fixation-index is assigned as an indication of element's ability to draw the user's initial fixation. Evidently, all the weighting strategies are directed towards the assignment of optimal representative fixation-index, analogous to majority-weighted-voting utilized in crowdsourcing [84, 85]. To note, all the weighting strategies result in the same effective FI in the presence of single unique FI. The proposed effective fixation-index (effective FI) prediction approach is described in the rest of the section for each weighting strategy.

4.5.2 Prediction Model

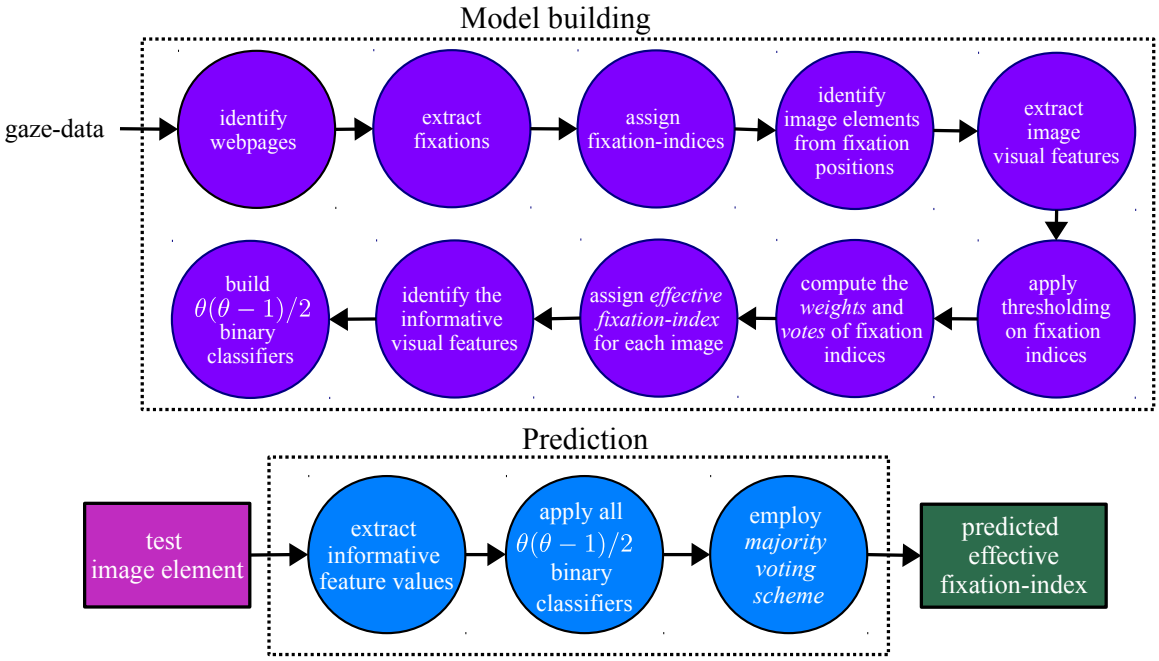


Figure 4.12: Proposed model flow-diagram

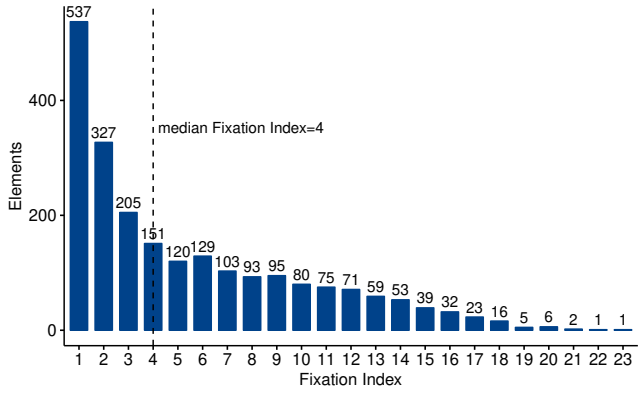
The prediction approach is shown in Figure 4.12. The aforementioned multi-class classification is utilized for attention prediction. That is, as there are θ classes, we built $\binom{\theta}{2} = \theta(\theta-1)/2$ binary classifiers to classify between every pair of possible classes. For this, the ensemble of data-points corresponding to both the classes are utilized from given d data-points. For a test image element I_t, the classification results are obtained from the constructed $\theta(\theta-1)/2$ classes and the class with the highest frequency is determined as the effective fixation-index e_t. That is, the *majority-voting-scheme* is employed towards the effective fixation-index prediction.

To analyze the performance of the proposed approach, the fixation data from eye-tracking Experiment-II (as described in Section 3.2) is utilized.

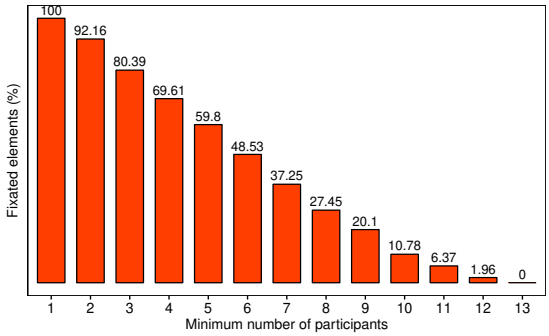
Preliminary Data Characteristics

The average number of fixated image elements per webpage is 6. On an average 5.54 participants attended to each image element. The percentage of attended image elements by varying minimum number of attended participants is shown in Figure 4.13b. Maximum number of participants attended to an image element is 12. Further, more than 50% of the images are fixated by at least five participants. This highlights the necessity of effective fixation-index assignment.

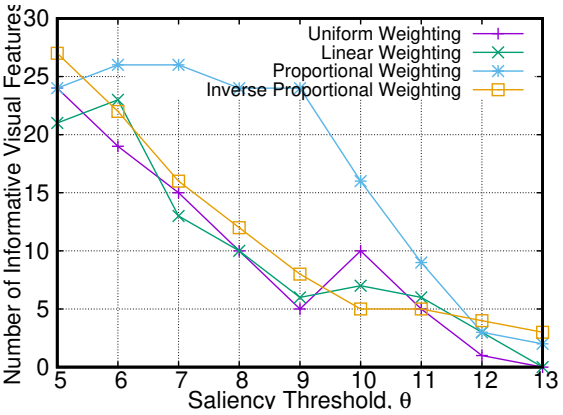
For the analysis, the median FI is considered for θ selection. That is, θ is set to 5 in Equation 4.11.



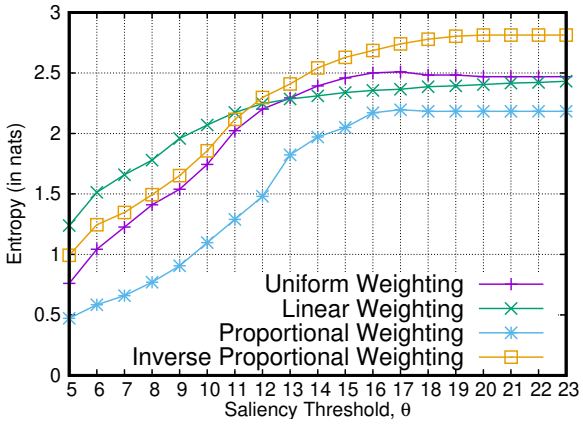
(a) Frequency distribution



(b) fixated images (%)



(c) Informative features



(d) Entropy

Figure 4.13: θ based variation in (c) number of information visual features, and (d) entropy, for four weighting strategies

Feature Selection

Feature selection helps to prune the redundant features and to select the informative features in explaining the effective visual attention. We considered the features with positive information gain scores for the analysis, analogous to [19]. The number of informative visual features are 24, 21, 24, 27 respectively for uniform, linear, proportional and inverse proportional weightings. Among them, 18 visual features were informative across the four weighting strategies. The informative features along with information gain scores are shown in Table 4.6.

Across the weighting strategies, the element's 'top position' (distance from the top edge of the webpage) obtained the predominant information gain highlighting its attention explaining ability. The HISTOGRAM features succeeded the POSITION feature and are highly informative across the weighting strategies. However, the HISTOGRAM features corresponding to the *pure* colors (first and eighth histogram bins) were not found to be informative across all weighting strategies. This is attributed to their wide usage as

Table 4.6: Information-gain scores of visual features at $\theta = 5$ for four weighting strategies. Color intensity represents the relative importance of a feature. The feature names starting with "hist" and "diff_hist" respectively denote the histogram features and contrast histogram features where the corresponding color component (R,G, B, gray) and bin number $(1, \ldots, 8)$ are suffixed. The "rect.top", "rect.left", "rect.bottom", "size" denote the rectangular image element's 'top distance', 'left distance', 'bottom distance', and 'area'.

Group	Feature	Uniform	Linear	Proportional	Inv. Pro.
POSITION	rect.top	0.1110	0.1380	0.0976	0.1186
HISTOGRAM	hist_R_6	0.1118	0.0998	0.0935	0.1043
HISTOGRAM	hist_gray_7	0.1086	0.0809	0.0981	0.1174
HISTOGRAM	hist_gray_5	0.1047	0.0876	0.1061	0.1217
HISTOGRAM	hist_G_6	0.1037	0.0998	0.0967	0.1084
HISTOGRAM	hist_G_5	0.1002	0.0773	0.0889	0.1101
HISTOGRAM	hist_gray_6	0.0997	0.1054	0.1067	0.0989
POSITION	size	0.0946	0.0967	0.0972	0.1193
HISTOGRAM	hist_B_5	0.0982	0.0799	0.0852	0.1030
HISTOGRAM	hist_R_5	0.0925	0.0842	0.0947	0.1103
HISTOGRAM	hist_R_4	0.0802	0.1389	0.0671	0.0962
HISTOGRAM	hist_B_3	0.0863	0.0869	0.0781	0.0851
HISTOGRAM	hist_B_6	0.0877	0.0707	0.0665	0.0816
HISTOGRAM	hist_B_4	0.0763	0.0692	0.0700	0.0985
POSITION	rect.left	0.0746	0.1014	0.0635	0.0919
HISTOGRAM	hist_gray_3	0.0720	0.0773	0.0675	0.0921
HISTOGRAM	hist_gray_4	0.0664	0.0710	0.0619	0.0858
HISTOGRAM	hist_B_2	0.0660	0.0743	0.0692	0.0916
HISTOGRAM	hist_R_7	0.0931		0.0775	0.1119
HISTOGRAM	hist_G_7	0.0887		0.0810	0.0980
HISTOGRAM	hist_G_3	0.0795		0.0617	0.0936
CONTRAST	diff_hist_G_5	0.0743		0.0776	0.0927
HISTOGRAM	hist_G_2		0.0695		0.0896
HISTOGRAM	hist_G_4	0.0679			0.0794
CONTRAST	diff_hist_R_3	0.0749		0.0614	
HISTOGRAM	hist_R_3		0.0684		0.0755
CONTRAST	diff_hist_B_2			0.0527	0.0549
POSITION	rect.bottom		0.0858		
HISTOGRAM	hist_gray_2				0.0799

background color (typically, white) across all the webpages, normalizing their attention explaining ability. Overall, the element's position ('top distance', 'left distance'), size and the de-saturated color histograms were found to be informative across all the weighting strategies. This signifies the prominence of image position, size, and intrinsic histogram visual features in explaining the free-viewing user attention.

Surprisingly, no COMPREHENSIVE feature was found to be informative in any weighting strategy and no CONTRAST feature (from Table 3.3) was found to be informative across all four weighting strategies. Though three CONTRAST features were informative in at least one weighting strategy, the associated information gain scores were lower. The non-prominence of CONTRAST features may be attributed to the user's inherent preference for the image elements on a webpage.

Number of Informative Features Vs. θ

To further understand the informative features, the number of informative visual features are computed with variation in θ as shown in Figure 4.13c. The number of informative features for uniform, linear, and inverse proportional weighting significantly decreased with an increase in θ. On the contrary, the number of informative features were relatively stable for proportional weighting up to $\theta = 9$ after which it followed the reduction pattern of remaining weighting strategies. Beyond $\theta = 12$, the number of informative visual features are approaching zero as other factors influence the attention which cannot be described by visual features.

Entropy Vs. θ

To understand the influence of θ on effective FI assignment (ground-truth), the entropy was computed for the four weighting strategies as shown in Figure 4.13d. At median θ, the linear weighting achieved the highest entropy among all strategies, and consistently remained till $\theta = 11$. However, the increment in the entropy gradually reduced as the latter FIs are less frequent and were further weighted lower. The inverse proportional weighting achieved the second highest entropy till $\theta = 11$ but outperformed the linear weighting beyond. The highest entropy achieved for the latter θ is attributed to the higher weight assignment for the less frequent latter FIs. The uniform weighting and proportional weighting followed a similar entropy pattern, however, the latter weighting strategy achieved relatively lesser entropy owing to the FI's frequency based weighting.
Observing the variation in number of informative visual features (Figure 4.13c) and the entropy (Figure 4.13d), we conclude that the distribution of effective fixation-indices varying significantly (higher entropy with increase in θ) which can not be explained by the visual features (lesser number of informative features with increase in θ). This also indicates, among the four strategies, proportional weighting strategy is a better candidate for modeling the effective visual attention with image visual features. However, to quantify and compare the prediction performance, metrics were computed for all the four strategies.

4.5.3 Prediction Performance

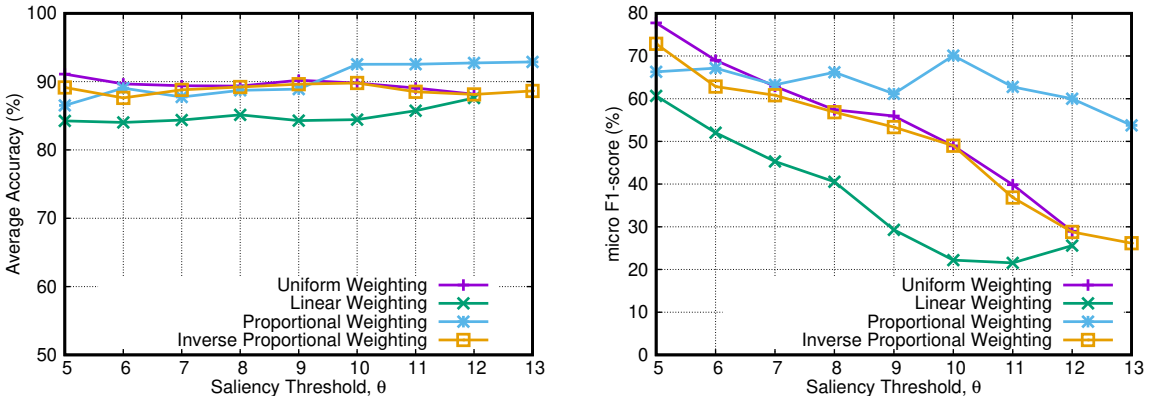
The prediction model was trained with 90% of data and tested with remaining 10% obtained from the application of each weighting strategy. At median FI, a total of $\theta(\theta-1)/2 = 10$ binary SVM classifiers were constructed with the majority-voting-scheme employed to predict the effective FI on test images. The performance is validated with *Repeated 10-fold cross-validation* with 5 repetitions. That is, data was randomly sampled after each repetition (in which 10-fold performance metrics were computed) to mitigate the influence of possible special structures emanating in splitting the data. The performance over all the iterations and repetitions were averaged to obtain the overall model performance.

Table 4.7: Prediction performance at median FI. Boldface indicates the best performance and underlined text indicates the second best performance.

Weighting	Average Accuracy			micro F1-score		
	Predicted	Baseline	↑ (%)	Predicted	Baseline	↑ (%)
Uniform	**91.10**	69.22	<u>31.61</u>	**77.75**	23.04	237.46
Linear	84.26	68.43	23.13	60.65	21.08	187.71
Proportional	86.52	67.25	28.65	66.30	18.14	<u>265.49</u>
Inverse Proportional	<u>89.14</u>	67.31	**32.43**	<u>72.85</u>	18.28	**298.52**

Towards the performance, the standard multi-class classification metrics, *Average Accuracy* and *micro F1-score* were computed.

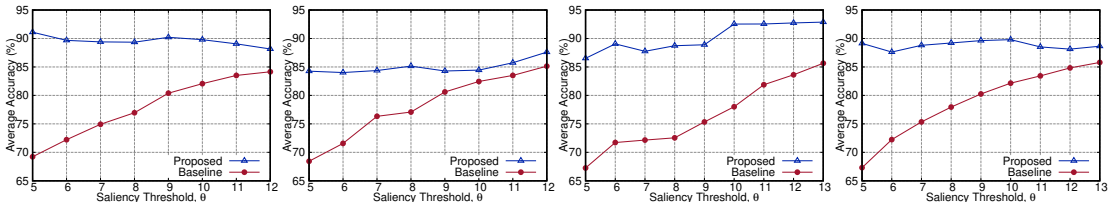
The performance metrics of the four weighting strategies are shown in Table 4.7. The visual features performed the best at uniform weighted effective visual attention prediction, followed by Inverse proportional weighting, proportional weighting, and linear weighting. The best performance of uniform weighting indicates that all the FIs (up to median FI) are equally prominent and the visual features can computationally predict the corresponding effective visual attention. Additionally, it is supported by the relatively poorer prediction of linearly weighted (weight decreased with increase in FI) effective visual attention. Further support is provided by the second best performing inverse proportional weighting. The weights introduced in the Inverse proportional strategy are comparable to the uniform weights (as shown in Figure 4.11) which significantly varies from the linear weights. The prediction performance of the proportional weighted effective visual attention is more closer to the uniform weighted and inverse proportional weighted strategies than the linear weighted strategy.



(a) Average accuracy

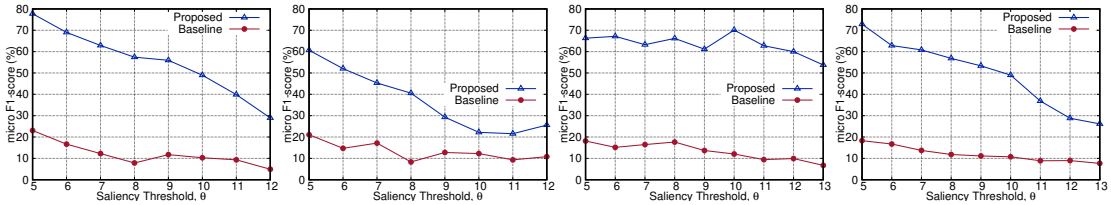
(b) micro F1 = micro Precision = micro Recall

Figure 4.14: Average Accuracy and micro F1-scores with variation in θ



(a) Uniform (b) Linear (c) Proportional (d) Inverse Proportional

Figure 4.15: Average accuracies of four weighting strategies compared with the baseline for varying θ



(a) Uniform (b) Linear (c) Proportional (d) Inverse Proportional

Figure 4.16: Micro F1-scores of four weighting strategies compared with the baseline for varying θ

Comparison with Baseline

The majority of the existing approaches are dichotomous (salient or not prediction) in nature and are limited in predicting the multi-level (FI) attention. As there is no baseline algorithm that exists for comparing with the proposed method, we follow the strategy proposed in [32, 118] by employing random prediction model. That is, to demonstrate the efficacy of visual features in predicting the effective visual attention, the performance was compared with *Random prediction (RP) model*— randomly predicts a effective fixation-index from $\{1, \ldots, \theta\}$ [107]. The RP model performance comparison with our proposed model is shown in Table 4.7. Evidently, all weightings outperformed the baseline across all metrics. The inverse proportional weighting outperformed the baseline with an improvement of 32.43% in average accuracy and an improvement of 298.52% in micro F1-score. The second best performance in average accuracy improvement is 31.61% (for uniform weighting) and in micro F1-score improvement is 265.49% (for Proportional weighting).

Performance with variation in saliency threshold θ

To understand the influence of θ, performance was measured at θ varying further after median, $\theta = 5, 6, \ldots$ as shown in Figure 4.14. The average accuracy of the four weighting strategies consistently remained high with more than 84% where linear weighting exhibited relatively poorer accuracy. In contrary, the differences in prediction performance were lucid in the micro F1-score computation as shown in Figure 4.14b. The proportional

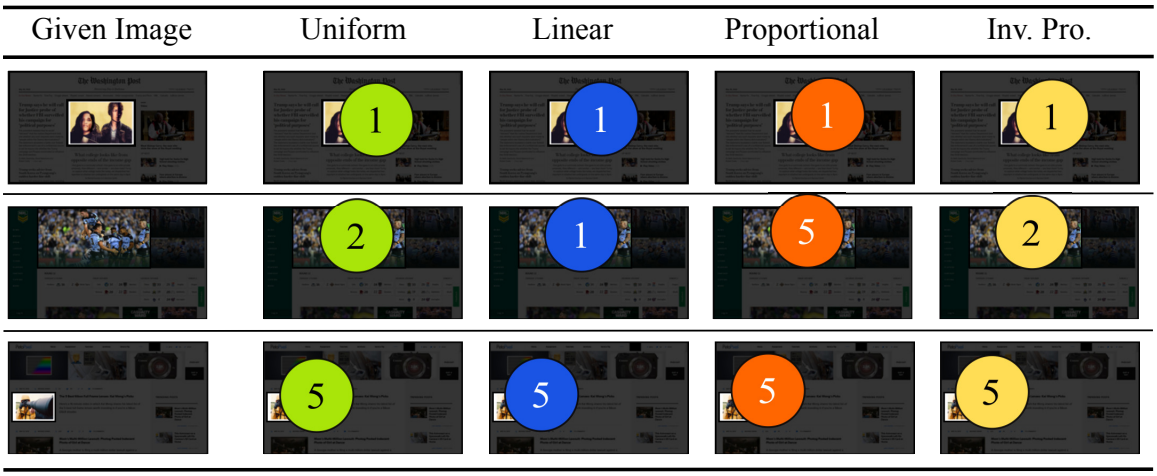


Figure 4.17: Example effective visual attention predictions. Left-most column represents the test image and subsequent columns denote the predictions from each weighting strategy

weighting consistently outperformed other weighting strategies with a micro F1-score of at least 60%. The uniform weighting and inverse proportional weighting demonstrated similar prediction with a gradual reduction in performance with an increase in θ. The linear weighting consistently demonstrated the poorer performance among all, with a significant reduction in performance with an increase in θ. Performance reductions of the uniform, linear and inverse proportional weighting strategies indicate that all FIs are not equally prominent and their prominence can neither be linearly weighted and nor their frequency of occurrence be ignored. Overall, the uniform weighting performs better for the initial FI prediction (up to effective FI=7) while proportional weighting performs the best for the latter FI prediction (effective FI=7 onwards) as shown in Figure 4.14. The performance comparison with the baseline are shown in Figure 4.15 for average accuracy and in Figure 4.16 for micro F1-score. With the increase in θ, the prediction performance approached towards the baseline. However, the proportional weighting significantly outperformed the baseline throughout the variation in θ (see Figure 4.15c and Figure 4.16c). The example predictions are shown in Figure 4.17.

4.6 Multi-users' Multi-level Attention (MUMLA) on Images

In Section 4.5, different weighting strategies are proposed to obtain a unique representative (aka effective) FI from multiple users' FI allocations. This section presents an approach to predict all the unique FIs associated with an image element without considering any weighting. Essentially, following research question is investigated.

R5 → How the image visual features perform in predicting the *free-viewing* Multi-users' Multi-level Attention (MUMLA)?

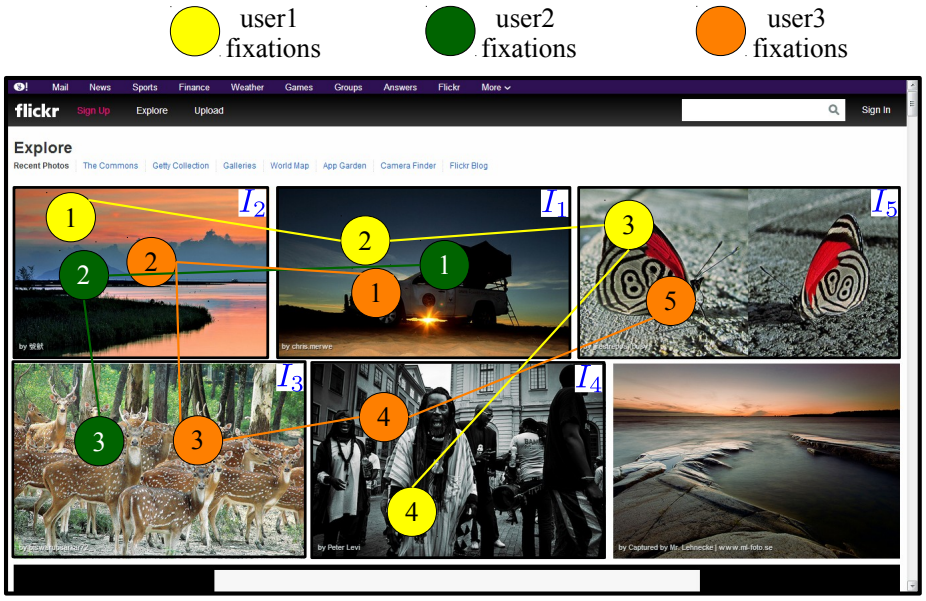


Figure 4.18: Illustrative example of multi-user attention on sample webpage.

4.6.1 Introduction and Motivation

Owing to the user idiosyncrasies [125], each image may either receive (i) *similar but different* FIs (levels) of user attention (say, first and second levels), or (ii) *diverse* FIs of user attention (say, first and seventh FIs), or (iii) same FI from all the users. The work by [155] is suitable for the last setting. In the first setting, assigning one of the similar levels (to web image) and overlooking others may result in inconsistencies as, the knowledge of the similar levels is not incorporated into the modeling. In the second setting, assigning a single multi-level attention may not be adequate as the image is associated with diverse levels (classes) with no specific level being dominant. Moreover, the multi-users' attention on web images may be an amalgamation of all the three settings which needs to be adequately modeled.

$\mathcal{D} \equiv$

Image	$FI = 1$	$FI = 2$	$FI = 3$	$FI = 4$	$FI = 5$
I_1	+1	+1	-1	-1	-1
I_2	+1	+1	-1	-1	-1
I_3	-1	-1	+1	-1	-1
I_4	-1	-1	-1	+1	-1
I_3	-1	-1	+1	-1	+1

In the setting of multi-user attention allocation, each image may receive either of three types of attention as described above. In Figure 4.18, the images I_1, I_2, I_3, I_4, I_5 respectively received the bag of fixation-indices $\wr 2, 1, 1\int, \wr 1, 2, 2\int, \wr 3, 3\int, \wr 4, 4\int, \wr 5\int$ where $\wr\int$ denotes a bag. Accordingly, the unique FIs associated with the images in Figure 4.18 are as tabulated in $\mathcal{D}$ above. The +1 indicates the presence of an attention-level and -1 indicates the absence. The images I_1, I_2 drew similar but different attention levels (FI=1,2 for each), images

I_3, I_4 drew same level of attention across users (FI=3 and FI=4 respectively), while I_5 drew diverse attention-levels (FI=3 and FI=5). Among the images, assigning a unique multi-level attention to I_1 and I_2 may not justify their attention-drawing ability as both are predominantly drawing users' initial attention but with different attention-levels. Further, for image I_5, assigning either of FI=3 and FI=5 levels may result in discarding of other diverse level with similar prominence.

Thus, to preserve the prominence of each attention-level and to predict the possible attention levels of a given image, we propose a Multi User Multi Level Attention (MUMLA) prediction approach as described in the following section.

4.6.2 MUMLA Prediction Approach

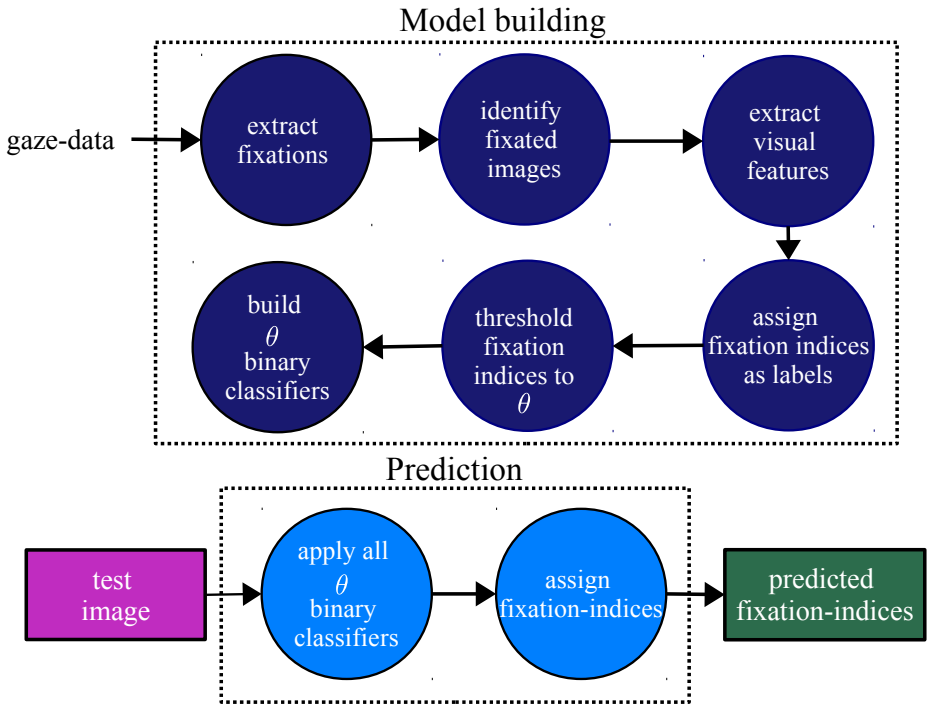


Figure 4.19: Proposed model flow-diagram

We formulated the MUMLA prediction as a multilabel classification problem. The model flow diagram of the proposed approach is shown in Figure 4.19.

To analyze the performance of the proposed MUMLA prediction approach, the fixation data from Experiment-III (described in Section 3.4) and Experiment-II (described in Section 3.3) are utilized.

4.6.3 MUMLA Prediction with Experiment-III

We demonstrate the performance of the proposed approach on open-source dataset [134] consisting of *Text*-rich webpages, *Mixed* webpages, and *Pictorial*-rich webpages (see the sample webpages in Figure 3.6). The rest of the section describes the data-processing and preparation followed by multi-user multi-level attention characteristics and prediction according to the proposed approach.

Data Preparation, Description & Characteristics

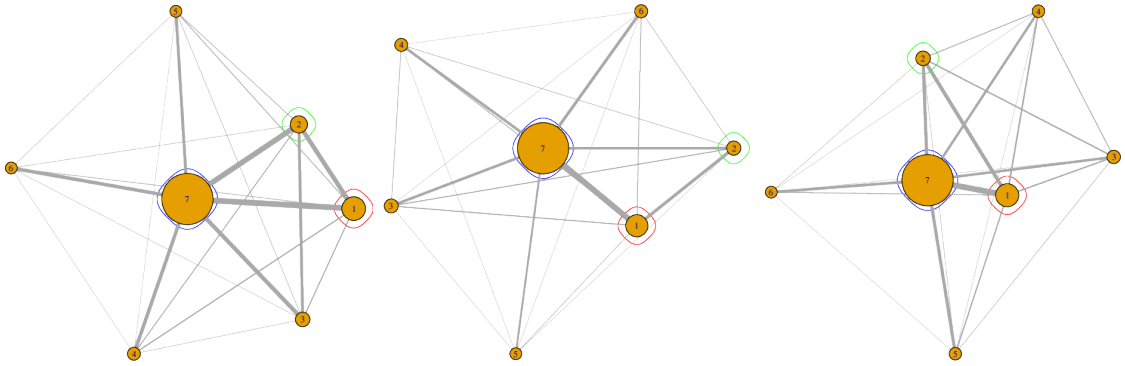
Table 4.8: Data Characteristics of the Three Categories

Characteristic	Text	Mixed	Pictorial
Number of Webpages	50	49	50
Number of Users	11	11	11
Number of unique fixations	1369	1866	2527
Median Attention-level	6	7	7
Label Cardinality, $LCard(\mathcal{D}) = \frac{1}{d}\sum_{i=1}^{d}\lvert L_i\rvert$ at $\theta = 7$	<u>2.0228</u>	1.947	**2.1426**
Label Density, $LDens(\mathcal{D}) = \frac{LCard(\mathcal{D})}{\theta}$ at $\theta = 7$	<u>0.289</u>	0.2781	**0.3061**
Label Diversity, $LDiv(\mathcal{D}) = \lvert\{L \mid \exists I : (I, L) \in \mathcal{D}\}\rvert$ at $\theta = 7$	71	<u>72</u>	**80**
Proportion of Label Diversity, $PLDiv(\mathcal{D}) = \frac{LDiv(\mathcal{D})}{d}$ at $\theta = 7$	**0.2023**	<u>0.1589</u>	0.1541

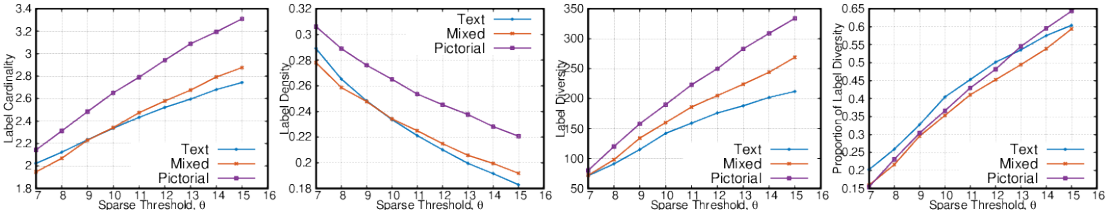
The frequency distribution of FIs for the three categories is shown in Figure 3.7. Evidently, web images on Pictorial-category webpages consistently received more fixations followed by Mixed-category and Text-categories. This is attributed to the presence of more number of image elements on Pictorial-category than its counterparts. However, across the categories, the frequency of FIs consistently reduced with increasing FI indicating the reduction in salient image elements.

θ Selection: The median FI of the three categories is 6,7,7 respectively for Text, Mixed, Pictorial categories. To segregate the prominent FIs from non-prominent FIs and to enable the comparison across categories, we consider maximum of the median fixation-indices, i.e., 7 as the representative *median saliency threshold* (median FI) θ. Accordingly, FIs $1, 2, \ldots, 6$ are prominent being the initial allocations and the FI 7 corresponds to non-prominent latter attention allocations.

MUMLA Data Characteristics: Considering each FI as a class label, the MUMLA data characteristics [13, 166] are summarized in Table 4.8. *Label Cardinality*— the average number of FIs associated with each image element; highlights the prominence of multi-label approaches. On an average, image elements received approximately two attention-levels across categories. Thus, it may not be good idea to apply the existing single-label assignment approaches as, two attention-levels may be similar but different allocations or diverse



(a) Co-occurrence graph: (from left to right) Text, Mixed, Pictorial



(b) Multilabel data characteristics: (from left to left) Cardinality, Density, Diversity, Proportional Diversity

Figure 4.20: Data characteristics: (a) Co-occurrence graph of MUMLA at median FI thresholding. The bubble size indicates the frequency of FI and the edge width indicates the co-occurring frequency of the connected fixation-indices. First two FIs and median FI (7) are highlighted for better interpretation; (b) Multilabel data characteristics with variation in sparse threshold θ.

attention allocations. Analogously, the *Label Density*— normalized Label Cardinality also followed the similar trend. The *Label Diversity*— distinct number of FI sets (where, each L_i is an attention-level set) in $\mathcal{D}$; further strengthened the necessity of multi-label approaches. At $\theta = 7$, $2^7 - 1 = 127$ unique FI sets (each $\subseteq \mathcal{L}$) are possible out of which 55.91%, 56.69%, and 62.99% are part of Text, Mixed, and Pictorial categories respectively. Further, Text-category images have more *Proportion of Label Diversity*—normalizing the Label Density with number of instances, than the remaining two categories. The relatively lower $LDiv$ value and the relatively higher $PLDiv$ value of Text-category indicates, the respective category images are associated with more *multi-labelness* though less number of image elements are present in this category. The counter discussion holds true for Pictorial-category.

To further understand the attention-levels' co-occurring characteristic, the co-occurrence (one attention-level occurring along with the another in L) graph [34] is plotted at median saliency thresholding as shown in Figure 4.20a. The Text-category followed by Pictorial-category images received relatively more 'similar but different' and 'diverse' levels of attention as indicated by the thicker connecting lines among first, second and seventh attention-levels. In contrary, the Mixed-category images received relatively more 'diverse' levels of attention allocations than 'similar but different' attention allocations as the line connecting first and

second attention-levels is relatively thinner.

MUMLA Data Characteristics with θ Variation: To understand the influence of θ selection on *multi-labelness*, the four characteristics are computed with increasing θ as shown in Figure 4.20b. The increase in Label Cardinality and the corresponding decrease in Label Density with θ indicates the latter attention-levels are sparse and are co-occurring with initial attention-levels which also contributed to increase in Label Diversity and Proportion of Label Diversity. Overall, the Pictorial-category demonstrated relatively more multi-labelness than its counterparts.

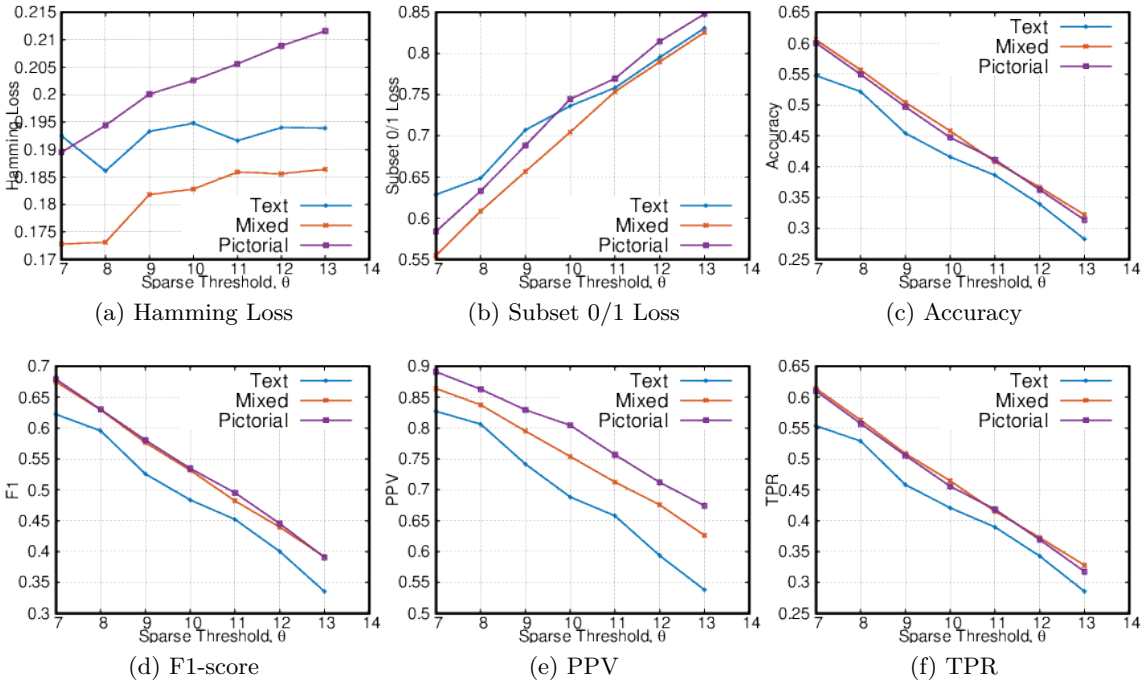
Prediction Performance

Table 4.9: Prediction performance at median saliency-thresholding ($\theta = 7$). Boldface: best performance; Underlined: second best performance; among the three categories.

Category	Metric	Predicted	Baseline	Outperformance (%)
Text	Hamming loss	0.1925	0.3172	64.78
	Subset 0/1 loss	0.6286	0.9077	44.40
	Accuracy	0.5476	0.3479	**57.40**
	F1-score	0.6223	0.4548	**36.83**
	PPV	0.8272	0.5052	**63.74**
	TPR	0.5537	0.5470	1.22
Mixed	Hamming loss	**0.1728**	0.2866	**65.86**
	Subset 0/1 loss	**0.5546**	0.8825	**59.12**
	Accuracy	**0.6057**	0.3862	56.84
	F1-score	0.6749	0.4958	36.12
	PPV	0.8641	0.5455	58.41
	TPR	**0.6141**	0.5966	**2.93**
Pictorial	Hamming loss	0.1895	0.3120	64.64
	Subset 0/1 loss	0.5838	0.8856	51.70
	Accuracy	0.6004	0.3878	54.82
	F1-score	**0.6792**	0.5029	35.06
	PPV	**0.8912**	0.5539	60.90
	TPR	0.6101	0.6027	1.23

Without loss of generality, the MUMLA dataset, $\mathcal{D}$ in each category is randomly split into 80:20 train, test ratio. θ binary-classifiers are constructed using the training data while the test-data is utilized towards prediction performance computation. Accordingly, *5-fold cross-validation* is performed with 10 iterations (to mitigate the possible occurrence of special structures during the random split). The performance metrics averaged across the iterations are obtained as overall performance.

The prediction performance at median saliency-thresholding is shown in Figure Table 4.9. To understand the quality of prediction, the performance is compared with *Random prediction model* [107]— randomly assign the FIs overlooking the constituting visual features. In each category, all the metrics outperformed the random prediction metrics demonstrating the



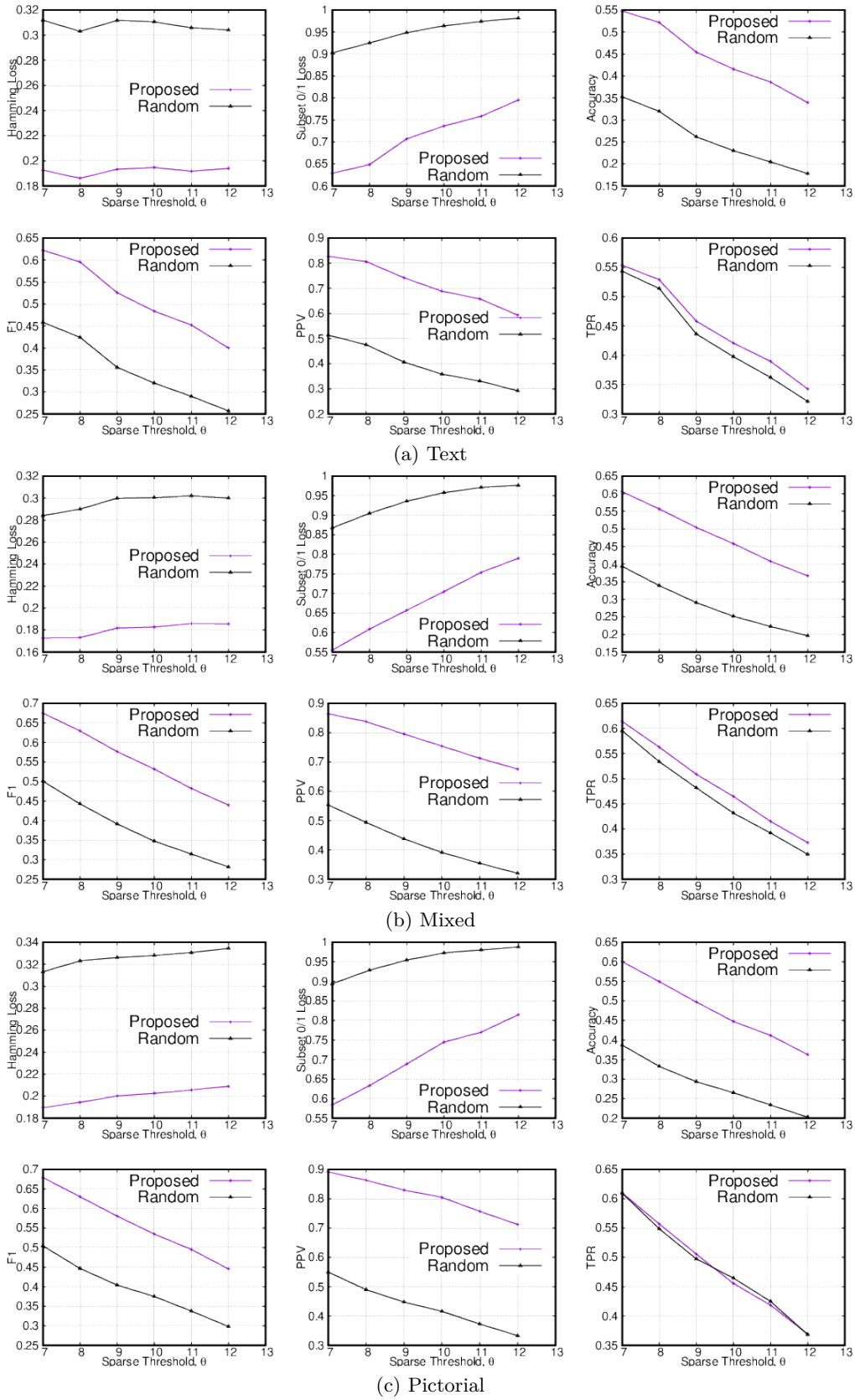
(a) Hamming Loss (b) Subset 0/1 Loss (c) Accuracy

(d) F1-score (e) PPV (f) TPR

Figure 4.21: MUMLA prediction performance with variation in saliency-threshold θ

efficacy of the proposed approach and the considered image visual features. To note, the True Positive Rate (TPR) (also called, recall) of the baseline is comparable to predicted performance. However, the Positive Predictive Value (PPV; also, called precision) and F1-score of the baseline are significantly lower than the predicted values highlighting the bias associated with random prediction. Overall, the MUMLA prediction on web images from Mixed-category and Pictorial-category performed better than the Text-category. However, the relative performance with respect to the random prediction is better in Text and Mixed categories than Pictorial category.

Prediction Performance with θ Variation: To further analyze the influence of θ on prediction, the performance metrics are computed for each increment in θ value for three categories. The progression of the performance is shown in Figure 4.21. Among all the metrics, the variation in θ has little influence on Hamming Loss which is consistently around 0.19 as shown in Figure 4.21a. Overall, the prediction performance for the Pictorial and Mixed categories are comparable and both are relatively better than the prediction for Text-category (see Figures 4.21c, 4.21d, 4.21e, 4.21f). For all the three categories, the prediction performance reduced with increasing θ indicating the influence of other factors (such as reduction in salient elements and possible influence of semantic features) which may not be explained by the considered visual features. However, the prediction performance consistently outperformed the baseline throughout the variation in θ as shown in Figure 4.22.



(a) Text

(b) Mixed

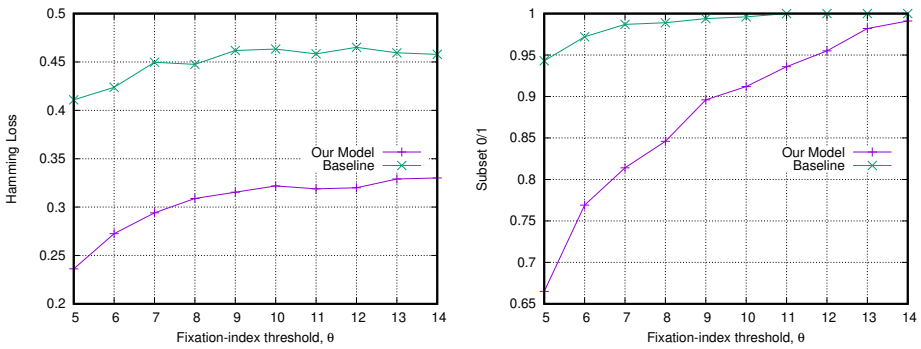
(c) Pictorial

Figure 4.22: Multilabel prediction performance comparison with baseline for three categories: (from left to right) Hamming loss, Subset 0/1 loss, Accuracy, F1-score, PPV, and TPR.

Table 4.10: Prediction performance at median fixation-index

Metric	Our Model (%)	Baseline (%)	Improvement (%)
Subset 0/1 loss	**66.50**	94.30	29.48 ↓
Hamming loss	**23.62**	41.08	42.50 ↓
Accuracy	**67.49**	47.24	42.87 ↑
Precision	**84.14**	66.30	26.91 ↑
Recall	**78.91**	71.92	09.72 ↑
F1-score	**76.57**	61.07	25.38 ↑

4.6.4 MUMLA Prediction with Experiment-II

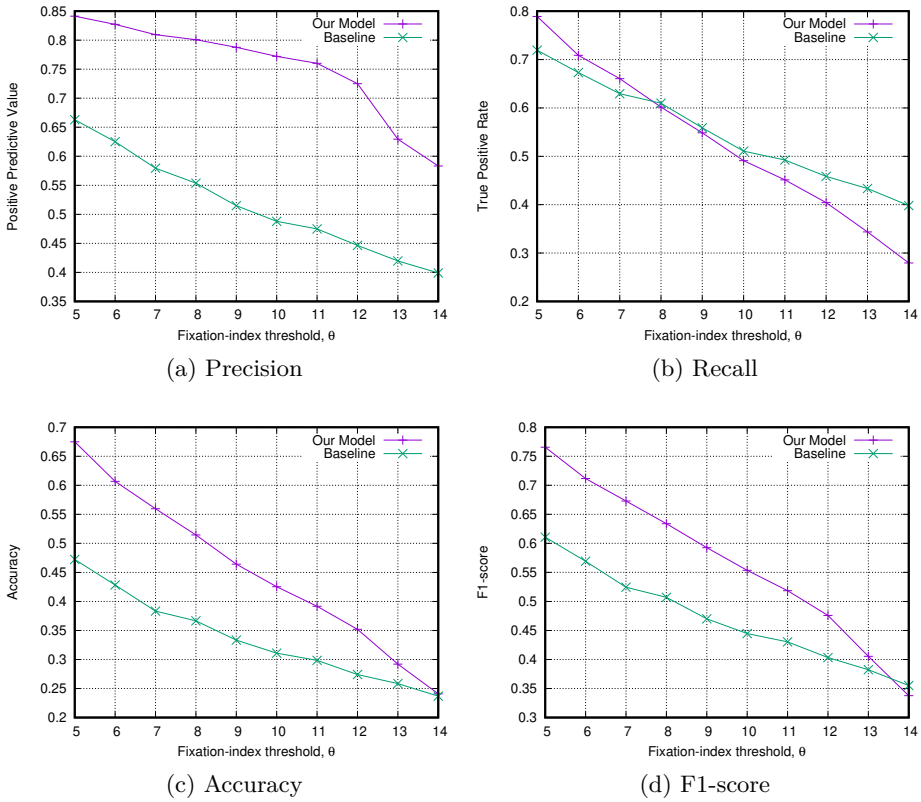


(a) Hamming Loss (b) Subset 0/1 Loss

Figure 4.23: Loss metrics with variation in fixation-index threshold

The computed performance metrics are summarized in Table 4.10. Clearly, our approach outperformed the baseline model for all the metrics. Among the loss functions, the subset 0/1 loss seems high with 66.50%. However, this is attributed to the nature of its computation where even a misclassification of the single label is considered as complete misclassification which was avoided in the computation of Hamming loss. To note, the subset 0/1 loss, hamming loss, accuracy, precision and F1-scores of our model performed significantly better than the baseline (more than 25% improvement). The recall of the baseline is comparable with our model indicating its bias towards a specific metric.

To further understand the influence of thresholding (θ value) on performance, metrics were computed with variation in θ from 5 to 14 as shown in Figure 4.23 and Figure 4.24. Our model consistently performed better than the baseline. However, the prediction performance reduced with increase in the threshold value, indicating the reduced influence of visual features in drawing the user attention. Especially, the baseline demonstrated relatively smaller but better performance than our model from eighth fixation-index thresholding. This performance is attributed to the nature of random predictions which tend to bias to a particular metric and perform significantly poorer with respect to the other metrics.



(a) Precision (b) Recall

(c) Accuracy (d) F1-score

Figure 4.24: Performance metrics with variation in fixation-index threshold

In summary, the performance metrics demonstrated the efficacy of our approach in predicting the MUMLA attention on webpage images.

4.7 Summary

In this chapter, primarily, position-based attention allocation was presented in terms of the fixation-indices. The analyses revealed that though the users predominantly allocate the initial attention to Middle region and the Top regions, the elements in Right and Bottom regions are not completely ignored. The observation may be attributed to the visual features of the elements. To understand the prominent visual features, the information-gain scores were computed between the visual features and fixation-indices. The analysis revealed that *space* and *font-size* visual features are informative for textual elements, and *Mid-level Color Histogram* visual features are informative for image elements, while `position` and `size` are informative for both the modalities. The subsequent computational attention modeling demonstrated the efficacy of informative visual features in predicting the user attention on web elements with 90% average accuracy and 70% micro-F1 score.

Owing to the prominence of image elements, respective modality-specific attention prediction models were proposed in the context of (i) non-linear association between visual features

and user attention, (ii) Weighted-Voting-based ground-truth attention preparation, and (iii) visual features predicting multiple levels of user attention (Multi User Multi Level Attention (MUMLA)) settings. The empirical results demonstrated that the image visual features outperform the random baseline in all the three settings. Though the Gaussian RBF kernel outperformed other kernels (91.64% average accuracy and 79.10% micro-F1 score), the comparable performance of the linear-kernel (89.04% average accuracy and 72.60% micro-F1 score) highlighted the linear association between image visual features and user attention. Towards the ground-truth preparation, the *uniform weighting* (up to effective FI of 7) is suitable for initial FI prediction while *proportional weighting* is best suited for latter FI prediction (from effective FI of 7 onwards).

5

Unification of Multi-Modalities

In Chapter 4, the prominence of text and image visual features are established in explaining the free-viewing attention. However, both the modalities in unison guide the user attention on bi-modal webpages. But, the intrinsic visual features of the modalities are different. This heterogeneity in visual features complicates the simultaneous user attention analyses and modeling on bi-modal webpages. This chapter presents a computational approach to unify the multi modalities and empirically analyzes the performance of the same.

The widely utilized Areas-Of-Interest (AOIs) are the prima facie of 'where on the screen' paradigm. That is, the AOI-based analyses explore *where* is the user interest (from the perspective of visual attention) on an interface in contrast to *what* interests the user on an interface. Further, the interpretations vary based on the manual selection of AOIs. Moreover, the interpretations are not generalizable as the underlying visual information is not associated with the user attention. To overcome these limitations, Data-Of-Interests (DOIs), a prima facie of the 'what on the screen' paradigm are introduced [3]. Additionally, [79] introduced a data-centric formalism (DOI-model) for eye-tracking analyses which helps in answering a broad set of questions corresponding to Roth's objective primitives— *identify, compare, rank, associate,* and *delineate* [127].

The predominant bottleneck in the DOI-based analyses is elements' *heterogeneity* on graphical user interfaces (GUIs). Especially, the prominent webpages consist of heterogeneous data modalities, including text and images. Elements of each modality are associated with respective intrinsic visual features, such as *brightness, intensity, color histograms* for images, and *font-size, color, font-weight* for text. As the visual features of one modality differ from the other modality, and both the modalities determine the attention in unison, performing the visual feature-based attentional analyses is complicated.

To overcome these limitations, we investigate the possible unification between heterogeneous modalities (text and images) on webpages. Especially, the prominence of user idiosyncrasies and interface idiosyncrasies are explored in the context of DOI formalism. The contributions are the investigations of the following research questions (as highlighted in Chapter 1) whose

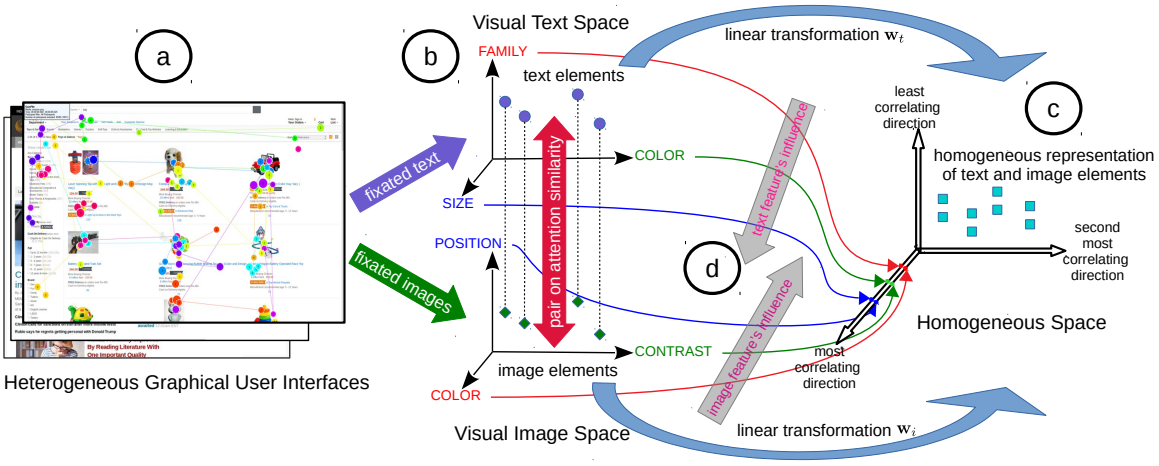


Figure 5.1: Flow diagram of the proposed approach: (a) Heterogeneous webpages with user fixations; (b) Represent DOIs (fixated webpage elements) in respective visual feature spaces and pair the heterogeneous DOIs based on *attention similarity* (measured with respect to fixation indices); (c) Learn a homogeneous space ($\mathbf{w}_t$s and $\mathbf{w}_i$s) through correlation maximization between paired DOI; (d) Compute the correlation between each of the heterogeneous visual features and obtain a maximally correlated homogeneous feature space to identify and compare the attention-influential cross-modal features; Compute the distances among DOIs in original feature space and the homogeneous space to quantify the delineation.

solutions are obtained through the approach shown in Figure 5.1:

R1: Are the text and image visual features correlated based on the free-viewing user attention allocation on bi-modal webpages? Do the user idiosyncrasies and the interface idiosyncrasies affect such correlations?

R2: Which cross-modal visual features are *comparable* with each other based on the free-viewing user attention allocation on bi-modal webpages? Do the user idiosyncrasies and the interface idiosyncrasies affect such comparisons?

R3: Can the text visual features *delineate* the free-viewing user attention on image visual features and vice-versa for the bi-modal webpages? Do the user idiosyncrasies and interface idiosyncrasies affect such delineations?

Concerning Roth's five task taxonomy [127], research question R1 indicates the objective primitive of *identify*. The research question R2 is directed towards answering the *compare*, *rank*, and *associate* objective primitives. Research question R3 is an insight eliciting research question whose results are based on overall observations of answers from R1 and R2; accordingly, it is under the *delineate* objective primitive of Roth's taxonomy.

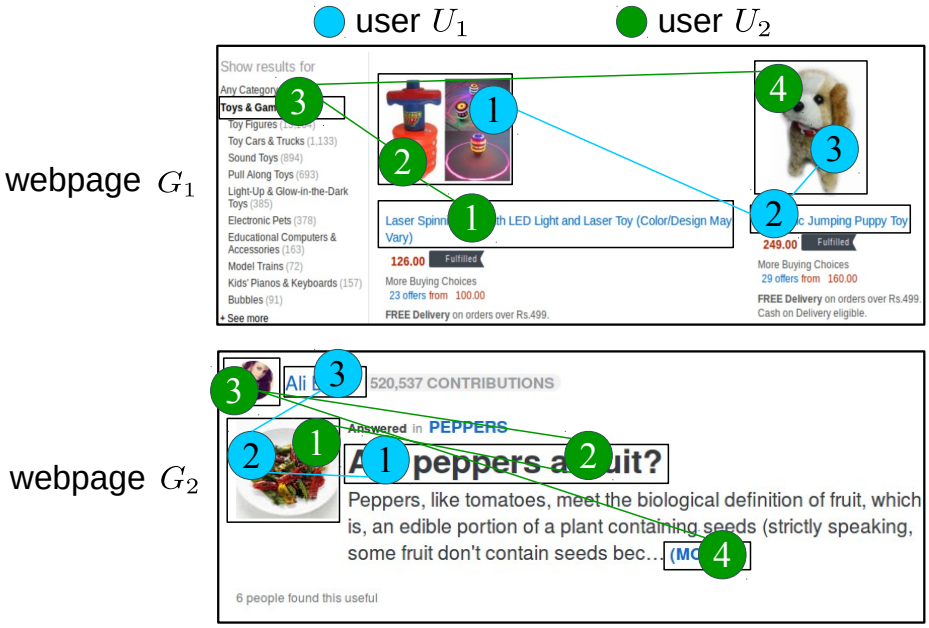


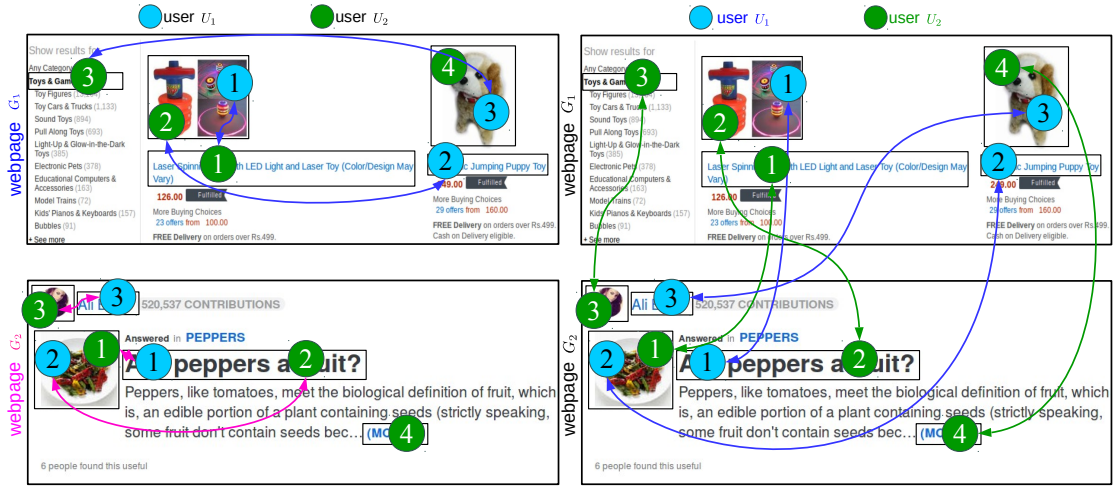
Figure 5.2: Example webpages (G_1 and G_2) and users (U_1 and U_2) considered for illustrating our proposed grouping criteria

5.1 Terminology

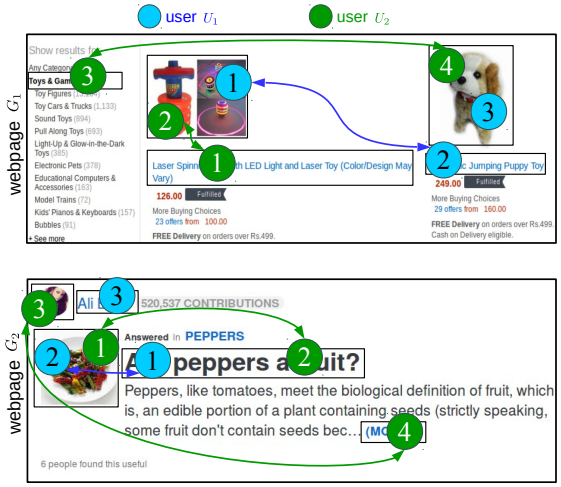
Consider, an interface consists of heterogeneous text and image elements. Each fixated element during the gaze session, aka data-of-interest (DOI), is associated with *data features, visual features, user features*, and *perceptual features*. The data features describe the identity of the DOI through its *modality* and constituting *interface*. The visual features (described in Section 3.5) represent the characteristics of the rendered elements. The user features describe the identity of the *user*. Accordingly, the perceptual features describe the fixation-related characteristics including *fixation-position*, *fixation-index*— the ordinal number of a fixation in a scanpath, and the *modality-fixation-index*— the modality-specific relative fixation-index on the DOI. To note, though the fixation-index is from standard terminology, we introduced modality-fixation-index to capture the relative prominence of modalities within a scanpath. For example, from Figure 5.2, the text DOI on the webpage G_1 with fixation-index numbered 3 from user U_2 received a modality-fixation-index of two as it is the second text-modality element to be fixated by U_2 on G_1. Similarly, the image DOI element on webpage G_1 with fixation-index numbered 2 from user U_2 received a modality-fixation-index of one as it is the first image-modality element to be fixated by U_2 on G_1.

5.2 Views of Attention and Grouping

Multiple text and image DOIs may associate with the same fixation-index, indicating the same attention drawing ability. Accordingly, the elements represent two views (one



(a) Webpage-oriented (WG) (b) User-oriented (UG)



(c) Webpage-and-User-oriented (WUG)

Figure 5.3: Pairings from each grouping for the attention shown in Figure 5.2.

corresponding to text and another corresponding to the image) of the same perceptual characteristic (analogous to views of the semantic concept in cross-modal information retrieval [32]). However, the perception is affected by two prominent factors namely, user idiosyncrasies and interface idiosyncrasies. Thus, we introduce three attention grouping strategies to constrain and marginalize their influence on user perception. Subsequently, text and image DOIs with a *similar* attention drawing ability are paired towards unification.

Webpage-oriented Grouping (WG)

The interface's idiosyncrasies influence the users' attention allocation [16, 164]. For example, users' attention allocation on text-rich webpages differ from image-rich webpages [110]. Consequently, we propose to constrain the influence of interface idiosyncrasies while marginalizing

the user idiosyncrasies. That is, *all users*' attention allocation on *each interface* is considered a group. Within the group, each text DOI is paired with every image DOI *sharing the same fixation-index*. The example pairs achieved using the WG strategy are shown in Figure 5.3a. On webpage G_1, the text element with FI=1 (achieved from user U_2) is paired to the image element with FI=1 (achieved from user U_1) as both the elements were able to draw the same level of user attention (FI=1) on the given webpage, G_1. Similarly, text elements with FI=2 and 3 were respectively paired to the image elements with FI=2 and 3. The same approach applied on the G_2 resulted in the pairs as shown in Figure 5.3a. To note, the work by [154] represents the WG based pairing.

User-oriented Grouping (UG)

Analogous to the interface, the user idiosyncrasies also determine the attention allocation on interfaces [23, 72, 108]. Consequently, we propose to constrain the influence of user idiosyncrasies while marginalizing the interface idiosyncrasies. *Each user*'s attention allocation on *all interfaces* is considered a group. Within the group, each text DOI is paired with every image DOI sharing the *same fixation-index*. The example pairs achieved using the UG strategy are shown in Figure 5.3b. For user U_1, the text element with FI=1 (achieved on webpage G_2) is paired to the image element with FI=1 (achieved on webpage G_1) as both the elements were able to draw the same level of user attention (FI=1) for the given user, U_1. Similarly, text elements with FI=2 and 3 were respectively paired to the image elements with FI=2 and 3. The same approach was applied for the user U_2 whose resulting pairs are shown in Figure 5.3b.

Webpage-and-User-oriented Grouping (WUG)

To account for both interface idiosyncrasies and user idiosyncrasies, we propose to constrain them simultaneously. That is, *each user*'s attention allocation on *each interface*, i.e. scanpath, is considered as a group. However, unlike WG and UG, no two DOIs will be associated with the same fixation-index within a group. Thus, we utilize the modality-fixation-index for pairing the text and image DOI. Within the group, pair the first fixated text with the first fixated image; second fixated text with the second fixated image; and so on. The example pairs achieved using the WUG strategy are shown in Figure 5.3c. For user U_1 attention on webpage G_1, the text element with FI=2 (modality-FI=1) is paired to the image element with FI=1 (modality-FI=1) as both the elements were quickest to draw the user attention among respective modality-specific web elements. Similarly, for user U_1 attention on webpage G_2, text element with FI=1 (modality-FI=1) was paired to the image elements with FI=2 (modality-FI=1). The remaining pairs obtained through the same approach were shown in Figure 5.3c.

5.3 Proposed DOI based Approach

5.3.1 (R1) Attentional Correlation Computation

Without loss of generality, consider $T \in \mathbb{R}^{m \times d}$ and $I \in \mathbb{R}^{n \times d}$ represent the visual features of the paired text and image DOI. That is, for n pairs, each text DOI with m features and each image DOI with n features represent a point in the *visual text space* and *visual image space* respectively, as shown in Figure 5.1b.

Towards answering the research question *R1*, consider, linear transformation of T along the direction $\mathbf{w}_t \in \mathbb{R}^m$ and linear transformation of I along the direction $\mathbf{w}_i \in \mathbb{R}^n$ maximizes the correlation (ρ) between text and image features as formulated below,

$$\rho_{\max} = \underset{\mathbf{w}_t, \mathbf{w}_i}{\operatorname{argmax}} \frac{\mathbf{w}_t^\top C_{ti} \mathbf{w}_i}{\sqrt{\mathbf{w}_t^\top C_{tt} \mathbf{w}_t \times \mathbf{w}_i^\top C_{ii} \mathbf{w}_i}} \tag{5.1}$$

where $C_{tt} = E[\widetilde{T}\widetilde{T}^\top]$, $C_{ti} = E[\widetilde{T}\widetilde{I}^\top]$ and $C_{ii} = E[\widetilde{I}\widetilde{I}^\top]$ are covariance matrices of paired data with $E[\cdot]$ being the expectation function and $\widetilde{T}$ and $\widetilde{I}$ being the mean centered text and image visual features, respectively.

Equation 5.1 is transformed to a constrained optimization problem of CCA [64, 68] and solved through a Generalized Eigenvalue Equation formulation [150]. The resulting $\mathbf{w}_t$ and $\mathbf{w}_i$ pairs, called *canonical variates* or *canonical directions*, are the Eigenvectors of $C_{tt}^{-1/2} C_{ti} C_{ii}^{-1} C_{it} C_{tt}^{-1/2}$ and $C_{ii}^{-1/2} C_{it} C_{tt}^{-1} C_{ti} C_{ii}^{-1/2}$, respectively. The obtained canonical directions (minimum of m and n) result in a space called *homogeneous space* or *maximally correlated space* (see Figure 5.1c). Accordingly, the obtained correlations help answer the research question **R1**.

5.3.2 (R2) Comparison of Visual Features

Even though the computed canonical directions help in establishing the correlation between text and images, the prominent cross-modal feature comparison (text features with image features) is not straightforward (answer to research question **R2**). The predominant reason is, $\mathbf{w}_t$ and $\mathbf{w}_i$ are obtained from the linear combination of features from respective visual spaces. Consequently, the identities of the individual visual features are concealed in the composite directions $\mathbf{w}_t$ and $\mathbf{w}_i$. To extract the prominence of individual features, we utilize the correlation-based approach [56]. That is, correlation between each of the visual feature and the corresponding canonical direction, called *canonical loading* [36, 56], is computed.

Considering each column of T and I represent a visual feature, i.e., $T = [T_{f_1} | T_{f_2} | \ldots | T_m]$

and $I = [I_{f_1}|I_{f_2}|\dots|I_n]$, the canonical loading is computed as

$$cor(T, \mathbf{w}_t^\top T) = E[\widetilde{T}(\mathbf{w}_t^\top \widetilde{T})^\top] = E[\widetilde{T}\widetilde{T}^\top]\mathbf{w}_t = C_{tt}\mathbf{w}_t \tag{5.2}$$

$$\text{similarly} \quad cor(I, \mathbf{w}_i^\top I) = E[\widetilde{I}(\mathbf{w}_i^\top \widetilde{I})^\top] = E[\widetilde{I}\widetilde{I}^\top]\mathbf{w}_i = C_{ii}\mathbf{w}_i \tag{5.3}$$

The canonical loading of each feature indicates the amount of its associated information (precisely, square root of variance) captured by the corresponding canonical variate. Thus, the text feature with the highest canonical loading is comparable to the image feature with the highest canonical loading when the canonical variates $\mathbf{w}_t$ and $\mathbf{w}_i$ are highly correlated (i.e. $\rho_{\max}$ is larger). To further account for the user idiosyncrasies and interface idiosyncrasies, the cross-modal feature comparison is repeated for the three grouping criteria, WG, UG, and WUG.

5.3.3 (R3) Cross-modal Delineation

The research question **R3** centers on the ability of one modality visual features (say, text) to represent the attention on another modality (say, image). In other words, we need to investigate if the text data and image data preserve their associated characteristics even after projected into the homogeneous space?

Considering the homogeneous space is constructed with h $(= \min(m, n))$ canonical directions, the data can be projected into the homogeneous space.

$$\text{Text projected into the homogeneous space } T^{'} = W_t^\top \times T \tag{5.4}$$

$$\text{Image projected into the homogeneous space } I^{'} = W_i^\top \times I \tag{5.5}$$

where $W_t = [\mathbf{w}_{t1}|\mathbf{w}_{t2}|\dots|\mathbf{w}_{th}]$ and $W_i = [\mathbf{w}_{i1}|\mathbf{w}_{i2}|\dots|\mathbf{w}_{ih}]$. That is, each column in W_t and W_i is a canonical direction horizontally stacked in the order of decreasing canonical correlations from left to right.

To answer **R3**, data characteristics in the original space and the projected space are inspected. That is, data characteristics of T are compared with $T^{'}$ and I are compared with $I^{'}$. For the quantitative comparison we employ the *Spearman's rho test* that measures the preservation of pair-wise distances between the data-points in both of the spaces [59, 139].

5.4 Results and Analysis

This section presents the empirical results for the three research questions. Among the considered features, the redundant constant features (variance=0) and the linearly dependent features are pruned. The resulting features (indicated as text rank and image rank in Table 5.1) are utilized for further analysis.

Table 5.1: Summary of empirical data, significant canonical directions (at the significance level $p = 0.01$), and achieved maximum correlation for three groupings

Parameter	**WG**	**UG**	**WUG**
No. of (Text, Image) pairings	14330	9057	3283
Text Features	71	71	77
Image Features	91	91	91
Text rank	52	52	59
Image rank	65	65	83
Significant directions ($p < 0.01$)	28	1	30
Maximum correlation ($\rho_{\max}$)	0.9948	0.3124	0.9946

5.4.1 Correlations between Text and Image DOIs

The application of the proposed approach (see subsection 5.3.1) resulted in the correlations between visualized text and image features as shown in Figure 5.4. The WG, WUG, and UG groups achieved the maximum correlation ($\rho_{\max}$) of 0.9948, 0.9946, and 0.3124 respectively. The number of correlation directions are equal to the minimum rank which is the rank of text data for three groupings. The second highest correlation for the three groupings is 0.9881, 0.9367, and 0.1876 respectively. A similar decreasing trend is followed for the correlations along remaining canonical directions. Though the highest $\rho_{\max}$ of WUG is comparable to WG, for the remaining directions, WUG outperformed the WG and UG with respect to the magnitude of correlation achieved between text and image features.

To investigate whether the achieved canonical correlations and the respective directions are significant or spurious, we employed Wilk's lambda test [160]. The test computes the amount of variance accounted by the canonical directions from that of the original feature variables. The higher the variance accounted the better is the significance of the canonical direction, with the ideal proportion of accounted variance being 1. The canonical variates that account for lower variance are rejected at the significance level of $p = 0.01$. The test resulted in 28, 30, and 1 significant canonical directions, respectively for WG, WUG, and UG as reported in Table 5.1.

Influence of User and Interface Idiosyncrasies The user-oriented grouping achieved a significantly lower $\rho_{\max}$ of 0.3124 and its value further reduced in each iteration of correlation computation. Though the UG grouping constrained the user idiosyncrasies (to note, the participants were young adults and all are experienced web users), still, the correlational association is not captured. In contrary, constraining the interface idiosyncrasies along with user idiosyncrasies (WUG grouping) achieved the correlation between text and images. Further, the sole constrainment of interface idiosyncrasies also achieved the high correlation between both of the modalities. The high correlation obtained for WG and WUG indicates that free-viewing attention drawing **text and image visual features are correlated when the interface idiosyncrasies are constrained** (answer to research question **R1**).

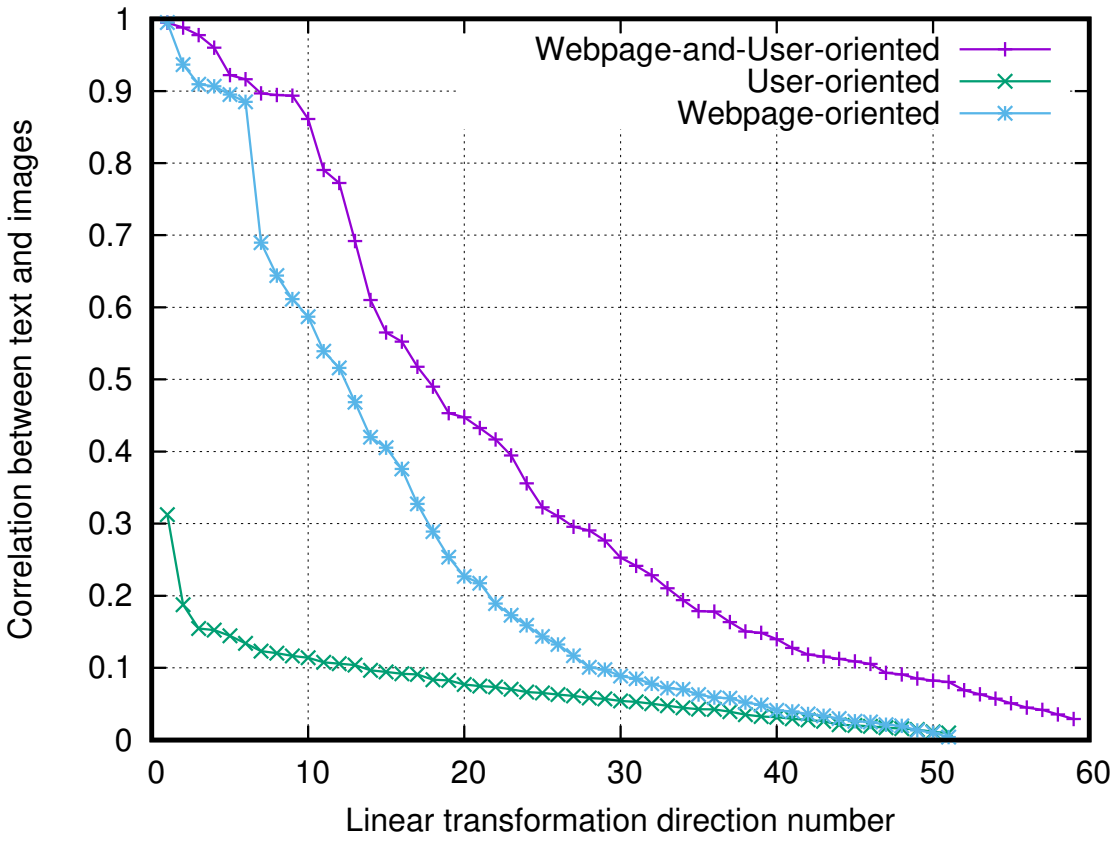


Figure 5.4: Correlations for the three grouping criteria

Additionally, the user idiosyncrasies have little influence on such a constrained correlation which is further supported by the lower correlation obtained for UG.

5.4.2 Comparative Text and Image Features

Among the three criteria, we utilized the high correlation achieving WG and WUG for cross-modal visual attribute comparison. The approach described in subsection 5.3.2 is utilized to obtain the comparative text and image features with respect to the free-viewing attention. Though, in total, the number of comparisons achieved is equal to the minimum feature count of text and image, the majority of them results in very low canonical loading. Consequently, we limit to compare the Top-3 visual features based on associated canonical loadings.

The *font-family* ('open sans') of the text is *comparable* with the color histogram (mid-level blue color component) of the image along the most prominent canonical direction for WG grouping as shown in Figure 5.5. The text feature achieved a canonical loading of 1.00 and the image feature achieved a canonical loading of 0.46. To note, in the histogram feature denoted as 'hist_B_6_182', 'B' (from R, G, B, gray) represents the blue color, '6' (from 1 to 8) represents the histogram bin number, and the number (182) denotes a

representative color component value ($\in [0, 255]$) for the respective bin. Further, the *font-family* ('arial') of text is comparable to the color histogram of the image (mid-level blue color histogram). In aggregation, the FAMLIY of the text is comparable to the HISTOGRAM of images in describing the free-viewing user attention on interfaces. This comparison is further strengthened by the very high squared $\rho^2_{\max} = 0.9896$ along the first canonical direction. With respect to the Top-2 canonical loadings (with $\rho^2_{\max} > 0.8$), the FAMILY, COLOR, SPACE features of text are as influential as HISTOGRAM, CONTRAST HIST, and POSITION features of the images in establishing the correlation between both the modalities.

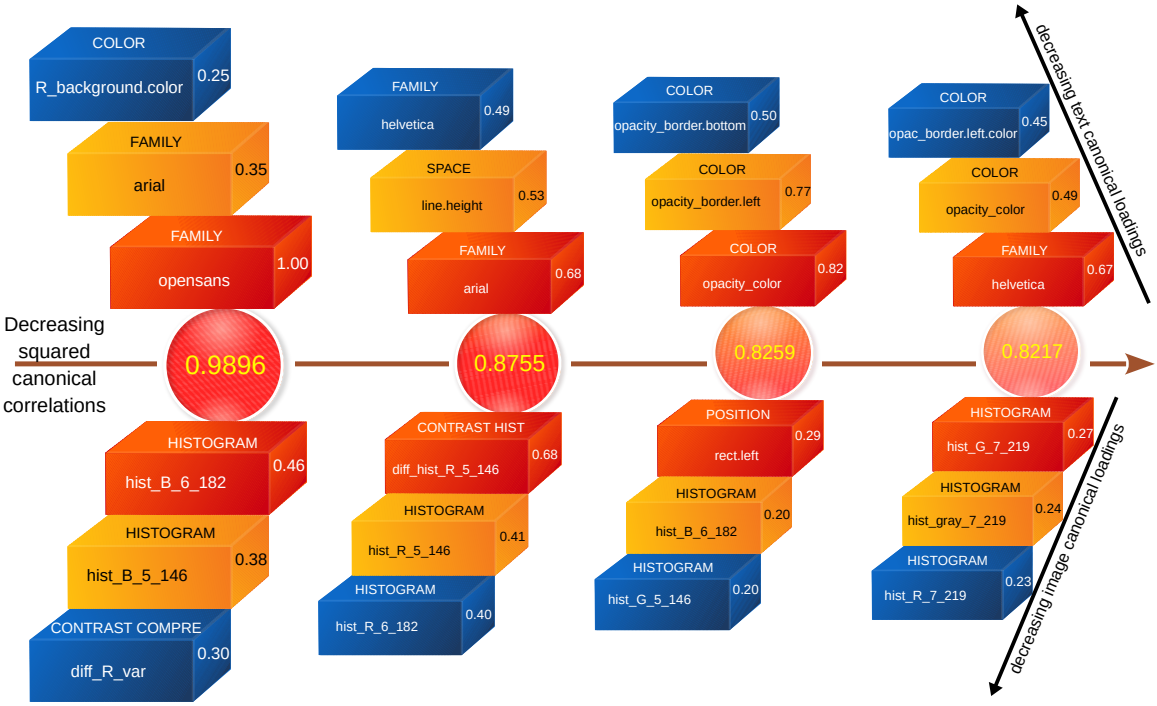


Figure 5.5: Interpretative visualization of WG canonical loadings. The decreasing intensity and size of the ball (moving from left to right) indicates the decreasing prominence (inscribed with $\rho^2_{\max}$) of the corresponding canonical direction. At each position, text feature cuboids are placed above and image feature cuboids are placed below the ball in the decreasing order of canonical loadings. Each cuboid consists of feature name (on front), feature group (on top), and corresponding canonical loading (on side).

As shown in Figure 5.6 for WUG, along the most prominent canonical direction, the FAMILY feature achieved the highest canonical loading of 1.00 while the CONTRAST COMPRE feature achieved 0.21. That is, a significant difference exists between both the canonical loading scores. On the other hand, along the second most canonical direction (with $\rho^2_{\max} = 0.9758$), the FAMILY group achieved comparable canonical loadings with that of the image's HISTOGRAM features. Similarly, FAMILY and HISTOGRAM features are comparable in achieving the correlation maximization along the third canonical direction. In contrast, FAMILY features are comparable to the CONTRAST HIST features in achieving

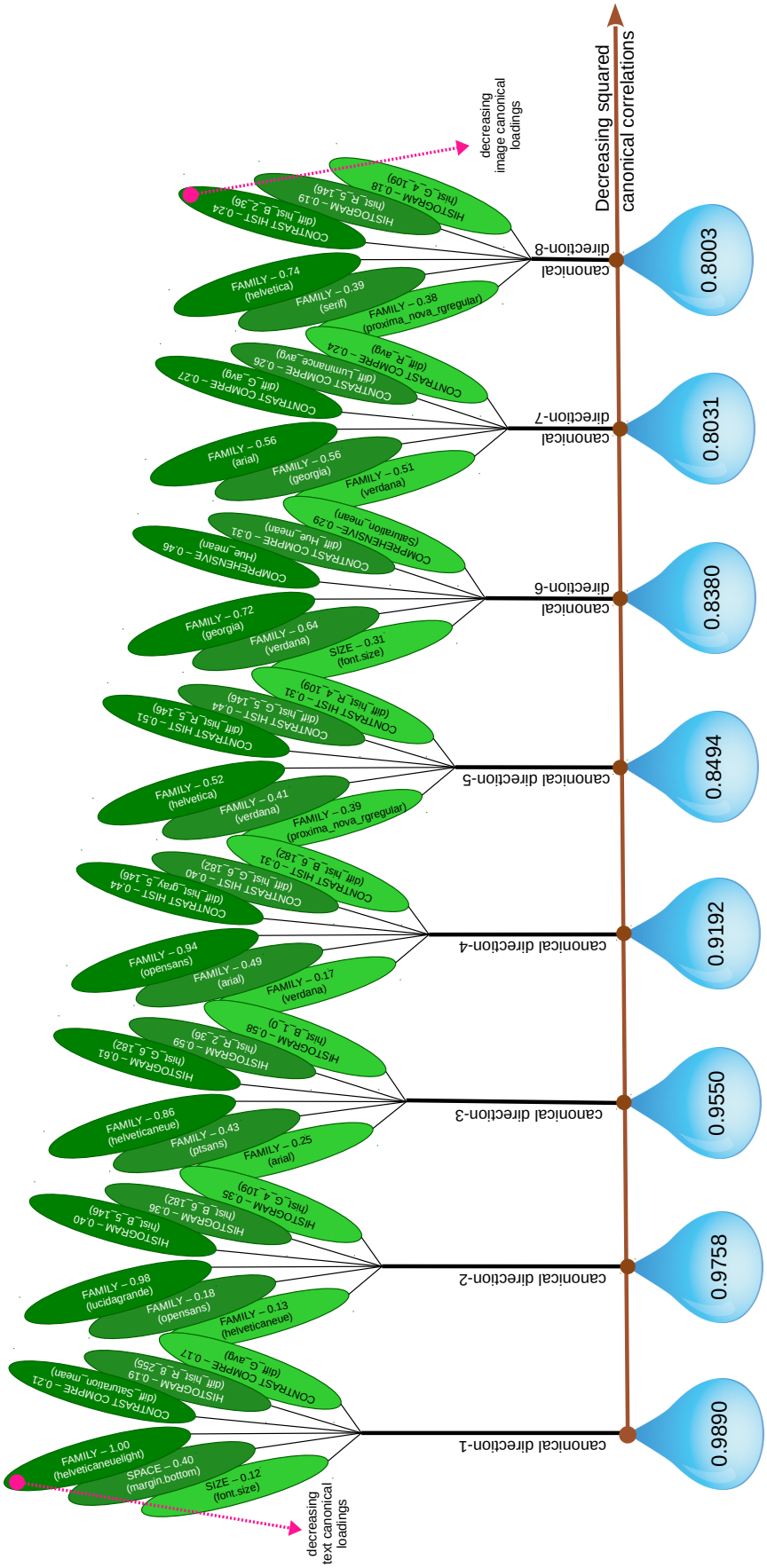


Figure 5.6: Interpretative visualization of WUG canonical loadings. The decreasing tree height (from left to right) indicates the decreasing prominence (indicated with $\rho^2_{\max}$ in the water balloon at the bottom) of the corresponding canonical direction. For each tree, text feature leaves are placed on the left side and image feature leaves are placed on the right side of the tree trunk from top to bottom in the decreasing order of canonical loadings. Each leaf consists of feature name, feature group, and corresponding canonical loading.

the fourth and fifth canonical direction. The latter feature comparison is of little interest as the corresponding $\rho^2_{\max}$ values as well as the canonical loadings are reducing. Overall, FAMILY features of text dominated the canonical loadings across the canonical directions with $\rho^2_{\max} > 0.8$ and are comparable to the attention associated with the HISTOGRAM (followed by CONTRAST HIST) features, as visualized in Figure 5.6.

In summary, from the interface idiosyncrasies constrained WG and WUG groupings, **FAMILY features of text are comparable to and as influential as HISTOGRAM and CONTRAST HIST features of images in explaining the free-viewing user attention** (answer to research question **R2**). The low canonical correlation obtained for UG grouping indicates that the proposed method cannot enable the comparison of text and image features for the marginalization of user idiosyncrasies.

5.4.3 Cross-Delineation of Attention

The squared canonical correlation explains the amount of variance along a text canonical direction from that of the corresponding image canonical direction, and vice-versa [56]. For example, the first canonical direction of WG explains the 0.9896 cross-modal variance and the first canonical direction of WUG explains 0.9890 cross-modal variance. Thus, the higher the squared canonical correlation, the higher is the delineation association between both of the modalities. The very low $\rho^2_{\max} = 0.0976$ of UG indicates that text and images cannot delineate the user attention on each other in the absence of interfaces' idiosyncratic marginalization. In contrast, the very high squared canonical correlations achieved for WG and WUG prefer the cross-modal delineation between both the modalities.

The squared structural correlation (or squared canonical loading) explains the amount of original features' variance captured by the respective canonical direction [56]. For example, the second canonical direction of WG (see Figure 5.5) captured $0.68 \times 0.68 = 0.46$ variance of a text FAMILY feature and a CONTRAST HIST image feature. The canonical loadings visualized in Figure 5.5 and Figure 5.6 highlight the prominence of obtained canonical directions in preserving the variance associated with the original features. The delineation is further supported by the significance verified by the Wilk's Lambda test.

To further verify the cross-modal delineation and the preservation of visual characteristics, the approach presented in subsection 5.3.3 was applied. The distance of each data point was computed from every other data point in the original space (text space for text data point and image space for the image data point) and the respective homogeneous spaces (constructed from the significant canonical directions). The sample distribution of the distances for the three groupings are shown in Figure 5.7. The Spearman's Rho (ρ_s) correlation was computed between the distances obtained in the original space and the homogeneous space. The ρ_s for the three grouping criteria are reported in Figure 5.8. For the WG and WUG criteria, our approach reflects a *very strong* quality in preserving the local neighborhood for both text and images with ρ_s values greater than 0.94 at $p < 0.01$. Between the two modalities,

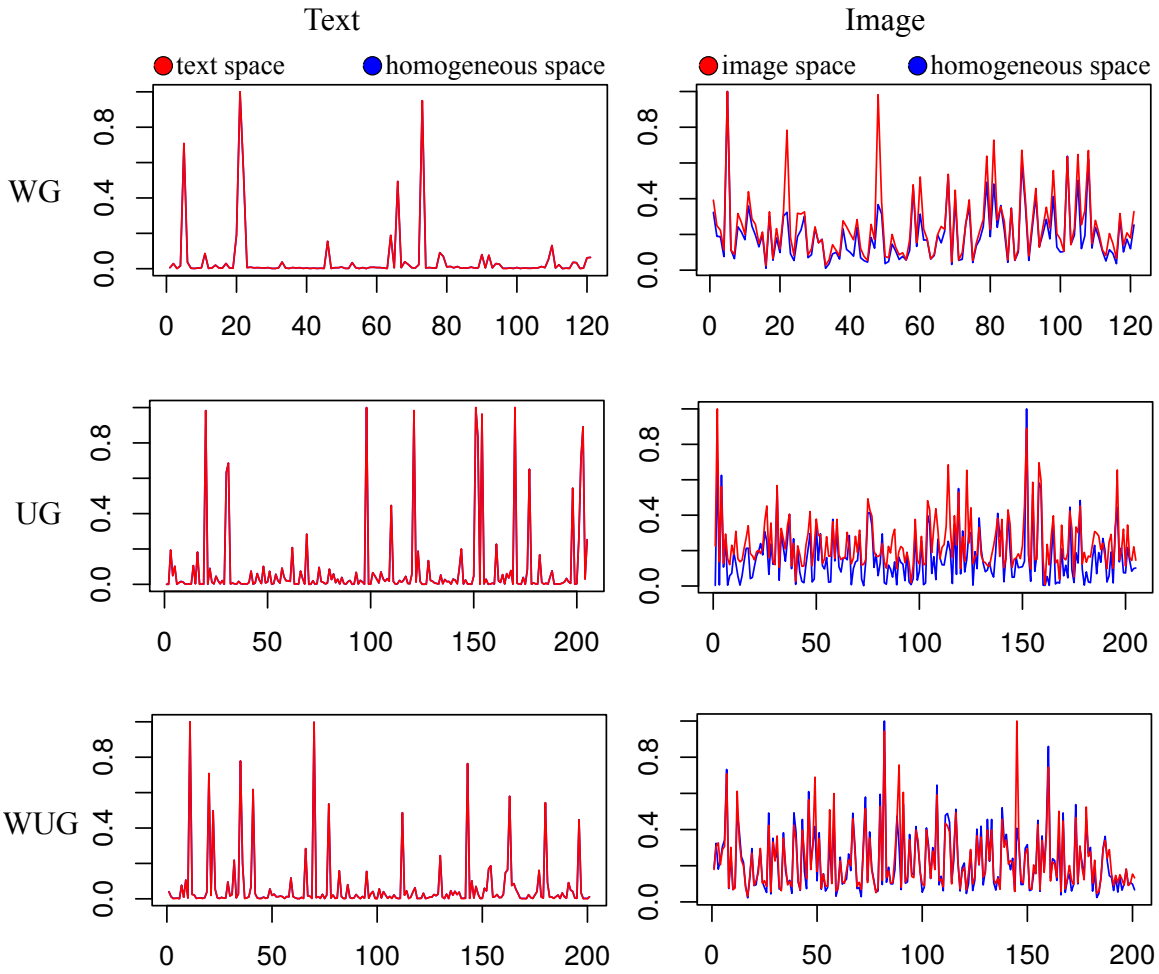


Figure 5.7: Distance of each DOI element from rest of the DOI elements in the original space and the homogeneous space for three groupings: The horizontal axis is the index of the DOI element and the vertical axis is the normalized distance (actual distance/maximum distance).

the neighborhood is relatively better preserved for the text than for the images. This is attributed to the relatively larger variation in the image visual features when compared with the text features. In user-oriented grouping, the local neighborhood preservation is *very strong* for the text and *strong* for the images. However, the Spearman's Rho differed significantly between text and images where such a difference is marginal in WG and WUG groupings. The variation in ρ_s for UG is attributed to the low homogeneity established between text and images as indicated by the respective canonical correlation. Further, the UG grouping preserves the local neighborhood along a single (significant canonical) direction, whereas WG and WUG preserve along 28 and 30 significant canonical directions, respectively. Overall, the corresponding distances as visualized in Figure 5.7 highlight that, for both text and images, distances are better preserved in WUG and WG groupings than in the UG grouping. In summary, **text and images delineate the attention on each other**

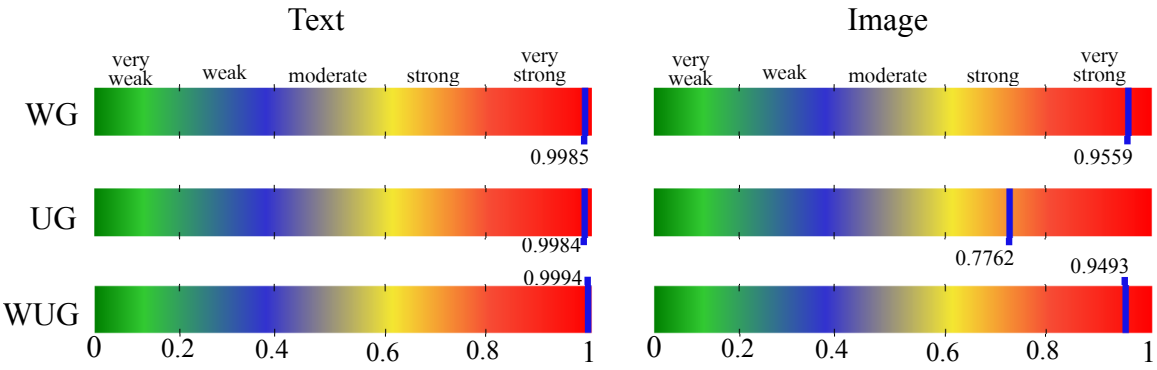


Figure 5.8: Spearman's Rho (ρ_s) for three groupings

when the interface idiosyncrasies are constrained (answer to research question **R3**).

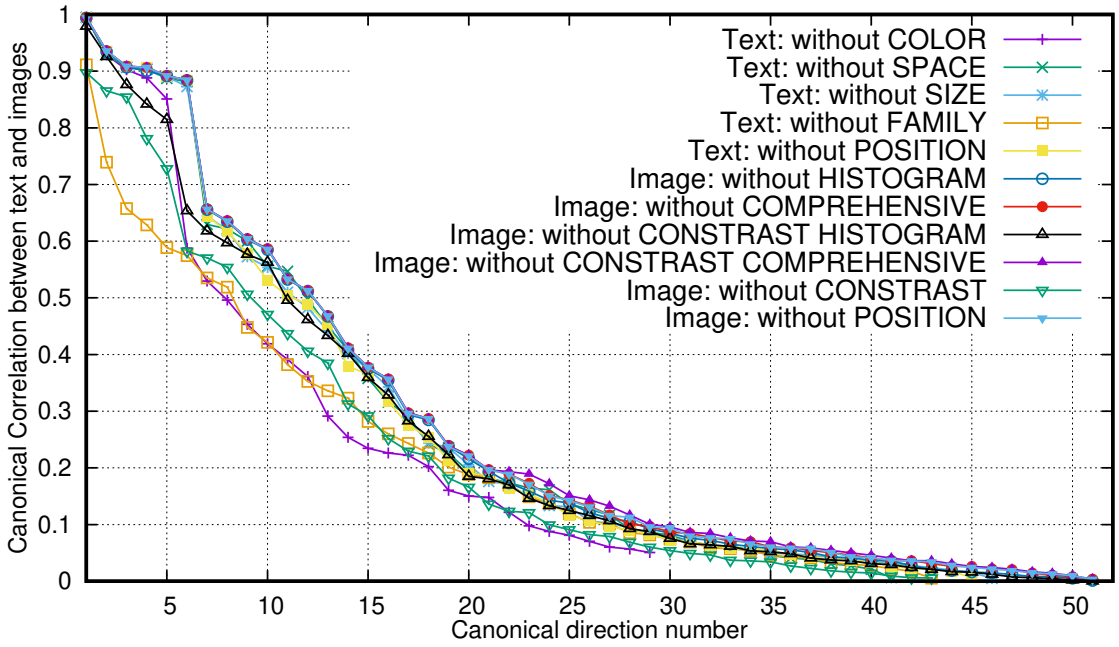
5.5 Prominent Feature Groups

To determine the prominent feature groups in unifying the text and image modalities for WG and WUG, we employ the robust *leave-one-out* strategy. That is, correlation computation procedure is repeated after excluding one feature group at a time.

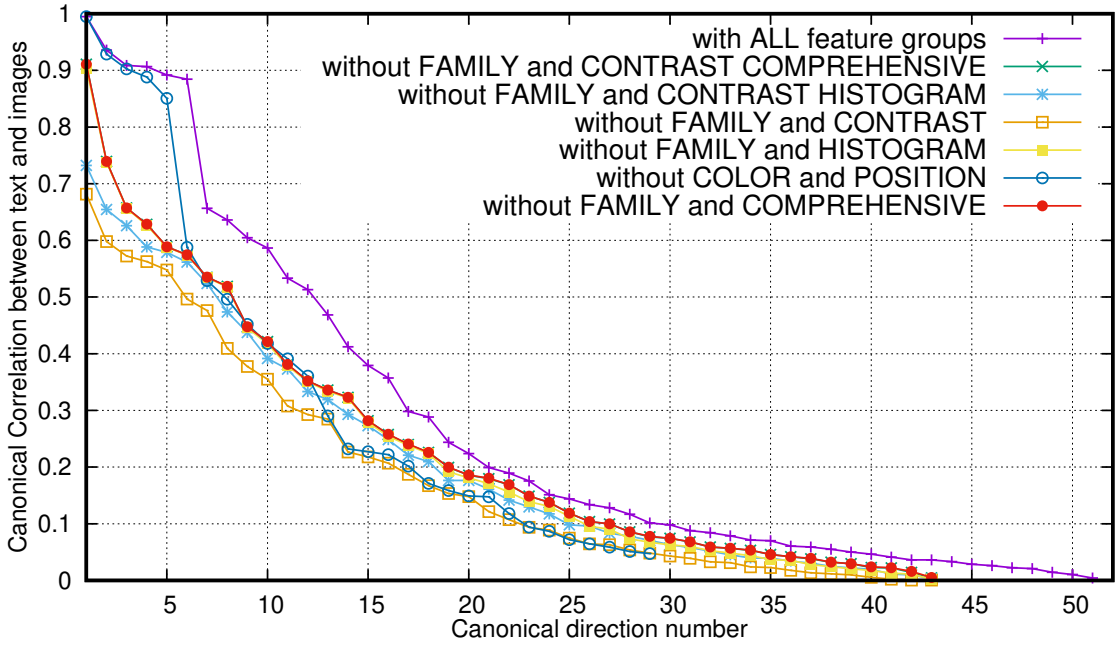
5.5.1 Webpage-oriented Grouping

The canonical correlation achieved through the *leave-one-out* strategy applied WG is shown in Figure 5.9a. The POSITION group (consists of element's position and size) which was earlier found to be predominant in drawing the user attention [61, 143] was not found to significantly influence the correspondence between text and images. The feature group's exclusion achieved a $\rho_{\max}$ of 0.9948 for both the modalities. The individual exclusion of text's SPACE and SIZE (font-size related features) groups, and image's COMPREHENSIVE, CONTRAST COMPREHENSIVE and HISTOGRAM feature groups did not significantly affect the canonical correlation. Though the exclusion of text COLOR features did not influence the initial canonical correlation, the later correlational correspondence was relatively affected by their absence (see Figure 5.9a). Between the two CONTRAST groups, the absence of CONTRAST HISTOGRAM reduced more correlation compared with the CONTRAST COMPREHENSIVE features.

The notable feature groups towards the text and image canonical correlations are the FAMILY of text and CONTRAST (combination of CONTRAST HISTOGRAM and CONTRAST COMPREHENSIVE) of images. The exclusion of FAMILY group resulted in the $\rho_{\max}$ of 0.9111 which further reduced progressively along the increasing canonical direction number. The CONTRAST group's exclusion resulted in $\rho_{\max}$ of 0.8972 which gradually reduced along other canonical directions. The text elements' font-family and image elements' contrast demonstrated the prominence in correlation establishment which were also shown to be



(a) Leave-one-out each feature group



(b) Leaving-out combination of feature groups, one from each modality

Figure 5.9: Canonical correlation between text and images for WG

prominent for the modality-specific controlled analyses [9, 99].

Owing to the predominance of FAMILY feature group, its combinatory correlational influence is studied with other image feature groups as shown in Figure 5.9b. The prominence of image CONTRAST features is further established as the exclusion of FAMILY and CONTRAST

features resulted in $\rho_{\max}$ of 0.6814 which further reduced along the subsequent canonical directions. Analogous to the leave-one-out strategy, the exclusion of COMPREHENSIVE, HISTOGRAM, and CONTRAST COMPREHENSIVE features along with the FAMILY had little influence on the correlation. Between the two CONTRAST groups, the absence of CONTRAST HISTOGRAM significantly reduced the $\rho_{\max}$ (to 0.7324) when compared with the absence of CONTRAST COMPREHENSIVE feature group ($\rho_{\max} = 0.9119$).
In summary, the FAMILY features of text and the CONTRAST (especially, the CONTRAST HISTOGRAM) features of image are influential in achieving the explicit correlation between both the modalities.

5.5.2 Webpage-and-User-oriented Grouping

The achieved correspondence between text and images as the unification correlation is shown in Figure 5.10.

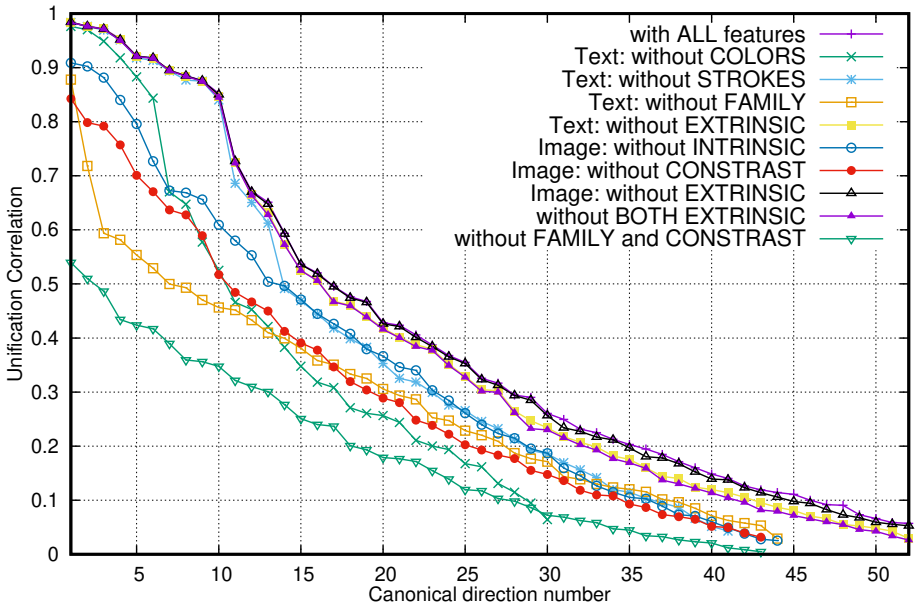


Figure 5.10: Unification correlation between images and text

The most influential features in establishing the cross-modal correlation are FAMILY features of text and CONTRAST features of image. The exclusion of the former feature-group resulted in correlations of 0.8779, 0.7182, 0.5939, and so on, while the exclusion of later feature-group achieved 0.8426, 0.7985, 0.7920, and so on correlations. The combined exclusion of both the feature groups further deteriorated the correlation to 0.5395, 0.5098, 0.4864, and so on.
The INTRINSIC (ensemble of HISTOGRAM and COMPREHENSIVE) visual features of image and the COLOR attributes of text are the second-most influential features in their respective modalities towards the unification. The quality along initial canonical directions was better determined by the INTRINSIC image features while the text COLOR

attributes dominate the determination of later canonical directions. The non-influential text's STROKES features indicate the prominence of other features over *font-size* and *spacing* associated with text elements.
The prominent *position* and element's *size* [61, 143] constituting EXTRINSIC group did not find to influence the cross-modal correlation. Even the exclusion of the feature-group from both the modalities resulted in 0.9838, 0.9763, 0.9712, and so on correlations (see Figure 5.10) indicating their limited prominence in fusing the text and images.

5.6 Summary

In this chapter, a computational approach was presented to correlated and unify the text and image visual features according to the user attention allocation. The empirical analysis revealed that text and image visual features are correlated when the interface idiosyncrasies are constrained. The *font-family* features of text are comparable to and as influential as HISTOGRAM and CONTRAST HISTOGRAM features of images in achieving the unification between both the modalities. Surprisingly, the prominent state-of-the-art features, including element's *position* and *size* [61, 143] (constituting EXTRINSIC group), did not find to influence the cross-modal correlation. However, the text and images delineate the attention on each other when the interface idiosyncrasies are constrained, that is, for Webpage-oriented Grouping (WG) and Webpage-and-User-oriented Grouping (WUG).

6

Element-based Attention Prediction

In Chapter 5, a computational approach has been presented, and found that Webpage-oriented Grouping (WG) and Webpage-and-User-oriented Grouping (WUG) groupings achieve unification between text and image modalities. In this chapter, the achieved unified representation is leveraged for predicting the user attention on webpages at the granularity of web elements. As opposed to the existing binary-level predictions (salient or not), the presented computational approach predicts a quantitative multi-level (each level is considered as a class) attention of the elements as indicated by the associated Fixation-Index (FI). In summary, we tried to answer the following research questions as poised in Chapter 1.

R1: Can attention on elements be predicted if all the elements are unified into a text modality?

R2: Can attention on elements be predicted if all the elements are unified into an image modality?

R3: How well the achieved unification perform to predict the attention on unseen data for both the above research questions?

6.1 Motivation

We derive the motivation from semantic feature-based *Cross-Modal Information Retrieval (CMIR)* [32, 118, 119] to propose the unified approach for attention prediction. The objective of CMIR is to retrieve the *semantically* equivalent text of a given image and vice-versa. Towards the CMIR, the semantic commonality between text and images is leveraged. As some text elements share the semantics with some image elements (for example, a paragraph describing a 'Tree' and a 'Tree' diagram itself), the cross modalities with shared semantics are paired and a common feature-space is learned using the Canonical Correlation Analysis (CCA) [68]. Towards the pairing, Rasiwasia et al. [118] utilized Wikipedia articles to pair

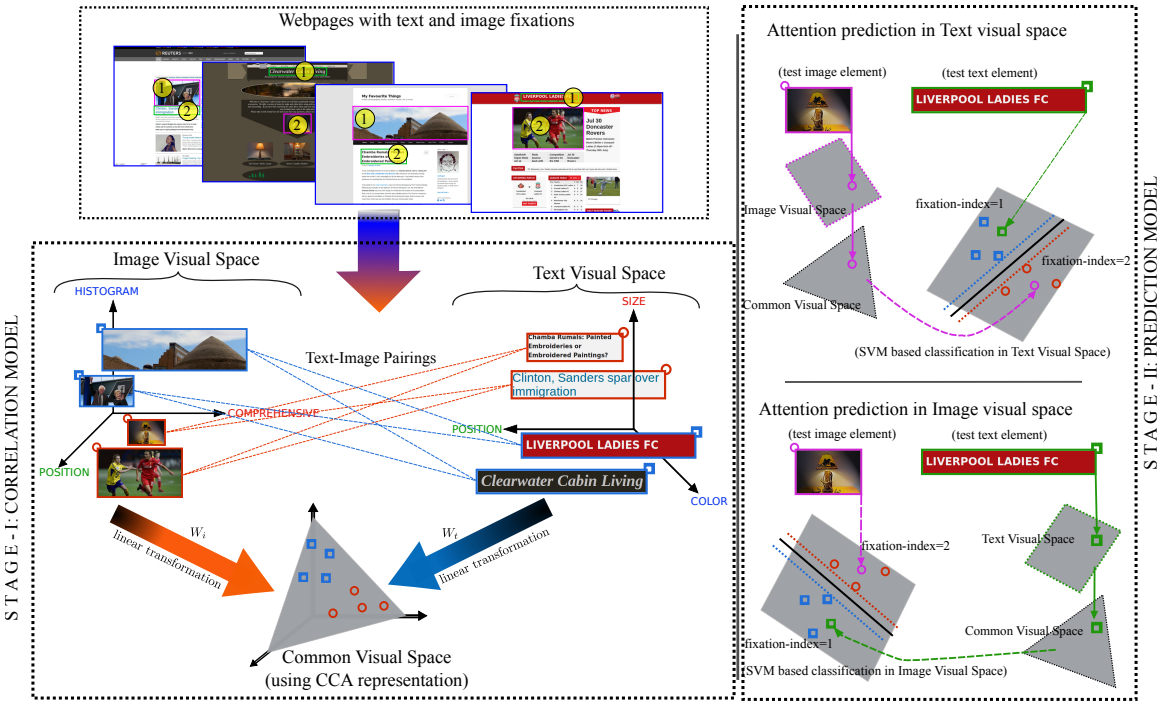


Figure 6.1: WG: Schematic of the proposed web element attention prediction approach. Left) Stage-I: Pair the text and image elements based on shared webpage and fixation-indices, subsequently, learn a maximally correlated Common Visual Space for both the natural visual spaces. Right) Stage-II: Project all cross-modal elements into either of natural visual space through the achieved Common Visual Space and employ SVM based classification for attention prediction.

each text paragraph with an accompanying image in the corresponding section. That is, text and corresponding image are considered as two views of the underlying semantic concept. Subsequently, *similarity* measurement techniques, such as k-nearest neighbor method, are employed in the common feature-space to achieve the cross-modal retrieval.

6.2 Proposed Approach

Text and image elements in tandem influence the user attention allocation on webpages. The multi-level attention prediction may be simplified if the text elements from text visual space are projected into the image visual space, or vice-versa. Unfortunately, no such natural correspondence exists between the respective spaces to achieve the unification. Consequently, we propose a two stage approach to predict the attention on web elements, as shown in Figure 6.1. In Stage-I, unification is achieved between text and images (as described in Chapter 5). In Stage-II, the multi-level attention prediction approach is delineated using the unified approach.

6.2.1 Stage-I: Correlation Model

Consider $T \in \mathbb{R}^{m \times d^{'}}$ and $I \in \mathbb{R}^{n \times d^{'}}$ represents the $d^{'}$ paired text and image elements with m text visual features and n image visual features. Accordingly, each text element and an image element represents a point in *text visual space* $\subseteq \mathbb{R}^m$ and *image visual space* $\subseteq \mathbb{R}^n$ respectively.

Let the linear transformation of text-data along $\mathbf{w}_t \in \mathbb{R}^{m \times 1}$ and image-data along $\mathbf{w}_i \in \mathbb{R}^{n \times 1}$ establishes the correspondence between them. Then, the correlation (ρ) between transformed text ($T\mathbf{w}_t$) and images ($I\mathbf{w}_i$) is,

$$\rho = \frac{\mathbf{w}_t^\top \Sigma_{ti} \mathbf{w}_i}{\sqrt{(\mathbf{w}_t^\top \Sigma_{tt} \mathbf{w}_t)} \sqrt{(\mathbf{w}_i^\top \Sigma_{ii} \mathbf{w}_i)}} \tag{6.1}$$

where, Σ_{tt} and Σ_{ii} are empirical co-variance matrices of text and image data respectively while Σ_{ti} denotes the empirical cross co-variance between text and image data. Now, the correlation is constrained to obtain the linear transformation directions $\mathbf{w}_t$ and $\mathbf{w}_i$ that maximize the correlation C as,

$$\rho_{max} = \max_{\mathbf{w}_t \neq 0, \mathbf{w}_i \neq 0} \frac{\mathbf{w}_t^\top \Sigma_{ti} \mathbf{w}_i}{\sqrt{(\mathbf{w}_t^\top \Sigma_{tt} \mathbf{w}_t)} \sqrt{(\mathbf{w}_i^\top \Sigma_{ii} \mathbf{w}_i)}} \tag{6.2}$$

where, ρ_{max} denotes the maximum correlation between text and images, accordingly, an indicator of quality of correspondence between text and images. The ideal value of ρ_{max} is 1.

Evidently, ρ_{max} in (Equation 6.2) is unperturbed by the scaling of $\mathbf{w}_t$ and $\mathbf{w}_i$ which results in numerous solutions corresponding to the scaled versions. To achieve the unique solution, the denominator is constrained to unity through the following reformulation [68].

$$\max_{\mathbf{w}_t, \mathbf{w}_i} \mathbf{w}_t^\top \Sigma_{ti} \mathbf{w}_i \tag{6.3}$$

$$\text{subject to} \qquad \mathbf{w}_t^\top \Sigma_{tt} \mathbf{w}_t = 1 \quad \text{and} \quad \mathbf{w}_i^\top \Sigma_{ii} \mathbf{w}_i = 1 \tag{6.4}$$

Subsequently, the constrained optimization problem is transformed into the unconstrained problem through the application of Lagrangian, L

$$\begin{aligned} L(\mathbf{w}_t, \mathbf{w}_i, \lambda_t, \lambda_i) = \mathbf{w}_t^\top \Sigma_{ti} \mathbf{w}_i + \\ \lambda_t(1 - \mathbf{w}_t^\top \Sigma_{tt} \mathbf{w}_t) + \lambda_i(1 - \mathbf{w}_i^\top \Sigma_{ii} \mathbf{w}_i) \end{aligned} \tag{6.5}$$

Partial derivatives are applied on L with respect to $\mathbf{w}_t$ and $\mathbf{w}_i$ towards the computation of

Algorithm 2: Multi-level attention prediction

Input:
test web element E_t,
saliency threshold θ,
binary classifiers $C_{i,j}$; for all $i = 1, \ldots, \theta - 1$; and $j = i + 1, \ldots, \theta$
Output: Predicted multi-level attention f_t

```
for k ← 1 to θ do                                  // iterate over each possible class
    v_k ← 0                                        // initialize the vote-count of kth class
end
for k ← 1 to θ − 1 do
    for l ← k + 1 to θ do
        j ← C_{k,l}(I_t)                           // binary-classify for k or l
        v_j ← v_j + 1                              // increment respective class vote
    end
end
f_t ← argmax_k(v_k)                                // assign the maximum voted class
return f_t
```

maximum correlation achieving directions.

$$\frac{\partial L}{\partial \mathbf{w}_t} = \Sigma_{ti}\mathbf{w}_i - 2\lambda_t \Sigma_{tt}\mathbf{w}_t = 0 \tag{6.6}$$

$$\frac{\partial L}{\partial \mathbf{w}_i} = \Sigma_{ti}^{\top}\mathbf{w}_t - 2\lambda_i \Sigma_{ii}\mathbf{w}_i = 0 \tag{6.7}$$

The above two equations are reformulated as below to indicate the formulation of generalized eigenvalue problem [91]

$$\begin{bmatrix} 0 & \Sigma_{ti} \\ \Sigma_{it} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{w}_i \end{bmatrix} = \lambda \begin{bmatrix} \Sigma_{tt} & 0 \\ 0 & \Sigma_{ii} \end{bmatrix} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{w}_i \end{bmatrix};\ \lambda = 2\lambda_t = 2\lambda_i \tag{6.8}$$

The iterative solution of the generalized eigen value problem indicates that the $\mathbf{w}_i$s are the eigen vectors of $\Sigma_{ii}^{\frac{-1}{2}}\Sigma_{it}\Sigma_{tt}^{-1}\Sigma_{ti}\Sigma_{ii}^{\frac{-1}{2}}$ while the eigen vectors of $\Sigma_{tt}^{\frac{-1}{2}}\Sigma_{ti}\Sigma_{ii}^{-1}\Sigma_{it}\Sigma_{tt}^{\frac{-1}{2}}$ gives the $\mathbf{w}_t$s [32]. In total, $h = \min(m, n)$ directions are possible for each modality. During the iterative process, the first pair of $\mathbf{w}_t$ and $\mathbf{w}_i$ transformation directions (also called *canonical directions*) achieves the highest correlation. Each of the subsequent iterations results in the decreasing correlations and the $\mathbf{w}_t$ and $\mathbf{w}_i$ pairs which are orthogonal to the existing transformation directions. The subspace constructed from the obtained linear transformation directions is called *Common Visual Space* $\subseteq \mathbb{R}^h$ in which both the modalities achieve unified representation as shown in Figure 6.1. Further, the elements of one modality can be transformed into the visual space of other modality through the Common Visual Space.

6.2.2 Stage-II: Prediction Model

Consider T and I represent the text and image elements (represented using the visual features) for which the multi-level attention needs to be predicted. Let, a total of h (minimum of m and n) linear transformation directions are obtained from the aforementioned correlation model. Arranging h directions into the transformation matrices respectively for text and images as $W_t = [\mathbf{w}_{t1}|\mathbf{w}_{t2}|\ldots|\mathbf{w}_{th}]_{m\times h}$, $W_i = [\mathbf{w}_{i1}|\mathbf{w}_{i2}|\ldots|\mathbf{w}_{ih}]_{n\times h}$.

$$\text{projected text in the image space } T^{'} = T \times W_t \times W_i^{+} \tag{6.9}$$

$$\text{projected images in the text space } I^{'} = I \times W_i \times W_t^{+} \tag{6.10}$$

where, W_i^{+} and W_t^{+} indicates the Moore-Penrose inverse or psuedo inverse of the W_i and W_t respectively.

The unified elements (ensemble of original feature space elements and the projected elements) in the text space are represented as $\left[\frac{T}{I'}\right]_{d\times m}$, and analogously, unified elements in the image space as $\left[\frac{T'}{I}\right]_{d\times n}$; where, d represents the number of unified elements. Considering ith row of either of the matrices as an element E_i, the associated multi-level attention is indicated as f_i.

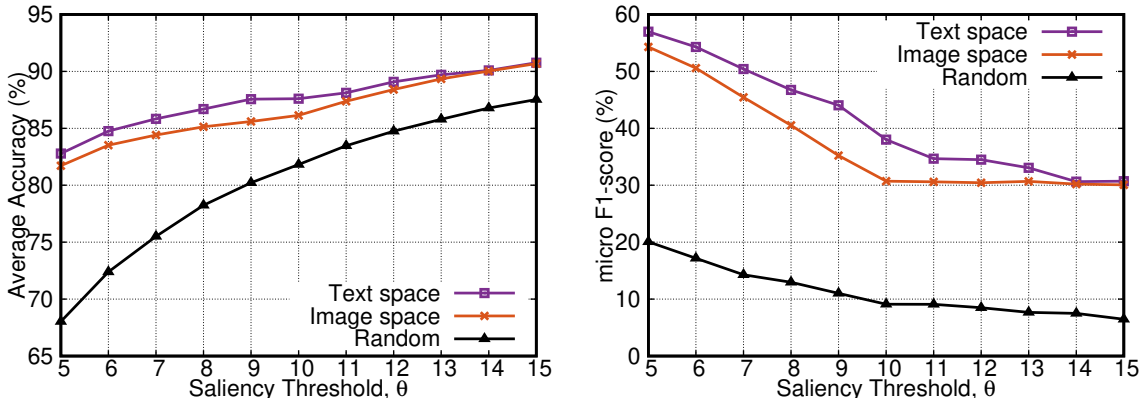
However, not all the levels of user attention is equally important. Especially, the later attention may be influenced by the factors apart from visual features, such as semantic features or exhaust in the salient elements, etc. Thus, we introduce thresholding on multi-levels to segregate and combine the non-prominent attention levels (the later fixation-indices). Analogous to the fixation-index, the increasing attention-level value indicates decreasing prominence. Thus, thresholding the maximum attention-level to θ (called *saliency-threshold*), results in θ classes each corresponding to an attention-level, i.e., $f_k \in \{1,\ldots,\theta\}$; $k = 1,\ldots,d$. Accordingly, the multi-level attention prediction on web elements transforms to a polychotomous classification problem with θ classes.

We utilize the multiclass support vector machine (multiclass SVM) based approach to solve this polychotomous prediction problem, as described in subsection 2.6.1.

We analyze the performance of the proposed approach on two real-world webpage datasets.

6.3 Prediction Performance using WG

W_t and W_i Selection: The inclusion of all features achieved the correlation of 0.9948, 0.9357, 0.9088, 0.9065, and so on respectively along first, second, third, and so on canonical directions. In total, 51 canonical directions and corresponding correlations are obtained as shown in Figure 5.9b. However, all correlations need not be significant, particularly, those correspond to the later canonical directions. The application of Wilks' Lamda [160] test identified 28 significant canonical directions at the significance level of 0.01 which are utilized in the multi-level attention prediction.



(a) Average Accuracy (b) Micro-F1 score

Figure 6.2: Experiment-I: Prediction performance metrics with variation in θ

6.3.1 Prediction Performance on Same Dataset

For the analysis in this section, the fixation-data from Experiment-I (described in Section 3.2) is used in both the stages.

Ground-truth Preparation: For each fixated web element, the fixation-indices from all users are obtained. The fixation-index resulting from the application *majority-voting-scheme* is assigned as the ground-truth multi-level attention. In case of conflict (multiple majority-voted fixation-indices), the fixation-index with the lower value is assigned as an indicator of element's attention drawing ability.

θ selection: The maximum of median fixation-indices is considered as the saliency-threshold in Equation 4.3. The median fixation-index of text and images (from Experiment-II) are 5 and 4 respectively. Accordingly, $\theta = 5$ is considered for the prediction performance analysis.

Procedure: The image elements are projected from Image Visual Space into the Text Visual Space (via the Common Visual Space) using Equation 6.10 where the significant canonical directions are utilized for W_t and W_i. On the unified data, the procedure described in subsection 6.2.2 is applied with a 5-fold cross-validation with 10 iterations. The micro prediction performance metrics (average accuracy and micro-F1 score) computed in the each iteration are averaged to obtain an overall prediction performance. Further, the whole procedure is repeated by projecting the text elements from Text Visual Space into the Image Visual Space via the Common Visual Space.

Baseline Selection: A very few of the existing approaches are centered on element-granular attention prediction and that to multi-level attention prediction. The location-based

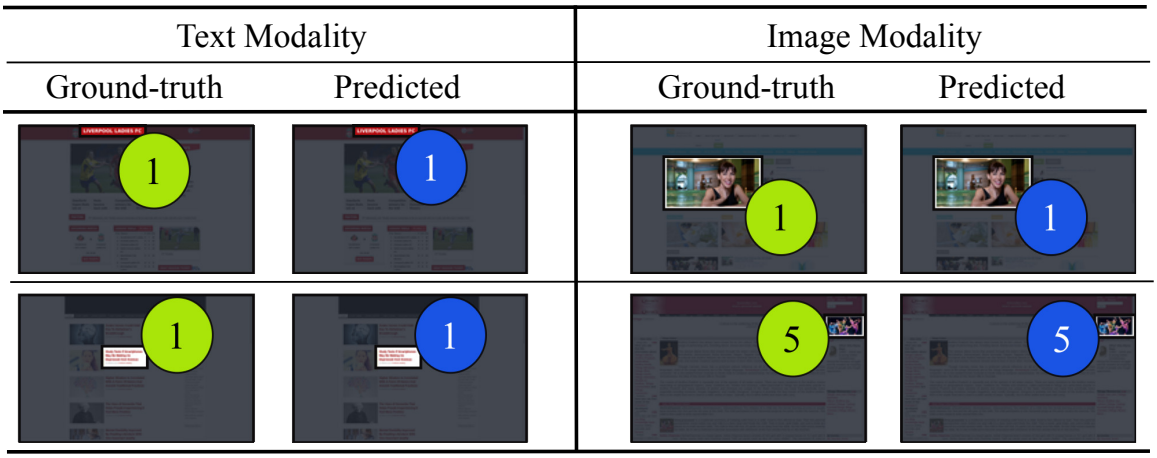


Figure 6.3: Experiment-I: Example multi-level attention predictions on elements

saliency-oriented approaches (though limited to binary-level predictions) cannot be utilized for the baseline as, two locations corresponding to the same element may indicate one as salient and the other as non-salient. On the other hand, *pattern-oriented* approaches are centered on eliciting an attention-pattern than prediction. Though the work in [19] considered limited visual features of web elements, the attention prediction is proposed for task-dependent settings and constrained to binary-level prediction. Thus, we utilize the random prediction as the baseline (analogous to [32, 118]) to comprehend the performance of the proposed approach.

At median saliency-threshold, the multi-level attention was predicted at an average accuracy of 82.79% and the micro-F1 score of 56.98% in the Text Visual Space. The attention prediction in Image Visual Space also achieved similar performance with the average accuracy of 81.72% and micro-F1 score of 54.29%. Both metrics outperformed the baseline random prediction which achieved an average accuracy of 68.03% and micro-F1 score of 20.07%. To further understand the influence of saliency-threshold, the performance metrics are computed with each unitary increment in θ as shown in Figure 6.2. The average accuracy gradually increased and micro-F1 score gradually reduced with θ towards saturation. Nevertheless, throughout the variation in θ, the approach outperformed the baseline. The contrasting variation in the performance metrics is attributed to the class-imbalance which is better accounted by the micro-metric F1-score. The prediction performance is consistent across the visual spaces. Thus, the research questions **R1** and **R2** are answered when the interface idiosyncrasies are constrained. However, the observed variation in prediction performance is attributed to the utilization of low dimensional (28 significant canonical directions) Common Visual Space for elements' unification where not all the dimensions achieved optimal correlation between text and images. Example predictions from the proposed approach are shown in Figure 6.3.

Table 6.1: Experiment-III: Prediction Performance at Median θ

Category	Metric	Text Space	Image Space	Baseline
Text	Avg. Acc.	86.45	85.23	75.56
	micro-F1	52.56	48.29	14.46
Mixed	Avg. Acc.	87.54	86.61	75.62
	micro-F1	56.40	53.14	14.66
Pictorial	Avg. Acc.	87.73	86.96	75.58
	micro-F1	57.05	54.35	14.52

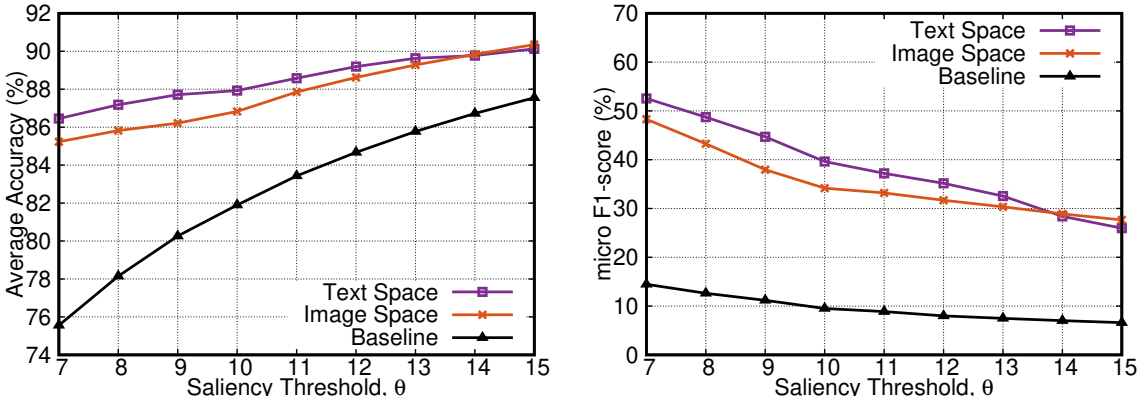
6.3.2 Prediction Performance on Unseen Dataset

For this analysis, we utilize the dataset of Experiment-I for Stage-I construction and dataset extracted from Experiment-III (described in Section 3.4) for Stage-II prediction.

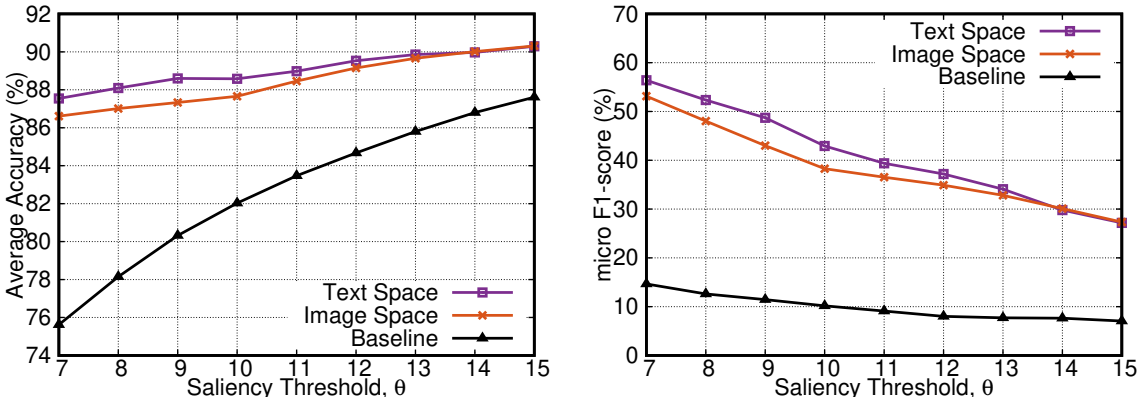
Procedure: The text data and linear transformation directions (W_t and W_i) from previous analysis are considered in combination with each category of image elements. The prediction performance at 5-fold cross-validation with 10 iterations is computed for each category in both the visual spaces. The saliency-threshold (θ) is considered as the maximum of median fixation-indices from three categories which is 7 (from Experiment-III).
The prediction performance at median saliency-thresholding is shown in Table 6.1. The prediction for the three categories outperformed the baseline and are comparable. The Figure 6.4 shows the performance variation with the unitary increments in θ. Analogous to analysis on same dataset, the average accuracy gradually increased, the micro-F1 score gradually decreased towards the saturation. The prediction was not significantly affected due to the utilized visual space. The example multi-level attention predictions are shown in Figure 6.5.
From both the above analyses, the reducing performance with the increasing θ indicates the decreasing prominence of visual features in predicting the user attention. Especially, the later multi-level attention allocations (larger fixation-index values) may be influenced by other factors such as the absence of attention-drawing elements or influence of semantic features, which are not explainable by the visual features alone. Thus, the proposed approach better predicts the user's prominent initial attention, which is guided by the underlying visual features. Thus, the research question **R3** is answered when the interface idiosyncrasies are constrained.

6.4 Prediction Performance using WUG

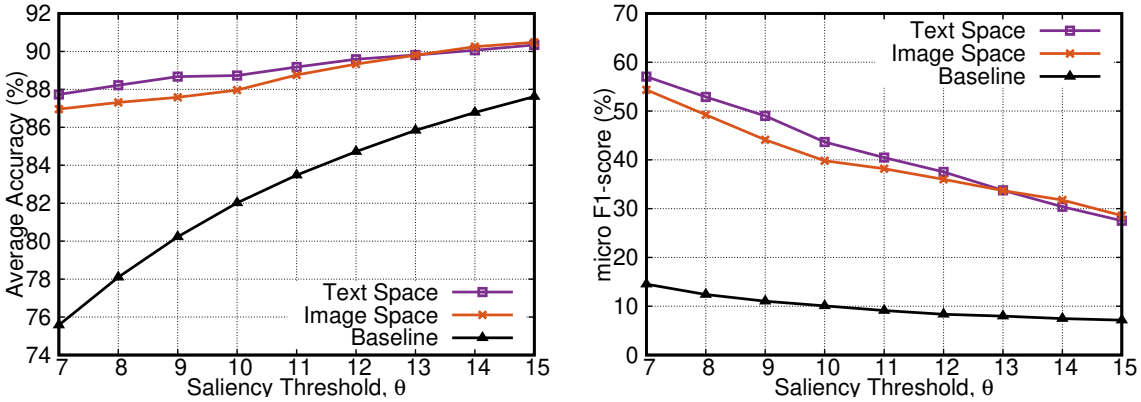
The approach presented in Section 6.3 is adopted for presenting the results using WUG. That is, in Stage-I, WUG grouping is utilized in place of WG grouping as shown in Figure 6.6.



(a) Text Category



(b) Mixed Category



(c) Pictorial Category

Figure 6.4: Experiment-III: Prediction performance metrics with variation in θ

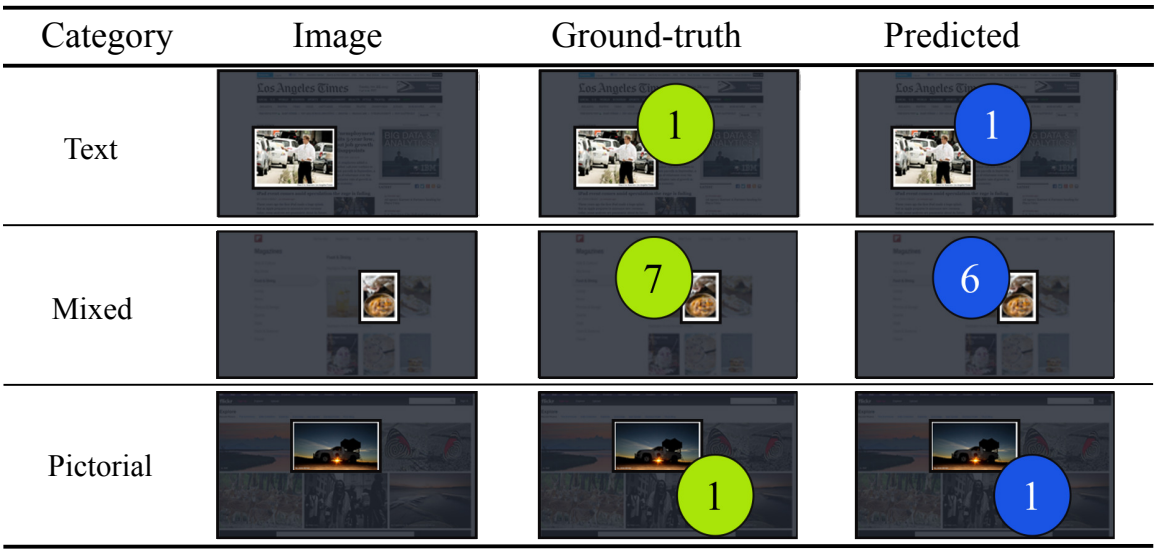


Figure 6.5: Experiment-III: Example multi-level attention predictions

W_t and W_i Selection: The inclusion of all features achieved the correlation of 0.9948, 0.9357, 0.9088, 0.9065, and so on respectively along first, second, third, and so on canonical directions. In total, 51 canonical directions and corresponding correlations are obtained as shown in Figure 5.9b. However, all correlations need not be significant, particularly, those correspond to the later canonical directions. The application of Wilks' Lamda [160] test identified 30 significant canonical directions at the significance level of 0.01 which are utilized in the multi-level attention prediction.

6.4.1 Prediction Performance on Same Dataset

For the analysis in this section, the fixation-data from Experiment-I (described in Section 3.2) is used in both the stages.

Ground-truth Preparation: For each fixated web element, the fixation-indices from all users are obtained. The fixation-index resulting from the application *majority-voting-scheme* is assigned as the ground-truth multi-level attention. In case of conflict (multiple majority-voted fixation-indices), the fixation-index with the lower value is assigned as an indicator of element's attention drawing ability.

θ selection: The maximum of median fixation-indices is considered as the saliency-threshold in Equation 4.3. The median fixation-index of text and images are 5 and 4 respectively. Accordingly, $\theta = 5$ is considered for the prediction performance analysis.

Procedure: The image elements are projected from Image Visual Space into the Text Visual Space (via the Common Visual Space) using Equation 6.10 where the significant

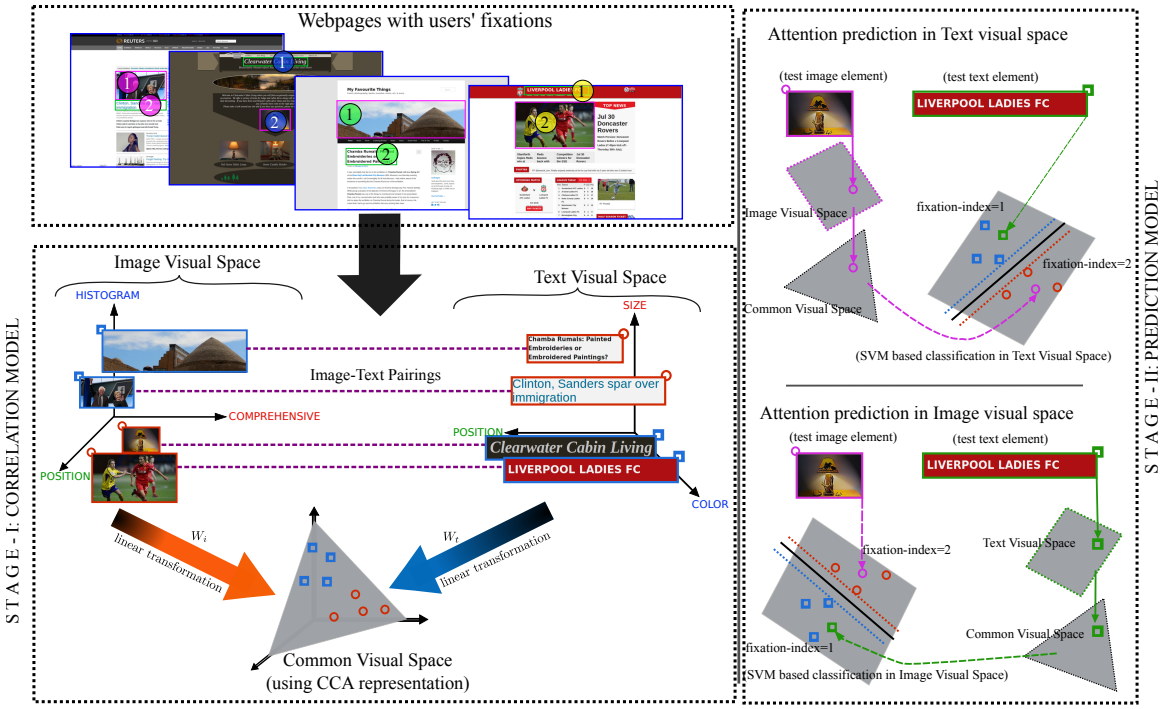
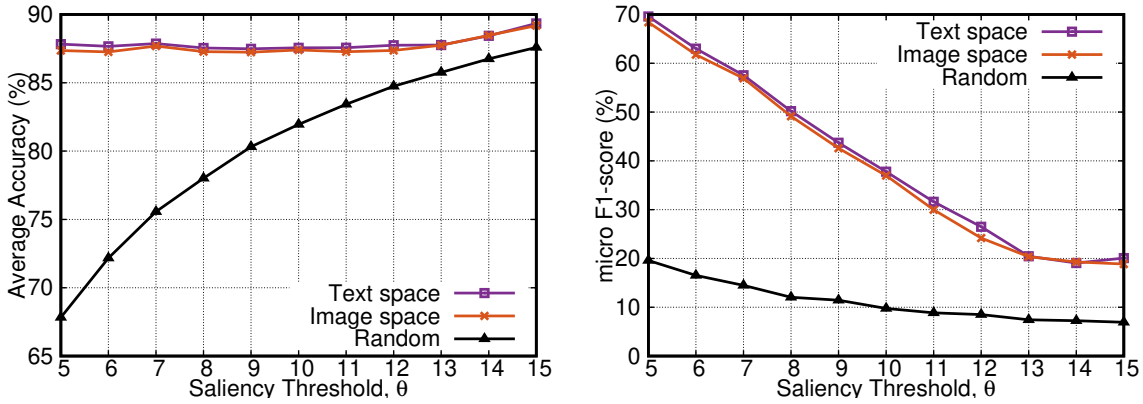


Figure 6.6: WUG: Schematic of the proposed web element attention prediction approach. Left) Stage-I: Pair the text and image elements based on shared webpage and fixation-indices, subsequently, learn a maximally correlated Common Visual Space for both the natural visual spaces. Right) Stage-II: Project all cross-modal elements into either of natural visual space through the achieved Common Visual Space and employ SVM based classification for attention prediction.

canonical directions are utilized for W_t and W_i. On the unified data, the procedure described in subsection 6.2.2 is applied with a 5-fold cross-validation with 10 iterations. The micro prediction performance metrics (average accuracy and micro-F1 score) computed in the each iteration are averaged to obtain an overall prediction performance. Further, the whole procedure is repeated by projecting the text elements from Text Visual Space into the Image Visual Space via the Common Visual Space.

Results: At median saliency-threshold (5, that is, average of text median FI=7 and image median FI=3), the multi-level attention was predicted at an average accuracy of 87.83% and the micro-F1 score of 69.58% in the Text Visual Space. The attention prediction in Image Visual Space also achieved similar performance with the average accuracy of 87.36% and micro-F1 score of 68.4%. Both metrics outperformed the baseline random prediction which achieved an average accuracy of 67.83% and micro-F1 score of 19.58%. To further understand the influence of saliency-threshold, the performance metrics are computed with each unitary increment in θ as shown in Figure 6.7. The average accuracy gradually increased and micro-F1 score gradually reduced with θ towards saturation. Nevertheless, throughout the variation in θ, the approach outperformed the baseline. The contrasting



(a) Average Accuracy (b) Micro-F1 score

Figure 6.7: Experiment-I: Prediction performance metrics with variation in θ

Table 6.2: Experiment-III: Prediction Performance at Median $\theta = 7$

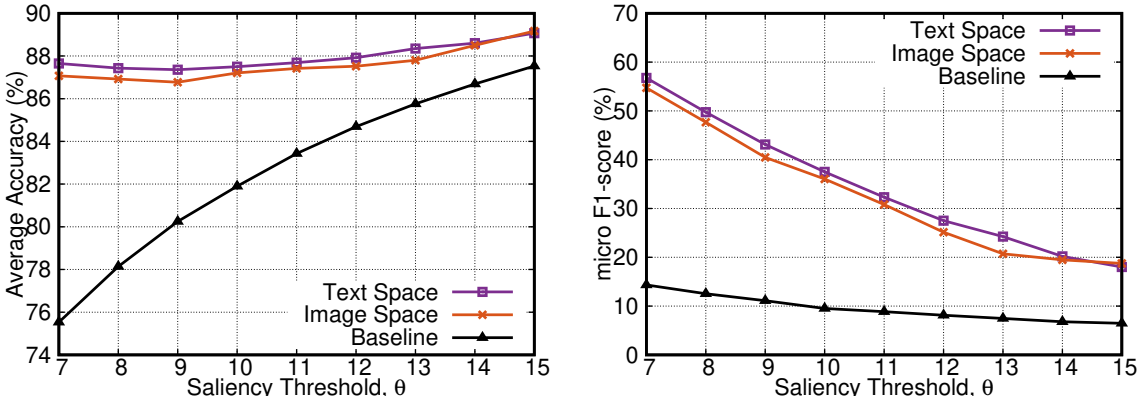
Category	Metric	Text Space	Image Space	Baseline
Text	Avg. Acc.	87.65	87.07	75.53
	micro-F1	56.76	54.73	14.34
Mixed	Avg. Acc.	88.55	87.98	75.59
	micro-F1	59.93	57.92	14.56
Pictorial	Avg. Acc.	88.7	88.53	75.5
	micro-F1	60.45	59.87	14.26

variation in the performance metrics is attributed to the class-imbalance which is better accounted by the micro-metric F1-score. The prediction performance is consistent across the visual spaces. Thus, the research questions **R1** and **R2** are answered when the user idiosyncrasies are constrained along with interface idiosyncrasies. However, the observed variation in prediction performance is attributed to the utilization of low dimensional (30 significant canonical directions) Common Visual Space for elements' unification where not all the dimensions achieved optimal correlation between text and images.

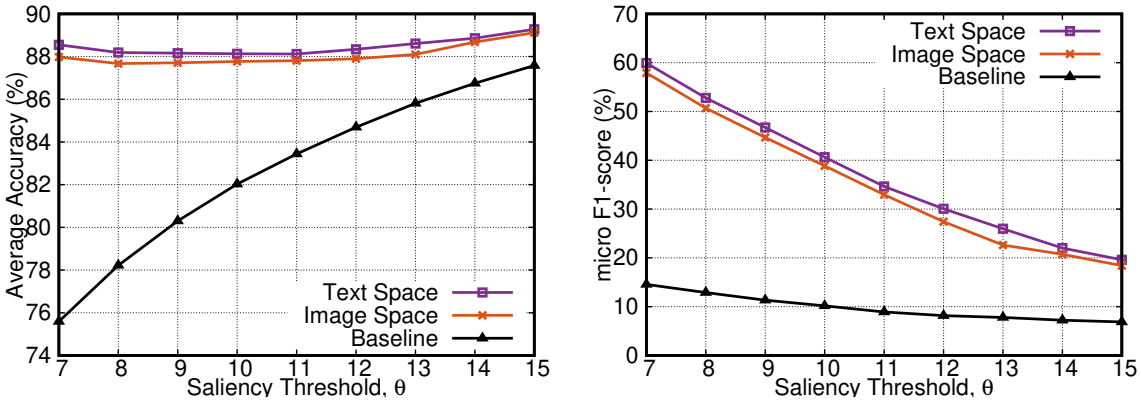
6.4.2 Prediction Performance on Unseen Dataset

For this analysis, we utilize the dataset of Experiment-I for Stage-I construction and dataset extracted from Experiment-III (described in Section 3.4) for Stage-II prediction.

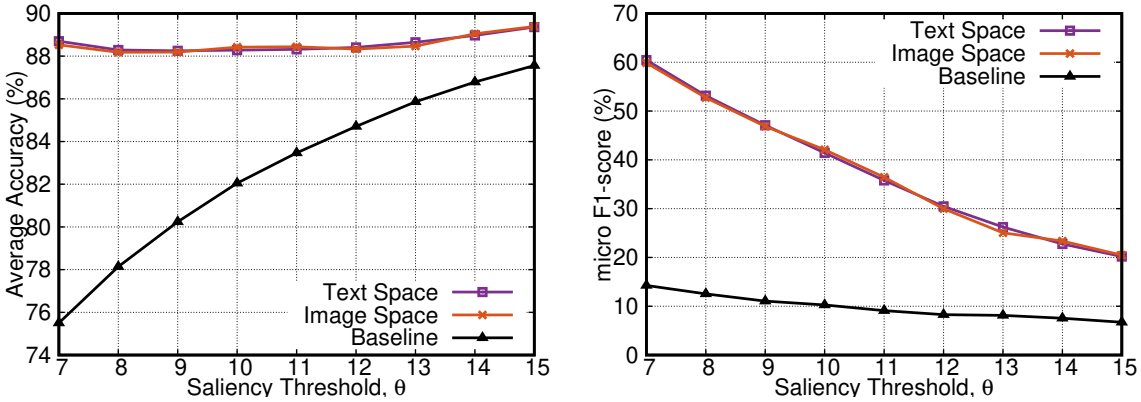
Procedure: The text data and linear transformation directions (W_t and W_i) from previous analysis are considered in combination with each category of image elements. The prediction performance at 5-fold cross-validation with 10 iterations is computed for each category in both the visual spaces. The saliency-threshold (θ) is considered as the maximum of median fixation-indices from three categories which is 7.



(a) Text Category



(b) Mixed Category



(c) Pictorial Category

Figure 6.8: Experiment-III: Prediction performance metrics with variation in θ

The prediction performance at median saliency-thresholding is shown in Table 6.2. The prediction for the three categories outperformed the baseline and are comparable. The Figure 6.8 shows the performance variation with the unitary increments in θ. Analogous to the attention prediction on same dataset, the average accuracy gradually increased, the micro-F1 score gradually decreased towards the saturation. The prediction was not significantly affected due to the utilized visual space.
From both the above analyses, the reducing performance with the increasing θ indicates the decreasing prominence of visual features in predicting the user attention. The later multi-level attention allocations (larger fixation-index values) may be influenced by other factors such as the absence of attention-drawing elements or influence of semantic features, which are not explainable by the visual features alone. Thus, the proposed approach better predicts the user's prominent initial attention, which is guided by the underlying visual features. Thus, the research question **R3** is answered when the user and interface idiosyncrasies are constrained together.

6.5 Summary

The text and image visual features predict the user attention on web elements for the WG as well as WUG groupings. The prediction performance outperforms the baseline through-out the variation in scanpath threshold θ. However, the visual features' predictive performance reduces with increase in θ as other factors such as absence of salient elements guide the latter attention which may not be explained by the considered visual features.

7

Scanpath Prediction on Webpages

In Chapter 6, element-granular attention prediction approach is presented utilizing the unification approach. In this chapter, we propose to extend the element-granular attention into a prominent scanpath prediction. In summary, we answer the following research questions as poised in Chapter 1.

R1: Can scanpath be predicted if all the elements are unified using WG?

R2: Can scanpath be predicted if all the elements are unified using WUG?

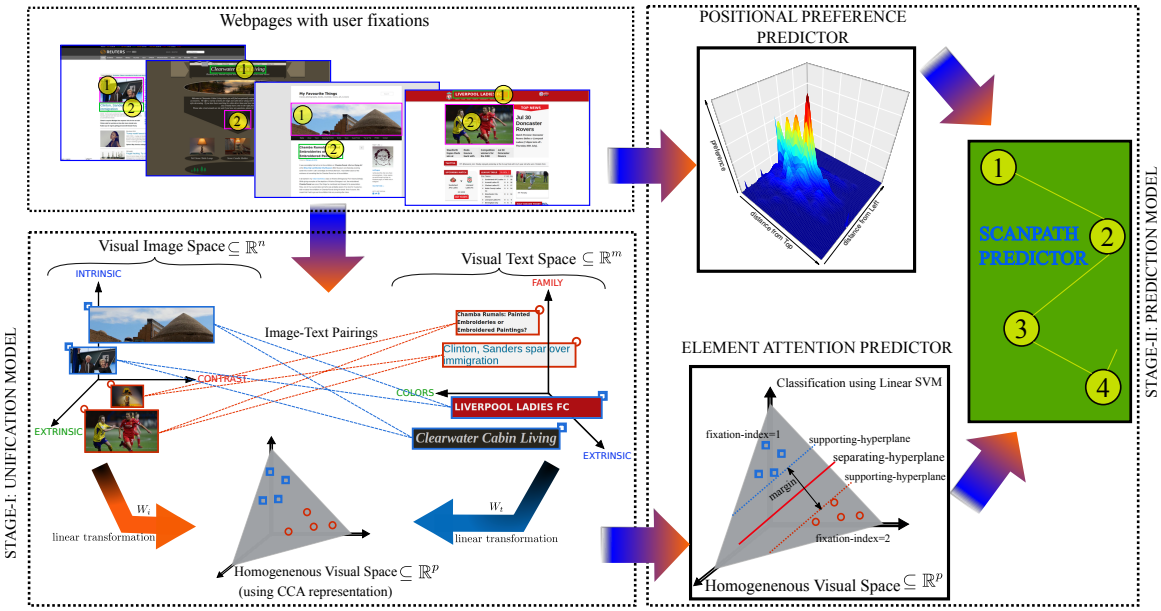
R3: How well the achieved unification perform to predict the scanpath on unseen data for both the above research questions?

7.1 Motivation

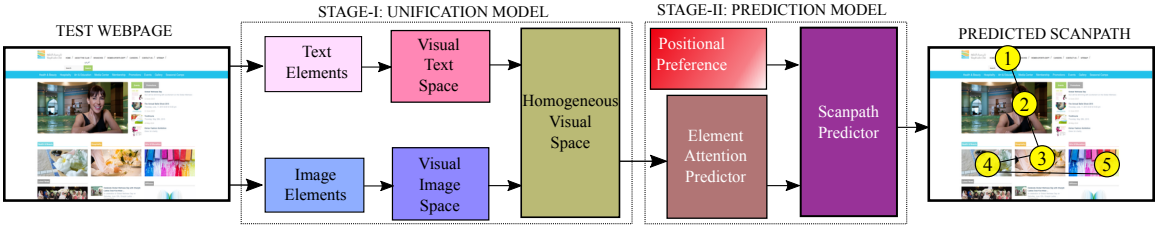
To note, though the *trending* scanpath prediction approaches exist in the literature [46, 47, 49], those approaches expect the users' scanpath data on a given test webpage. That is, the existing approaches are pattern-eliciting than scanpath predictive. Especially, collecting users' fixation data on multiple webpages, as well as multiple times on the same webpage (after every modification during web-designing) is not feasible due to money, time, and user availability constraints. In contrast to the existing approaches, our work determines the prominent scanpath through the selection of most attention drawing elements.

7.2 Prediction Model

The schematic of the proposed approach is shown in Figure 7.1 for WG grouping. Comparing with the two-stage element attention prediction in Chapter 6, the first stage is considered as it is, while the second stage is adapted to the scanpath prediction.
The Prediction model consists of Element Attention Predictor (EAP), Positional Preference Predictor (PPP), and Scanpath Predictor modules. The EAP helps to quantitatively



(a) Model Training



(b) Model Testing

Figure 7.1: WG: Schematic of the proposed approach. UNIFICATION MODEL achieves homogeneous representation for text and images. ELEMENT ATTENTION PREDICTOR utilizes the homogeneous representation to build classifiers for attention prediction on web elements. SCANPATH PREDICTOR arranges the elements into scanpath where the positional-bias captured in POSITIONAL PREFERENCE PREDICTOR helps in resolving the conflicts among elements.

determine the user attention (fixation-index) on web elements while the PPP provides generic positional preference. The Scanpath Predictor module utilizes the results of EAP and PPP to designate the scanpath.

Scanpath-threshold θ: A webpage may contain large number of elements, but only some of those elements prominently draw the user attention. To note, on an average users allocate up to five fixations on a webpage (described later). Thus, to incorporate this user behavior we introduce a *Scanpath-threshold*, θ, that defines the scanpath length (number of elements incorporated into the scanpath). Rest of the section describes the construction of Stage-II modules to predict the scanpath of length θ.

Element Attention Predictor (EAP)

Consider T and I be the overall collection of text and image elements (represented using visual features) fixated by the users. Then, their corresponding unified feature representation is as achieved

$$\text{text projected into homogeneous space } T^{'} = T \times W_t \tag{7.1}$$

$$\text{images projected into homogeneous space } I^{'} = I \times W_i \tag{7.2}$$

Where, the h canonical directions are horizontally stacked into the matrices $W_t = [\mathbf{w}_{t1}|\mathbf{w}_{t2}|\dots|\mathbf{w}_{th}]_{m\times h}$ and $W_i = [\mathbf{w}_{i1}|\mathbf{w}_{i2}|\dots|\mathbf{w}_{ih}]_{n\times h}$.

Accordingly, the ensemble of unified elements in the homogeneous visual space are represented as $E = \begin{bmatrix} T^{'} \\ I^{'} \end{bmatrix}_{d\times p}$. That is, each row of matrix E acts as a point in h dimensional homogeneous space (see EAP module in 7.1a). In addition, each data point is associated with a fixation-index (class-label or attention level). Together we notate every row of E along with the associated fixation-index as a set given as $\{(E_1, f_1), (E_2, f_2), \dots, (E_d, f_d)\}$.

However, not all the fixation-indices are equally important. Especially, the fixation-indices denoting the latter attention allocation may be influenced by factors apart from visual features (such as semantic features or exhaust in the salient elements, etc.) and may be difficult to predict from the visual-features alone. Thus, we utilize the scapth-threshold, θ, to segregate the prominent fixation-indices from the non-prominent (later) fixation-indices towards the attention prediction. The θ thresholding is applied to every f_k $(k = 1, \dots, d)$ and the revised dataset is re-written as $\{(E_1, min(f_1, \theta)), (E_2, min(f_2, \theta)), \cdots, (E_d, min(f_d, \theta))\}$. The attention prediction on web elements transforms to a polychotomous classification problem with θ classes. We employ the multiclass classification procedure described in subsection 2.6.1 to solve this classification problem.

Positional Preference Predictor (PPP)

The EAP helps in estimating the user attention on web elements. However, some elements on the same webpage may be similar in drawing the user attention, say, half of the users fixated first on one element while the remaining half users fixated on another element. Accordingly, both elements may achieve the same FI from EAP. In contrary, the scanpath prediction should follow *zero-sum-rule*— no two elements can obtain the same fixation-index. That is, no two elements can be associated with the same fixation-index. Thus, to resolve these conflicts and to abide by the *zero-sum-rule*, we utilize the users' overall positional preference (or, *bias*) on webpages.

Primarily, a uniform rectangular-grid is constructed with the dimension of the considered webpages. Each grid-cell size is set to the minimum size among all the elements. This consideration overcomes the possibility of mapping multiple elements into the same cell during

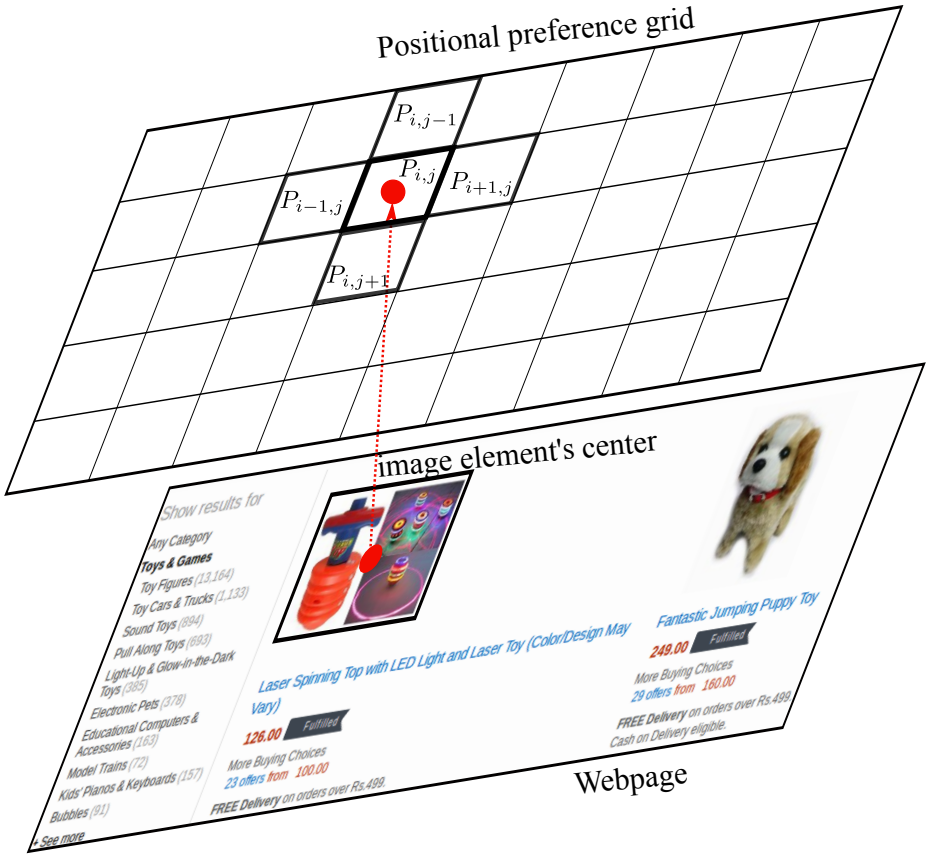


Figure 7.2: Illustration of element's mapping to a positional preference grid-cell

conflict-resolve. Now, for each fixated element ($E_k \in E;\ k = 1, \ldots, d$), the element's center is mapped to the underlying grid-cell (as shown in Figure 7.2) towards the accumulation of *positional preference*. In the grid-cell, the lower the associated FI value, the higher is the attention-drawing ability. Thus, to accordingly designate the prominence for each fixation, we considered the *likelihood* of associated fixation-index. The prominence added to the corresponding grid-cell is as computed.

$$\text{prominence of } f_k(\in \{1, \ldots, \theta\}) = \frac{frequency(f_k)}{\sum_{f'=1}^{\theta} frequency(f')} \tag{7.3}$$

The Equation 7.3 assigns highest prominence for FI=1 (as they are the most frequent FIs), second highest prominence for FI=2, and so on (see Figure 3.4 and Figure 3.5).

As elements may spread across multiple grid-cells, the prominence of the current cell needs to be propagated to the adjacent grid-cells. Towards this, we utilize the matrix smoothing formula presented for oceanographic data [87].

$$P_{i,j} = \frac{P_{i-1,j} + P_{i,j-1} + 2P_{i,j} + P_{i+1,j} + P_{i,j+1}}{6} \tag{7.4}$$

where $P_{i,j}$ denotes the preference accumulated in (i, j)th grid-cell, as shown in Figure 7.2.

Scanpath Predictor

The scanpath predictor module utilizes the results of EAP module and the PPP module to assign the scanpath on a given test webpage, as shown in Figure 7.1b. Given a test webpage, obtain and segregate the text elements and image elements. Compute the respective visual features for each of these modalities. Map every element into the homogeneous visual space using Equation 7.1 and Equation 7.2. Compute the fixation-index of each element through the application of multiclass classification in EAP module. Obtain the θ most attention-drawing elements. Arrange the θ elements into a scanpath based on predicted fixation-index and utilize the positional preference from PPP (assign lower FI value to the element with higher positional preference) to enforce the zero-sum-rule.
Towards the analysis of the proposed approach, the relatively smaller (in the number of stimuli webpages), second experiment (Experiment-II), is utilized to demonstrate the quality of unification (Stage-I) towards the scanpath prediction on unseen data (Stage-II). That is, attention-data from Experiment-I is utilized for Stage-I while the attention-data from Experiment-II is utilized for Stage-II prediction.

7.3 Prediction Performance using WG

7.3.1 Prediction Performance on Same Dataset

The dataset from Experiment-I is utilized in both the stages of the predictive model.

Positional Preference: The approach presented in the PPP is applied to the users' attention obtained from Experiment-I. The grid-cell size is set to 14 px, equivalent to the minimum sized fixated element. The smoothing procedure is repeated for ten times for the positional prominence propagation across grid-cells. The achieved positional preference is visualized in Figure 7.3.

Scanpath-threshold θ selection: The maximum of median fixation-indices is considered for the scanpath-threshold θ. The median fixation-index of text and images are 6 and 3. Accordingly, $\theta = 6$ is considered for the prediction performance analysis.

Ground-truth Preparation: For each fixated web element, the fixation-indices from all users are obtained. The fixation-index resulting from the application *majority-voting-scheme* is assigned as the ground-truth element-granular attention. Towards the scanpath assignment, the elements are sorted based on the element-granular attention with the utilization of positional preference towards resolving the conflicts.

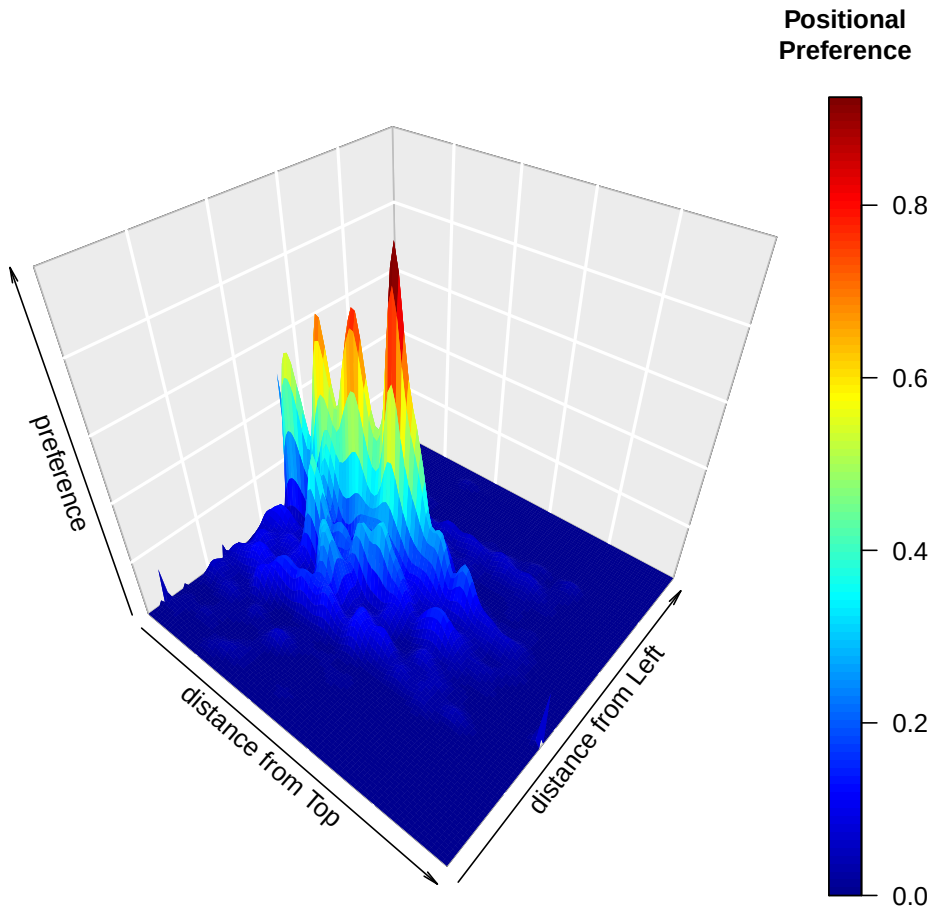


Figure 7.3: Computed Positional Preference

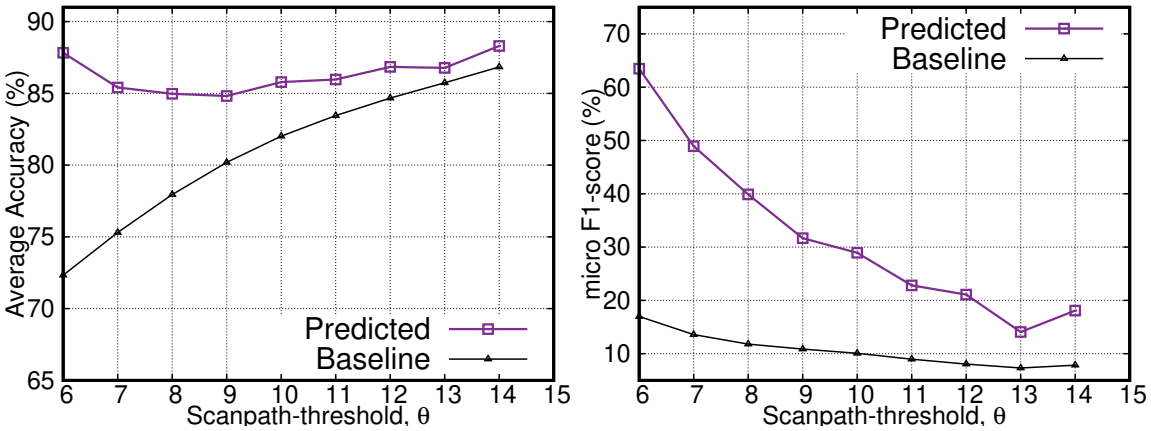
Feature selection: Primarily feature pruning is performed on both the modalities to remove the redundant (constant and linearly dependent) features. Subsequently, text and image canonical directions are computed, and Wilk's Lambda test [160] is applied to determine the significant directions. A total of 28 canonical directions were obtained at the significance-level of 0.01.

Procedure: The elements were projected into the homogeneous space using Equation 7.1 and Equation 7.2, where the significant canonical directions are utilized for W_t and W_i. On the unified data, the presented EAP approach is employed with robust *leave-one-out strategy*. That is, elements of one webpage are used for attention prediction while all the remaining elements' data is utilized in training the model. The procedure is iterated for each webpage as the test webpage and others as training data. The same approach is repeated for scanpath prediction with the utilization of positional preference shown in Figure 7.3. As the classes (fixation-indices) are imbalanced, the micro metrics (average accuracy and micro-F1 score) were computed in each iteration and are averaged to obtain an overall prediction performance.

Table 7.1: Experiment-I: Element-granular attention and Scanpath prediction performance at median thresholding ($\theta = 6$)

Granularity	Metric	Predicted	Baseline	↑ (%)
Scanpath	Avg. Acc.	**87.83**	72.33	21.43
	micro-F1	**63.48**	16.99	273.63

Baseline Selection: A very few of the existing approaches are centered on web elements limiting the baseline comparison. Though the work in [19] considered limited visual features of web elements, the attention prediction is proposed for task-dependent settings and constrained to binary-level prediction. Thus, we utilize the random prediction [107] as the baseline (analogous to [32, 118]) to comprehend the performance of the proposed approach.



(a) Scanpath (b) Scanpath

Figure 7.4: Experiment-I prediction performance with variation in thresholded value

Results: The prediction performance at median scanpath-threshold is shown in Table 7.1. Evidently, the presented approach outperformed the baseline for both metrics and at both granularities (individual elements and the whole scanpath). Between the granularities, the prediction performance is relatively lower for the scanpath than the individual elements. This is attributed to the *zero-sum-rule*. Even though the element's attention-drawing ability (associated fixation-index) is correctly predicted at the element-level, different FIs must be assigned to these elements in case of conflicts resulting in performance reduction. Comparing the metrics, the average accuracy is relatively higher than the micro F1-score for both granular predictions. However, the relatively higher accuracy is observed for the baseline as well. This is attributed to the class imbalance of fixation-indices (initial FIs are more frequent and the latter FIs are less frequent), which results in the bias of average accuracy (towards more frequent FI). The bias-mitigating micro F1-score is the better indicator of the model's performance.

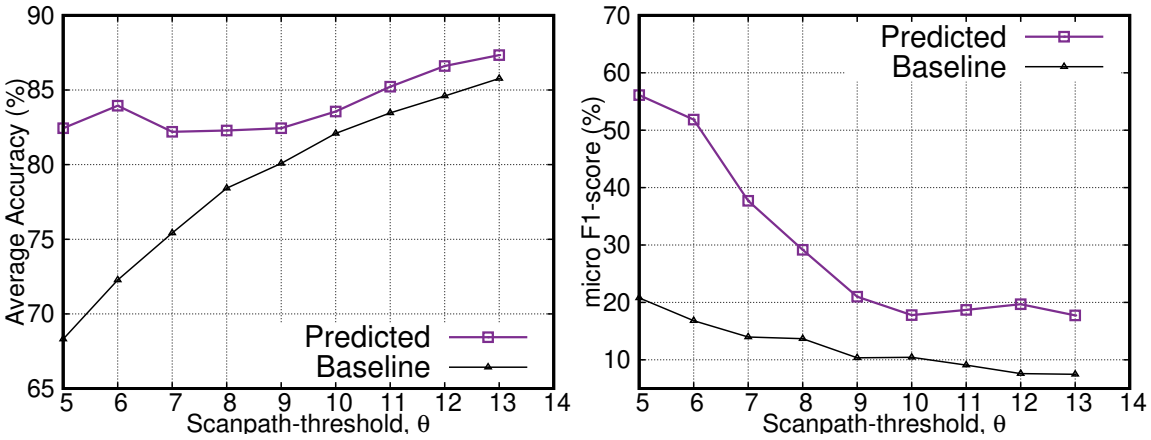
Table 7.2: Experiment-II: Element-granular attention and Scanpath prediction performance at median thresholding ($\theta = 5$)

Granularity	Metric	Predicted	Baseline	↑ (%)
Scanpath	Avg. Acc.	**82.44**	68.31	20.69
	micro-F1	**56.11**	20.78	170.02

To understand the prominence of scanpath-threshold on the performance prediction, the aforementioned procedure is repeated for each increment in the scanpath threshold (θ) value. The presented approach consistently outperformed the baseline, as shown in Figure 7.4. However, the prediction performance reduced with increasing θ indicating the limitation of the considered visual features in predicting the latter attention. Thus, the research question **R1** is answered.

7.3.2 Prediction Performance on Unseen Dataset

To further analyze the efficacy of unification on scanpath prediction, the computed W_t and W_i are utilized to predict the scanpath on unseen webpages, i.e., webpages from Experiment-II (described in Section 3.3).



(a) Scanpath (b) Scanpath

Figure 7.5: Experiment-II prediction performance with variation in thresholded value

The frequency distribution of the FIs is shown in Figure 3.4. In total, text elements received 5726 unique fixations and image elements received 2223 unique fixations. The distribution of the fixation-indices is shown in Figure 3.5. Text elements were fixated with median FI of five and image elements were fixated with median FI of four. Accordingly, $\theta = 5$ is considered for median thresholding.

Procedure: The text data and image data are projected into the homogeneous space using W_t, and W_i respectively. Subsequently, the aforementioned *leave-one-out* strategy is

utilized to predict the user attention on individual web elements. The positional preference was computed and utilized towards the scanpath prediction.

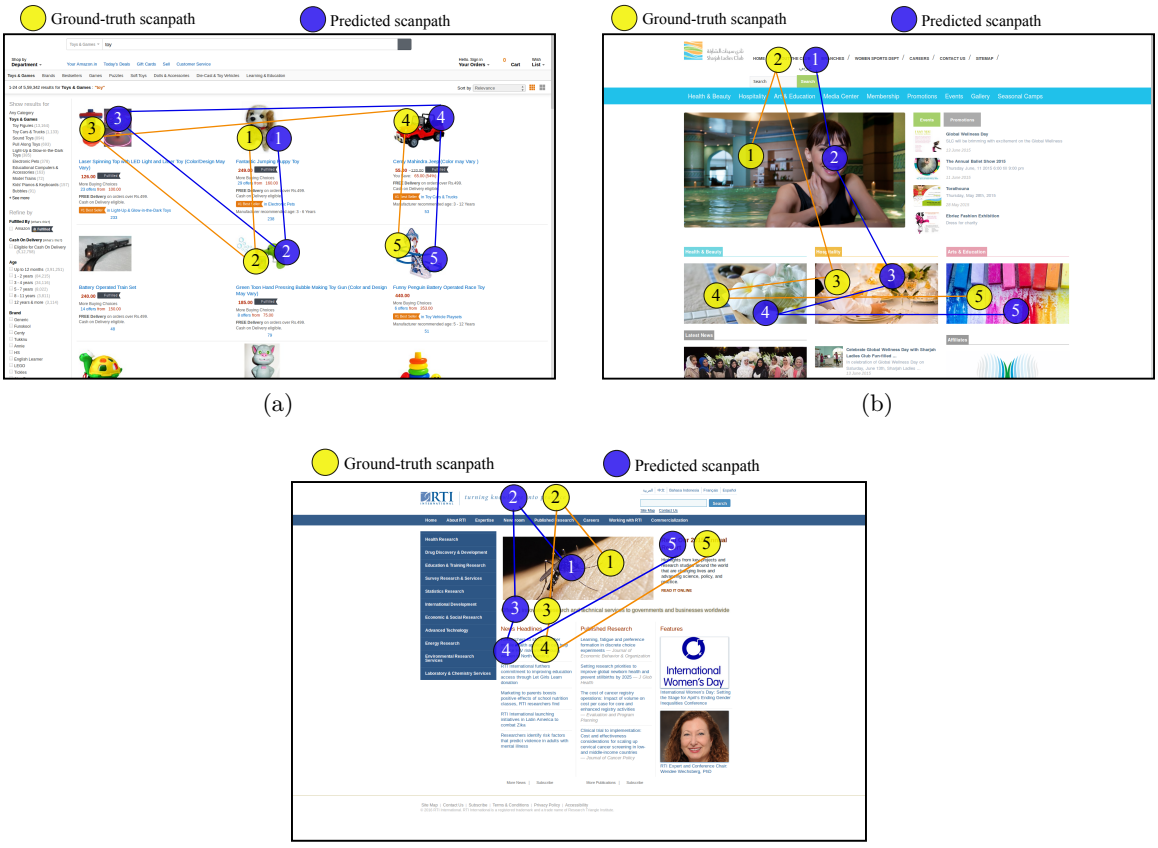
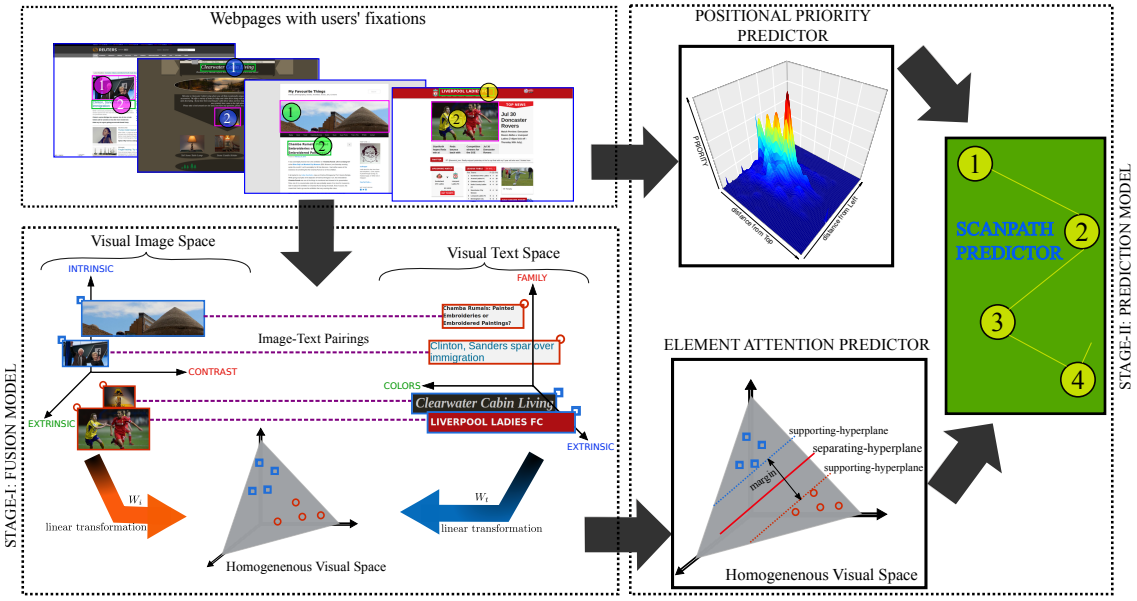


Figure 7.6: Example scanpath predictions

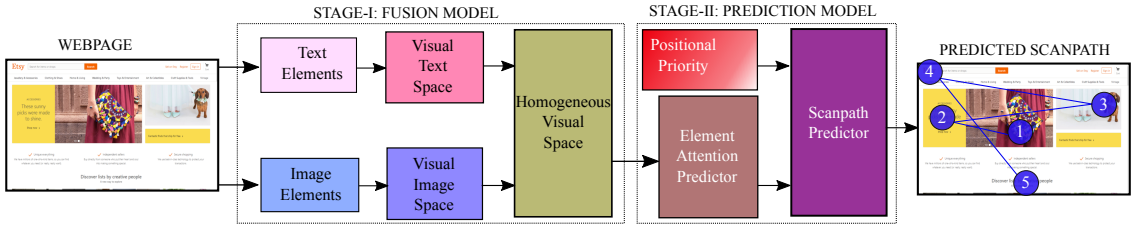
The prediction performance at median scanpath-thresholding is shown in Table 7.2. Analogous to Experiment-I, the model outperformed the baseline for both the metrics, and relatively more average accuracy was achieved than the micro F1-score due to the imbalance in frequency of FIs. Both the element-granular attention prediction and scanpath prediction performance was comparable, but relatively lower than that of Experiment-I. In between both the predictions, the element granular attention was better predicted than the scanpath owing to the adherence to zero-sum-rule. The Figure 7.5 shows the performance variation with the unitary increments in θ. The performance variation was analogous to the analysis of Experiment-I, where the prediction performance reduced with increasing θ. The example scanpath predictions are shown in Figure 7.6. From both the above analyses, the reducing performance with increasing θ indicates the reducing prominence of visual features in predicting user attention. However, the proposed approach better predicts the user's initial attention and corresponding scanpaths. Thus, the research question **R3** is answered for WG.

7.4 Prediction Performance using WUG

The aforementioned prediction approach is applied with the WUG grouping applied in Stage-I, as shown in Figure 7.7.



(a) Model Training



(b) Model Prediction

Figure 7.7: WUG: Schematic of the proposed approach. FUSION MODEL employs canonical correlation analysis (CCA) to achieve the homogeneous representation for text and images. ELEMENT ATTENTION PREDICTOR applies multiclass classification on the achieved homogeneous representation to predict the user attention on web elements. SCANPATH PREDICTOR ordinates the scanpath sequence through zero-sum rule where the positional-preference capturing POSITIONAL PRIORITY PREDICTOR is utilized towards resolving the conflicts among elements.

7.4.1 Prediction Performance on Same Dataset

The dataset from Experiment-I is utilized for training and testing with appropriate data splitting.

Feature Selection: Considering all the features and removal of constant and linearly dependent features resulted in 57 canonical directions. However, the correlations along the

Table 7.3: Experiment-I: Attention prediction performance at median thresholding ($\theta = 6$) for the granular element and scanpath (webpage)

Granularity	**Metric**	**Predicted**	**Baseline**	**↑ (%)**
Scanpath	micro-F1	51.85	15.56	233.23
	Avg. Acc.	83.95	71.85	16.84

latter canonical directions are significantly lower than that along the initial canonical directions. Thus, we employed Wilks' Lamda [160] test to determine the 30 significant canonical directions at the significance level of 0.01. That is, the dimension of the homogeneous visual space is $h = 30$. The significant canonical directions were utilized in the EAP module as well as the Scanpath Predictor module towards the attention prediction as described in rest of the paper.

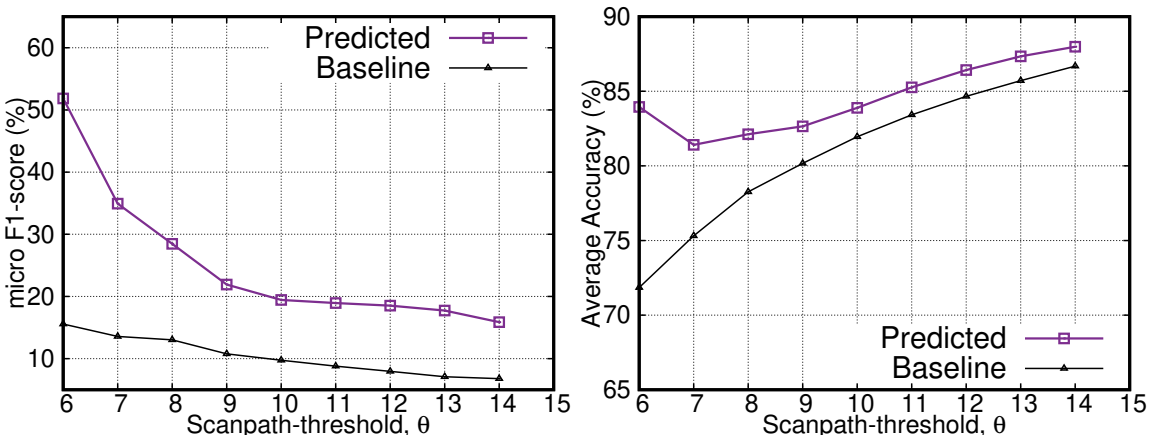
Positional Priority: The positional priority is computed from the user attention data of Experiment-I. Each grid-cell dimension is set to 14px, the size of the minimum sized fixated element. The smoothing procedure is repeated for ten times and the resulting positional priority is visualized in Figure 7.3 which we further utilized for scanpath prediction on webpages of Experiment-I.

Ground-truth Preparation: *Majority-voting-scheme* is employed to assign the ground-truth for individual web elements. That is, for each fixated web element, the frequency of associated fixation-indices is computed and the fixation-index with the highest number of votes is designated as the ground-truth attention for that element. In case of conflicts, the lower FI is assigned as an indicator of element's attention drawing ability. Towards the scanpath, the most attention-drawing θ elements are selected and the scanpath is assigned in the order of associated FIs. In case of conflicts in achieving the *zero-sum rule*, the PPP corresponding to the element's center is utilized. The element with higher positional priority obtains the lower FI value in the arrangement of scanpath.

Scanpath-threshold θ selection: The text modality median FI is 6 and the image modality median FI is 3. The maximum of these medians, 6 is considered as the scanpath-threshold (θ) in Equation 4.3. Accordingly, the prediction performance at $\theta = 6$ is denoted as the performance of the proposed model.

Procedure: All the text and image elements from their natural visual spaces were projected into the homogeneous visual space using Equation 7.1 and Equation 7.2 respectively. Here, significant canonical directions were considered in W_t and W_i. In the homogeneous visual space, EAP module is executed with *leave-one-out strategy*— train the model with elements of all webpages except one webpage which is utilized for prediction, and repeat the procedure

for all webpages, one at a time. Towards the scanpath prediction, the θ most attention-drawing elements (as determined by the FI values) are selected for each webpage and are ordered according to zero-sum-rule with the positional priority (see Figure 7.3) resolving the conflicts. As the FI classes are imbalanced, micro-metrics (micro F1-score and Average Accuracy) are computed to analyze the overall prediction performance.



(a) Scanpath (b) Scanpath

Figure 7.8: Experiment-I prediction performance with variation in scanapth-threshold θ

Baseline Selection: Our approach accounts for both the visual attention and element-granular visual features simultaneously. Unfortunately, most of the existing approaches are limited in their simultaneous incorporation. Thus, we consider the random prediction [107] as a baseline (analogous to [118]) to understand the proposed model's prediction performance. The prediction performance at median scanpath-thresholding is shown in Table 7.3. Clearly, the model outperforms the baseline in predicting the attention at the granularity of element as well as the webpage (that is, scanpath). Comparatively, the attention was better predicted at the granularity of web elements than the scanpath. This is especially attributed to the zero-sum-rule which enforces every element to take different FIs even if their attention-drawing ability is same.

To further understand the influence of scanpath-threshold selection on prediction performance, the performance was computed with each increment in θ as shown in Figure 7.8. Consistently, the element-granular attention was better predicted than the scanpath prediction. However, for both the predictions, the model outperformed the baseline. In between the two metrics, the average accuracy remained approximately consistent across the θ variation whereas the micro F1-score reduced with increase in θ value. The consistency in average accuracy may be attributed to the class imbalance, where the accuracy is tend to bias towards the most frequent FI. The micro F1-score better indicates the model's performance. Overall, the model performance approached the baseline for the larger θ values indicating the limitation

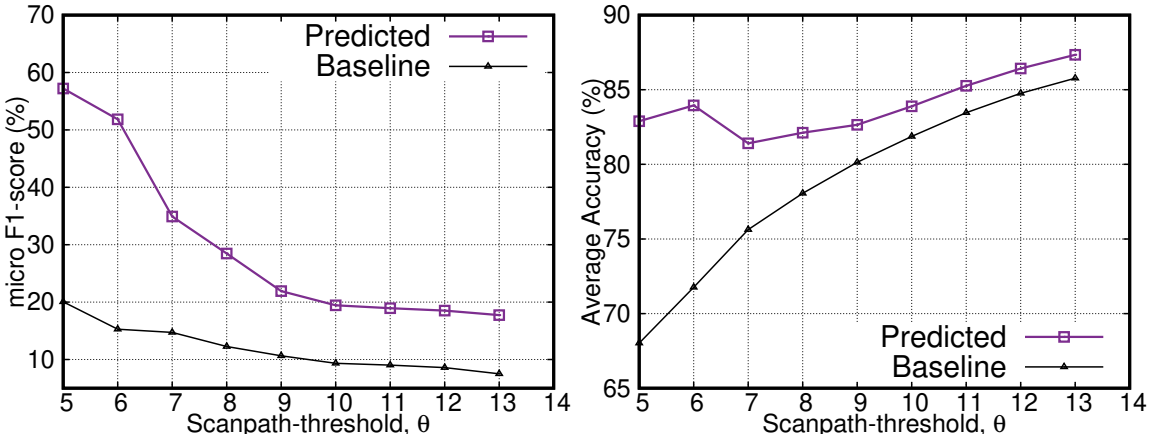
Table 7.4: Experiment-II: Attention prediction performance at median thresholding ($\theta = 5$) for the granular element and scanpath (webpage)

Granularity	Metric	Predicted	Baseline	↑ (%)
Scanpath	micro-F1	**57.22**	20.06	185.24
	Avg. Acc.	**82.89**	68.02	21.86

of visual features' ability in describing the users' latter visual attention. Thus, the research question **R2** is answered.

7.4.2 Prediction Performance on Unseen Dataset

To further demonstrate the efficacy of the two-stage model, the W_ts and W_is from the Experiment-I, are utilized to predict the element-granular and scanpath prediction on new dataset, Experiment-II (see Table 3.1).



(a) Scanpath (b) Scanpath

Figure 7.9: Experiment-II prediction performance with variation in scanpath-threshold θ

The visual features are computed for text and images analogous to the Experiment-I. The text and image elements are projected into the homogeneous space. The EAP module, PPP module and the scanpath predictor module are applied as reported in aforementioned section. The median FI of text is five and the median FI of image is four. Thus, scanpath-threshold θ is set to the maximum of medians, i.e., five.

Results: The prediction performance at median scanpath-thresholding is shown in Table 7.4. The prediction performance was found to be similar to that of Experiment-I. The Figure 7.9 shows the model's performance with increasing θ values. The element-granular attention was better predicted than the scanpath and micro F1-score better indicated than model performance than the average accuracy which remained approximately constant

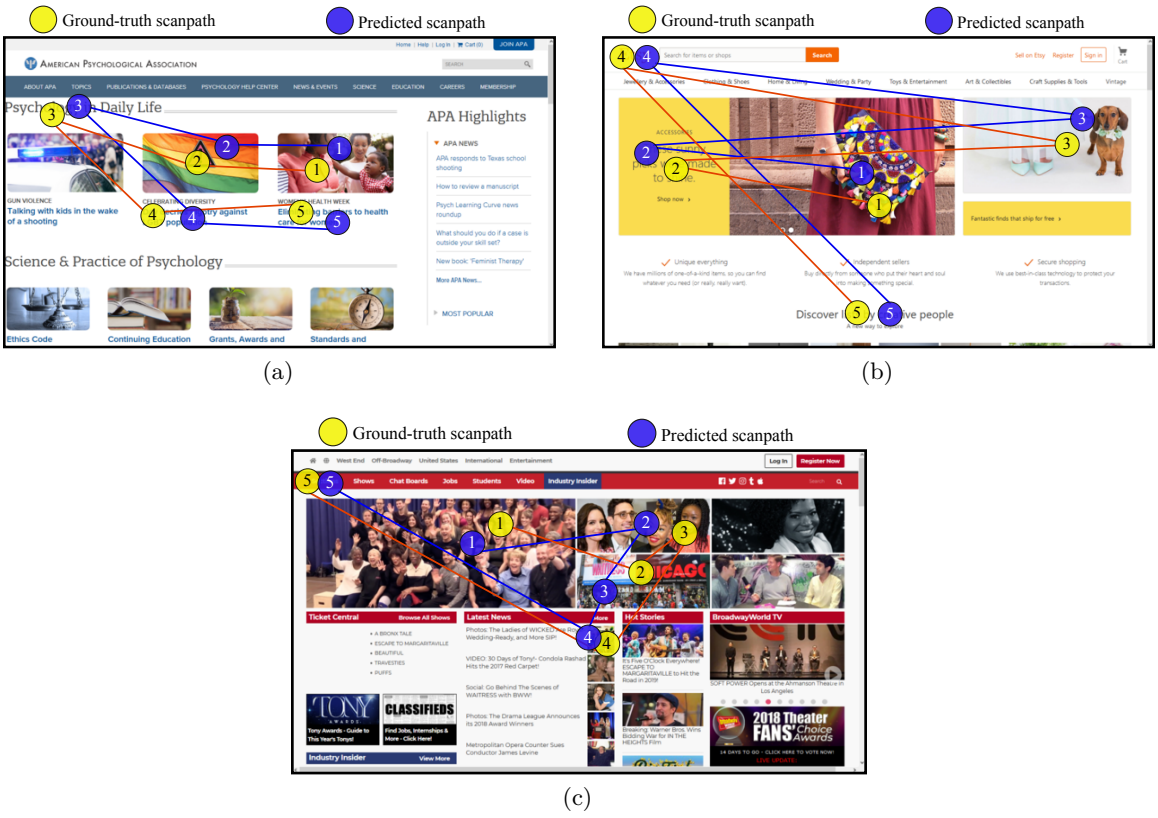


Figure 7.10: Example scanpath predictions

throughout the variation in θ. Overall, the model performance approached the baseline with increasing θ. In summary, the results demonstrate the efficacy of the proposed approach on predicting the scanpaths on webpages. The example scanpath predictions from the proposed approach are shown in Figure 7.10. Thus, the research question **R3** is answered for WUG.

7.5 Summary

The text and image visual features predict the users' scanpath on webpages for the WG as well as WUG groupings. The prediction performance outperforms the baseline through-out the variation in scanpath threshold θ. However, the visual features' predictive performance reduces with increase in θ as other factors such as absence of salient elements guide the latter attention which may not be explained by the considered visual features.

8

Conclusions and Future Perspectives

The user attention modeling on webpages benefits the web designers, content authors, individual bloggers, and advertisers. The evolution of the Web, Internet, and Internet users (used interchangeably with *Web users*) drawn research community's interest towards attention modeling on webpages.

The attention modeling evolved from user attention-based observational models (such as F-pattern) to the visual feature-based computational models (based on image processing techniques). The former approaches are limited in applicability due to the constraints such as user availability, time, and money involved in the experimentation. The latter approaches are limited in the interpretability due to the non-direct incorporation of user attention and the limited granularity (salient or not determination). This thesis aims to overcome the limitations of both the approaches while leveraging the best of both worlds (user attention from UAB approaches and visual features from VFB approaches).

8.1 Summary of Thesis

The objective of the thesis is to consider the webpage as an ensemble of elements, and determine the prominence of element as a whole. This processing is essential as the designers synthesize the whole element and then incorporate into a webpage during designing. The findings of element-based attentional approaches are as summarized.

In Chapter 4, the prominence of elements' position and visual features is presented. The analyses revealed that though the users predominantly allocate the initial attention to Middle region and the Top regions, the elements in Right and Bottom regions are not completely ignored. The observation may be attributed to the visual features of the elements. To understand the prominent visual features, the information-gain scores were computed between the visual features and fixation-indices. The analysis revealed that *space* and *font-size* visual features are informative for textual elements, and *Mid-level Color Histogram* visual features are informative for image elements, while `position` and `size` are informative for both the

modalities. The subsequent computational attention modeling demonstrated the efficacy of informative visual features in predicting the user attention on web elements with 90% average accuracy and 70% micro-F1 score.

To leverage the prominence of visual features in explaining the free-viewing user attention, the Chapter 5 presented a computational approach to correlate and unify the text and image modalities. The empirical analysis revealed that text and image visual features are correlated when the interface idiosyncrasies are constrained. The *font-family* features of text are comparable to and as influential as HISTOGRAM and CONTRAST HISTOGRAM features of images in achieving the unification between both the modalities. The text and images delineate the attention on each other when the interface idiosyncrasies are constrained, that is, for Webpage-oriented Grouping (WG) and Webpage-and-User-oriented Grouping (WUG).

In Chapter 6 and Chapter 7, the established unification is extended to predict the user attention on web elements and scanpath on a webpage respectively. The analysis found that text and image visual features predict the user attention on web elements and scanpath on webpages for the WG as well as WUG groupings. The prediction performance outperforms the baseline through-out the variation in scanpath threshold θ. However, the visual features' predictive performance reduces with increase in θ as other factors such as absence of salient elements guide the latter attention which may not be explained by the considered visual features.

8.2 Applications

Our computational attention modeling and prediction approach is centered on web-design oriented visual features. Thus, designers may manipulate the visual features directly and estimate the user attention on individual web elements or the whole webpage. The model also helps in the repeated computation of user's attention after every alteration to webpage elements during their synthesis and rendering. This overcomes the difficulty in finding the time, money, and especially users to perform the repetitive eye-tracking experiments for understanding the attention.

In Chapter 4, *space*-related and *font-size*-related features were found to be informative for text, *Mid-level Color Histograms* are found to be informative, while element's *position* and *area* was found to be informative for both modalities. Thus, to achieve the desired attention on an element, the corresponding informative visual features may be varied. For example, (i) element's area on a webpage may be increased and the corresponding attention may be predicted. Accordingly, the area may be further increased or decreased until the desired attention is achieved. In case the area is constrained (say, to incorporate multiple other elements), the element's position may be varied to achieve the desired attention. (ii) For text elements, besides position and size, the remaining informative

CSS features including `column-gap, line-height, padding, font-size, font-weight` may be varied to achieve the desired user attention. These features would be useful in case the element's position and size cannot be modified. Say, increasing the `font-size` and `font-weight` will make the text appear in more bolder font making it more (quick) attention-drawing than in earlier settings. (iii) For image elements, besides position and size, the intrinsic color features need to be manipulated towards achieving the desired user attention. Using the *purer* colors (with color component values around 0 and around 255) for images may assimilate them with the background which is majorly either white in color or sometimes black in color. In contrast, using the *Mid-level Color Histograms* may highlight the image elements, especially from the background. Thus, color component values may be manipulated towards mid-level color components (say, from 50 to 200) to achieve the desired user attention.

In Chapter 5, we have established that text and image visual features are unifiable concerning the user's attention. This finding is especially useful in cross-modal element replacement on a webpage. That is, to replace an image element with a text element and vice-versa. The motivation for the application is described as follows.

The Internet bandwidth is limited in the developing and least developed countries (LDCs) when compared with the developed countries. As per United Nation's Information and Communication Technology's (ICT's) Facts and Figures [130], in the year 2016, the international Internet bandwidth per user is 6 kilobits per second (kbps) for LDCs and 53 kbps for developing countries while it is 140 kbps for the developed countries. It means, for example, to visualize a complete 7.5kB (or 60 kb) image on the webpage, users from LDCs have to wait for 10 seconds. Such high latency causes user dissatisfaction besides affecting their psychological and cognitive processing. Thus, apart from other solutions that address limited bandwidth problem, web designers can choose to render the content with low bandwidth cost modalities (text as opposed to images) to serve the information needs of the progressive 41.3% of people from developing countries (Internet users) [130]. However, the unimodal conversion of the multimodal webpages, through replacement of high-cost modalities (images) with the low-cost modalities (text), should achieve equivalence from two representational aspects of the modalities: 1. *semantic features*— to represent the semantic information associated with the elements, 2. *visual features*— to represent how users visually perceive the elements presented on the interface. Towards this, the semantic cross-modal analyses (given a text fetch the semantically equivalent image and vice-versa) [157] helps to obtain the equivalent semantic low-cost modality of a given modality. However, such an equivalent modality cannot be used for direct replacement as user attention needs to be preserved even after the replacement. The preservation is desired for both the user and the website owner as web elements are designed so as to draw user's desired attention. The homogeneous visual space constructed in Chapter 5 will be useful to find a similar attention-drawing cross-modal element [129].

In Chapter 6, we have presented a computational approach to predict the user attention on individual web elements and scanpath on a webpage. The element-based attention prediction finds application in improving the user's perceived webpage load time [88]. Typically, users perceive a webpage as quick loading if the user's information of interest is rendered before others. Thus, our approach can be employed to identify the prominent (quicker) attention-drawing elements (say, elements predicted to achieve first and second fixations), and render them a priori to other elements. This approach improves the user's perceived webpage load time and thereby satisfaction. Especially, with the techniques available to capture the webpage screenshot in advance from the user-end [37, 60, 145], the rendering policies are implementable at the web servers, as well.

In Chapter 7, we have presented a computational approach to predict the user scanpath on webpages. This approach finds applications in web designing as well as rendering. The size of the Web (in terms of number of webpages) is increasing progressively with the time [131], and users are provided with numerous alternatives for each need. For example, multiple e-commerce websites are available to purchase the same goods, and multiple online-platforms are available to learn the same subject. In this competitive environment, it is essential to draw the users' attention to their desired entity (displayed on a webpage) as quick as possible, to prevent the churn. Our scanpath prediction approach helps to determine the sequence of quickly attention-drawing elements. This overcomes the possibility of assigning multiple elements with same attention-drawing ability as per the approach presented in Chapter 6. For the web-designer, the predicted scanpath indicates whether the desired entity (say, a new product on an e-commerce webpage) is able to quickly draw the user attention or not (say, among first three fixations). If not, the visual features of the corresponding element may be modified to achieve the desired attention in the scanpath. Accordingly, the desired entity may be rendered before the other elements (without the necessity to resolve the conflicts among multiple elements with is likely as per the approach presented in Chapter 6). Additionally, the predicted scanpath can be utilized to enhance the attention-drawing ability of the semantically prominent elements through the manipulation of their corresponding visual features. This helps to achieve the user satisfaction from information as well as visual perception aspects.

8.3 Limitations

Our thesis work helps to overcome the predominant bottleneck associated with multi modalities on webpages. However, like any other research work, it is worth noting the limitations of our thesis work. The limitations concern the three aspects namely, user, attention, and the interface (precisely webpages). Regarding the *user*, the user attention may be affected by the age, gender, culture, etc. Regarding the *attention*, the user goal, such as searching, browsing, and free-viewing, may determine the user's attention allocation

besides the influence of visual features. Regarding the *interface*, the user interface on which the users view the webpage may determine the user attention allocation, such as mobile devices, laptops, palmtops, desktops, etc. As addressing the affect of each factor involves a dedicated experimental setup with elaborate analyses and multiple user participation, which is beyond the scope of our thesis, this section highlights the prominent limitations to guide the future research direction.

- The intrinsic visual feature influence on free-viewing attention was analyzed for prominent Generation Y (age 18–31 [38]) users. However, further investigations are solicited for the influence of other user characteristics such as culture [35, 97, 124] and age [148] on the intrinsic visual feature-based attention allocation. As participants in our study were experienced web users, extending the analyses to novice users help in enhancing the model's applicability for diverse users.

- For the current thesis work, webpage images were identified using the image-related HTML tags (`<img>` and `<picture>`), analogous to [19]. Accordingly, text embedded into the images is processed as an image though users may perceive it as text; However, extracting the text along with its visual features from images is not feasible as these images (such as logos) were designed using graphics editors, representing them using the web-designer oriented CSS features may not be practical.

- Though the performance of our model is demonstrated for predicting the free-viewing attention, further investigations are required for task-dependent settings [10] such as task cognitive demand [6], task motivation [111], searching [58, 109], etc.

- In our thesis work, motion elements such as animations were discarded from the analysis. Their presence may influence the attention allocation. However, their inclusion would complicate the analysis due to the incorporation of motion related characteristics and synchronizing the user attention with motion frame.

8.4 Future Research Directions

As described in Chapter 1, our thesis work is based on five key research components, namely, attention modeling, webpages, interface, users, and eye-tracking. The scope for future research expands to each of the component as described below.

Attention Modeling: The presented approach explored for the linear association between text and images (in Stage-I), further approaches are desired to explore the possible non-linear association. In Stage-II, multiple other prediction approaches may be employed for improved performance. Especially, with the advancement in machine learning techniques, more sophisticated approaches such as deep learning and reinforcement learning techniques

may be applied. However, it is worth to note the limited data availability due to the costs involved in experimentation.

Eye-tracking: The prominent future direction is to extend and analyze the performance of the proposed approach for task-dependent settings. In the current setting, the *task-at-the-hand* and *user expectational-bias* are constrained. In their presence and variation, respective factor-dependent components need to be incorporated towards the pairing and attention modeling. Accordingly, the stimuli preparation and gaze parameter selection needs to be performed. For example, in search tasks, the gaze-duration (an indicator of cognitive load) on the web elements is important and may need to be considered besides the fixation-index.

Interface: In the current setting, we limited the analysis to regular desktop monitors. However, the approach can be extended to analyze the proposed model performance on other interfaces such as mobile phones and other devices such as palm-tops. Especially, such investigations help reveal the prominence of elements' display size or proportion of elements' occupancy on a given interface on the user attention.

Users: The proposed views of human attention and the corresponding pairing of text and images assume that all the user attention allocations on a webpage are similar. However, the idiosyncratic characteristics (such as age, gender, web experience) may influence the user attention allocation and accordingly need to be incorporated towards the cross-modal pairing in their respective presence.

Webpages: The current work is limited to static webpages and category-specific webpage features are not considered. That is, the webpages intended for e-commerce may be designed differently in comparison to the webpages intended for news portals. The proposed approach in our thesis may be analyzed for different categories for the applicability. Further, the evolving eye-tracking technologies may be applied on the dynamic webpages to analyze and model the element-driven attention while catering to the evolution of the Web [1]. Even for the static webpages, the work may be extended to develop the browser-plugins so that user attention can be predicted on-fly.

Publications

From Thesis

1. **Sandeep Vidyapu**, Vijaya Saradhi Vedula & Samit Bhattacharya, "Investigating and Modeling the Web Elements' Visual Feature Influence on Free-viewing Attention", *ACM Transactions on the Web (TWEB)*, 2020.

2. **Sandeep Vidyapu**, Vijaya Saradhi Vedula & Samit Bhattacharya, "Weighted-Voting-based Effective Visual Attention Prediction on Web Image Elements", *Interacting with Computers (IWC)*, 2020.

3. **Sandeep Vidyapu**, Vijaya Saradhi Vedula, Michael Burch & Samit Bhattacharya, "Attention-based Cross-Modal Unification of Visualized Text and Image Features", *ACM Symposium on Eye Tracking Research & Applications (ETRA), ETWEB, Stuttgart, Germany*, 2020.

4. **Sandeep Vidyapu**, Vijaya Saradhi Vedula & Samit Bhattacharya, "Attention Prediction on Webpage Images using Multilabel Classification", *IEEE International Conference on Systems, Man, and Cybernetics (SMC), Bari, Italy*, 2019.

5. **Sandeep Vidyapu**, Vijaya Saradhi Vedula & Samit Bhattacharya, "Quantitative Visual Attention Prediction on Webpage Images Using Multiclass SVM", *ACM Symposium on Eye Tracking Research & Applications (ETRA), ETWEB, Denver, USA*, 2019.

6. **Sandeep Vidyapu**, Vijaya Saradhi Vedula & Samit Bhattacharya, "Fixation-Indices based Correlation between Text and Image Visual Features of Webpages", *ACM Symposium on Eye Tracking Research & Applications (ETRA), Warsaw, Poland*, 2018.

7. **Sandeep Vidyapu**, Vijaya Saradhi Vedula & Samit Bhattacharya, "From Multimodal to Unimodal Webpages for Developing Countries", *Neural Information Processing Systems (NeurIPS), ML4D Workshop, CA, USA*, 2017.

8. **Sandeep Vidyapu**, Vijaya Saradhi Vedula & Samit Bhattacharya, "CCA: Attention based Cross-Modal Retrieval on Webpages", *International Conference on Machine Learning (ICML), IML Workshop, Sydney, Australia*, 2017.

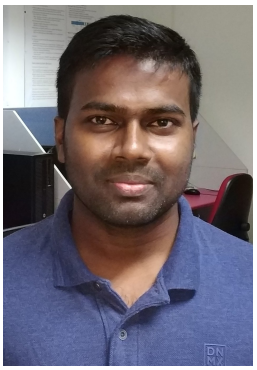
Outside Thesis

1. Awnish Kumar, **Sandeep Vidyapu**, Vijaya Saradhi Vedula & Tamarapalli Venkatesh, "A Multi-view Subspace Learning Approach to Internet Traffic Matrix Estimation", *IEEE Transactions on Network and Service Management (TNSM)*, 2020.

2. Sai Anirudh Kondaveeti, **Sandeep Vidyapu** & Samit Bhattacharya, "Improved Gaze Likelihood Based Web Browsing", *Indian Conference on Human Computer Interaction (India HCI), Mumbai, India*, 2016.

3. **Sandeep Vidyapu**, Samit Bhattacharya & Vijaya Saradhi Vedula, "Study on the Effect of Multimodality and Gender on Visual Attention of Webpages and Design Principles" (poster), *Indian Conference on Human Computer Interaction (India HCI), Mumbai, India*, 2016.

Brief Vitae



Vidyapu Sandeep joined the Dual-Degree (M.Tech + Ph.D) program at the Department of Computer Science and Engineering (CSE) of Indian Institute of Technology (IIT) Guwahati, India in July 2013. Prior to the joining, he received his Bachelor of Technology (B.Tech) degree in Electronics and Communications Engineering (ECE) from Jawaharlal Nehru Technological University, Hyderabad, India.

In January 2020, he joined as an Assistant Professor in the Department of Computer Science and Information Systems of Birla Institute of Technology and Science Pilani (BITS Pilani), Hyderabad.

He has keen interest in the interdisciplinary research, especially, at the intersection of Human-Computer Interaction (HCI) and Applied Machine Learning (Applied ML). His current research interests include computational modeling of user attention on graphical user interfaces (precisely, webpages), and user behavior analysis through empirical studies.

Contact Information

Email: sandy.apj911@gmail.com; s.vidyapu@iitg.ac.in
Address: H. No. 3-5-887/1-4-1, Volkswagon Showroom, Vittalwadi, Hyderabad-500029, Telangana, India

---

Bibliography

[1] Sareh Aghaei, Mohammad Ali Nematbakhsh, and Hadi Khosravi Farsani. Evolution of the world wide web: From web 1.0 to web 4.0. *International Journal of Web & Semantic Technology*, 3(1):1, 2012. doi:10.5121/ijwest.2012.3101.

[2] M Elgin Akpınar and Yeliz Yesilada. Vision based page segmentation algorithm: Extended and perceived success. In *International Conference on Web Engineering*, pages 238–252. Springer, 2013.

[3] Sayeed Safayet Alam and Radu Jianu. Analyzing eye-tracking information in visualization and data space: from where on the screen to what on the screen. *IEEE Transactions on visualization and computer graphics*, 23(5):1492–1505, 2017. doi:10.1109/TVCG.2016.2535340.

[4] Sutcliffe Alistair and Namoun Abdallah. Predicting user attention in complex web pages. *Behaviour & Information Technology*, 31(7):679–695, 2012. doi:10.1080/0144929X.2012.692101. URL `https://doi.org/10.1080/0144929X.2012.692101`.

[5] Ioana Bacivarov, Mircea Ionita, and Peter Corcoran. Statistical models of appearance for eye tracking and eye-blink detection and measurement. *IEEE Transactions on consumer electronics*, 54(3):1312–1320, 2008.

[6] Hyejin Bang and Bartosz W Wojdynski. Tracking users' visual attention and responses to personalized advertising based on task cognitive demand. *Computers in Human Behavior*, 55:867–876, 2016. doi:10.1016/j.chb.2015.10.025.

[7] Jae Won Bang, Eui Chul Lee, and Kang Ryoung Park. New computer interface combining gaze tracking and brainwave measurements. *IEEE Transactions on Consumer Electronics*, 57(4), 2011.

[8] Shabab Bazrafkan, Anuradha Kar, and Claudia Costache. Eye gaze for consumer electronics: Controlling and commanding intelligent systems. *IEEE Consumer Electronics Magazine*, 4(4):65–71, 2015.

[9] Michael L Bernard, Barbara S Chaparro, Melissa M Mills, and Charles G Halcomb. Comparing the effects of text size and format on the readibility of computer-displayed times new roman and arial text. *International Journal of Human-Computer Studies*, 59(6):823–835, 2003. doi:10.1016/S1071-5819(03)00121-6.

[10] Torsten Betz, Tim C Kietzmann, Niklas Wilming, and Peter Koenig. Investigating task-dependent top-down effects on overt visual attention. *Journal of Vision*, 10(3): 15–15, 2010. doi:10.1167/10.3.15.

[11] Sergey Bezryadin, Pavel Bourov, and Dmitry Ilinih. Brightness calculation in digital image processing. In *International symposium on technologies for digital photo fulfillment*, volume 2007, pages 10–15. Society for Imaging Science and Technology, 1 2007. doi:10.2352/ISSN.2169-4672.2007.1.0.10.

[12] Charles Bigelow and Donald Day. Digital typography. *Scientific American*, 249(2): 106–119, 1983. doi:10.1038/scientificamerican0883-106.

[13] Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M Jones. mlr: Machine learning in r. *Journal of Machine Learning Research*, 17(170):1–5, 2016.

[14] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013. doi:10.1109/TPAMI.2012.89.

[15] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, New York, NY, USA, 1992. ACM. doi:10.1145/130385.130401.

[16] Steven Bradley. 3 design layouts: Gutenberg diagram, z-pattern, and f-pattern. `http://vanseodesign.com/web-design/3-design-layouts/`, 2011. Online; accessed 09 Aug 2016.

[17] A Buchner and N Baumgartner. Text–background polarity affects performance irrespective of ambient illumination and colour contrast. *Ergonomics*, 50(7):1036–1063, 2007. doi:10.1080/00140130701306413.

[18] Michael Burmester and Marcus Mast. Repeated web page visits and the scanpath theory: A recurrent pattern detection approach. *Journal of Eye Movement Research*, 3(4):5, 2010.

[19] Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30, NY, USA, 2009. ACM. doi:10.1145/1518701.1518705.

[20] Michael D Byrne, John R Anderson, Scott Douglass, and Michael Matessa. Eye tracking the visual search of click-down menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 402–409, New York, NY, USA, 1999. ACM. doi:10.1145/302979.303118.

[21] Julie A Bzostek and Michael S Wogalter. Measuring visual search time for a product warning label as a function of icon, color, column and vertical placement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 43, pages 888–892. Sage Publications Sage CA: Los Angeles, CA, 1999. doi:10.1177/154193129904301603.

[22] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. doi:10.1145/1961189.1961199.

[23] Chien-Hsiung Chen and Shu-Ying Chiang. Effects of screen resolution and column ratio on search performance and subjective preferences. *Displays*, 33(1):28–35, 2012. doi:10.1016/j.displa.2011.10.004.

[24] Kang-A Choi, Chunfei Ma, and Sung-Jea Ko. Improving the usability of remote eye gaze tracking for human-device interaction. *IEEE Transactions on Consumer Electronics*, 60(3):493–498, 2014.

[25] Tim Chuk, Antoni B Chan, and Janet H Hsiao. Understanding eye movements in face recognition using hidden markov models. *Journal of vision*, 14(11):8–8, 2014.

[26] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi:10.1177/001316446002000104.

[27] World Wide Web Consortium et al. *Document Object Model (DOM) Level 3 Core Specification*. W3C, 2004. URL https://www.w3.org/TR/2004/REC-DOM-Level-3-Core-20040407/.

[28] World Wide Web Consortium et al. Web content accessibility guidelines (wcag) 2.0. *World Wide Web Consortium*, 2008.

[29] World Wide Web Consortium et al. *Cascading style sheets level 2 revision 1 (css 2.1) specification*. W3C, 2011. URL https://www.w3.org/TR/2011/REC-CSS2-20110607/.

[30] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995. doi:10.1007/BF00994018.

[31] Paulo Cortez. *rminer: Data Mining Classification and Regression Methods*. R Foundation for Statistical Computing, 2016. URL https://CRAN.R-project.org/package=rminer. R package version 1.4.2.

[32] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014. doi:10.1109/tpami.2013.142.

[33] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42(3): 692–700, 2010.

[34] Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

[35] Dianne Cyr, Milena Head, and Hector Larios. Colour appeal in website design within and across cultures: A multi-method evaluation. *International Journal of Human-Computer Studies*, 68(1-2):1–21, 2010. doi:10.1016/j.ijhcs.2009.08.005.

[36] Arthur Pentland Dempster. *Elements of continuous multivariate analysis*. Addison-Wesley Publishing Co., 1969.

[37] Sanjeev Dhanda and Yi Zhang. Testing the rendering of visual content, Mar 2015. US Patent 8,971,642.

[38] Soussan Djamasbi, Marisa Siegel, and Tom Tullis. Generation y, web design, and eye tracking. *International Journal of Human-Computer Studies*, 68(5):307–323, 2010. doi:10.1016/j.ijhcs.2009.12.006.

[39] Gautier Drusch and JMC Bastien. Analyzing visual scanpaths on the web using the mean shift procedure and t-pattern detection: a bottom-up approach. In *Proceedings of the 2012 Conference on Ergonomie et Interaction homme-machine*, page 181. ACM, 2012.

[40] Robert L Duchnicky and Paul A Kolers. Readability of text scrolled on visual display terminals as a function of window size. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 25(6):683–692, 1983.

[41] Andrew Duchowski. *Eye tracking methodology: Theory and practice*, volume 373. Springer Science & Business Media, 2007.

[42] Andrew T Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, 34(4):455–470, 2002.

[43] Mary C Dyson. How physical text layout affects reading from screen. *Behaviour & Information Technology*, 23(6):377–393, 2004. doi:10.1080/01449290410001715714.

[44] Mary C Dyson and Mark Haselgrove. The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, 54(4):585–612, 2001.

[45] Mary C Dyson and Gary J Kipping. The effects of line length and method of movement on patterns of reading from screen. *Visible Language*, 32(2):150, 1998.

[46] Sukru Eraslan and Yeliz Yesilada. Patterns in eyetracking scanpaths and the affecting factors. *J. Web Eng.*, 14(5-6):363–385, 2015.

[47] Sukru Eraslan, Yeliz Yesilada, and Simon Harper. Identifying patterns in eyetracking scanpaths in terms of visual elements of web pages. In *Int. Conf. Web Eng.*, pages 163–180. Springer, 2014. doi:10.1007/978-3-319-08245-5_10.

[48] Sukru Eraslan, Yeliz Yesilada, and Simon Harper. Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. *Journal of Eye Movement Research*, 9(1), 2015.

[49] Sukru Eraslan, Yeliz Yesilada, and Simon Harper. Scanpath trend analysis on web pages: Clustering eye tracking scanpaths. *ACM Transactions on the Web (TWEB)*, 10(4):20, 2016. doi:10.1145/2970818.

[50] Sukru Eraslan, Yeliz Yesilada, and Simon Harper. Engineering web-based interactive systems: Trend analysis in eye tracking scanpaths with a tolerance. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, EICS '17, pages 3–8, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5083-9. doi:10.1145/3102113.3102116. URL http://doi.acm.org/10.1145/3102113.3102116.

[51] Pete Faraday. Visually critiquing web pages. In *Multimedia'99*, pages 155–166, Vienna, 2000. Springer, Available at http://facweb.cs.depaul.edu/cmiller/faraday/Faraday.htm.

[52] James Fogarty, Scott E Hudson, Christopher G Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C Lee, and Jie Yang. Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(1): 119–146, 2005. doi:10.1145/1057237.1057243.

[53] Adrian Ford and Alan Roberts. Colour space conversions. *Westminster University, London*, 1998:1–31, 1998.

[54] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*. R Foundation for Statistical Computing, 2018. URL https://CRAN.R-project.org/package=caret. R package version 6.0-79.

[55] William J Gibbs and Ronan S Bernas. Visual attention in newspaper versus tv-oriented news websites. *Journal of Usability Studies*, 4(4):147–165, 2009. URL http://uxpajournal.org/visual-attention-in-newspaper-versus-tv-oriented-news-websites/.

[56] Robert Gittins. *Canonical analysis: a review with applications in ecology*, volume 12. Springer Science & Business Media, 2012.

[57] Joseph H Goldberg and Jonathan I Helfman. Scanpath clustering and aggregation. In *Proceedings of the 2010 symposium on Eye-Tracking Research & Applications*, pages 227–234. ACM, 2010.

[58] Joseph H Goldberg, Mark J Stimson, Marion Lewenstein, Neil Scott, and Anna M Wichansky. Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, pages 51–58, NY, USA, 2002. ACM. doi:10.1145/507072.507082.

[59] Antonio Gracia, Santiago González, Victor Robles, and Ernestina Menasalvas. A methodology to compare dimensionality reduction algorithms in terms of loss of quality. *Information Sciences*, 270:1–27, 2014.

[60] Jason Graves and Ben Haley. Method and system for transforming an integrated webpage, Oct 2013. US Patent 8,549,395.

[61] Rebecca Anne Grier. *Visual attention and web design*. PhD thesis, University of Cincinnati, 2004. URL http://rave.ohiolink.edu/etdc/view?acc_num=ucin1092767744.

[62] Antonio Gulli and Alessio Signorini. The indexable web is more than 11.5 billion pages. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, pages 902–903, New York, NY, USA, 2005. ACM. doi:10.1145/1062745.1062789.

[63] Stephen R Gulliver and George Ghinea. Stars in their eyes: What eye-tracking reveals about multimedia perceptual quality. *IEEE Trans. Syst., Man, Cybern. A*, 34(4): 472–482, 2004.

[64] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16 (12):2639–2664, 2004.

[65] Prateek Hejmady and N Hari Narayanan. Visual attention patterns during program debugging with an ide. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 197–200. ACM, 2012.

[66] Helene Hembrooke, Matt Feusner, and Geri Gay. Averaging scan patterns and what they can tell us. In *Proc. 2006 Symp. Eye Track. Res. & Appl.*, pages 41–41. ACM, 2006.

[67] John Heminghous and Andrew T Duchowski. icomp: a tool for scanpath visualization and comparison. In *ACM SIGGRAPH 2006 Research posters*, page 186. ACM, 2006.

[68] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. doi:10.1093/biomet/28.3-4.321.

[69] Yu-Chen Hsieh and Kuo-Hsiang Chen. How different information types affect viewer's attention on internet advertising. *Computers in Human Behavior*, 27(2):935–945, 2011. doi:10.1016/j.chb.2010.11.019.

[70] Yu-Chen Hsieh, Kuo-Hsiang Chen, and Min-Yuan Ma. Retain viewer's attention on banner ad by manipulating information type of the content. *Computers in Human Behavior*, 28(5):1692–1699, 2012. doi:10.1016/j.chb.2012.04.008.

[71] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002. doi:10.1109/72.991427.

[72] Shih-Miao Huang. A study of reading time and viewers' preferences for a variety of combinations of character-background chromaticity for small traditional chinese characters. *Perceptual and motor skills*, 103(3):887–895, 2006.

[73] Iztok Humar, Mirko Gradis, et al. The impact of color combinations on the legibility of a web page text presented on crt displays. *International journal of industrial ergonomics*, 38(11):885–899, 2008.

[74] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

[75] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001. doi:10.1038/35058500.

[76] Laurent Itti, Christof Koch, and Ernst 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. doi:10.1109/34.730558.

[77] Robert JK Jacob. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 11–18. ACM, 1990.

[78] Ananya Jana and Samit Bhattacharya. Design and validation of an attention model of web page users. *Advances in Human-Computer Interaction*, 2015:1, 2015. doi:10.1155/2015/373419.

[79] Radu Jianu and Sayeed Safayet Alam. A data model and task space for data of interest (doi) eye-tracking analyses. *IEEE Trans. Vis. Comput. Graphics*, 24(3):1232–1245, 2018. doi:10.1109/TVCG.2017.2665498.

[80] Marcel A Just and Patricia A Carpenter. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329, 1980. doi:10.1037/0033-295X.87.4.329.

[81] Young Sik Kang and Heeseok Lee. Understanding the role of an it artifact in online service continuance: An extended perspective of user satisfaction. *Computers in Human Behavior*, 26(3):353–364, 2010. doi:10.1016/j.chb.2009.11.006.

[82] Ziho Kang and Steven J Landry. An eye movement analysis algorithm for a multielement target tracking task: Maximum transition-based agglomerative hierarchical clustering. *IEEE Transactions on Human-Machine Systems*, 45(1):13–24, 2015.

[83] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL http://www.jstatsoft.org/v11/i09/.

[84] David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961, 2011. URL http://dl.acm.org/citation.cfm?id=2986459.2986677.

[85] David R Karger, Sewoong Oh, and Devavrat Shah. Efficient crowdsourcing for multi-class labeling. *ACM SIGMETRICS Performance Evaluation Review*, 41(1):81–92, 2013. doi:10.1145/2465529.2465761.

[86] Keith S Karn. "saccade pickers" vs."fixation pickers": the effect of eye tracking instrumentation on research. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, pages 87–88, New York, NY, USA, 2000. ACM. doi:10.1145/355017.355030.

[87] Dan Kelley and Clark Richards. *oce: Analysis of Oceanographic Data*, 2019. URL https://CRAN.R-project.org/package=oce. R package version 1.1-1.

[88] Conor Kelton, Jihoon Ryoo, Aruna Balasubramanian, and Samir R. Das. Improving user perceived page load time using gaze. In *Proceedings of the 14th USENIX Conference on Networked Systems Design and Implementation*, NSDI'17, pages 545–559, Berkeley, CA, USA, 2017. USENIX Association. ISBN 978-1-931971-37-9. URL http://dl.acm.org/citation.cfm?id=3154630.3154675.

[89] Raymond M Klein and Jason Ivanoff. Inhibition of return. *Scholarpedia*, 3(10):3650, 2008.

[90] Oleg V Komogortsev, Denise V Gobert, Sampath Jayarathna, Do Hyong Koh, and Sandeep M Gowda. Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, 57(11):2635–2645, 2010.

[91] Alan J Laub. *Matrix analysis for scientists and engineers*. Siam, 2005. ISBN 978-0-89871-576-7. doi:10.1137/1.9780898717907.

[92] Hyeon Chang Lee, Duc Thien Luong, Chul Woo Cho, Eui Chul Lee, and Kang Ryoung Park. Gaze tracking system at a distance for controlling iptv. *IEEE Transactions on Consumer Electronics*, 56(4), 2010.

[93] JooWon Lee and Jae-Hyeon Ahn. Attention to banner ads and their effectiveness: an eye-tracking approach. *International Journal of Electronic Commerce*, 17(1):119–137, 2012.

[94] Jian Li, Li Su, Bo Wu, Junbiao Pang, Chunfeng Wang, Zhe Wu, and Qingming Huang. Webpage saliency prediction with multi-features fusion. In *IEEE International Conference on Image Processing (ICIP)*, pages 674–678, Phoenix, AZ, USA, Sept 2016. IEEE. doi:10.1109/ICIP.2016.7532442.

[95] Hsuan Lin, Yu-Chen Hsieh, and Fong-Gong Wu. A study on the relationships between different presentation modes of graphical icons and users' attention. *Computers in Human Behavior*, 63:218–228, 2016. doi:10.1016/j.chb.2016.05.008.

[96] Jonathan Ling and Paul van Schaik. The influence of line spacing and text alignment on visual search of web pages. *Displays*, 28(2):60–67, 2007. doi:10.1016/j.displa.2007.04.003.

[97] Sebastian Linxen, Silvia Heinz, Livia J. Müller, Alexandre N. Tuch, and Klaus Opwis. Mental models for web objects in different cultural settings. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 2557–2562, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2474-8. doi:10.1145/2559206.2581209.

[98] Cewu Lu, Li Xu, and Jiaya Jia. Real-time contrast preserving decolorization. In *SIGGRAPH Asia 2012 Technical Briefs*, page 34, New York, NY, USA, 2012. ACM. doi:10.1145/2407156.2407174.

[99] Yu-Fei Ma and Hong-Jiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proc. 11th ACM Int. Conf. Multimed.*, pages 374–381. ACM, 2003. doi:10.1145/957013.957094.

[100] Christopher M Masciocchi and Jeremiah D Still. Alternatives to eye tracking for predicting stimulus-driven attentional selection within interfaces. *Human–Computer Interaction*, 28(5):417–441, 2013. doi:10.1080/07370024.2012.731332.

[101] Marcus Mast and Michael Burmester. Exposing repetitive scanning in eye movement sequences with t-pattern detection. *IADIS IHCI*, 11:137–145, 2011.

[102] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012. doi:10.11613/BM.2012.031.

[103] Tom M Mitchell et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37): 870–877, 1997.

[104] A. H. Munsell. A pigment color system and notation. *The American Journal of Psychology*, 23(2):236–244, 1912. ISSN 00029556. doi:10.2307/1412843. URL http://www.jstor.org/stable/1412843.

[105] Jakob Nielsen. F-shaped pattern for reading web content. Alertbox: Current Issues in Web Usability (retrieved on 13 May 2015 from https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content-discovered/), 2006.

[106] EAB Over, ITC Hooge, BNS Vlaskamp, and CJ Erkelens. Coarse-to-fine eye movement strategy in visual search. *Vision Research*, 47(17):2272–2280, 2007.

[107] Manos Papagelis and Dimitris Plexousakis. Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents. *Engineering Applications of Artificial Intelligence*, 18(7):781–789, 2005. doi:10.1016/j.engappai.2005.06.010.

[108] Siegmund Pastoor. Legibility and subjective preference for color combinations in text. *Human Factors*, 32(2):157–171, 1990. doi:10.1177/001872089003200204.

[109] Robert Pearson and Paul van Schaik. The effect of spatial layout of and link colour in web pages on performance in a visual search task and an interactive search task. *International Journal of Human-Computer Studies*, 59(3):327–353, 2003. doi:10.1016/S1071-5819(03)00045-4.

[110] Kara Pernice. F-shaped pattern of reading on the web: Misunderstood, but still relevant (even on mobile). retrieved on 14 July 2019 from https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/, 2017.

[111] Rik Pieters and Luk Warlop. Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of Research in Marketing*, 16(1):1–16, 1999. doi:10.1016/S0167-8116(98)00022-6.

[112] Rik Pieters and Michel Wedel. Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of Marketing*, 68(2):36–50, 2004. doi:10.1509/jmkg.68.2.36.27794.

[113] Alex Poole and Linden J Ball. Eye tracking in hci and usability research. In *Encyclopedia of Human Computer Interaction*, pages 211–219. IGI Global, 2006. doi:10.4018/978-1-59140-562-7.ch034.

[114] Philipp Probst, Quay Au, Giuseppe Casalicchio, Clemens Stachl, and Bernd Bischl. Multilabel classification with r package mlr. *arXiv preprint arXiv:1703.08991*, 2017.

[115] Qingzhu, Prasenjit Dey, and Parvez Ahammad. Perceived performance of top retail webpages in the wild. *SIGCOMM Comput. Commun. Rev.*, 47(5):42–47, October 2017. ISSN 0146-4833. doi:10.1145/3155055.3155062.

[116] Laurence R. Young and David Sheena. Survey of eye movement recording methods. *Behavior Research Methods & Instrumentation*, 7:397–429, 09 1975.

[117] Kari-Jouko Räihä. Some applications of string algorithms in human-computer interaction. In *Algorithms and Applications*, pages 196–209. Springer, 2010.

[118] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proc. 18th ACM Int. Conf. Multimed.*, pages 251–260. ACM, 2010. doi:10.1145/1873951.1873987.

[119] Nikhil Rasiwasia, Dhruv Mahajan, Vijay Mahadevan, and Gaurav Aggarwal. Cluster canonical correlation analysis. In *AISTATS*, pages 823–831, 2014.

[120] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998. doi:10.1037//0033-2909.124.3.372.

[121] Keith Rayner, Caren M Rotello, Andrew J Stewart, Jessica Keir, and Susan A Duffy. Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of experimental psychology: Applied*, 7(3):219, 2001. doi:10.1037/1076-898X.7.3.219.

[122] Keith Rayner, Xingshan Li, Carrick C Williams, Kyle R Cave, and Arnold D Well. Eye movements during information processing tasks: Individual differences and cultural effects. *Vision research*, 47(21):2714–2726, 2007.

[123] Keith Rayner, Brett Miller, and Caren M Rotello. Eye movements when looking at print advertisements: The goal of the viewer matters. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22 (5):697–707, 2008. doi:10.1002/acp.1389.

[124] Matías Recabarren and Miguel Nussbaum. Exploring the feasibility of web form adaptation to users' cultural dimension scores. *User Modeling and User-Adapted Interaction*, 20(1):87–108, 2010. doi:10.1007/s11257-010-9071-7.

[125] Evan F Risko, Nicola C Anderson, Sophie Lanthier, and Alan Kingstone. Curious eyes: Individual differences in personality predict eye movement behavior in scene-viewing. *Cognition*, 122(1):86–90, 2012. doi:10.1016/j.cognition.2011.08.014.

[126] Edward Rosbergen, Rik Pieters, and Michel Wedel. Visual attention to advertising: A segment-level analysis. *Journal of consumer research*, 24(3):305–314, 1997.

[127] Robert E Roth. An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. *IEEE transactions on visualization and computer graphics*, 19(12):2356–2365, 2013. doi:10.1109/TVCG.2013.130.

[128] Dario D Salvucci and Joseph H Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000. doi:10.1145/355017.355028.

[129] Vidyapu Sandeep, V Vijaya Saradhi, and Samit Bhattacharya. From multimodal to unimodal webpages for developing countries. In *ML4D, NeurIPS (NIPS), arXiv preprint arXiv:1711.02068*, 2017.

[130] Brahima Sanou. *ICT facts and figures 2017 (retrieved on 14 July 2018 from https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx)*. ITU, 2017.

[131] Barry Schwartz. Google: We know about 30 thousand trillion urls on the web but... (june 3, 2015), 2015. URL https://www.seroundtable.com/google-30-trillion-urls-20387.html.

[132] Michael W Schwarz, William B Cowan, and John C Beatty. An experimental comparison of rgb, yiq, lab, hsv, and opponent color models. *ACM Transactions on Graphics (TOG)*, 6(2):123–158, 1987. doi:10.1145/31336.31338.

[133] Wei Shan, Guangling Sun, Xiaofei Zhou, and Zhi Liu. Two-stage transfer learning of end-to-end convolutional neural networks for webpage saliency prediction. In *International Conference on Intelligent Science and Big Data Engineering*, pages 316–324, Cham, Switzerland, 2017. Springer. doi:10.1007/978-3-319-67777-4_27.

[134] Chengyao Shen and Qi Zhao. Webpage saliency. In *Computer Vision–ECCV 2014*, pages 33–46. Springer, Cham, Switzerland, 2014. doi:10.1007/978-3-319-10584-0_3.

[135] Chengyao Shen, Xun Huang, and Qi Zhao. Predicting eye fixations on webpage with an ensemble of early features and high-level representations from deep network. *IEEE Trans. Multimedia*, 17(11):2084–2093, 2015. doi:10.1109/TMM.2015.2483370.

[136] Y. Shin, K. Choi, S. Kim, and S. Ko. A novel single ir light based gaze estimation method using virtual glints. *IEEE Transactions on Consumer Electronics*, 61(2): 254–260, May 2015. ISSN 0098-3063. doi:10.1109/TCE.2015.7150601.

[137] Sav Shrestha and Kelsi Lenz. Eye gaze patterns while searching vs. browsing a website. *Usability News*, 9(1):17, 2007.

[138] Linda E Sibert and Robert JK Jacob. Evaluation of eye gaze interaction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 281–288. ACM, 2000.

[139] Sidney Siegel. *Nonparametric statistics for the behavioral sciences.* McGraw-hill, 1956.

[140] Alvy Ray Smith. Color gamut transform pairs. *ACM Siggraph Computer Graphics*, 12 (3):12–19, 1978. doi:10.1145/800248.807361.

[141] César R Souza. Kernel functions for machine learning applications. *Creative Commons Attribution-Noncommercial-Share Alike*, 3:29, 2010.

[142] Jeremiah D Still. Web page attentional priority model. *Cognition, Technology & Work*, 19(2-3):363–374, 2017. doi:10.1007/s10111-017-0411-9.

[143] Jeremiah D Still. Web page visual hierarchy: Examining faraday's guidelines for entry points. *Computers in Human Behavior*, 84:352–359, 2018. doi:10.1016/j.chb.2018.03.014.

[144] Jeremiah D Still and Christopher M Masciocchi. A saliency model predicts fixations in web interfaces. In *5th International Workshop on Model Driven Development of Advanced User Interfaces (MDDAUI 2010)*, page 25, Atlanta, Georgia, USA, 2010. Citeseer, ACM. URL http://ceur-ws.org/Vol-617/.

[145] Boleslaw Szewczyk. Rendering approximate webpage screenshot client-side, Jul 2012. US Patent 8,225,197.

[146] Barbara G Tabachnick, Linda S Fidell, and Steven J Osterlind. *Using Multivariate Statistics*. Allyn and Bacon Boston, 2001.

[147] Haruhiko Takeuchi and Yoshiko Habuchi. A quantitative method for analyzing scan path data obtained by eye tracker. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*, pages 283–286. IEEE, 2007.

[148] Shuto Tamura and Keiko Sato. Age-related differences in visual search for color targets manipulated based on cone-contrast model. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2992–2997, Miyazaki, Japan, 2018. IEEE, IEEE. doi:10.1109/SMC.2018.00508.

[149] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*. R Foundation for Statistical Computing, 2018. URL `https://CRAN.R-project.org/package=rpart`. R package version 4.1-13.

[150] Robert M Thorndike. Canonical correlation analysis. *Applied multivariate statistics and mathematical modeling*, pages 237–263, 2000. doi:10.1016/B978-012691360-6/50010-0.

[151] Hoi Ying Tsang, Melanie Tory, and Colin Swindells. eseetrack&# 8212; visualizing sequential fixation patterns. *IEEE Trans. Vis. Comput. Graphics*, 16(6):953–962, 2010.

[152] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007. doi:10.4018/jdwm.2007070101.

[153] Nicholas Vanderschantz, Claire Timpany, David Whitehead, and Wendy Diane Carss. A small scale study into the effect that text & background colour has on processing and self-correction rates for childrens' on-screen reading. *International Journal of the Book*, 7(4):75–88, 2010. doi:10.18848/1447-9516/CGP/v07i04/36841.

[154] Sandeep Vidyapu, V Vijaya Saradhi, and Samit Bhattacharya. Fixation-indices based correlation between text and image visual features of webpages. In *Proc. 2018 Symp. Eye Track. Res. & Appl.*, ETRA '18, pages 53:1–53:5, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5706-7. doi:10.1145/3204493.3204566.

[155] Sandeep Vidyapu, Vijaya Saradhi Vedula, and Samit Bhattacharya. Quantitative visual attention prediction on webpage images using multiclass svm. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, pages 90:1–90:9, NY, USA, 2019. ACM. ISBN 978-1-4503-6709-7. doi:10.1145/3317960.3321614.

[156] Jingwei Wang, Ali Borji, C-C Jay Kuo, and Laurent Itti. Learning a combined model of visual saliency for fixation prediction. *IEEE Transactions on Image Processing*, 25 (4):1566–1579, 2016. doi:10.1109/TIP.2016.2522380.

[157] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[158] Michel Wedel and Rik Pieters. A review of eye-tracking research in marketing. In *Review of marketing research*, pages 123–147. Emerald Group Publishing Limited, 2008. doi:10.1108/S1548-6435(2008)0000004009.

[159] Julia M West, Anne R Haake, Evelyn P Rozanski, and Keith S Karn. eyepatterns: software for identifying patterns and similarities across fixation sequences. In *Proc. 2006 Symp. Eye Track. Res. & Appl.*, pages 149–154. ACM, 2006.

[160] Samuel S Wilks. On the independence of k sets of normally distributed statistical variables. *Econometrica*, pages 309–326, 1935. URL https://www.jstor.org/stable/1905324.

[161] Ben J Winer. *Latin squares and related designs*. McGraw-Hill Book Company, New York, NY, US, 1962. doi:10.1037/11774-000.

[162] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016. doi:10.1016/c2009-0-19715-5.

[163] Seong Wook Chae and Kun Chang Lee. Exploring the effect of the human brand on consumers' decision quality in online shopping: An eye-tracking approach. *Online Information Review*, 37(1):83–100, 2013. doi:10.1108/14684521311311649.

[164] Chaitra Yangandul, Sachin Paryani, Madison Le, and Eakta Jain. How many words is a picture worth?: attention allocation on thumbnails versus title text regions. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, page 3, New York, NY, USA, 2018. ACM. doi:10.1145/3204493.3204571.

[165] Panayiotis Zaphiris and Sri Kurniawan. *Human Computer Interaction Research in Web Design and Evaluation*. IGI Global, 2007. ISBN 1599042460. doi:10.4018/978-1-59904-246-6.

[166] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on knowledge and data engineering*, 26(8):1819–1837, 2014. doi:10.1109/TKDE.2013.39.