Implicit image annotation by using gaze analysis

Hajimirza, S. Navid

Implicit Image Annotation by using Gaze Analysis

S. Navid HAJIMIRZA

Queen Mary
University of London

Department of Electronic Engineering and Computer Science

PhD Thesis
London 2012

ABSTRACT

Thanks to the advances in technology, people are storing a massive amount of visual information in the online databases. Today it is normal for a person to take a photo of an event with their smartphone and effortlessly upload it to a host domain. For later quick access, this enormous amount of data needs to be indexed by providing metadata for their content. The challenge is to provide suitable captions for the semantics of the visual content. This thesis investigates the possibility of extracting and using the valuable information stored inside human's eye movements when interacting with digital visual content in order to provide information for image annotation implicitly. A non-intrusive framework is developed which is capable of inferring gaze movements to classify the visited images by a user into two classes when the user is searching for a Target Concept (TC) in the images. The first class is formed of the images that contain the TC and it is called the TC+ class and the second class is formed of the images that do not contain the TC and it is called the TC- class. By analysing the eye-movements only, the developed framework was able to identify over 65% of the images that the subject users were searching for with the accuracy over 75%. This thesis shows that the existing information in gaze patterns can be employed to improve the machine's judgement of image content by assessment of human attention to the objects inside virtual environments.

Acknowledgement

First and foremost, I would like to express my gratitude to my supervisor, Professor Ebroul Izquierdo, for his continuous support and invaluable advices which have helped me reach my full potential.

My special thanks go to Dr. Michael Proulx, my co-supervisor, for his friendship, guidance, and constructive feedback throughout my PhD.

I would like to thank the Electronic Engineering Department at Queen Mary, University of London, my colleagues in the Multimedia and Vision Group whom I had the pleasure to share the last few years with.

This research has also been conducted within the framework of the European Commission funded Network of Excellence PetaMedia.

I am grateful to my beloved parents and brother for their on-going support and encouragement.

I would like to give a very special thanks to my beloved wife Parisa for her patience, motivation and unconditional love which has enabled me to complete this work.

Lastly I would like to say a special thanks to my bundle of joy, my baby daughter Elana who has brought light in to my life.

Table of Contents

List of Figures

List of Tables

Nomenclature

AoI: Area of Interest

$\boldsymbol{\alpha}$**:** Lagrange multipliers in SVM

b: Vector of the hyperplane for SVM

C: Set of coefficients to be found for a Least Square Estimation system

D: Set of input data points with given corresponding outputs for inputs

δ_j: Error of node j in neural networks

E: Error of the system

$\bar{\varepsilon}$: A coefficient to specify the stop condition of the algorithm

F1: F1 measure

FIS: Fuzzy Inference System

FV: Feature Vector

fp: False Positive - The images that belonged to the TC+ class but they were classified mistakenly as TC-fn: False Negative - The images that belonged to the TC- class but they were classified mistakenly as TC+

K<x,y>: Kernel function of SVM

L: Lagrangian

LSE: Least Square Estimation

MF: Membership Function

NN: Neural Networks

$\mu_A(x)$: Membership degree of x in relation to the set A

γ: Learning rate that has the effect on the convergance speed and stability of the weights during the learning process of a NN

pr: Precision

P_n^k: The k-th calculated potential value of being a cluster centre for the n-th point

P_n^*: Potential of then-th cluster centre respectively

r_a: The influence factor of the cluster centre

r_b: A positive constant that defines the radius of the neighbourhood that the effective potential reduction happens in subtractive clustering

rc: recall

RVM: Relevance Vector Machines

ρ: SVM margin

Sail: Saliency

SP: Scan Path

Simm: Similarity of the main concept

Simb: Similarity of the background

SVM: Support Vector Machines

TC: Target Concept

TC+: Images that belong to the Target Concept class

TC-: Images that do not belong to the Target Concept class

tp: True Positive - The images that belonged to the TC+ class and they were classified correctly as TC+

tn: True Negative - The images that belonged to the TC- class and they were classified correctly as TC-

TSK-FIS: Takagi Sugeno Kang Fuzzy Inference System

ω_c: Firing strength of the c-th rule in fuzzy logic

ω: Vector of the hyper plane in SVM

X^*: Set of found cluster centres by subtractive clustering algorithm

$\mathrm{x_n}|_{\mathrm{in}}$: The n-th input data point for training an intelligent system.

$\mathrm{x_n}|_{\mathrm{out}}$: The n-th output data point for training an intelligent system

1 Introduction

1.1 Problem and motivation

Location tags and low level features including colour and edge histograms are examples of the metadata that are currently stored along with the visual content. These metadata are extracted automatically by the machines without any human contribution. However there still remains the challenge to provide appropriate captions for the contents of visual contents like images. This problem stems from the Semantic gap which is the difference between the understanding of human and machine from the contents of the images. In order to bridge this gap there needs to be a common ground for both that in collaboration the machines can achieve a closer interpretation of images to that of the humans. At the current stage of the technology this is not possible without the contribution of human in the process of interpretation of the visual content semantics for machine. However this is an expensive and time consuming task.

In order to exploit the human knowledge for the benefit of the machines in understanding the visual concepts of the images in an inexpensive manner (both financially and cost of time) the implicit methods are developed. In these methods the machine gets the help from the human but the human does not elaborate to the task. For example when a person is searching for a specific image of a car, they are constantly providing different forms of feedbacks to the retrieved images by clicking on them, looking at them, ignoring them etc. The goal of implicit annotation is for the machines to collect these data for better understanding of the images that the user had direct or indirect interaction with.

This study investigates the possibility of extracting and using the valuable information stored in human's eye movements when interacting with visual digital data in order to annotate images implicitly. In this thesis, the development of a non-intrusive framework is described which is capable of inferring human eye movements in order to classify the visited images by the user into two classes. One class represents the images that the user's eye movements reveal that the user is searching for them and the other class represents the images that the user's eye movements do not reveal any specific perceptual information about them.

1.2 Research objectives

As a bridge that connects the two sides (human and machine) of the semantic gap, this research tries to find a non-intrusive solution for annotating huge image databases in a fast and truthful manner implicitly by interpreting the human eye-movements and separating the images that the user was interested in them from the rest of the visited images. As a result the following objectives are addressed in this study:

1) To find an approach for empirically explanation of one's attention to an image on the screen in form of gaze features
2) To discover and extract visual features as a form of description of gaze movements that can be interpreted by machine for the task of classification
3) To study the effect of different factors (i.e. saliency of an image on the screen and its similarity to other images) on one's visual attention that could affect the gaze features
4) To tailor the best scenario for the experiment that records the eye-movements for the purpose of this study to classify the images based on gaze data into two different classes 1) the images that the user is searching for 2) The images that the user is not looking for them
5) To select the best feature set for the classification purpose

1.3 Thesis Structure:

Chapter 2 - State of the Art

In this chapter first a detailed review of the current sate-of-the-art of image annotation is provided. It is discussed how implicit models can contribute to make the process cost effective, fast and precise. Next a brief history of gaze-tracking, different methods of eye-tracker development and implementations of the technology in different fields are presented. Then the nature of human eye and different eye movements are introduced. Finally the state-of-the-art of this thesis is discussed.

Chapter 3 - Background theory

This chapter briefly introduces the mathematical models and algorithms that are used in this research to process the visual perception data. This includes, Least Square Estimation, Back propagation Neural Networks, Support Vector Machines, Fuzzy expert systems, ANOVA, Correlation Coefficient, etc.

Chapter 4 - Framework Structure

In this chapter first it is explained how gaze-tracking technology gives us the access to visual perception and how it can be used for implicit image annotation. Next there is an introduction to the eye-tracking equipment that is used in this research and the experiments designed to investigate gaze behaviour. The developed user interface is presented with detailed review of the scenarios that the users experiences. Finally the general structure of the processing system that analyses the gaze data is illustrated.

Chapter 5 - Gaze Feature Analysis

This chapter is dedicated for discussion of gaze features. Two feature vectors, namely Scan Path and Area of Interest are explained in detail. All of the features in these vectors are discussed and it is explained how their data are extracted from the raw gaze movement data. By using ANOVA test it is shown that how the values of these features can be affected by the saliency, similarity and Target Concept[*] class factors. Moreover the results of finding correlation coefficient of every feature with each of the three factors are illustrated which shows the grade by which they can affect the gaze features.

Chapter 6 - Results

This chapter embodies the performance of different classifiers in order to distinguish the Target Concept in a scenario from the rest of the images. Also the stored information in different feature is studied to choose the best subset from the investigated features.

[*] Target Concept (TC) is the key concept (i.e. a specific animal like lion, an object like Car, etc.) that the user is assigned to searching for during an experiment. This concept appears in a fraction of the images that are shown to the user during the experiment. Once a user is assigned to search for a TC, the TC does not change for the rest of the experiment. The images that contain the Target Concept belong to the Target Concept class and are denoted as TC+ images.

2 State of the Art

At the time of writing this thesis 300 million photos are uploaded to Facebook everyday with 526 million daily active users [**1**]. At peak times 28 photos per second (more than 100,000 per Hour) are uploaded to Flickr only [**2**]. Because of the massive volume of the media that are stored in their databases and the diverse queries that they receive for retrieval, it is impossible for media storage companies to hire people for indexing purposes and at the current stage of technology the machines are still incapable of understanding the high level semantics (human perception level) of the contents of the media. To improve their process of indexing and retrieval, the providers of media browsing facilities are looking constantly for new and innovative methods which can give them competitive advantage in the massive market of multimedia. One approach is to use the power of the population that stores these media in order to annotate their own content. In this study we investigate the possibility of using the eye-movement of a normal media user for measurement of their interest in the visual content and exploiting it for the mentioned purposes.

The recorded history of eye-tracking goes back to the 19s century when the developed methods were intrusive and based on ocular observations [**3**] [**4**] mostly to study the reading process [**5**] [**6**]. Nowadays, the eye-tracking technology has advanced to non-invasive products that can identify one’s gaze point in 3D space [**7**] [**8**] just by monitoring their eyes with video cameras. In [**9**] [**10**] [**11**] [**12**] [**13**] [**14**] some application of eye-tracking technology can be found. These show that this new technology is opening its way from the laboratories into industrial applications where it can be used in everyday life of a person.

2.1 A literature review of implicit annotation and gaze movement as implicit feedback

Image annotation [**15**] , also called Image tagging, is one of the methods that helps the indexing process. It is a process for assigning metadata, formed of captions and keywords, to the images in the databases [**16**] [**17**]. The provided metadata are used later for quick access to the images for retrieval in response to a search query [**18**]. By increasing the size of the visual databases specifically in distributed environments the necessity has risen to annotate and organize them with an undemanding, inexpensive and truthful approach. There are three approaches for image annotation [**19**] [**20**]:

1- Manual image annotation is when humans are hired to do the annotation job. This approach is accurate but expensive, time consuming and exhaustive. LabelMe [**21**] is an example of such a method.

2- Automatic image annotation [**22**] [**23**] is when computers are used to do the annotation job. This approach is prompt and cheap but accuracy remains an important issue.

3- Semi-Automatic image annotation [**24**] [**25**] is performed by interaction between human and computer with a higher accuracy than automatic method and cheaper approach than manual method.

If the annotation is automatic [**26**] [**27**] [**28**] [**29**] [**30**] the machines conduct the process by inspection of the low-level features of the images [**19**] Their classification [**31**] and optimization of the classification results. However, regardless of how well the classification is performed, the semantic gap [**32**] [**33**] [**34**] remains a problem. This is the gap between machine's understandings and human's understandings of the information and the concepts that the images store. As a result, at the current state of the technology the human contribution to the process of annotation seems to be unavoidable. Consequently different studies started to investigate the semi-automatic algorithms [**35**].

These methods are normally formed of extensive machine processes accompanied by human judgment [**19**]. It is called implicit image annotation in case the feedback of the human factor is provided unconsciously [**36**] [**37**] [**38**].

Implicit semi-automatic annotation techniques, unlike the other semi-automatic annotation methods, do not need the users to spend any effort and pay any attention to the process of the annotation. This is the responsibility of the machines to monitor people reactions and interaction when they come across visual data in their normal life and annotate or classify them accordingly. Implicit image annotation by playing games [**39**] [**40**], monitoring the brain waves by Electroencephalography (EEG) [**41**] and studying users' gaze movement* [**42**] [**43**] [**44**] [**45**] can be mentioned as three approaches in this category which are currently under research. In [**46**] Ntalianis et al. combined the clickthrough data gained from informed users with visual concept models to annotate images implicitly. Compared to use of games, eye tracking is more feasible as it can be used in normal life every day and there is no need to motivate the users for using the interface. Furthermore eye tracking is less intrusive than EEG, in that the equipment does not have to be in contact with the user, thus allowing the user to behave more normally.

As humans we use selective visual attention [**47**] in order to interact with our environment which holds an unlimited amount of visual information. Psychologists have divided our attention to these visual contents into top-down and bottom-up models [**48**]. The former is the deliberate variation in attention's focus and the latter is unintentional shift in attention [**49**] due to the appearance factors that make a gaze target emergent compared to its surroundigs.

Most of the studies that employ the eye-tracking technology in the image processing field focus on image retrieval [**50**]. For example by using the ranking SVM model Pasupa et al. [**13**] tried to improve the online learning of the ranking images for retrieval. This method tried to combine the content base features of the images and the eye-movement data for

* Overt attention as indexed by eye movements reveals what one is looking for [42] , and thus can be used as a means of implicitly annotating images

this purpose. Later in another study Hardoon and Pasupa [**11**] found it “unrealistic as no eye movement will be presented a-priori for new images”. Consequently they tried to improve their model by training their system with the images that had gaze data and just used the low level features of the images to rank them for retrieval. Later, Auer et al. [**51**] used the finding of the previous two approaches for their system, Pinview, which combines the gaze feedbacks to the mouse clicks as a form of relevance feedback for the prupose of Content-Based Image retrieval.

Buscher et al. [**52**] tried to automatically predict relevance from gaze movements. In their approach they proposed a method based on gaze data which is able to differentiate between the situations when a person is reading and a person is skimming a piece of text. Furthermore based on the results they tried to determine the relevance of a document. Later in [**53**] they showed that by personalising the gaze measures a clear relationship can be defined between them and the relevance of the documents. In addition they used gaze-based feedback to personalise text search and showed that this personalisation can improve the web search experience.

In [**54**] Faro et al. proposed a method for implicit relevance feedback to improve the results of Content Based Image Retrieval (CBIR) systems by using gaze data. In they approached they used eye movements to re-rank the retrieved images. In their proposed method they computed the most salient regions of the retrieved images by montitoring the users’ gaze movement. Their presented work shows that 87% of their users are more satisfied when the retrieved images are ranked again by gaze data. Zhang et al. [**55**] developped an experiment that the users corrected image tags. Like [**54**] they tried to improve the CBIR performance using gaze data. In their experiment they added speech recognition as another form of implicit feedback from the users. They concluded that there is a potential for improvement in the CBIR performance for some users by integrating CBIR with gaze movement and speech recogniction.

In [**56**], Florian et al. used gaze information to develop user adaptive web content. In their application the web content is customizable based on the gaze data including length of fixation, number of fixation and dwell time. In their approach each web page is produced based on analysis of the gaze movements for the previous page. They showed that there is a significant increase in attention of the users using their framework compared to the randomly chosen content.

Klami et al. [**57**] developed a framework that demonstrates whether a screen that shows four images, contains a Target Concept and if the answer is yes, which of the images belong to the TC+ class. They used Linear Discriminant Analysis for this purpose. Kozma et al. [**12**] developed a gaze base interface for image browsing and search, called GaZIR, by using Logistic Regression model. This is a scenario dependant framework where 5 features out of the 17 are limited to the GaZIR interface and cannot be used in other scenarios.

Pantic and Vinciarelli in [**58**] investigated Implicit Human-Cantered Tagging (IHTC) that exploits the information from user feedbacks that are not provided by the user verbally while interacting with multimedia. This includes "facial expressions like smiles or head gestures like shaking". They indicated that this is in contrast with explicit tagging paradigm in which a visual content is tagged if the user elaborated to the tag intentionally. They stated that these IHTC tags are expected to be more robust from the explicitly provided tags in terms of generality and statistical reliability. In [**58**] they reviewed implicit tagging and its purposes such as assessing the correctness of explicit tags, assigning new tags and user profiling (e.g. "based on the user's implicit feedback, an implicit tag could be associated with this data indicating that the user in question favours less a particular website, facilitating lower ranking of the target data next time").

Soleymani and Pantic in [**59**] reviewed the definition and applications of implicit human-cantered tagging. They reviewed the publicly available relevant databases and annotation methods for implicit annotation and the challenges in the field. In this article they discussed the techniques for gathering implicit annotation based on implicit reactions. They studied that how different state-of-the-art techniques build a ground truth database and the challenges they meet. This includes the discussion that emotional self-reporting is done in free-response or forced response formats where in the former the participants are free to express their emotions in their own words and in the latter format they are asked specific questions regarding their emotions. Furthermore they introduce four publicly available databases in [**59**] for the purpose of HCIT namely MAHNOB HCI, DEAP, Pinview and LDOS-PerAff-1.

Recently Koelstra and Patras [**60**], presented a multi-modal approach which tries to generate affective tags based on facial features and electroencephalography (EEG) signals. Their main focus was “classification and regression in the valance and arousal space”. In their research they showed that the both modalities have complementary effect on each other where it is possible to improve the results when using both of them for affective annotation.

2.2 A review of eye-tracking technology

The history of eye-tracking goes back to the late 1800s where mechanical uncomfortable eye-trackers used a bite-bat to ensure the participants do not move their heads. Even in some experiments anaesthetise were used on the eyeballs in order to attach a Paris ring to the eyes [**3**].

In 1970s Yarbus et a [**6**] developed the first eye trackers that could record detailed and precise eye movements. For a long time electromagnetic coil systems were considered to be the most accurate measurements of eye-movements [**3**] [**61**] [**62**]. In this method the eye movements were measured by placing a silicon contact lenses plate in the anaesthetized eye and measuring the electromagnetic induction [**63**]. In early times of the last decade a trend started to build eye-trackers by advance image processing techniques and some scientists started to use the reflection of the Infra-Red light to detect the position of the pupils [**64**]. In 2004, Yu et al. [**65**] used a head mounted eye-tracker to reconstruct the gaze-point based on four point correspondence in two view. Zhu et al [**66**] tried to solve the problem of the low tolerance of the eye-trackers to the head movements. They developed a dynamic computational head compensation model to remove the effect of head movements as a result the old chin-rests could be removed from the eye-trackers and there was no need for head-mounted eye-trackers anymore which made them non-intrusive equipment.

As the computational power of the computers increased the cost of eye-trackers [**67**] kept decreasing while their efficiency started improving. Magee et al. [**68**] developed EyeKeys, a low cost eye-tracker, which is a non-intrusive interface that can run on a normal computer using a USB camera independent from the lighting condition. Their system detects a subject's face using multi-scale template correlation and uses the symmetry between both eyes to check the direction that the person is looking. Hennessey and Lawrence [**69**] used a binocular set of cameras in order to estimate the point-of-gaze (POG) of a subject in 3-D environment by employing the vergence of the eyes. Rantanen et al. [**70**] built a light-weight, wearable and wireless eye-tracker that integrated the video-based gaze tracking with capacitive facial movement detection which can send commands by the combination of gaze point and face gestures.

The calibration phase has been a part of the most of the eye-trackers which deliver high precision. Normally in this phase every new user is asked to look at multiple reference points in the space or 9 points on the screen (3x3 starting from top left corner to bottom right corner) in case the experiment involved visual objects on screen. Villanueva et al. [**71**] developed a system with just one calibration point by improving the geometric analysis of the pupils an eyeballs. However they still insist that by increasing the calibration points one can achieve less average error.

Generally the modern eye-trackers are formed of a set of cameras that record the eye-events and send the resulting video images to a separate processing unit in order to extract the gaze direction. However recently Kim and Han [**72**] developed "a smart sensor for an eye tracker using pixel-level analogue image processing" that with improvement in efficiency, it performs a part of the image processing task which is computationally expensive and has a high power consumption.

At the current pace of the technology it is not unreal to consider a day that almost all of the digital devices that need to interact with visual attention of human subjects will be equipped with eye-trackers[*]. As a result, inspired from the psychological studies in the

[*] Recently Samsung has released its latest smartphone (Galaxy S III) product armed with an eye-tracker to prevent the phone from going to the energy saving sleep mode while there exist eyes watching it.

topics of eye-tracking, gaze-movement and visual attention [**73**] [**74**] [**75**] [**76**] [**77**] [**78**] [**79**] , the researchers in the multimedia and human computer interaction fields started to develop numerous applications for this recently grown technology [**80**] [**81**] [**82**] [**83**] [**84**] [**85**] [**86**].

The eye movements:

Table 1: Fixation duration and saccade size in regular visual activities [5]

Task	Mean Fixation Duration (ms)	Mean saccade size (degree)
Silent Reading	225	2(8 Letters)
Oral Reading	275	1.5(6 Letters)
Visual Search	275	3
Scene Perception	330	4
Music Reading	375	1
Typing	400	1(4 Letters)

According to Rayner [**5**] different forms of eye movements can be divided into the following categories:

Fixation: The main visual processing by brain happens during the fixations when the eyes focus on a single object for the further development of data recognition by brain. This is the condition when the eyes are motionless which normally takes around 200-300 milliseconds long. It should be mentioned that the eyes are never still and while they appear to be motionless they are having three miniature types of movement called nystagmus, drifts and micro saccades that their subject is out of the scope of this thesis.

Saccade: Saccade movements happen when we are reading, watching, driving, etc. These are quick eye movements happen between each two fixations. These rapid movements reach up to 500 degrees per second, and the distance they cover influences their duration.

Table 1 shows the duration and distance of the saccadic movements in different regular life activities. As it can be figured when more cognitive process and response to the visited concept is required the fixation length increases. This is like typing and music reading, which possess the maximum fixation length. It should be mentioned, “Saccades need to be distinguished from the other three types of eye movements”:

Pursuit: When eyes are following a moving object. This type of movement is slower than saccades and is formed of sequential smaller saccades.

Vergence: When both eyes are moved inward to fixate on a nearby object.

Vestibular: When the eyes rotate to compensate for head or body movement.

2.3 Contribution of the thesis

Recently an increasing number of studies are trying to exploit the eye-movement information for the task of implicit annotation of images.

In [**87**] Ajanki et al. used the eye-movements in order to infer implicit information retrieval query for retrieval of new documents. They tried to show that when there is no prior information about the user’s searching topic, the available eye-movements can improve the search results significantly. In this article their objective is to “predict from term-specific eye movements a query vector that can be used to evaluate the relevance of yet unseen documents”. Their main purpose is to construct a query function, $g_w(d)$, where d is the Bag-of-Words (BoW) representation of a new document and g is a two-class classifier that predicts the relevance of d. The main task is to provide the classifier with the implicit query (w) that makes it capable of classifying according to the user’s interest. For this purpose they extract 26 gaze features while the users are asked to read the provided documents and searching for the ones that have similar content to a previously assigned topic of interest. Most of the features that they used for studying the eye-movements are text specific features and cannot be generalised for finding the existence of user’s interest to other types of objects. The reasons are there are no distracting effects caused from the appearance of the different objects with variety of textures and colours on the screen. Also when reading the text, the majority of the eye movements are limited to the direction that a user is reading the text. Hence not all of these features can be used for other types of visual content that cause an undefined direction of eye-movements.

Jiao and Pantic [**38**] studied an implicit tagging technique to “annotate multimedia based on the user’s spontaneous non-verbal reactions”. For this purpose they used the user’s facial expression for collecting the user feedback when interacting with multimedia. In this work they investigate that whether the user depicts any different emotion in their face when they encounter a correct tag. After extracting feature vectors from positions of facial points using Hidden Markov Models they tried to classify the visited images by the user into the correctly tagged class and incorrectly tagged image. They conclude that the facial expression can be used for this task with the accuracy of over 72%. Although they achieve promising results, their framework is limited to visual contents that are already annotated and the relevance of annotation to the accompanying tag. This method is incapable of associating a new key concept to the visited visual content.

A similar approach to [**40**] using different modalities was incorporated by Soleymani et al. [**88**]. They recorded the multimodal MAHNOB-HCI with the goal of emotion recognition and implicit tagging. They recorded “face videos, audio signals, eye gaze data and peripheral/central system psychological signals”. In the course of the data collection their participants attended two experiments where in the first one they watched emotional videos and reported their emotions with keywords such as arousal and valance and in the second experiment short videos and images were shown to the participants once without tags and once with a tag that could be correct or incorrect. In this experiment for the purpose of implicit tagging the users faced 18 images and 14 videos where each of them was associated with a tag. They extracted 19 features from the gaze data and used the Hidden Markov Models associated with Adaboost [**40**] for classification of the gaze data. At the best situation they achieved 73% of prediction rate by combining 4 classifiers for the eye data. Although they reach an acceptable prediction rate but their database is limited to just 28 images and there is no task of hunting for images in their scenario. The images do not appear on the screen together and like [**40**] their approach can only predict whether an associated tag with an image is relevant to the image content. Their database is not suitable for seperating the images that the user is hunting for and the rest of the images as proposed in this thesis.

Subramanian et al. [**89**] employed a similar approach to [**38**] and [**88**] for 110 social scenes. In their study they compared the explicit feedback of verbal description and implicit feedback of eye-movements during their experiment. They concluded that there is a strong correlation beween the explicit and implicit inputs. In their work they mainly focused on the contents within an image by using previously recorded gaze data for subset of images from MIT and NUSEF databases, and presented them to new users to record their understanding of the image explained verbally. In their work they tried to explain how eye movements can be used to tell apart interactive and non-interactive social scenes, mild and high intensity of facial expressions portraits of clothed persons versus nudes. Like the previously mentioned studies this study tries to confirm the contents of an image based on the eye movements that happened inside the image rather than comparing the attention of the user to different images while they appear simultaniousely on the screen.

Walber et al. in [**90**] tried to find different objects inside the images by studying the gaze data. They took a similar approach to [**38**], [**88**] and [**89**] by trying to find the correct and incorrect relationship between an image and a provided tag. In addition they tried to divide the images into different regions using the gaze information.

Verochidis et al. [**91**] studied the possibility of automatic annotation of video scenes during video retrieval by exploiting eye movements. They used a classifier during the retrieval session to identify shots of interest for new users. They clustered the new shots submitted by new users when they were searching for a topic in order to classify then as relevant or non-relevant to the topics by the classifier. In their study they obtained the ground truth by recieveng the submitted images from user by physical interaction (i.e. mouse click). In their work they show statistically acceptable results with F-scores between 0.5 and 0.55. However their study relies on the images that the user had direct physical contact with them which can have direct effect on visual attention of the users. This results in uncertain output for the scenarios in which the images are relevant to a specific topic but are not physically attended by the user. For example when the output of a search engine shows the results of a query for Japan and the user is looking for images relevant to the nature of Japan and they only click on few number of the relevant images.

In contrast with above state-of-the-aret studies, This thesis focuses on implicit annotation in situations that a user is faced with multiple images on the screen. The goal is to classify the images into two groups of with-target-concept (TC+) and without-targe-concpt (TC-), where the Target Concept is the key concept that the user is hunting for when looking at the images. This classification takes place only based on eye movements regardless of the fact that the user had physical interaction with the image which includes interaction by mouse or keyboard. The major contributions of the thesis include:

- 27 gaze features have been defined, extracted and studied carefully to investigate which ones are the informative features for the purpose of classification of the images based on eye-movements.
- Three scenarios where carefully tailored for the purpose of data collection to plan the most appropriate experiment scenario that mimics a real life situation which is suitable for development of the proper classifier for the task of image classification.
- The effect of different factors, i.e. saliency[*] and similarity[†], other than the semantic content of the images on the visual attention are studied to investigate whether they bear any negative impact on the classification results.
- The performances of 5 different classifiers were compared on the recorded gaze data to identify the best.

[*] Saliency is the visual factor that guides one's attention involuntary. Like a red spot on a white background.

[†] Similarity is the factor that shows to what extent an image is similar to the rest of the images that appear on the screen at the same time.

3 Background Theory

Different systems were trained with the gaze data and their outputs were studied to investigate how they can provide proper classification of the image database. These systems include Fuzzy Inference Systems (FIS), Neural Networks (NN), Support Vector Machines (SVM), etc..

In this chapter the backbone mathematics and logics of these models are studied. Also we discuss the Subtractive clustering algorithm which helps us in construction of one of our FIS which is called Takagi-Sugeno-Kang Fuzzy Inference System (TSK-FIS).

3.1 Least Square Estimation

The Least Square Estimation (LSE) has been used as a fast and practical model to “reduce the influence of errors when fitting models to given observations” [**92**]. This model is a frequent approach to solve the Over-determined systems of equations [**93**]. In an Over-determined system the number of equations exceeds the number of unknowns. As a result there will be more than one solution. Consequently instead of solving the equations exactly it is tried to minimise the sum of residuals.

Consider for a system with m-1 inputs and 1 output as follows:

$$output = \sum_{i=1}^{m} input_i \times c_i$$

where $input_1$ is 1 by default and C=$\{c_1$,c_2 … $c_m\}$ is the set of coefficients that define the properties of the system by which the output of the system can be calculated from the inputs. The goal is to find the C set based on the sample data points to best fit the given set of data points by minimising the error for those points. Let D=$\{d_1,\ d_2, \ldots, d_n\}$ be the set of n recorded sample data points that provide the corresponding output values based on inputs where $d_j = \{1, input_{j,2}, \ldots, input_{j,m-1}$:$output_j\}$.

To find the C coefficients we can form a system of linear equations, $D_{in}X = D_{out}$,as follows:

$$\begin{bmatrix} 1 & input_{1,2} & \cdots & input_{1,m-1} \\ 1 & input_{2,2} & \cdots & input_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & input_{n,2} & \cdots & input_{n,m-1} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} output_1 \\ output_2 \\ \vdots \\ output_n \end{bmatrix}$$

The problem arises when we have an overwhelmed system with $n > m$ and more than one solution for *C*. As a result the objective is to find a set of *C* that minimises the error (*E*) of the system with the given data points [**94**] [**95**] [**96**], where:

$$E = \sum_{j=1}^{n} (output_j - \sum_{i=2}^{m} input_{n,\mathrm{i}-1} \,.\, c_i - c \quad)^2$$

When the n equations of the n data points are linearly independent, it can be shown that there is a solution for *C* with optimised values for error:

$$D_{in}{}^{T} D_{in} C = D_{in}{}^{T} D_{out} \implies C = \left(D_{in}{}^{T} D_{in}\right)^{-1} D_{in}{}^{T} D_{out}$$

3.2 Fuzzy Logic

The processing heart of the developed framework is formed of algorithms based on fuzzy logic inference approaches. Following advantages of the Fuzzy Inference Systems (FIS) were the motivations to use them as the main processing unit of the framework:

- **Computationally low cost training phase:** Because Subtractive Clustering is used for training, the computation cost increases linearly with respect to the increase of the number of the data points used at the training phase [**97**].
- **Interpretable developed systems:** The lingual variables, which form the inference system, that are used in the structure of the fuzzy rules of a FIS have made it easy to interpret for human.
- **Interpretable final outputs:** The FISes are capable of fuzzy output generation and classification. This means rather than showing whether an element is part of a class they calculate the membership grade to all of the classes.
- **Manually adjustable for testing:** The interpretability of the FISes for human gives them the direct access to the internal inference system properties for further investigation, testing and observation.
- **Flexible to new added feature vectors:** The structure of FISes is flexible to accept additional features in the input vectors with minimum effort.
- **Auto adjustable to features:** When the subtractive clustering algorithm performs the system identification task the developed FIS is automatically optimized for the training feature vector and there is no need for system optimization.

The ability of fuzzy logic "to deliver a satisfactory performance in the face of uncertainty and imprecision" [**98**] has made it one of the favourite models to design the systems that have to cope with unpredictable input/output data sets. Unlike the classical crisp logic that an object is either a member or not a member of a set, in fuzzy logic a membership degree is assigned to the object that defines its level of membership to a set.

Let X be the universal set of objects that its generic elements are shown as x. Then A which is a subset of X can be shown as:

$$A=\{(x,\mu_A(x)), x \in X\}$$

where, μ_A(x) is the Membership Function (MF) that defines the membership status of x in relation to A. If A is a crisp set. So:

$$\mu_A(\mathrm{x}) = f(x) = \begin{cases} 1, & \text{iff x } \in \text{ A} \\ 0, & \text{otherwise} \end{cases}$$

On the other hand in a fuzzy set μ_A(x) maps X into the membership space M=[0,1] [**98**] [**99**].

Fuzzy inference systems:

The behaviour of a fuzzy inference system (FIS) is governed by its If-Then rules. For example one of the rules for the fan speed of a freezer can be:

- *If temperature is high and stored food amount is large then fan speed is very high.*

These descriptive rules can be divided into two parts:

1) Premise

2) Consequent.

The former part is formed of the lingual variables as inputs (temperature, stored food amount), the input sets (high temperature, large amount) and descriptive T-norm operators [**100**] ('and' operator) that are applied on these inputs.

Depending on the type of the FIS the consequent part of the rules can be either lingual like the sample rule above (high speed) or a mathematical function. The final output of the FIS is the aggregation result of all of the consequent parts. It should be noted that the aggregation method varies according to the type of the system. Figure 1 shows the diagram of a FIS with two rules that its premise part is formed of the two lingual inputs of x and y and the input sets of {A1, A2} and {B1, B2} for the lingual inputs respectively. With 'z' as the consequent, the lingual form of the two rules of this FIS are as follows:

- R1: If x is A1 and y is B1 then z1 is {c1 (type 1), c1 (type 2), ax+by+c (type 3)}
- R2: If x is A2 and y is B2 then z2 is {c2 (type 1), c2 (type 2), px+qy+r (type 3)}

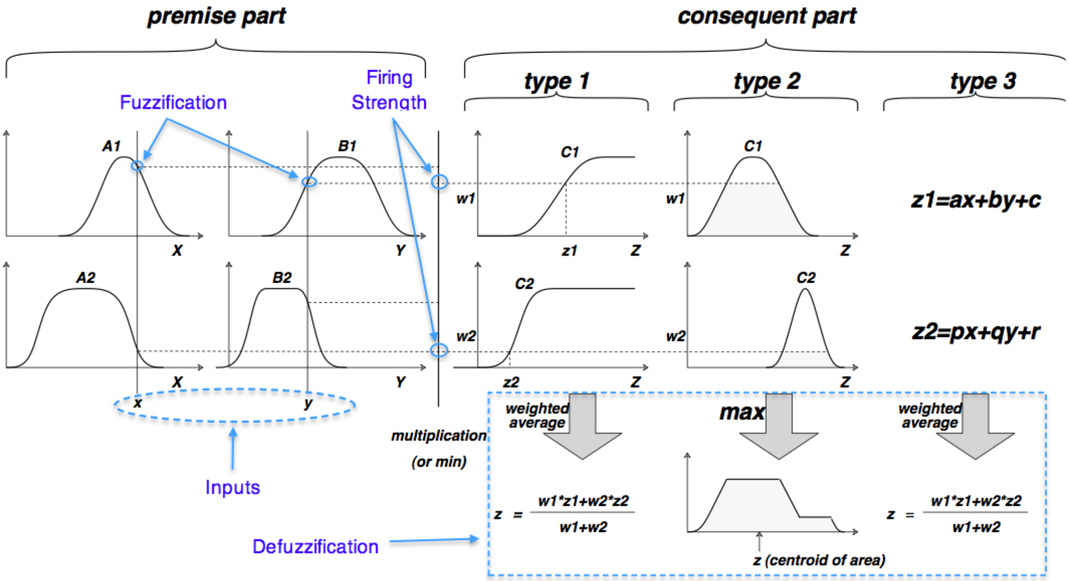


Figure 1: A sample FIS and three possible types of the consequent part (adopted from [101])

In general fuzzy reasoning of a FIS is formed of 4 steps [**101**]:

Step 1: Also called fuzzification step. In this step the membership degree of the inputs are calculated by the MF of the corresponding input set (e.g. for “x is A1” the system calculates the membership degree of x which is $\mu_A(x)$)

Step 2: In this step the T-norm operators are applied on the inputs. These operators are usually:

1. And
2. Or

They operate as multiplication or minimum in case of the former one and maximum in case of the latter (e.g. "$\mu_{A1}(x)$ and $\mu_{B1}(x)$"= min($\mu_{A1}(x)$, $\mu_{B1}(x)$). The output result of this step is called the firing strength (weight) of the rule shown with ω_i where ‘i’ is the number of the rules.

Step 3: In this step the consequent of each rule is calculated. The method to calculate the consequent part of the rules in a FIS is the main difference between the three types of the FISes. The consequent can be either crisp (type 1 and type 3 in Figure 1) or fuzzy (type 2 in Figure 1). The main factor in calculation of the consequent part is the firing strength of each rule.

Step 4: Also called Defuzzification step. In this step the consequent part of all of the rules are aggregated to form the final crisp output of the FIS. Figure 1 shows the three most popular methods of FIS defuzzification.

In Figure 1 three of the most popular fuzzy inference systems are demonstrated:

- Type 1 - In this type the final output is the weighted average of the output of each rule and it can be determined by calculating the membership degree of the firing strength of the rule on the monotonically non-decreasing output membership functions.
- Type 2 - This type is called the Mamdani FIS [**102**]. In this type first the output of each rule is calculated by applying the firing strength as the α −cut [**103**] over the corresponding output MF. Then the results are aggregated usually by max operation to form the fuzzy output. Finally the fuzzy output is converted to the crisp output by calculating the centroid of the area of the fuzzy output.
- Type 3 - This type is called TSK-FIS and is the type that is used mainly in the course of conducting this research. In this type of FIS the output of each rule is calculated by the “linear combination of input variables plus a constant term and the final output is the weighted average of the output of each rule” [**101**].

3.3 Subtractive Clustering

Normally in order to develop a FIS, human experts with the knowledge about inputs and outputs of the system are required. The experts are able to design and evaluate the MFs and the necessary variables for the system. To reduce the role of human in the development process of the FISes, algorithms are introduced which automatize the process of identifying the FIS structure by fuzzy clustering, generating rules for each cluster and developing corresponding MFs.

Fuzzy clustering is a form of clustering in which every element belongs to all clusters with a membership degree in [0 1] interval. The popular methods for fuzzy clustering are:

1) Fuzzy C-means [**104**]: In this method the number of clusters and initial cluster centres should be defined explicitly then the algorithm tries to find the cluster centre iteratively. This results in the clustering quality to be highly dependable on the initial values.

2) Mountain clustering [**105**]: This method can find the number of clusters and their corresponding centres but in this approach by increasing the dimension of the input vector the computation cost grows exponentially.

3) Subtractive clustering [**106**] [**107**]: This method is an enhanced form of the second method with the difference that "the computation is simply proportional to the number of data points" [**97**].

Having mentioned the different fuzzy clustering algorithms, it can be concluded that the subtractive clustering does not depend on the initial conditions that makes it consistent in the outputs. Also it is not an exhaustive process like Mountain clustering that makes it suitable for real-time systems. Consequently to generate the MFs for the input sets and find the number of rules for our TSK-FIS subtractive clustering was used in this research.

3.3.1 Finding clusters by Subtractive Clustering

Let X = $\{x_1, x_2, \dots, x_N\}$ be a vector of normalized (i.e. in a hypercube) N data points in an M-dimensional space. Each data point x_n is considered as a potential cluster centre with the first potential value of:

$$P_n^1 = \sum_{n=1}^{N} \frac{1}{e^{(\frac{2\|\mathrm{x}_n - \mathrm{x}_i\|}{r_a})^2}} \tag{1}$$

where, r_a is a positive constant that defines the radius of the effective neighbourhood that the surrounded data points have a greater influence on the potential cluster centre. After the calculation of potential for all of the data points the one with the highest potential is selected as the first cluster centre. Let x_1^* and P_1^* be the coordinates and potential of the first cluster centre respectively. Next the potential of all of the data points is updated by the following formula:

$$\mathrm{P}_1^* = \mathrm{P}_\mathrm{n}^* - \frac{P_1^*}{e^{\left(\frac{2\|\mathrm{x}_\mathrm{n} - x_1^*\|}{\mathrm{r}_\mathrm{b}}\right)^2}}$$

where, r_b is a positive constant that defines the radius of the neighbourhood that the effective potential reduction happens. From here the algorithm continues iteratively by:

1- Choosing the data point with the maximum potential as the new cluster centre

2- Reducing the potential point of the remaining data points by the following general formula after the k-th iteration:

$$\mathrm{P}_1^{\mathrm{k}+1} = \mathrm{P}_\mathrm{n}^\mathrm{k} - \frac{\mathrm{P}_\mathrm{k}^*}{e^{\left(\frac{2\|\mathrm{x}_\mathrm{n} - \mathrm{x}_\mathrm{k}^*\|}{\mathrm{r}_\mathrm{b}}\right)^2}}$$

Finding new cluster centres continues until $\mathrm{P}_\mathrm{k}^* \leq \bar{\varepsilon} P_1^*$ where $\bar{\varepsilon}$ is a coefficient to specify the stop condition of the algorithm. When the algorithm reached this condition it checks if $\mathrm{P}_\mathrm{k}^* \leq \underline{\varepsilon} P_1^*$ is true in which case rejects x_k^* and ends the process otherwise checks if:

$$\frac{\mathrm{d}_{\mathrm{min}}}{\mathrm{r}_\mathrm{a}} + \frac{\mathrm{P}_\mathrm{k}^*}{\mathrm{P}_1^*} < 1$$

where, r_a is the same as the one used in equation (1) and indicates the influence factor of the cluster centre. If it is true, the algorithm rejects x_k^*, sets its potential to 0 and continues with the next iteration. The final output of the subtractive clustering with C cluster centres will be:

$$X^* = \{\mathrm{x}_1^*, \mathrm{x}_2^*, \dots, \mathrm{x}_\mathrm{C}^*\} \tag{2}$$

3.3.2 Developing TSK-FIS by Subtractive Clustering

TSK-FIS is the most commonly used FIS for system modelling [**106**]**.** This type of FIS uses the lingual form and mathematical functions for the premise and consequent parts of the fuzzy rules of a fuzzy system respectively [**107**]. This structure of TSK-FIS makes it interpretable both for human and machine and adopts flexibility to the developed systems. We used the first order TSK-FIS for our framework. It is called first-order TSK-FIS because first-order polynomials form the output of each rule as the consequent part.

Higher order systems increase complexity and computational cost with little merit [**108**]. A generic form of the k-th rule of a first-order TSK-FIS with M inputs and one output can be denoted as follows:

R_k: If x_1 is $A_{k,1}$ and x_2 is $A_{k,2}$ and ... and x_M is $A_{k,M}$ then $z_k = b_{k,1}x_1 + b_{k,2}x_2 + ... + b_{k,M}x_M + b_{k,M+1}$

where Xm, $A_{K,m}$ and z_k are the m-th input, m-th input-set and the output of the k-th rule respectively.

Every discovered cluster centre by subtractive clustering (SC) is an example of a “characteristic behaviour of the system” [**97**] which results in a new rule for the to-be-constructed TSK-FIS.

Let X = $\{\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n}\ \}$ be N data pairs that are used for constructing TSK-FIS. Each data pair (x_n) is formed of an input vector and a scalar output:

$$\mathrm{x_n} = [\mathrm{x_n}|_{\mathrm{in}} \vdots\ \mathrm{x_n}\ |_{\mathrm{out}}]$$

Also let X^* be the set of C cluster centres discovered in X and introduced by equation (2) with the corresponding $\alpha = \frac{4}{r_a}$. A TSK-FIS can be constructed with C rules that:

$$\omega_c = \ e^{\alpha\|\mathrm{x}|_{in} - x_c^*\|} \tag{3}$$

where ω_c is the firing strength of the c-th rule. Based on the introduced generic form of the rule R_k earlier, the output of each rule of the TSK-FIS can be shown as:

$$z_c = B_c[x|_{in} \vdots 1]^T = B_c\bar{x}|_{in}^T \tag{4}$$

where $B_c = [b_1 \dots b_{M+1}]$ shows the coefficient matrix of the output membership function for the c-th rule and the system is identified by finding these coefficients. By using equation (3) and equation (4) and considering:

$$\bar{\omega}_c = \frac{\omega_c}{\sum_{c=1}^{C} \omega_c}$$

The final output of the TSK-FIS can be shown as follows:

$$z = [\bar{\omega}_1 \bar{x}|_{in} \cdots \bar{\omega}_C \bar{x}|_{in}] \begin{bmatrix} B_1^T \\ \vdots \\ B_C^T \end{bmatrix}$$

Now for N data pairs of $[x|_{in} \vdots\ x\,|_{out}]$ we can write:

$$\begin{bmatrix} x_1|_{out} \\ \vdots \\ x_N|_{out} \end{bmatrix} = \begin{bmatrix} \bar{\omega}_{1,1}\bar{x}_1|_{in} & \cdots & \bar{\omega}_{1,C}\bar{x}_1|_{in} \\ \vdots & \ddots & \vdots \\ \bar{\omega}_{N,1}\bar{x}_N|_{in} & \cdots & \bar{\omega}_{N,1}\bar{x}_N|_{in} \end{bmatrix} \begin{bmatrix} B_1^T \\ \vdots \\ B_C^T \end{bmatrix} \tag{5}$$

With a closer look it can be seen that the equation (5) has the form of B = AX which can be solved simply by using the popular Least-Square Estimation [**109**] method for system identification.

3.4 Support Vector Machines (SVM)

Using Support Vector Machines (SVM) [**110**] is one of the most popular methods for classification problems. SVMs are not only capable of performing linear classification, by using the Kernel trick they can also implement non-linear classification by mapping the input data into high dimensional feature spaces implicitly. A SVM defines a set of hyperplanes in a high dimensional space and tries to find the best hyperplane that has the largest distance from the training data points of all classes. The mapping that is used by SVM is dot products that are defined in terms of Kernel Functions K<x,y>. Depending on the problem at hand, different kernel function can be selected for mapping the data. SVM is designed based on the structural risk minimisation principle [**111**].

In a binary separable learning problem, the set of indicator functions for defining separating hyperplanes can be represented as:

$$(\boldsymbol{\omega}.\boldsymbol{x_i}) + \ b = 0, \quad \boldsymbol{\omega} \in \mathbb{R}^n, \quad b \in \mathbb{R}, \quad i = 1, 2, 3, \dots, n$$

where vector ω and scalar bias b define the actual location of the hyperplan. The margin is the minimal distance of samples for different classes from the decision hyperplane. In case of separable data, $\boldsymbol{\omega}$ and b are rescaled so that the closest points of training data to the hyper plane satisfy the following condition [**112**]:

$$|\boldsymbol{\omega} \cdot \ \boldsymbol{x_i} + b\,| = 1$$

In this case a canonical hyperplane representation is obtained in the following form [**20**]:

$$y_i\big((\boldsymbol{\omega} \cdot \boldsymbol{x_i}) + \ b\big) \geq 1, \quad i = 1, 2, 3, \dots, n$$

The optimal hyperplane is the one that separates all vectors without error and at the same time maximizes its distance from the closest data points to it.

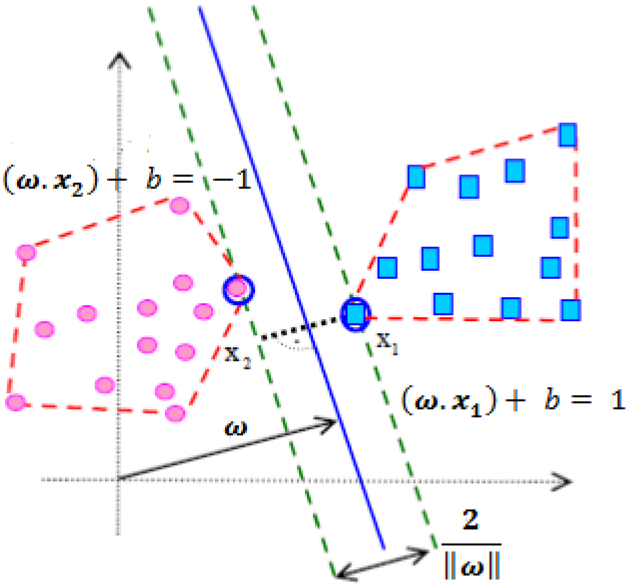

$(\boldsymbol{\omega}.\boldsymbol{x_2}) + \ b = -1$
x2
x1
$(\boldsymbol{\omega}.\boldsymbol{x_1}) + \ b = \ 1$
$\boldsymbol{\omega}$
$\frac{2}{||\boldsymbol{\omega}||}$

Figure 2: Separation margin in a binary classification problem [20]

The distance $d(\boldsymbol{\omega}, b; \boldsymbol{x})$ of a point x from the hyperplan $(\boldsymbol{\omega}, \mathrm{b})$ is:

$$d(\boldsymbol{\omega}, b; x) = \frac{|\boldsymbol{\omega} \cdot \ x_i + b\,|}{||\boldsymbol{\omega}||}$$

From Figure 2 we can see that the optimal hyperplane has the same orthogonal distance from the data points resting on the two convex hulls. This hyperplane is constructed by maximising the margin ρ:

$$\rho(\boldsymbol{\omega}, \mathrm{b}) \ = \min_{x_i : y_i = -1} (\boldsymbol{\omega}, b; x_i) + \min_{x_i : y_i = 1} (\boldsymbol{\omega}, b; x_i)$$

$$= \min_{x_i : y_i = -1} \frac{|\boldsymbol{\omega},\ x_i + b\ |}{||\omega||} + \min_{x_i : y_i = 1} \frac{|\omega,\ x_i + b\ |}{||\omega||}$$

$$= \frac{1}{||\boldsymbol{\omega}||}\left(\min_{x_i : y_i = -1} |\boldsymbol{\omega}, x_i + b| + \min_{x_i : y_i = 1} |\boldsymbol{\omega}, x_i + b|\right)$$

$$= \frac{2}{||\boldsymbol{\omega}||}$$

It can be concluded the optimal hyperplane satisfies the following equation:

$$\min(\frac{1}{2}\ \|\boldsymbol{\omega}\|^2) \tag{6}$$

Subject to the constraint:

$$y_i\big((\boldsymbol{\omega} \cdot \boldsymbol{x_i}) +\ b\big) \geq 1, \qquad i = 1, 2, 3, \ldots, n \tag{7}$$

The VC dimension [**113**], h, of the set of canonical hyperplanes in n dimensional space is bounded by:

$$h \leq \min[R^2 A^2, n] +\ 1$$

where R is the radius of a hypersphere enclosing all the data points. Hence minimising (6) is equivalent to minimising the upper bound on the VC dimension. The solution to this problem finds the circled points on the two convex hulls in Figure 2. To solve the above minimisation problem we can use the Lagrangian multipliers for each inequality in the above equation where:

$$L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) = \frac{1}{2}\,\|\boldsymbol{\omega}\|^2 - \sum_{i=1}^{1} a_i(y_i|(\boldsymbol{\omega}, \boldsymbol{x_i}) + \ b|-1)$$

where $\boldsymbol{\alpha}$ values are the Lagrange multipliers. Here, the objective is to minimise the Lagrangian with respect to $\boldsymbol{\omega}, b$ and maximised with respect to $\boldsymbol{\alpha} \geq 0$.

$$\max_{\alpha \geq 0}\ (\min_{\boldsymbol{\omega}, b} L(\boldsymbol{\omega}, b, \boldsymbol{\alpha})) \tag{8}$$

The minimum with respect to $\boldsymbol{\omega}$ and b of the Lagrangian L, is given by [**20**]:

$$\frac{\partial L}{\partial b} = 0\ \Rightarrow \sum_{i=1}^{n} \alpha_i\, y_i = 0 \tag{9}$$

$$\frac{\partial L}{\partial \omega} = 0\ \Rightarrow \omega = \sum_{i=1}^{n} \alpha_i\, y_i\, x_i = 0 \tag{10}$$

$$y_i\left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) - 1 \geq 0, i = 1, . . , n$$

$$\alpha_i \geq 0, \forall i$$

$$\alpha_i\left(y_i\left((\boldsymbol{w} \cdot \boldsymbol{x_i}) + b\right) - 1\right) = 0, \forall i$$

From equations (8), (9) and (10) the dual problem can be shown as:

$$max \sum_{k=1}^{n} \alpha_k \; - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \left(x_i, \, x_j\right)$$

$$\alpha^* = \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \left(x_i, \, x_j\right) - \sum_{k=1}^{n} \alpha_k \; = 0$$

When:

$$\alpha_i \geq 0 \qquad\qquad i = 1,2, \ldots, n$$

The Lagrangian multipliers are non-zero when the constraint at (7) is met. Only the points on 12 have $\alpha_i \geq 0$. It is from these points that we can define the margin. These points are called the support vectors. As a result we can write the hyperplane decision function as follows:

$$f(x) = \; sgn \left(\sum_{i=1}^{n} \; \alpha_i \, y_i \; . \; K\left(x_i, x_j\right) + \; b \right)$$

If the data cannot be separated at the current dimension mapping the data into a higher dimension can be of help to separate them. As a result the linear dot product can be replaced by inner product in Hilbert space [**20**].

$$K(\mathbf{x}_i , \mathbf{x}_j) = \; \langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle = \varphi(\mathbf{x}_i). \varphi(\mathbf{x}_j)$$

Table 2 shows examples of some of the most commonly used kernels in SVM.

Table 2: Most commonly used kernels in SVM

Kernel	Function
Linear kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = \langle x_i, x_j \rangle$
Polynomial	$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\langle x_i, x_j\rangle + r)^{\mathrm{d}}, \gamma > 0$
Radial Basis Functions	$K(\mathbf{x}_i, \mathbf{x}_j) = xp(-\gamma\lVert x_i - x_j\rVert^2),$ $\gamma = \frac{1}{2\alpha^2}, \alpha > 0$
Sigmoid kernel	$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma\langle x_i, x_j\rangle + r)$

3.5 Back Propagation Neural Networks

A back propagation neural network BPNN is composed of interconnected neurons that each of them is a nonlinear function that produces single output from multiple inputs. This type of neural network is one of the most popular types of classifiers for prediction [**114**]. These functions of neurons (also denoted as node) are called the Activation Function or the Transfer Function. These functions can be implemented in various forms including the popular signum, sigmoid and hyper-tangent functions:

Signum function:

$$sgn(x) = \begin{cases} 1, & if\ x \geq 0 \\ 0, & otherwise \end{cases}$$

Sigmoid function:

$$\mathrm{Sig(x)} = \frac{1}{1 + \exp(-x)}$$

Hyper-tangent function:

$$\tanh(\mathrm{x}) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$$

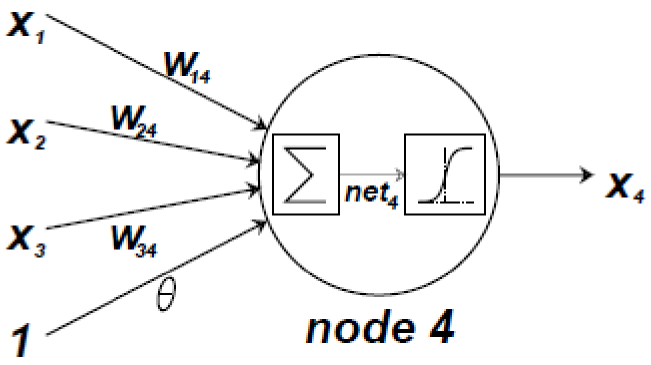


Figure 3: A sample node of the Back Propagation Neural Networks

Figure 3 shows a sample of a node for Back Propagation Neural Networks. A threshold added to the sum of the weighted incoming inputs to a node form the net input of the node. For example for node j we have:

$$net_j = \sum_i w_{ij} x_i + \theta_j$$

where x_i is the input signal from node I (output of node i), w_{ij} is the weight associated to output of node i and input of node j for x_i and θ_j is the threshold on node j.

The output of node j is:

$$x_j = y(net_j)$$

where y(x) can be one of the mentioned transfer functions earlier.

In reality the weights (w_{ij}) are the internal property of each neuron. Consequently by changing the weights the behaviour of the neuron changes which results in the change of the behaviour of the whole network.

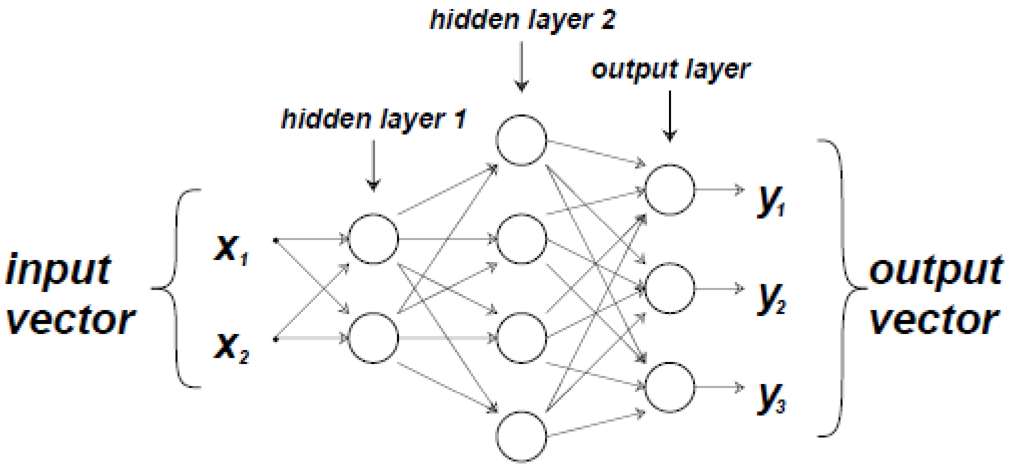


Figure 4: A BPNN with two hidden layers, two inputs and three outputs

Figure 4 shows a BPNN with two hidden layers, two inputs and three outputs. For simplicity this network is denoted as a 2-2-4-3 BPNN. The hidden layers are the group of neurons that are located at the same order of input and are not either input layer of output layer. The learning algorithm of the NN defines its type where the back propagation indicates that the network is iteratively trained by a gradient decent algorithm to minimize the mean squared error between the actual outputs and the desired outputs. This error can be shown as:

$$E = \|d(\boldsymbol{x}) - NN(\boldsymbol{x})\|^2$$

where $\boldsymbol{x}$ is the input vector to BPNN, NN $(\boldsymbol{x})$ is the output vector of the BPNN and d $(\boldsymbol{x})$ is the actual desired output. The process of training a network is continuously updating the weight of the neurons in order to minimize the error above. As a result for node j an error is defined as follows:

$$\delta_j = \begin{cases} -2\left(d_j(\boldsymbol{x}) - NN_j(\boldsymbol{x})\right) sig'\left(net_j\right), & If\ node\ m\ is\ in\ output\ layer \\ sig'\left(net_j\right) \sum_m \delta w_{jm}\,, & Otherwise \end{cases}$$

where w_{jm} is the weight of the connection from node j to node m, $d_m(\boldsymbol{x})$ is the j-th component of $d_j(\boldsymbol{x})$, $NN_j(\boldsymbol{x})$ is the j-th component of $NN(\boldsymbol{x})$ and $sig'(net_j)$ is the derivative of the Sigmoid function**.** Consequently the update amount of weight w_{jm} is:

$$\Delta w_{jm} = -\gamma \delta_j x_i$$

where γ is a learning rate that has the effect on the convergance speed and stability of the weights during the learning process. BPNNs are one of the often used machine learning structures for various applications including speech recognition, pattern recognition, signal processing, data compression, etc.

3.6 Two-way ANOVA

The analysis of variance (ANOVA) provides a statistical test to compare the means of two or more data sets, where each data set contains an independent sample of mutually independent observations [**93**] [**115**]. Along with other outputs, the test returns the p and the F values under the null hypothesis that all samples are taken from the populations with the same mean. The F value (called F statistic) is the ratio of the between variability and within group variability and the p value is the probability. When p is very small it sheds doubt on the null hypothesis and proposes that at least one mean of one of the samples is significantly different from the other sample means. Common significant levels for p-value are 0.05 and 0.01.

The following assumptions are made when using ANOVA.

- All sample populations are normally distributed.
- All sample populations have equal variance.
- All observations are mutually independent.

“The ANOVA test is known to be robust with respect to modest violations of the first two assumptions” [**93**].

Two-way ANOVA is using the ANOVA test on two dimensional data samples. Consider:

$$D = \begin{array}{cc} \begin{array}{cc} O1 & \quad O2 \end{array} & \\ \begin{bmatrix} d_{111} & d_{121} \\ d_{112} & d_{122} \end{bmatrix} & User\ 1 \\ \begin{bmatrix} d_{121} & d_{221} \\ d_{122} & d_{222} \end{bmatrix} & User\ 2 \end{array}$$

where D is a data collection and it is formed of the observation O1 and O2 which is made by monitoring two users (User 1 and User 2). The two-way ANOVA test compares the means of the columns and rows of D. The difference in the columns indicates the alteration in observation factor and the difference in the rows shows the alteration made by the users.

For two-way ANOVA there are three null hypotheses:

1- That all samples from Observation factor (i.e. O1 and O2) are extracted from the same population
2- That all samples from User factor (i.e. User 1 and User 2) are extracted from the same population
3- That the effects due to factors Observation and User are *additive* (i.e., that there is no interaction between them)

As a result, for each hypothesis we will have one F statistic and one p-value where the former indicates whether the expected values of a quantitative variable within several pre-defined groups differ from each other and the latter indicates the significance probability of this value.

For a detailed discussion of ANOVA test the readers are addressed to [**115**].

3.7 Correlation coefficient

Correlation coefficient is a value between +1 and -1 indicating the linear dependence between two vectors [**116**] which is widely used in science (e.g. [**117**]).

It is the covariance between two variables divided by the product of their standard deviations. Consider the two vectors of V1 and V2, their correlation coefficient can be calculated as follows:

$$Correlation\ \text{Coefficient} = \frac{E[(V1 - \mu_{V1})(V2 - \mu_{V2})]}{\sigma_{V1} \times \sigma_{V2}}$$

where E is the covariance between the vectors, μ shows the mean of a vector and σ denotes the standard deviation of a vector.

4 Framework Structure and Experiments

In this study a framework is developed to present the images in the database to the user, record their corresponding eye movements and interpret those eye movements for the purpose of classification of the images into two classes of TC+ and TC-. This classification is based on the fact that whether the image contains the key concept in the user mind that they are hunting for or not. This framework is formed of two main interfaces of User Interface and Processing Interface.

The user interface is responsible for interaction with the user by providing the graphical objects on the monitor screen, accepting commands from the user, recording the user's gaze data throughout the experiment and sending all the data to the processing interface. The Processing Interface is responsible for receiving all of the data from the User Interface, analysing them and finally classifying them.

Because simply the input of the framework is raw gaze data and the images and the output are the classified images, it makes it effortless to incorporate this framework in different systems with diverse goals to use it as a black box for not only evaluation of the user's interest to the images but also to other visual objects on screen. In the context of this thesis, the user's interest in an image only and only means the user's attention to the image when the image contains the concept that the user is looking for.

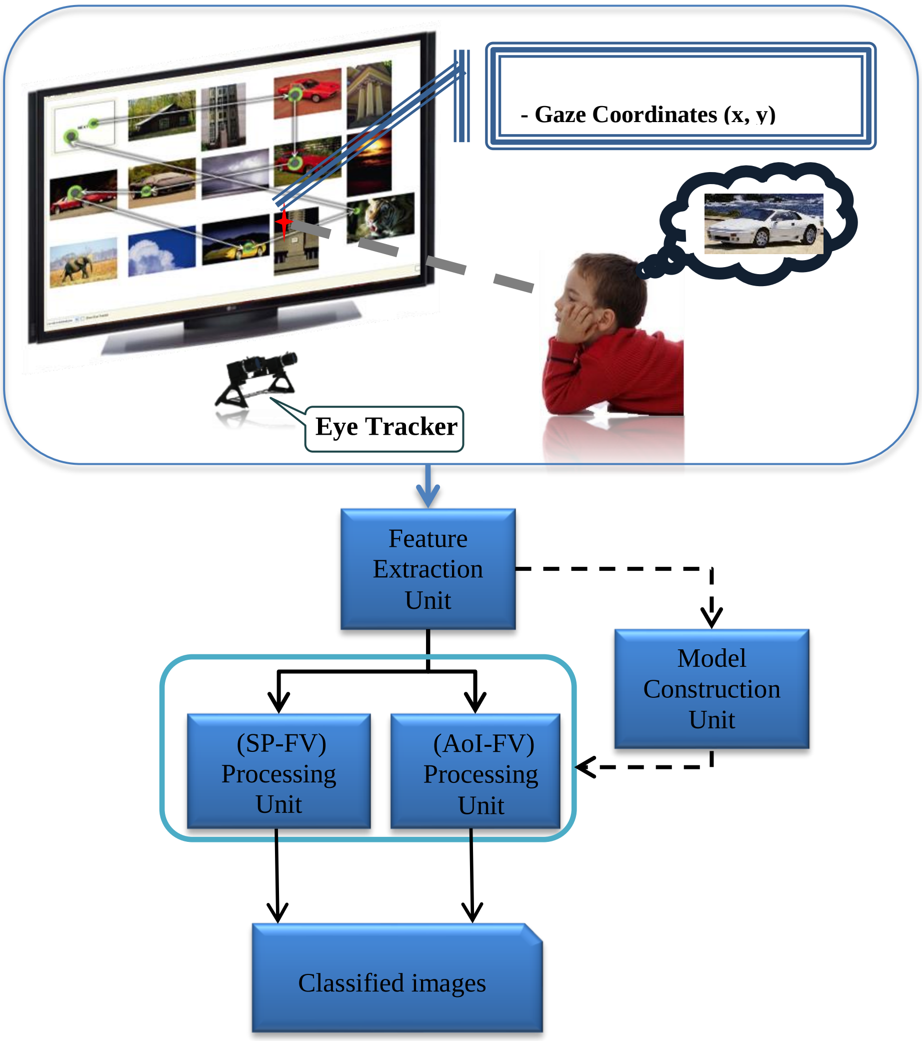


Figure 5: Block diagram of the general structure of the developed framework (adopted from [50])

Figure 5 shows the general structure of the developed framework. When a user is looking at the screen, the coordinates of the point on the screen that they are looking at, gaze point, along with the corresponding time that it happens are extracted by the eye-tracker as the raw gaze data. Two vectors are created by the sequence of these gaze data form the gaze movement information. These vectors are sent to the feature extraction unit where 27 gaze features, which will be introduced in detail in Chapter 5, are extracted in two feature vectors of Scan Path Feature Vector (SP-FV) and Area of Interest Feature Vector (AoI-FV). The former is formed of the features regarding the dynamics of the eye movements and the latter is formed of the features about the visual attention of the user to a region of interest, in here it is the area of an image on the screen.

After the features are extracted, if no gaze analysis inference system is developed, the two feature vectors are sent to the model construction unit to train an inference system for each feature vector. When the inference systems are trained, each of them classifies the images separately. In the Results chapter the classification of the images based on the trained classifiers for each feature vector and their agreement are discussed further.

4.1 Application of eye-tracking in this research

The reading process is not the smooth movement of the eyes over every single word in the text lines, but it is formed of a combination of sequential fixations and saccades [**5**]. When reading, one's brain automatically selects the next word to read which is not necessarily adjacent to the current word that was read right now. The rest of the words in between the two words that are selected to be read are comprehended by the side view during the current fixation [**118**]. Figure 6 shows how we read a line of text in normal reading action. This nature of eye movement is used in this research to differentiate between the images that the user showed interested in them and the rest of the images.

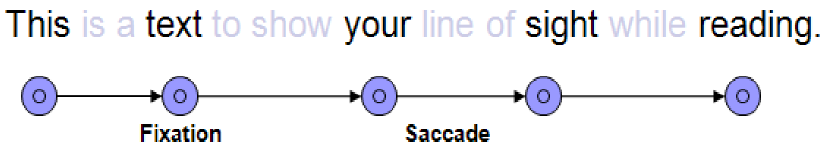


Figure 6: The way we look at the words of a line during the reading process

Figure 7 shows that how rows of images can be similar to lines of text. The positions of the circles show the fixation position of the eyes, which are the gaze points, and the different sizes of the circles show the fixation duration. In this scenario if it is considered that the concept in the user's mind is ***CAR,*** during exploration in the images the user fixates for a longer time on the images containing car. This is due to the process time that the brain needs to match the image content and the concept in the mind to make sure that they agree. Normally the users do not fixate on images without car that they already have identified by their side view.

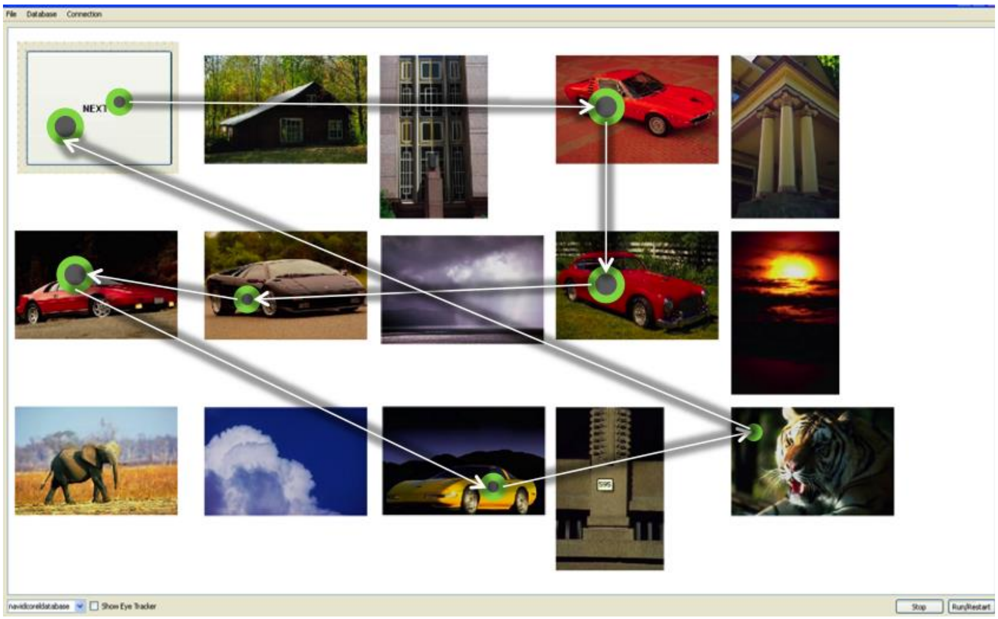


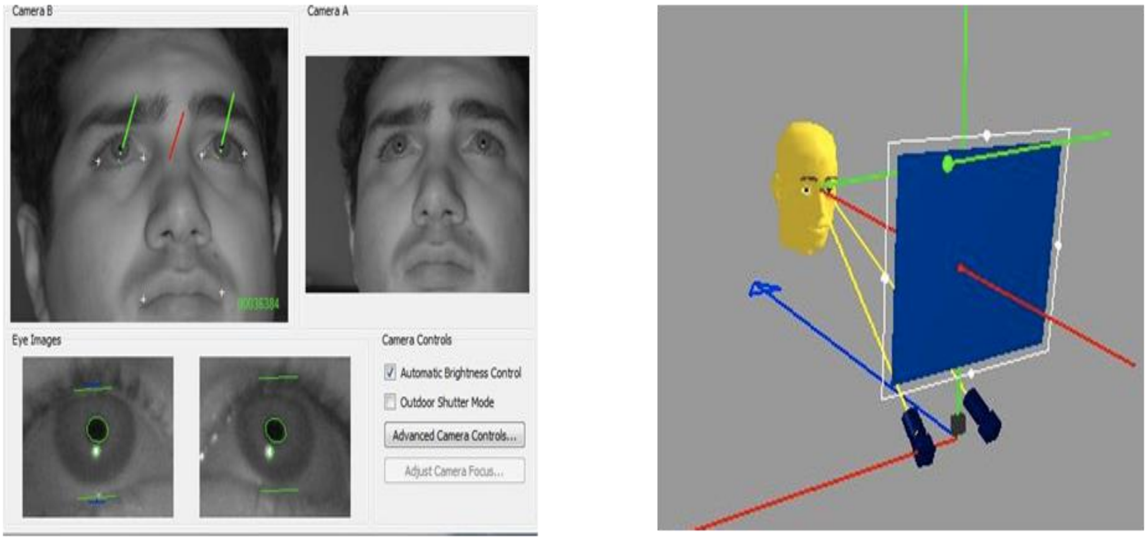
Figure 7: An example of exploring images similar to reading process

As can be seen sometimes there are images that do not contain the Target Concept but a user is attracted to them like the ***Tiger*** in Figure 7. Various reasons could cause this behaviour. This includes the saliency of the image which triggers the bottom up process of attracting one's visual attention due to its distinctive appearance. Another reason is the personal interest of the use in the visited subject. Also similarity of the image to the concept under pursue can determine one's shift of attention to a non-Target-Concept image. These types of images are the main challenges for classifying the visited set into TC+ and TC- groups as they are attended very similar to the way that TC+ images are paid attention to.

4.1.1 Technology

The basic of the eye-tracking technology is to identify and locate the pupils of both eyes and calculate the point that the subject is looking at by tracking the position of the pupils. According to Zhu et al. [**14**] the easiest way to detect the pupil is to detect the reflection of Infra-Red light in the pupils. When eyes are under radiation of IR, the resulting reflections appear as the brightest spots in the taken images. Figure **8**-a shows a sample image of such an event. This reflection belongs to the *corneal* area inside the eyes which is responsible for human central sharp vision and is used for driving, watching TV, reading, etc. [**10**].

After locating the pupils, the eye tracker starts to extract eye movements and rotation by implementing complex image processing techniques on the video images received from the cameras. To do this, different eye-trackers employ different methods however most of them use a manually assigned point in the real world as the Origin and find the rotation of each eye with respect to it. The accuracy of the tracking system highly depends on the image quality of these cameras. Finally the gaze intersection with objects and target screen can be tracked by knowing their coordinates and spatial angles with respect to the conventional origin. The shortcoming of this method is the effect of the lighting condition of the environment where if there is too much sunlight in the room it with interfere with the reflection of the IR source in the eyes.



☑ Automatic Brightness Control
□ Outdoor Shutter Mode

(a) (b)

Figure 8: (a) A snapshot of the video images of the eye-tracker (b) The corresponding 3D reconstruction

4.1.2 The faceLAB interface

The faceLAB interface and the eye trackers from Seeing Machines Company were used for conducting the presented research. The reasons that faceLAB is chosen over other products are its higher technology and non-user-intrusive approach. Unlike other eye-tracking devices that are head mounted or need the users to keep their head in a position motionless, faceLAB is a standalone portable system that gives the users the ability to move their heads in a wide range of area.

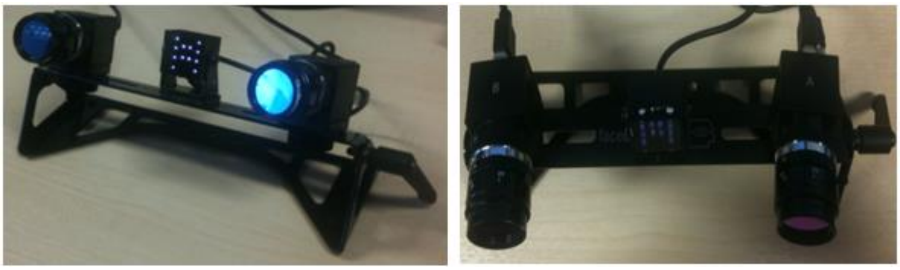


Figure 9: FaceLAB Stereo Head cameras and the IR Pod

The structure and user interface of faceLAB are as follows:

1- **Stereo Head Cameras**: There are two cameras A (right) and B (left) that their position cannot be exchanged. These cameras are responsible for providing the video images of user's face. They capture the images at the rate of 60Hz and are connected to the dedicated faceLAB PC by FireWire. After taking every image, they stamp it with the capture time of the image and its frame number starting from zero at the beginning of the tracking process. These cameras should be located in a way that both of them can see the user's face from the bottom of the lips up to the top of the eyebrows. The both eyes should be positioned in the range of the two cameras to be tracked. This will give the users a freedom space to move their heads approximately inside a 30x30x30 cm box at a normal distance from monitor. Wherever the stereo head is located, the origin of the coordinate system will be set at the middle point of the two cameras.

2- **Infrared Pod:** The Infrared Pod provides infrared illumination at a safe level that is invisible to the naked eyes. It is better to locate the IR pod in the middle of the both cameras as shown in Figure 9, however in case of limitations it can be placed at any other place where it can radiate the IR directly into the both eyes.

3- **Calibration of the stereo head:** At the very beginning of eye tracking the stereo head should be calibrated. This is to figure out the distance and angle of the both cameras with respect to each other. Then the position of the IR Pod is measured manually with respect to the origin of the coordinate system and entered to the program. These data are used for later position calculations. This calibration is not required any more until the position of the cameras changes.

4- **User Calibration:** After the stereo head and IR pod are calibrated this is the time for the user to be calibrated. The user face is calibrated automatically; however another calibration is required to fine adjust the gaze point detection of the user. In the later calibration the user is simply asked to follow a black cantered bright circle in a dark screen over a grid with nine points including the four corners of the screen.

5- **Providing data:** When the user calibration is finished faceLAB starts to put the tracking results on the TCP/IP port of the faceLAB PC. The data could be provided in two structures:

   a. **Real-Time:** In this structure limited but necessary types of data are provided in Real-Time. It contains information regarding head position, head rotation, gaze point coordinate on the screen, etc. This is the data structure that is used in this research.

   b. **Accurate:** In this data structure extra information are extracted from the raw gaze inputs which include blink occurrence and saccade identification. Due to the required extra computation of the raw information these data are provided with a 2-second delay.

6- **World 3D Reconstruction:** The faceLAB GUI provides the facility of presenting the real environment in a virtual 3D reconstructed environment. As can be seen in Figure **8**-b the eye-trackers, the screen and the human subject appear in this virtual 3D world. Here the position of the first two is defined manually and the system finds the position of the third one automatically. Additionally the system finds the gaze intersection point and the head direction of the users and shows them with the green and red vectors respectively.

These provided data can be received by a client computer, which can use them for its applications. In this research Real-Time data structure is used due to the real-time nature of the project and the atomic functions that need real-time data. Table 3 shows the detailed specification of faceLAB.

4.1.3 Hardware connection

The annotating system is formed of two computers and two monitors. Due to the high load over the PC that handles the eye-tracking process, it is not possible to put more processing load on it. Consequently it sends the resulting eye-tracking data to a LAN network with the TCI/IP protocol. On the other side a second PC receives these data and uses them to control the user interface. The eye-tracker cameras work at 60HZ frequency, consequently the time difference between each two successive frames is 1000 ms/60 = 16.66 ms. To calculate the time difference between two coordinates of gaze, it is enough to subtract the Frame stamp on each and multiply it by 16.66. This means that the resolution of the system is up to 16.66ms and the system cannot measure events that occur in a shorter time. As a result in this thesis the terms time and frame number are used interchangeably where every frame is equal to 16.66 ms.

Table 3: faceLAB specifications

Head data	Head position, Head rotation
Eye data (each eye)	Eye position, Eye rotation, Eye gaze position against screen, Eye gaze position against world model, Pupil diameter, Eye vergence distance, Saccade events
Eyelid data (for each eyelid)	Blink events, Blink frequency, Blink duration, Eyelid aperture, PERCLOS fatigue metric
Facial feature data	Eyelid behaviour, Lip behaviour
Timing data	Experiment frame number
Head Rotations	Tracking and recovery up to +/- 90° around the y-axis. Tracking and recovery up to +/- 45° around x-axis.
Gaze Rotations	Eye rotations of up to +/- 45° around the y-axis. Eye rotations of up to +/- 22° around the x-axis.
Head Box	Horizontal tracking range up to 0.35m (13.8"), vertical range up to 0.23m (9"), distance range up to 0.6m (23.6")
Tracking Accuracy	Typical static accuracy of head measurement within +/- 1mm of translational error and +/- 1° of rotational error. Typical static accuracy of gaze direction measurement 0.5-1° rotational error, depending on selected field-of-view

4.2 Data transfer synchronisation

To make sure that the received data from the eye trackers are transferred between the machines in real-time, following measures are taken into consideration for the development of the user interface:

- The connection between the computers is set up as a TCP connection which is a data transfer protocol that guaranties all the data packages are transferred intact.
- The experiments are formed of different pages of images that appear to the user one by one when the user requests for a new page by mouse click. When the user is visiting a specific page, the extracted visit times are based on the interval between the time that the user's gaze entered the area of the interest of an image and the time that the gaze left the area of interest of that image. As a result of this design, because the time difference is calculated, any asynchronous data communication is ineffective within the page.
- When the user requests for a new page by clicking on an image or on the next button the receiver machine sends a signal to the eye-tracker to stop recording and any remaining information in the communication data buffer is read and emptied.
- Throughout the recording period a program loop in the receiver machine checks the communication buffer exactly every 10 milliseconds which is quicker than the eye-tracker's resolution (17ms). At the start of every page the connection between the two machines is re-established immediately after the images of the new page are loaded on the screen. At this point the receiver machine signals the eye-tracker machine to start recording gaze data and if the program loop does not receive the first package of data within the third round of the mentioned loop, within 30ms[*], an error is thrown and the program terminates.

4.3 Dataset properties

For Implicit image annotation using gaze information it is important that the acquired gaze information have some specific properties. These properties are explained as follows:

[*] Initially the permitted delay was setup to 20 milliseconds where the first package had to be received within the 2nd round of the loop. However because the resolution of the eye-tracker is more than 16ms and there are overhead process for re-establishing the connection, the program kept terminating frequently. After increasing the waiting time to 30 ms the program terminated just once due to a heavy load on the eye-tracker's machine by other programs. This was prevented during the recording of the main experiments by running only the eye-tracking process on the machine.

- Type of visual content: Although it is possible to generalize the findings of this thesis to any type of visual contents, it is important to use images as stimuli for studying the gaze movements for implicit image annotation.
- Area of interest: For image annotation it is important that the users search between images not within an image which could change the main task to image segmentation [**119**]. As a result it is desirable to have multiple images on the screen and investigate gaze movement between them rather than having a single image on the screen and study the gaze movements inside that image.
- Eye movement: To imitate a real image search scenario it is important that the users be free in terms of direction of the eye movements rather than having a predictable direction like reading texts from left to right.
- Physical interaction: The goal of implicit image annotation is to study the non-conscious user reactions to the images. As a result it is desirable that the users have minimal physical interaction (e.g. mouse click) with the visual content.
- Database diversity: A diversified database is desired so it is possible to study different gaze behaviours that include images with similar content and images with different content.
- Similarity to real life environment: To simulate a real searching scenario, it is important that the instructions for the users during the experiments for data acquisition are made around the intuitive interactions that the user would normally show in real life.
- Users' bias: It is possible that the user's knowledge about the main purpose of the experiments (implicit image annotation based on gaze data) prior to conducting it can affect their gaze behaviour during data acquisition. So it was important that the users are kept from detailed information until the end of the experiment.
- Users' class: The profession class of the users can bias them towards the task, for example people with computer science background can guess the main purpose of the experiments which results in guiding their visual attention in favour of the experiment or vice versa. So it was desirable to choose the human subject from the group of people without such backgrounds.

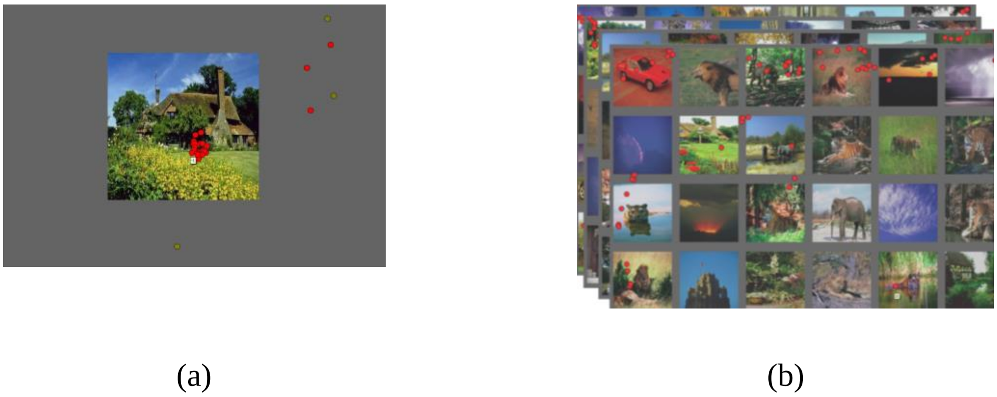
- Gaze data: The purpose of this study is to implicitly annotate images based on gaze data. As a result it is important that the information of gaze location on the screen is provided for the study.

Table 4 illustrates the properties of the datasets in 5 other eye-movement studies. From this table it is observable that except for the data set used for Vrochidis et al. [**91**] all other studies cannot completely fulfil the necessary properties of the required dataset for this study. In addition we can see in Vrochidis et al. [**91**] the user interaction with the contents is high, clicking on multiple images in short periods of time and constructing the ground truth based on the submitted clicks, and the users are biased researchers in computer science. Consequently as will be explained later in this chapter an experiment was set up to acquire a dataset that holds all the mentioned properties earlier in this section.

Table 4: Properties of used datasets in other studies

	Ajanki et al. [87]	**Jiao and Pantic** [38]	**Soleymani t al.** [88]	**Ramanathan et al.** [89]	**Vrochidis et al.** [91]
Type of visual content	Text	Image	Video and Image	Image	Image
Area of interest	N/A	Within image	Within video/images	Within image	Between images
Eye movement	Guided	Free	Free	Free	Free
Physical interaction	Minimal	Minimal	Minimal	Minimal	High
Database diversity	Diverse	Diverse	Diverse	Diverse	Diverse
Users' bias	Yes	Not available	No	Not available	Yes
Users' class	Senior researchers in computer science	Not available	Minimum degree: Undergraduate	Not available	Researchers in computer science
Gaze data	Available	Not available	Available	Available	Available

4.4 User Interface



(a) (b)

Figure 10: User interface. a) Start page that shows the Target Concept b) Appearance of the images in the screen in 6 columns and 4 rows

In the scenario of this framework the sets of images appear on the screen simultaneously in four rows and six columns, where every set is referred to as a 'page' throughout this thesis.

Similar User Interfaces (UI) with different scenarios are used in order to investigate different aspects of the behaviour of the user’s attention which can change. In each experiment it was tried to study how the scenario would affect the user’s strategies while looking at the images and as a consequence how the visual attention changes to be compatible to the strategy.

The introduced Graphical User Interface (GUI) in this thesis was experienced on a 32cm by 20cm screen with 1280 by 800 pixels. A subset of 700 images from the Corel database [**120**] was used as the ground truth database. This subset was formed of seven key concepts of sky, lion, tiger, elephant, building, car and vegetation with 100 images belonging to each category. These images and their associated key concept which is defined by two men aged 28 and 29 with computer science degree are provided in APPENDIX A. The original size of all of the images was 384 by 256 pixels in either landscape or portrait. To avoid the saliency effect of the shape of the images on the user's attention, the aspect ratio of all of the images were stretched from 3:2 or 2:3 into 1:1.

In all scenarios the raw data are gathered in three vectors that are sent to the processing unit in real-time for further analysis. Each data point in these vectors includes:

1- X coordinate of gaze
2- Y coordinate of gaze
3- Time Stamp that this coordination occurred

Best Choice, First 100 and All in Page are the names of the three principal scenarios that were used in this frame work.

Figure 10 shows the user interface as it appears in front of the users. In subfigure (a) the start page is demonstrated which appears to the user at the beginning of the experiment to show the concept to search for. In this figure the start page is the garden of a house which is considered as vegetation. The image that is chosen to show the Target Concept is chosen randomly from the database of all of the images to prevent any priory bias regarding the Target Concept.

After visiting the start page by clicking on a provided 'Next' button at the bottom right side of the screen the scenario starts. All of the images are chosen randomly from the database and appear just once in every page. In every page 5 images are chosen randomly from the Target Concept class. Depending on the scenario there are different conditions that by meeting them a user can proceed to the next page of the images or the scenario ends.

4.4.1 Best Choice scenario

The real-world case of this scenario is when a user has a specific concept, Target Concept (TC), in mind and they are exploring a database of images to find the ones that contain similar concept to the TC. This can take place when using a search engine or exploring the images of an on-line photo gallery such as Flickr or Facebook.

In these situations the users normally click on a limited number of images to enlarge them, read their corresponding comments, etc. However there remain many images that appeared on the screen which the user noticed but did not click on for various reasons resulting in no information for the system. The aim of this scenario is to simulate such situations.

In this scenario every time that the user clicks on an image while looking at it they are provided with the next page of 24 images. For performance measurement of the framework the users were told to imagine that they are responsible for selecting an image for the cover of a magazine from the images that appear on the screen. This image had to contain the same concept as a randomly selected image that appeared at the beginning of the experiment in the start page. The key concept of this image considered to be the TC in the user's mind.

The users were instructed to find the most similar images to the TC and click on it to proceed to the next page. To prevent the users from clicking on the first similar image that they face on each new page they were told that there is a point based system which considers how they look at other similar images and scores them. However the users were strongly reminded that their main goal is not to gain higher scores but to simulate an image search scenario. The point based system also helped keep the users motivated throughout the experiment.

4.4.2 First 100 scenario

This scenario tried to simulate the situation when the users are looking for a TC however it is not necessary to find all of them. So they might not look at every image to see whether it contains the Target Concept. This can take place when the users want to quickly find a product from the search results of a shopping website. In these situations users normally pass through the images and they do not mind missing some of the TC as they are provided with enough results and they are sure that they can find it quicker by a random scan path compared to when looking at every single image.

In this scenario every time that the user clicks on a provided Next button he/she is provided with the next page of 24 images. For performance measurement of the framework the users were told to freely look at the images and proceed to the next page as they wish. They were instructed that the scenario would continue until they had looked at 100 Target Concept images. The goal was set implicitly as naturally their intention was to finish the experiment as soon as possible.

4.4.3 All-in-Page scenario

This scenario tried to simulate the situation when the users are looking for a TC however it is necessary to find all of them and they are not allowed to miss any TC. So they might have to look at every image to see whether it contains the TC. This can take place when the users want to select the images that they appeared in from the photo album of an event.

In these situations users normally pass through the images and they try not to miss any of the TCs. Consequently the scan is normally more deliberate and less casual.

In this scenario every time that the user clicks on a provided Next button he/she is provided with the next page of 24 images. However the Next button is disabled until the user has looked at all of the 5 TCs in the page. To prevent the users from continuously checking the Next button to see whether it is enabled, the screen would notify the user that all of the Target Concepts had been visited by freezing with hashed lines over the page allowing the user to proceed to the next page.

As the true intention of the subjects is to finish the experiment as soon as possible, naturally in this scenario they start to look at the images one by one in the same order for all of the pages which is not desirable. As a result, to prevent the users from following a simple line by line predetermined path to finish the experiment, one TC randomly was chosen to be visited twice and the users were told that the machine somehow understand if their visit was just a passing by and it does not count the visits that are not based on natural gaze behaviour.

4.5 Experiments

In this experiment, 12 females and 4 men, aged between 18 and 36, volunteered to attend. None of the volunteers were studying computer science. Out of the 16 volunteers 12 of them disclosed their ethnicity or their country of origin which are: 1 Iranian, 1 Catalan, 1 mixed German and Nigerian, 1 British, 2 Bangladeshi, 1 Indian, 1 mixed Arab and British, 1 Australian, 1 Arab, 1 from Kashmir born in the UK, 1 Bangladeshi born in the UK. Because of the extreme diversity of the participants, it was not possible to conclude any result based on the ethnicity of the participants.

On average it took every participant 25 minutes to finish all four scenarios including the dedicated time for instructions of each scenario. All of the attendees were paid £5 for their contribution and they signed the Consent form of Queen Mary University of London with QMREC protocol number of QMREC2009/09. At the end of the experiment every volunteer filled a questionnaire and were asked about the conditions of the experiments including General Discomfort, Fatigue, Eye Strain, Difficulty in Focusing, Nausea, Difficulty in Concentration, Blurred Vision, Dizzy, Entertainment Level, Enjoyable, Easy to Learn and Easy to Use. They were asked to give a grade between 1 and 10 rating the strength of each category from very week to very strong respectively. Table 5 shows the average of the ratings of the users to these categories and a breakdown of the percentage of the users who gave their ratings within a specific range.

We can see most of the users, 93%, gave a rating more than 10 for ease of use and ease of learn to the interface and very few of them had problems with negative categories like discomfort and fatigue. Also more than half of the users were neutral about the experiment being enjoying and entertaining.

Table 5: The ratings of the 16 users to the 12 categories in the questionnaire, the values in the first three rows indicate the percentage of the users who gave a rating within the corresponding range of the column (All of the values except the ones for the Average Score field are in per-cent %)

Score range	**General Discomfort**	**Fatigue**	**Eye Strain**	**Difficulty in Focusing**	**Nausea**	**Difficulty in Concentration**	**Blurred Vision**	**Dizzy**	**Entertainment Level**	**Enjoyable**	**Easy to Learn**	**Easy to Use**
Less than 3	81.25	62.5	56.25	62.5	93.75	75	93.75	100	6.25	0	0	0
Between 3 - 7	12.5	31.25	18.75	25	6.25	18.75	6.25	0	50	56.25	6.25	6.25
More than 7	6.25	6.25	25	12.5	0	6.25	0	0	43.75	43.75	93.75	93.75
Average Score	2.25	2.81	3.75	3.25	1.37	2.50	1.93	1.25	5.75	6.62	8.75	8.56

5 Analysis of Gaze Features

This chapter studies the 27 gaze features that are extracted from the raw gaze movements in detail. These features empirically indicate different measurements of gaze movements. It is investigated how the factors that can shift one's volitional visual attention affect these features. Saliency of an image compared to the other images that appeared with it on the screen, Similarity of the image to the images that appeared with it on the screen and the Target Concept class variable of the image are the three factors that their effect on visual attention are examined.

The number of the features that have been used in most of the studies in the field of eye-tracking has been limited between 3 to 6 features. In most cases they include number of fixations, duration of fixations and its derivatives and saccades. Greene t al. [**121**] noted that the static scan paths and fixations are not enough to classify the user's task and Target Concept. Kozma et al. [**12**] used 17 features however 5 of them are specific to their scenario as they are calculated based on eye movements over circles that are formed of images. The most comprehensive feature extraction in this field was reported by Ajanki et al. [**87**] for the reading process with 22 features, but most of them (14 features) are not applicable in the image processing domain. Pasupa et al. [**13**] and Hardoon and Pasupa [**11**] used the same 33 features in two consecutive approaches emphasizing mostly on the spatial properties of the gaze movements and their features calculated based on within image fixations independent from the rest of the images on screen. All of these studies focused on the gaze features for individual objects [**87**] [**13**] [**11**].

In this study we extracted two different feature vectors (in total 27 features) where one of them contains the properties of eye-movements regardless of the underlying objects (Scan Path Feature Vector) and the other one contains information about the properties of gaze intersection with the image (Area of Interest Feature Vector).

In this chapter the features and their effect on classification quality are investigated with two types of statistical analysis: Analysis of Variance (ANOVA) and Correlation Coefficient. Because the sampling rate of the cameras of the faceLAB's eye tracker is at 60Hz rate, in place of measuring the time in milliseconds (ms), it is measured in number of frames where each frame is approximately 16.6 ms.

5.1 Feature vectors

As stated earlier there are two feature vectors that are extracted from gaze movements namely Area of Interest Feature Vector (AoI-FV) and Scan Path Feature Vector (SP-FV). The former is formed of the features that hold data regarding the attention of the user to a specific area on the screen which is the area of an image on the screen and the latter is formed of the features that represent the dynamics of the gaze movement.

5.1.1 Area of Interest Feature Vector (AoI-FV):

This feature vector is formed of the features that represent the attention of the user to a specific area; in the experiments of this thesis the area covered by an image is an AoI. The data in these features are collected or calculated after the user is finished with one page and requested for a new page. For each page the number of data points for AoI-FV is equal to the number of images that appeared on the page. The list of features in this feature vector and their explanation are as follows:

1- iAd: This feature measures the total time that a user spent on an image on the same page in different visits.

2- iAdp: This feature shows the proportion of iAd over the total time that the user spent on the page that the image appeared in.

3- iAdr: This feature shows the order of iAd value of an image compared to the iAd values of other image that appeared on the same page.

4- iAv: This feature measures the average time that a user spent on an image on the same page in different visits.

5- iAvp: This feature shows the proportion of iAv over the total time that the user spent on the page that the image appeared in.

6- iAvr: This feature shows the order of iAv value of an image compared to the iAv values of other image that appeared on the same page.

7- iFt: This feature measures the time at first visit that a user spent on an image in a specific page.

8- iFtp: This feature shows the proportion of iFt over the total time that the user spent on the page that the image appeared in.

9- iFtr: This feature shows the order of iFt of an image compared to the iFt of other image that appeared on the same page.

10- iMx: This feature measures the maximum time that a user spent on an image on the same page in different visits.

11- iMxp: This feature shows the proportion of iMx over the total time that the user spent on the page that the image appeared in.

12- iMxr: This feature shows the order of iMx value of an image compared to the iMx values of other image that appeared on the same page.

13- iVn: This feature counts the number of times that the user's gaze entered the area of an image from another image.

14- iSn: This feature shows the number of times that an image was in the path of the gaze movement but it was skipped by the user.

5.1.2 Scan Path Feature Vector (SP-FV):

This feature vector is formed of the features that represent the properties of gaze as scan path. The data in these features are collected in a stack form that by moving the gaze point over the images they are added at the top of the stack. Along with the features the unique ID of every image in the scan path is stacked to keep a record for further investigation on which gaze behaviour belongs to which image or pair of visited images successively - the one at the start (pre-movement image) and the one at the end (post movement image) of a gaze movement from image to image. For each page the number of data points for SP-FV is equal to the total number of visits (including revisits) that the user paid to different images on the page. The list of features in this feature vector and their explanation are as follows:

1- tAngle: This feature shows the angle by which the gaze entered the area of an image.
2- tPnPSpd: This feature measures the total visit length of the pre-movement and post-movement images divided by the distance between the two images.
3- tPoSpd: This feature measures the total visit length of the post-movement image divided by the distance between the post and pre movement images.
4- tPrDist: This feature holds the value of the distance between pre-movement and post-movement images.
5- tPrPTime: This feature shows the ratio between the visit length of pre-movement image over the visit length of the post-movement image.
6- tPrSpd: This feature measures the total visit length of the pre-movement image divided by the distance between the post and pre movement images.
7- tTime: This feature shows the visit length of every image in the scan path.
8- tPreGrad: This feature shows the difference between the visit length of post-movement and pre-movement images.
9- tMnGrad: For every image this feature shows the average of the tPreGrad values when the image is the post-movement image and when the image is the pre-movement image.
10- t2reg: If a visit to an image is happening for the second or more time this feature shows how long has passed since the previous visit.
11- tVn2reg: If a visit to an image is happening for the second or more time this feature shows how many other images were visited since the previous visit.
12- lmn: This feature shows zero or the visit length of an image if it is a local minimum compared to the visit length of the previous visited and next visited images.

13- lmx: This feature shows zero or the visit length of the current image if it is a local maximum compared to the visit length of the previous and next visited images.

5.1.3 Gaze features used in other studies

Table 6 shows the list of the features that are used in this thesis and their equivalent names or numbers in other studies. The features in Soleymani et al. [**88**] and Ramanathan et al. [**89**] were also considered however because these studies focus on within image gaze behaviour compared to between image gaze behaviour in this study there cannot exist similar features providing similar information. In Table 6 we can see that 17 features are not used in any other studies. Out of these new features, as we will see in the Feature Selection section of Results chapter, iMxr, iAdr, iAvp and iMxp are the features that bear more information for classification than other features.

Table 6: The gaze features from this thesis that are used in other studies with a different name or number

Features	**Kozma et al.** [12]	**Ajanki et al.** [87]	**Pasupa et al.** [13]	**Vrochidis et al.** [91]
iAd	SumLength	17	totalFixLen	2
iAdp	RatioTotal			
iAdr				
iAv	MeanLength			3,4
iAvp				
iAvr				
iFt	FirstVisit	6		
iFtp		21		
iFtr				
iMx				
iMxp				
iMxr				
iVn	RevisitCount	1	nJumpsFix	1,5
iSn		22		
tAngle				
tPnPSpd				
tPoSpd				
tPrDist		9,10		
tPrPT				
tPrSpd				
tTime		5, 8		
tPreGrad				
tMnGrad				
t2reg		19		
tVn2reg				
Lmn				
Lmx				

In each of the studies in Table 6 the gaze features that are not used in this thesis hence did not appear in the table are listed as follows. The reasons that these features are not used include:

- The scenario and user interface that were used for the study were different in a way that it was not possible to use the same features. For example most of the features in Kozma et al. [**12**] were dependant on the ring of images that appeared on the screen.
- The record format of these studies was different from the record format in this thesis. For example Pasupa et al. [**13**] uses number of fixations within an image however in this thesis the within image data were not recorded or the gaze coordinates inside the image because these data could not be accurate because of the 0.5 degrees error of the eye-tracker.

The following are the features in Kozma et al. [**12**] which are not used in this study:

- IsRandom: Whether the images were shown in the first stage (before any user feedback)
- FixTimeSpread: Standard deviation of fixation occurrance times
- MaxContView1: Maximum continuous viewing time without viewing other images
- MeanContView: Mean length of continuous viewing sessions of image
- MaxContView2: Maximum continuous viewing time without fixating outside the image
- RatioRing: Proportion of total viewing time over total viewing times in the same ring.
- MeanSaccLength: Mean Length of saccade before fixation
- MeanPrevImage: Proportion of times when previous fixation also over this image
- MeanPrevEmpty: Proportion of times when previous fixation over empty space
- MeanPrevRing: Proportion of times when previous fixation over the same ring
- FirstVisitIndex: How many images viewed on this ring before the first fixation on image
- PrevDist: Average distance from previously viewed image on the same ring

The following are the features in Ajanki et al. [**87**] which are not used in this study:

- 2: fixations to the word when the word is first encountered
- 3: Did a fixation occur when the line that the word was in was encountered for the first time?
- 4: Did a fixation occur when the line that the word was in was encountered for the second time?
- 7: Sum of durations of fixations to a word when it is first encountered
- 11: Distance (in pixels) between the fixation preceding the first fixation on a word and the beginning of the word
- 12: Distance (in pixels) of the first fixation on the word from the beginning of the word
- 13: Distance (in pixels) between the last fixation before leaving the word and the beginning of the word
- 18: Did a regression initiate from the following word?
- 20:Mean pupil diameter during fixation

The following are the features in Pasupa et al. [**13**] which are not used in this study:

- numFix: total number of fixations
- meanFixLen: mean length of fixations
- fixPrct: percentage of time spent in fixations
- maxAngle: maximal angle between two consecutive saccades
- landXFix: x-coordinate of the first fixation
- landYFix: y-coordinate of the first fixation
- exitXFix: x-coordinate of the last fixation
- exitYFix: y-coordinate of the last fixation
- xSpreadFix: difference between largest and smallest x-coordinate
- ySpreadFix: difference between largest and smallest y-coordinate
- elongationFix: ySpreadFix/xSpreadFix
- firstFixLen: length of the first fixation
- firstFixNum: number of fixations during the first visit
- distPrev: distance to the fixation before the first

- durPrev: duration of the fixation before the first

5.2 Finding the saliency and similarity scores

For every image that appeared on the screen a saliency score and two similarity scores, Background Similarity and Main-concept Similarity, are empirically calculated. The assigned values help better understand the effect of these two factors on one's variation of visual attention in different situations. Later in section 6.2.1 we use this information to investigate whether selecting the features based on their maximum dependency on Target Concept and minimum dependency on saliency and similarity can improve the results or performance.

5.2.1 Similarity

In this thesis the similarity is considered the human level likeliness of the content of two images that can shift one's visual attention. For example regardless of low level features like colour histogram, edge histogram, etc. when an image contains a Tiger and another image contains a lion how similar the contents of two images are according to a human observer.

Lepetit et al. in [**122**] developed a key-point based method by matching them from an input image against a target object. In their method they find different key points on the target object and find the corresponding points on the input image. In this way they can decide the similarity between two images. This study was used for 3D object detection and pose-estimation. Jing et al. in [**123**] proposed a model for top-down visual attention based on similarity distance. In this study by the help of similarity distance they adjust the weights of intensity, colour and orientation in order to compute visual expectation. Xiong et al. in [**124**] segmented and clustered the visual objects in an image and calculate a probability vector to associate the image objects to keyword. Then for every pair of images they calculate their similarity by correlating the probability vector of the images. Recently Chalom et al. in [**125**] summarised 4 different approaches for measuring image similarity that include "pattern recognition, comparison of frames in a video sequence, image stabilization using homographic transformation and using image feature points to compute similarities and generate an image mosaic”. All of the above methods use machines (mainly by using low level features or guessing the image contents by classification methods) for calculating the similarity between the images; However, as stated earlier the similarity score that is considered in his study needs human level evaluation in order to have a real understanding of the effect of this factor on gaze features.

To calculate the similarity score, the www.axemoon.com webpage was developed where the users could assign similarity scores to each appeared pair of images during their session. Because it was not feasible to compare every single image to all other images in the database[*], from each of the seven concept classes (sky, lion, tiger, elephant, building, car and vegetation) 10 images were chosen to represent the class in the web-evaluation experiment. To do so first the 100 images in every class where divided into 2 groups based on their similarity by three people. Then a subset of images that all the three groupings agreed was formed from each group. Next, from every subset 5 images were selected randomly. The 10 selected images from the two groups formed the images that represented their corresponding concept class in the web-evaluation experiment. As a result there were 2485 pairs of images (including comparison of an image to itself) for the web trial. With 7 different classes there are 21 class pairs (excluding the comparison of the classes to themselves with total 55 image pairs for each) that in each class pair there are 100 image pairs.

Figure 11 shows a screen shot of the developed web evaluation page. When a user enters the page, 28 of the image pairs that have the minimum number of trials are selected from the database as a session and are rated by them. The web evaluation is one page that is divided into four sections:

1- Image frame: This section is located at the top centre of the page and the to-be-rated images of an image pair are presented to the user inside a grey frame.

[*] With 700 images in the database there are 245350 unique pairs of images that need to be rated. Considering every user is patient enough to rate 100 images in every session, around 2500 users are required to rate all of the image pairs.

2- User information form: This is the top left corner of the page where the users can enter their gender and age information. If the session is completed without completing this form the user is recorded as ‘anonymous’.

3- The rating section: This section is at the bottom centre of the page. This is where the users can assign a Main-concept Similarity and Background Similarity to the pair of the images with the provided rings. There are two red and blue 5-ring groups by which the users can assign the similarity. To help the users understand the level of the similarity they are assigning, two boxes show the grade of the selected circles with the values of very low, low, similar, high and very high as the users select the larger rings respectively. In this section there is a provided ‘Next’ button that becomes available when both Main-concept Similarity and Background Similarity ratings are completed and browses the next pair of images for rating in the grey frame.

4- Score board: To keep the users motivated to make ratings more similar to the moving average, finish the session and try new sessions, a score board is provided that shows the calculated score of the top ten users. Every user’s score is calculated by comparing their class similarity ratings to the average of the previous users.

Overall 57 users contributed to the 98 tried sessions of the web evaluation and rated the 2485 image pairs in the database. Out of these users 12 of them participated anonymously and the maximum number of sessions that a user tried was 5 sessions. The details of all of the participants are shown in Table 7. For every class pair the average of the similarity scores of the entire image pairs in the group is calculated. This average represents the similarity grade of the two image classes. In this thesis we use the average similarity of the classes of two images as the similarity score between them. For example if the average similarity of the web evaluated images that represent Tiger and Lion classes is 3 out of five for all Tiger-Lion pair of images, in the eye-tracking experiment for every Tiger-Lion pair of images the similarity is considered 3 out of 5.

For the SP-FV as the user’s gaze moves from an image to another image the calculated averages of the Background Similarity and Main-concept Similarity between the classes of the two sequentially visited images are stacked at the top of two similarity vectors where, as will be explained later, they are used for the ANOVA test and correlation coefficient calculation. For example when a user looks at a Tiger first then looks at the lion next; the similarity of 3 is pushed at the top of the stack.

For the AoI-FV it is impossible to use the mentioned values because there is no sequence of visits available. Consequently for every appeared image on the screen the overall Background Similarity and Main-concept Similarity scores of the image in the page are calculated based on the similarity of the image to every image on the page and their corresponding distances. It should be noted that because we are investigating the effect of similarity on guiding one's visual attention, and because distance has a negative effect on this effect [**126**] [**127**] (i.e. when two images are further from each other their similarity has less effect in attracting one's gaze) we reduced the similarity factor between the two images by a factor called distance coefficient noted as $dcoef_{mn}$.

Let $dcoef_{mn}$ be the distance coefficient of the m_th and n_th images (of the 24 appeared images on the screen) that decreases exponentially on a normal curve as d_{mn}, the distance between the two images, increases:

$$dcoef_{mn} = e^{-(\frac{d_{mn}^2}{4})}$$

The overall similarity of the m-th image, noted as SIM_m, can be calculated via equation where sim_{mn} indicates the average similarity score that is assigned to the classed of the m_th and n_th images. It should be noted that in this equation the similarity effect of an image on itself is removed.

$$SIM_m = \frac{\sum_{n=1}^{24} sim_{mn} \times dcoef_{mn} - sim_{mm}}{\sum_{n=1}^{24} dcoef_{mn} \ - 1}$$

For example to calculate the similarity score of the 5th image (SIM_5) on the screen for the AoI-FV, $sim_{5,10}$ shows the similarity between the classes of the 5th image and the 10th image on the screen and $dcoef_{5,10}$ is the calculated distance coefficient between the two images.

This similarity score shows that overall how similar is an image to the rest of the images on the screen with more emphasis on the similarity of the image to its close neighbourhood [**128**]. So if there are many images on the screen that the calculated average similarity of their class is high, the similarity factor of all of them increases if they are located close to each other. For example when we have 2 images with Tigers and 3 images with lions on the same screen and the Tiger-Tiger, Lion-Tiger and Lion-Lion class pairs all have high value of similarity based on the web evaluation, these images gain higher similarity score on the screen compared to the situation that there is only one image from the Tiger or Lion class.

Table 7: Details of participants in the web evaluation

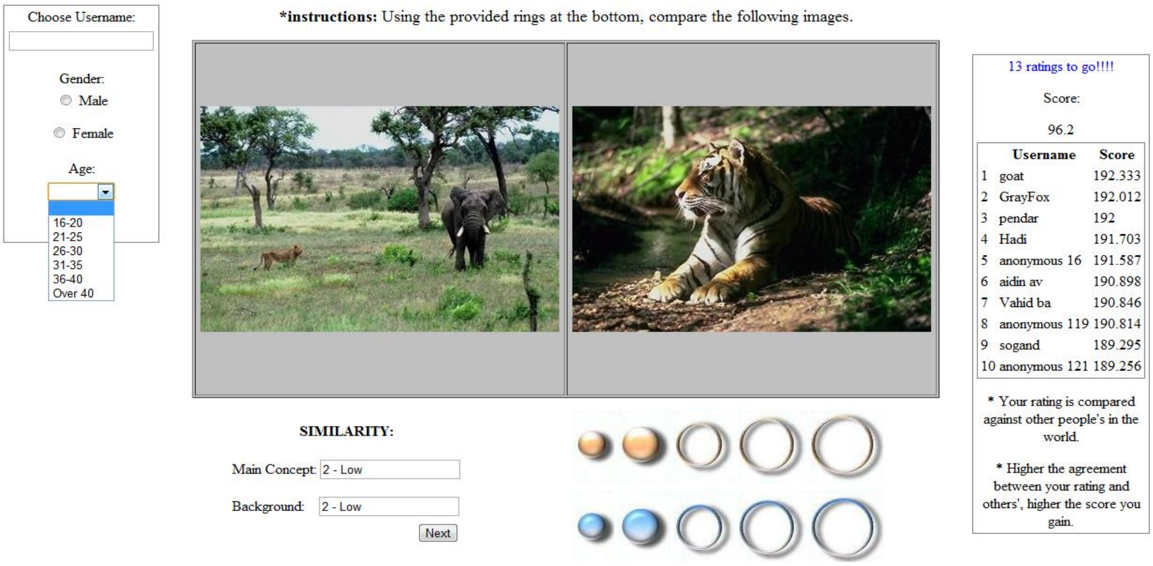
Age	16-20	21-25	26-30	31-35	36-40	Over 40	Not filled	Total
Female	2	6	6	2	1	1	1	19
Male	1	5	13	4	0	2	1	26

5.2.2 Saliency

In this thesis the saliency score shows the possibility grade that an image is fixated on before or for a longer period than the rest of the images on the same screen. In the visual attention field, saliency is defined [**129**] as the effect that can attract one’s visual attention in the bottom-up process because it makes an area more apparent to the eyes compared to the rest of neighbouring area. An example of saliency is a red circle in the middle of a white paper which can quickly attract one’s visual attention.

Itti and Koch in [**129**] modeled the visual attention computationally by considering five main factors. Thses include, 1) Dependance of perceptual saliency on the surrounding context, 2) “A unique ‘saliency map’ that topographically encodes for stimulus conspicuity over the visual scene has proved to be an efficient and plausible bottom-up control”. 3) Inhibition of return, which is equal as inhibition of regression in this thesis and means prevention of returning of the visual attention to a previously attended area. 4) The interaction between attention and eye-movement. 5) Object recognition limits the selection of attended location for saliency. It is based on Itti and Koch in [**129**] that Harel et al. [**130**] develop their work which is developing a graph based visual saliency map for the given images.

In this thesis saliency is calculated by the help of the Graph-Based Visual Saliency (GBVS) model [**129**] [**131**] [**130**] which tries to imitate the bottom-up process resulting from inadvertent shift of attention towards a visually emergent area of the screen. To calculate the saliency score of the images over the course of experiment, the saliency map of every constructed page is generated by the GBVS model, Figure 12, by evaluating the snap shot of the whole page by the GBVS model. The saliency score of an image is the average of the grey level values of the upper quartile pixels in an image area. This is because if an area of an image is salient to the eyes and attracts the gaze point, it attracts the gaze point to the entire image regardless of the saliency of the rest of the image.



Choose Username:

Gender:
○ Male
○ Female

Age:
16-20
21-25
26-30
31-35
36-40
Over 40

***instructions:** Using the provided rings at the bottom, compare the following images.

SIMILARITY:

Main Concept: 2 - Low

Background: 2 - Low

Next

13 ratings to go!!!!

Score:

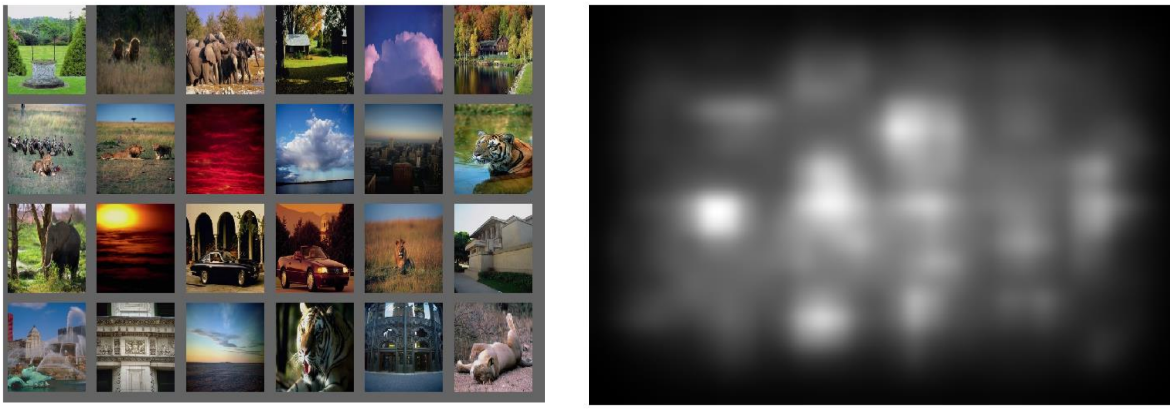
96.2

	Username	Score
1	goat	192.333
2	GrayFox	192.012
3	pendar	192
4	Hadi	191.703
5	anonymous 16	191.587
6	aidin av	190.898
7	Vahid ba	190.846
8	anonymous 119	190.814
9	sogand	189.295
10	anonymous 121	189.256

* Your rating is compared against other people's in the world.

* Higher the agreement between your rating and others', higher the score you gain.

Figure 11: Screenshot of one of the web evaluation sessions. Completion of the small form at the top left corner is optional for the users. In case it is left blank, the users can contribute to the experiment anonymously



(a) (b)

Figure 12: Finding the saliency map of a page by the GBVS model a) The page as it appeared in front of the user b) Constructed saliency map based on the image in (a)

This is because throughout this research an image is considered to be the Area of Interest and as a result the attention to a sector of the AoI is considered to be the attention to the whole area. Consequently, total saliency of the AoI is reflected by the upper 25 percentile of the pixels of the image. Furthermore as can be seen in Figure 12 these pixels are adjacent to each other and make a larger area salient.

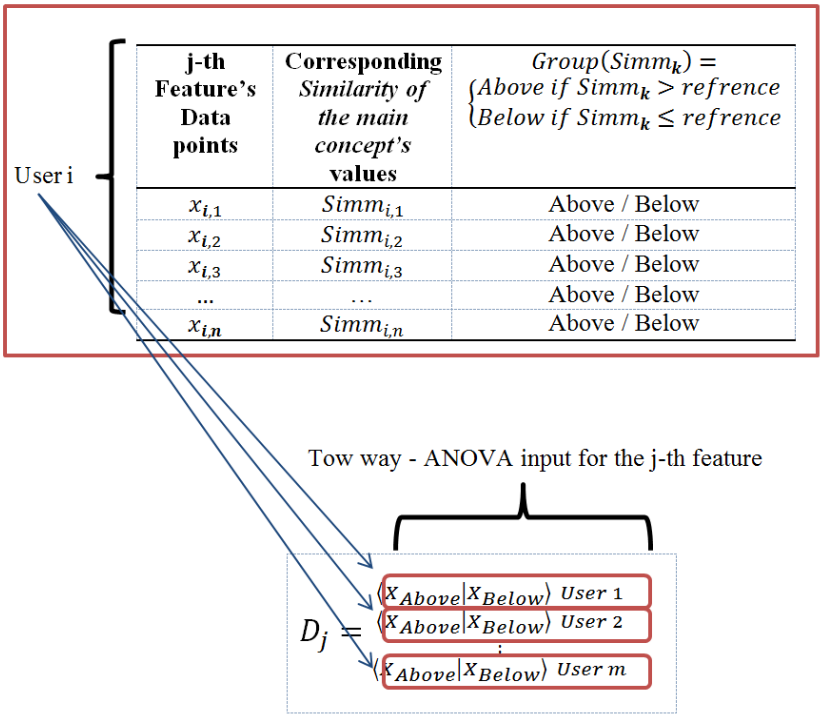
5.3 ANOVA test

In this section the properties of the gaze features that are extracted during the different scenarios are studied. The effect of the factors of saliency of the images in a page, similarity between the images in a page and being in the Target Concept class on the feature values are investigated.

For every feature the entire recorded data in different pages for all of the users who tried the same scenario are put together and divided into two groups depending on the factor under investigation. For the saliency effect the data points of the features are divided into two groups of 'Above' and 'Below' based on their saliency being above or below the median saliency of every data point. For the similarity factor in the AoI-FV the reference is defined as the median of the data points in the feature. Also for the SP-FV the reference is defined 2.5 which is the middle value of similarity scores assigned to the images. This is because there is a chance that the reference point might be set to zero[*] and all of the images are grouped in the 'Above' class. For the Target Concept class variable factor the data of a feature are divided into TC+ and TC- classes depending on them belonging to the Target Concept class.

Figure 13 shows an example of dividing the data points in a feature for the i-th user into two groups of Above and Below based on the Simm values and integrating the grouped data for all users in order to run the Two-Way ANOVA test that tests the effect of different users and the two groups on the values of the j-th feature.

[*] In the SP-FV every data point shows the properties of gaze intersection with a new image. The data points of the similarity are the similarities between the concept classes of every two consecutive visited images and a user might look at a large number of images with zero similarity.



j-th Feature's Data points	Corresponding *Similarity of the main concept's* values	$Group(Simm_k) = \begin{cases} Above\ if\ Simm_k > refrence \\ Below\ if\ Simm_k \le refrence \end{cases}$
$x_{i,1}$	$Simm_{i,1}$	Above / Below
$x_{i,2}$	$Simm_{i,2}$	Above / Below
$x_{i,3}$	$Simm_{i,3}$	Above / Below
…	…	Above / Below
$x_{i,n}$	$Simm_{i,n}$	Above / Below

User i
Tow way - ANOVA input for the j-th feature
$D_j =$
$\langle x_{Above} | x_{Below} \rangle$ User 1
$\langle x_{Above} | x_{Below} \rangle$ User 2
$\langle x_{Above} | x_{Below} \rangle$ User m

Figure 13: Dividing the data of features into two groups (Above and Below), integrating the data from different users and constructing the data structure (D_j – Data structure for the j-th feature) for the ANOVA test

After dividing the features into two groups, the two-way ANOVA test was used to investigate whether there are any differences between the values in the two groups of a feature. The two-way ANOVA produces three groups of result which are the effect of the users on the values of the features, the effect of the saliency, similarity or Target Concept class factors and the mutual effect of the factors and the users.

The main interest of this thesis is in the effect of the saliency, similarity or Target Concept class which are shown in Table 8 to Table 11, the results of the two other types of effects are presented in APPENDIX B.

5.3.1 Investigating saliency effects on gaze patterns by ANOVA

Table 8 shows the ANOVA results of the saliency effect for the Best Choice, the First 100 and the All-in-Page scenarios. The values in this table are the F statistics and their corresponding p-values of the saliency effect. The User effect and mutual effect of user and saliency are provided in APPENDIX B.

In this table for all of the scenarios for the majority of the features the saliency has a great effect with very high F statistic values and low corresponding p-values. There are between two to four features in each scenario (highlighted with dark shading in the table) that are not affected by the saliency and these features are not the same for any two scenarios except for iAdp, iAdr, and iMxp. This is a proof of the dependency of the saliency to the scenario. These features for every scenario are: iAdp, iAdr, iMxp and tPrDis for Best Choice scenario; tMnGrad, t2reg, tVn2reg for First 100 scenario; iAdr, iFtp, iMx, iMxp for All-in-Page scenario.

Table 8: Results of Two-Way ANOVA based on the effect of the saliency factor

	Best Choice		First 100		All-in-Page	
Features	F Stat.	P value	F Stat.	P value	F Stat.	P value
iAd	0.9701	0.3247	5.7582	0.0164	2.0712	0.1501
iAdp	0.0113	0.9152	2.7255	0.0988	8.2406	0.0041
iAdr	9.9756	0.0016	14.4650	0.0001	0.0015	0.9693
iAv	36.8663	0.0000	49.6766	0.0000	21.2137	0.0000
iAvp	11.5435	0.0007	32.3430	0.0000	4.5264	0.0334
iAvr	29.2572	0.0000	22.9606	0.0000	17.9932	0.0000
iFt	18.4832	0.0000	21.7818	0.0000	7.3311	0.0068
iFtp	6.4452	0.0111	13.1762	0.0003	0.1517	0.6970
iFtr	32.2975	0.0000	28.9948	0.0000	16.6166	0.0000
iMx	10.8694	0.0010	14.6273	0.0001	0.8473	0.3573
iMxp	2.8867	0.0894	8.3718	0.0038	0.9837	0.3213
iMxr	19.6538	0.0000	21.4984	0.0000	3.0832	0.0791
iVn	52.9472	0.0000	34.5621	0.0000	82.6966	0.0000
iSn	84.6233	0.0000	29.5657	0.0000	54.8621	0.0000
tAngle	147.3352	0.0000	25.7203	0.0000	40.8323	0.0000
tPnPSpd	28.9116	0.0000	18.5548	0.0000	27.6998	0.0000
tPoSpd	23.2484	0.0000	14.0987	0.0002	29.6792	0.0000
tPrDist	7.2305	0.0072	20.4398	0.0000	9.0326	0.0027
tPrPT	6.1982	0.0128	9.5696	0.0020	2.5379	0.1112
tPrSpd	33.4155	0.0000	17.1059	0.0000	23.9712	0.0000
tTime	27.5328	0.0000	26.6431	0.0000	20.4102	0.0000
tPreGrad	48.8826	0.0000	26.9828	0.0000	42.3212	0.0000
tMnGrad	1.7226	0.1894	0.0304	0.8616	12.8143	0.0003
t2reg	3.9026	0.0483	0.1197	0.7294	46.3377	0.0000
tVn2reg	7.4170	0.0065	0.0039	0.9501	38.4655	0.0000
Lmn	10.4231	0.0013	4.2934	0.0384	8.8633	0.0029
Lmx	4.1934	0.0406	2.0019	0.1572	2.0406	0.1532

The low F statistic value and the high p values for these features show that we don’t have enough evidence to show there is a significant difference between the values of the data points in these features that belong to images with high saliency score and the ones that belong to images with low saliency score.

5.3.2 Investigating similarity effects on gaze patterns by ANOVA

Table 9 and Table 10 show the ANOVA results of the main concept and Background Similarity effects for the Best Choice, First 100 and All-in-Page scenarios. The values in these tables are the F statistics and their corresponding p-values of the similarity effects. The user effect and mutual effect of user and similarity are provided in APPENDIX B.

Table 9: Results of Two-Way ANOVA based on the effect of the Main-concept Similarity factor

	Best Choice		First 100		All-in-Page	
Features	F Stat.	P value	F Stat.	P value	F Stat.	P value
iAd	118.2493	0.0000	105.6076	0.0000	248.1071	0.0000
iAdp	107.8283	0.0000	79.5760	0.0000	441.4963	0.0000
iAdr	248.0594	0.0000	64.1566	0.0000	332.2571	0.0000
iAv	109.4175	0.0000	76.4210	0.0000	260.0217	0.0000
iAvp	55.1831	0.0000	43.7692	0.0000	269.5564	0.0000
iAvr	92.2566	0.0000	34.6474	0.0000	156.5398	0.0000
iFt	67.1008	0.0000	56.8808	0.0000	173.0875	0.0000
iFtp	35.2330	0.0000	41.3203	0.0000	198.1483	0.0000
iFtr	89.4793	0.0000	44.7836	0.0000	165.5459	0.0000
iMx	134.8884	0.0000	96.4127	0.0000	311.0526	0.0000
iMxp	88.7053	0.0000	61.9219	0.0000	375.2684	0.0000
iMxr	230.4131	0.0000	68.1186	0.0000	323.8898	0.0000
iVn	62.5389	0.0000	13.4521	0.0002	35.9531	0.0000
iSn	1.1961	0.2741	1.1889	0.2756	0.8666	0.3520
tAngle	9.1294	0.0025	0.0499	0.8232	4.9224	0.0265
tPnPSpd	63.3265	0.0000	24.8062	0.0000	31.5444	0.0000
tPoSpd	52.9662	0.0000	25.1240	0.0000	33.5663	0.0000
tPrDist	9.2897	0.0023	5.7867	0.0162	0.0148	0.9030
tPrPT	27.5467	0.0000	15.6806	0.0001	26.6697	0.0000
tPrSpd	79.8465	0.0000	30.1855	0.0000	39.1146	0.0000
tTime	107.5363	0.0000	32.7841	0.0000	58.1593	0.0000
tPreGrad	51.7466	0.0000	24.6699	0.0000	48.5901	0.0000
tMnGrad	11.0077	0.0009	39.1066	0.0000	127.0344	0.0000
t2reg	4.7378	0.0295	8.3670	0.0038	7.4103	0.0065
tVn2reg	3.3792	0.0661	10.2755	0.0014	7.3732	0.0066
Lmn	146.0940	0.0000	18.3155	0.0000	40.4607	0.0000
Lmx	60.0348	0.0000	9.7410	0.0018	17.7007	0.0000

Table 10: Results of Two-Way ANOVA based on the effect of the Background Similarity factor

	Best Choice		First 100		All-in-Page	
Features	F Stat.	P value	F Stat.	P value	F Stat.	P value
iAd	59.6713	0.0000	54.4308	0.0000	140.9708	0.0000
iAdp	48.2000	0.0000	15.9785	0.0001	235.0284	0.0000
iAdr	143.6991	0.0000	31.7614	0.0000	234.8299	0.0000
iAv	55.5554	0.0000	33.4490	0.0000	150.6449	0.0000
iAvp	25.0584	0.0000	3.9786	0.0461	137.2158	0.0000
iAvr	48.6406	0.0000	12.8602	0.0003	123.1045	0.0000
iFt	33.4003	0.0000	31.1049	0.0000	86.1885	0.0000
iFtp	15.1867	0.0001	6.1089	0.0135	89.7360	0.0000
iFtr	44.7965	0.0000	22.5133	0.0000	101.6135	0.0000
iMx	67.2888	0.0000	49.4507	0.0000	178.9083	0.0000
iMxp	40.2338	0.0000	9.6439	0.0019	199.1541	0.0000
iMxr	126.8698	0.0000	31.7951	0.0000	207.2149	0.0000
iVn	38.6893	0.0000	6.2628	0.0124	30.0913	0.0000
iSn	0.3123	0.5763	0.0396	0.8422	0.7270	0.3939
tAngle	5.5711	0.0183	0.0820	0.7746	5.7475	0.0165
tPnPSpd	59.0671	0.0000	19.6416	0.0000	38.4070	0.0000
tPoSpd	51.4740	0.0000	20.8137	0.0000	38.8602	0.0000
tPrDist	7.8952	0.0050	8.1255	0.0044	0.0000	0.9948
tPrPT	33.6042	0.0000	16.1384	0.0001	22.9096	0.0000
tPrSpd	79.9547	0.0000	27.3704	0.0000	51.2366	0.0000
tTime	90.8078	0.0000	33.0610	0.0000	84.5068	0.0000
tPreGrad	47.4880	0.0000	27.0984	0.0000	65.4671	0.0000
tMnGrad	7.9487	0.0048	44.5735	0.0000	160.6296	0.0000
t2reg	6.5817	0.0103	10.7625	0.0010	9.2496	0.0024
tVn2reg	3.5032	0.0613	10.6614	0.0011	10.1606	0.0014
Lmn	139.3026	0.0000	27.2351	0.0000	45.7828	0.0000
Local maximum visit (lmx)	52.7282	0.0000	12.0368	0.0005	32.6754	0.0000

Like saliency for all of the scenarios for the majority of the features the Main-concept Similarity has a great effect with very high F statistic values and low corresponding p-values. We can see that there is no feature in the Best Choice scenario that is not affected by the Main-concept Similarity effect and there are very few features that are not affected by the same effect (highlighted by dark shading). We can also see that the only feature that is not impacted by one of the effects in all of the scenarios is the skip number, iSn. These results show that in the Best Choice, First 100 and All-in-Page scenarios, almost all of the gaze movement features are affected by the similarity effects.

5.3.3 Investigating the effect of Target Concept class on gaze patterns by ANOVA

Table 11 shows the ANOVA results of the Target Concept class effects for the Best Choice, First 100 and All-in-Page scenarios. The values in this table are the F statistics and their corresponding p-values of the similarity effects. The User effect and mutual effect of the user and the Target Concept are provided in APPENDIX B.

Table 11: Results of Two-Way ANOVA based on the effect of the Target Concept class factor

	Best Choice		First 100		All-in-Page	
Features	F Stat.	P value	F Stat.	P value	F Stat.	P value
iAd	3116.3688	0.0000	1886.1467	0.0000	1692.1953	0.0000
iAdp	3739.1578	0.0000	1914.4860	0.0000	3598.3498	0.0000
iAdr	2466.7453	0.0000	1513.1079	0.0000	1659.3135	0.0000
iAv	2781.6623	0.0000	1410.3656	0.0000	1465.6768	0.0000
iAvp	1706.3187	0.0000	1144.8940	0.0000	1542.6794	0.0000
iAvr	839.7420	0.0000	824.4711	0.0000	633.0889	0.0000
iFt	1534.1091	0.0000	1103.8316	0.0000	1210.7458	0.0000
iFtp	1084.1812	0.0000	943.5327	0.0000	1468.7992	0.0000
iFtr	916.3436	0.0000	794.5083	0.0000	876.9811	0.0000
iMx	3348.8389	0.0000	1932.2837	0.0000	2209.1789	0.0000
iMxp	2752.6724	0.0000	1545.3923	0.0000	2710.8857	0.0000
iMxr	2426.0917	0.0000	1525.2577	0.0000	1837.9401	0.0000
iVn	623.0772	0.0000	133.7967	0.0000	178.4285	0.0000
iSn	3.9516	0.0469	0.7973	0.3720	0.7808	0.3769
tAngle	7.5355	0.0061	0.3329	0.5640	24.0409	0.0000
tPnPSpd	714.3968	0.0000	226.1173	0.0000	416.2207	0.0000
tPoSpd	475.4991	0.0000	194.0532	0.0000	386.8654	0.0000
tPrDist	0.2433	0.6218	8.6165	0.0033	2.1390	0.1436
tPrPT	587.4281	0.0000	110.6793	0.0000	284.8045	0.0000
tPrSpd	740.0673	0.0000	227.4385	0.0000	385.6793	0.0000
tTime	2956.7377	0.0000	1147.3720	0.0000	2042.4809	0.0000
tPreGrad	1952.1625	0.0000	787.9474	0.0000	1091.1252	0.0000
tMnGrad	3.7644	0.0524	16.3436	0.0001	23.8999	0.0000
t2reg	71.2447	0.0000	20.3980	0.0000	31.8357	0.0000
tVn2reg	32.8203	0.0000	13.4871	0.0002	11.6521	0.0006
Lmn	157.3224	0.0000	44.9918	0.0000	68.8848	0.0000
Lmx	1100.6864	0.0000	386.9540	0.0000	699.6464	0.0000

In this table the F statistic values are significantly greater than the previous effects. As highlighted by dark shading in the table, there are very few features that are not affected by the Target Concept class. There are only one feature for each of the Best Choice scenario and All-in-Page scenario that is not influenced by the Target Concept class which are Distance from Previous (tPrDist) and Skip Number (iSn) respectively. Also iSn and entrance angle (tAngle) are the only two features that we do not have any evidence they are affected by the TC class in the First 100 scenario. Like Background Similarity effect, iSn has very low values of F.

5.4 Correlation coefficient between gaze features and the three factors

The ANOVA results showed saliency, similarity and Target Concept class have an impact on the values of the gaze feature. The magnitude of this impact was investigated by measuring the correlation coefficient between the features and these factors.

It should be reminded that for every data point in a feature there is a corresponding value in the mentioned factors where the values in the saliency factor show the saliency score of the visited image that the feature’s value belongs to, similarity factor values show the similarity score of the images and Target concept vector shows whether the values of a data point in the features belong to an image with target concept or not. The graphs in Figure 14 show the magnitude of linear correlation of all features and these three factors.

To do so, the data-points of every feature across all pages and their corresponding values of saliency, similarity or Target Concept class variable were integrated in two large vectors and their correlation coefficient was calculated for every user. Figure 14 shows the average correlation coefficient of all users between all of the features and saliency, similarity and Target Concept for all scenarios. In this figure the colour assigned to each pair of <effect-feature> shows the magnitude of the correlation coefficient between them. The corresponding standard deviations are provided in APPENDIX C.

In Figure 14, the sub figures a, c; e and g show the correlation coefficient of the effects with the features in the AoI-FV for the First 100, Best Choice and All-in-Page scenarios respectively. They show that compared to the other effects the saliency effect has the lowest correlation coefficient with these features where it is generally below 0.05 for all of the features except Visit Number (iVn) and Skip Number (iSn) that it reaches around 0.14 and 0.25 respectively. In the same graphs we can see Background Similarity and Main-concept Similarity have very similar behaviour where almost for all of the features they have very close correlation coefficient values with the same trend. For these two effects the correlation coefficients with the AoI-FV hardly exceed o.3 however for most of the features they are not below 0.15.

In these charts it is easily noticeable that compared to the Saliency effect, Main-concept Similarity and Background Similarity, the Target Concept effect has significantly higher correlation coefficient values. This result is in accordance with results we had from the ANOVA analysis where we had very high F statistic values. For most of the features the correlation coefficient value with the Target Concept effect is between 0.45 and 0.6. These results show that there is enough information in the AoI-FV gaze features for identifying the Target Concept. We can see that except for Saliency effect, the Skip Number (iSn) has the minimum correlation coefficient with all other effects in all of the scenarios. As we also observed this in the ANOVA analysis this was the feature that either had very low F statistic or was not impacted by all of the effects except Saliency effect. The highest correlation coefficients with Target Concept effect belong to the Proportion of total visit (iAdp) and Proportion of maximum visit duration (iMxp) features. Also we can see in all of the scenarios the highest correlation coefficient values belong to the features from the total visit family (iAd, iAdp, iAdr) and maximum family (iMx, iMxp, iMxr). It can be noticed the Background Similarity generally has slightly higher correlation coefficient values with the features in the AoI-FV compared to Main-concept Similarity. This suggests that the background of the images have a rather greater impact on the gaze behaviour compared to the main concept of the images.

In Figure 14, the sub figures b, d, f and h show the correlation coefficient of the effects with the features in the SP-FV for the First 100, Best Choice and All-in-Page scenarios respectively. We can see that generally for all of the effects the magnitude of the correlation coefficient with the features in the SP-FV is less than AoI-FV and in most cases they are not considerable. Similar to the AoI-FV, in the SP-FV the highest correlation coefficient values of features are with the Target Concept effect. The Visit time (tTime), local maximum visit (lmx) and Difference of visit time with the previous visit (tPreGrad) have the maximum correlation with the Target Concept effect. This shows that these features have stored more information about the TC compared to the rest of the features in the SP-FV.

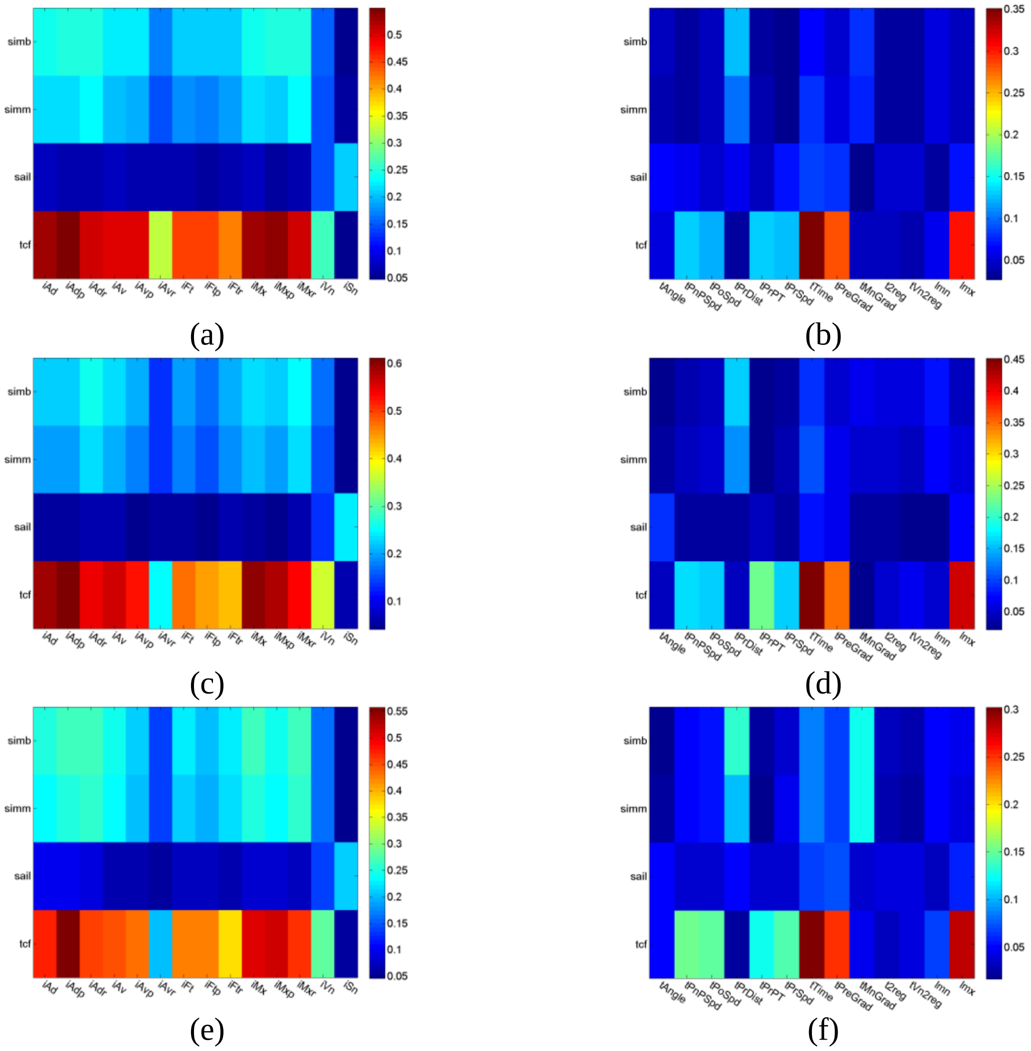


Figure 14: Average Correlation Coefficient between the features and Background Similarity (simb), Main-concept Similarity (simm), Saliency Score (sail) and Target Concept (tcf) vectors for all Users. The colours show the magnitude of correlation coefficient. a) First 100 scenario, AoI-FV b) First 100 scenario, SP-FV c) Best Choice scenario, AoI-FV d) Best Choice scenario, SP-FV e) All-in-Page scenario, AoI-FV f) All-in-Page scenario, SP-FV

5.5 Dependency of the gaze features

Most of the extracted features in this thesis are not mutually independent from each other. Many of them share part of their information with other features. The relationships between the used features are by some means complicated. To simplify the explanation of dependency of the features on each other, the following relationships are defined:

- Parent (P): Feature A is the parent of feature B if in anyway the values in A are used to calculate the values in B. For example in AoI-FV the values of iFt and iMx are used to calculate the values in iAd along the other visit durations for an image in a page hence iFt and iMx are Parents of iAd.
- Child (C): Feature A is the child of feature B if the values in A are somehow calculated by the values in B. For example iAv is a child of iAd and iVn because its values are formed of dividing the values in iAd by the values in iVn.
- Sibling (S): Two features are siblings if they share the same parent for example iMxp and iMxr are siblings because they are both derived from iMx.
- Grand Parent (GP): Feature A is grandparent of feature B if the parent of B is the child of A.
- Great Grand Parent (GGP): Feature A is great grandparent of B if the grand parent of B is the child of A.
- Grand Child (GC): Feature A is grandchild of feature B if the child of B is the parent of A.
- Great Grand Child (GGC): Feature A is great grandchild of B if the grand child of B is the parent of A.
- Related(R): Two features are related if they have a common feature that they are related to by any of the above relationships.

Table 12 and Table 13 show how the features in both AoI-FV and SP-FV are related to each other with the above definitions.

For the AoI-FV:

- iSn is the only feature that is not related to any other feature
- iVn, iFt and iMx are the only features that are not children of other features. This means that the values in all other features are somehow calculated from the values in these three features.
- iAvp and iAvr are the most dependant features where each has two great grandparents, two grandparents and one parent. They are siblings of each other and they are related to all other features.

For the SP-FV:

- tAngle, t2reg, tVn2reg are the feature without any relationship to other features.
- There are no grandparent or grandchild relationships.
- Amongst the features with relationship tPrDist and tTime are the features without any parent which means all other features are somehow derived from them.
- The dominant relationship between the features is Sibling that means most of the features share the same parent with each other.

When looking at the correlation coefficient graphs we can see that the independent features (i.e. features without any relationship or features without any parent) are not necessary the most correlated ones to the Target Concept. For example iAvp bears more correlation with TC than its great grandparent iFt. This shows that there is the chance that the features with Child relationship might bear more information than their parents. This will be more investigated in the Results chapter where the redundant features or the features without sufficient information for classification of TC are removed from the feature vectors.

Table 12: Dependency of the features of the AoI-FV on each other. The features in rows have the following relationship to the features in columns: P = Parent, GP=Grand Parent, GGP = Great Grand Parent, S=Sibling, R=Related, C=Child, GC=Grand Child, GGC = Great Grand Child

	iAd	iAdp	iAdr	iAv	iAvp	iAvr	iFt	iFtp	iFtr	iMx	iMxp	iMxr	iVn	iSn
iAd	X	P	P	P	GP	GP	C	S	S	C	S	S	-	-
iAdp	C	X	S	S	R	R	GC	R	R	GC	R	R	-	-
iAdr	C	S	X	S	R	R	GC	R	R	GC	R	R	-	-
iAv	C	S	S	X	P	P	GC	R	R	GC	R	R	C	-
iAvp	GC	R	R	C	X	S	GGC	R	R	GGC	R	R	GC	-
iAvr	GC	R	R	C	S	X	GGC	R	R	GGC	R	R	GC	-
iFt	P	GP	GP	GP	GGP	GGP	X	P	P	-	-	-	-	-
iFtp	S	R	R	R	R	R	C	X	S	-	-	-	-	-
iFtr	S	R	R	R	R	R	C	S	X	-	-	-	-	-
iMx	P	GP	GP	GP	GGP	GGP	-	-	-	X	P	P	-	-
iMxp	S	R	R	R	R	R	-	-	-	C	X	S	-	-
iMxr	S	R	R	R	R	R	-	-	-	C	S	X	-	-
iVn	-	-	-	P	GP	GP	-	-	-	-	-	-	X	-
iSn	-	-	-	-	-	-	-	-	-	-	-	-	-	X

Table 13: Dependency of the features of the SP-FV on each other. The features in rows have the following relationship to the features in columns: P = Parent, GP=Grand Parent, GGP = Great Grand Parent, S=Sibling, R=Related, C=Child, GC=Grand Child, GGC = Great Grand Child

	tAngle	tPnPSpd	tPoSpd	tPrDist	tPrPTime	tPrSpd	tTime	tPreGrad	tMnGrad	t2reg	tVn2reg	lmn	lmx
tAngle	X	-	-	-	-	-	-	-	-	-	-	-	-
tPnPSpd	-	X	S	C	S	S	C	S	S	-	-	-	-
tPoSpd	-	S	X	C	S	S	C	S	S	-	-	-	-
tPrDist	-	P	P	X	-	P	-	-	-	-	-	-	-
tPrPTime	-	S	S	-	X	S	C	S	S	-	-	-	-
tPrSpd	-	S	S	C	S	X	C	S	S	-	-	-	-
tTime	-	P	P	-	P	P	X	P	P	-	-	P	P
tPreGrad	-	S	S	-	S	S	C	X	S	-	-	-	-
tMnGrad	-	S	S	-	S	S	C	S	X	-	-	-	-
t2reg	-	-	-	-	-	-	-	-	-	X	-	-	-
tVn2reg	-	-	-	-	-	-	-	-	-	-	X	-	-
lmn	-	S	S	-	S	S	C	S	S	-	-	X	S
lmx	-	S	S	-	S	S	C	S	S	-	-	S	X

5.6 Choosing the best scenario

As presented earlier, three scenarios were developed to record the gaze data. Not all of these scenarios are suitable for classification of TC+ and TC- images. It is crucial to choose the best scenario for the task of classification of Target Concept. Because all parts of the framework are developed around the chosen scenario and the calculations, feature selections and conclusions are made based on the recorded data for the chosen scenario. When looking at the correlation coefficient graphs in Figure 14 we can see that the Best Choice bears the maximum correlation coefficient with Target Concept class variable vector with at least 5% difference from the other two scenarios for both AoI-FV and SP-FV. This shows that the gaze features in this scenario are highly correlated with the attention that the users paid to the TC+ images. Also when we look at the F statistic values in Table 11 that shows the effect of the Target Concept factor on the gaze feature by calculating the variance between groups of the variance within groups, we can see that the F statistic values for the Best Choice scenario are much higher compared to the other two scenarios. So based on computational investigations we can see that the Best Choice is the best scenario for classification of the Target Concept.

If we review the implementation of the three scenarios we can see that for the First100 scenario we have the situations that the user accidentally visited a TC+ image and that image was counted towards the 100 visited TC+ images without any intention from the user. Also for the All-in-Page scenario there are many situations that a gaze intersection with a TC+ might happen without any intentional look from the user which results in activation of the Next button. These intersections with TC+ images without any intention from the user resulted in the mentioned differences in the F statistic and correlation coefficient values of the TC effect for the Best Choice scenario and the other two scenarios.

6 Results

As discussed in Chapter 4 a framework is developed that extracts 27 features from gaze movements for the purpose of classification of visited images into two classes of TC+ and TC-. TC+ is the class of the images that contain a specific concept, called Target Concept (TC), that the user is searching for and TC- is the class of the images that do not contain the TC. In the previous chapter the stored data in the 27 extracted features were studied carefully and the effect of three factors on the values of these features was investigated. These factors include Saliency, Similarity, and TC class factors. As mentioned in section 5.6 because of providing more useful data for classification of TC+ and TC- images, from the three available scenarios, the Best Choice scenario is chosen for more investigation.

In this chapter it will be explained how first 5 different classifiers are used for classification of the images as explained earlier based on the gaze features. Out of these five classifiers, 3 of them are chosen for further investigation. Then the classifiers' performance will be tested by different features and the redundant features and the features that have negative effect on the classification results will be removed from the feature vectors. Next the performance of the classifiers will be investigated when they are trained to classify the records of one user when they are trained by the records from other users. Finally the proposed method will be tested by another dataset.

To measure the performance of the classifiers, their classification results are compared against the class variable of the test images. The test images are the images that the corresponding gaze features are used in order to classify them by the classifiers after training the classifiers. To calculate the performance of a classifier, the properties of the classification quality namely True Positive, False Positive, True Negative and False Negative are defined as follows:

1. True Positive (tp): The images that belonged to the TC+ class and they were classified correctly as TC+
2. False Positive (fp): The images that belonged to the TC+ class but they were classified mistakenly as TC-
3. True Negative (tn): The images that belonged to the TC- class and they were classified correctly as TC-
4. False Negative (fn): The images that belonged to the TC- class but they were classified mistakenly as TC+

By using these properties we can calculate the classification performance measures [**132**] [**133**] namely Recall, Precision and F1-measure as follows:

Recall: Shows a classifier's performance quantitatively. This measure shows the ratio of the correctly classified TC+ images to all of the images belonging to the TC+ class. It is calculated as follows:

$$recall = \frac{t_p}{t_p + f_n}$$

Precision: Shows the classifiers performance qualitatively. This measure shows the ratio of the number of the images that are correctly classified as TC+ to the number of all of the images that were classified as TC+. It is calculated as follows:

$$precision = \frac{t_p}{t_p + f_p}$$

F1 measure: Shows the overall performance of a classifier which includes both quantitative and qualitative performances. The values of F1-measure over 0.5 or 50% show that there is more than 0.5 probability that the current performance of the classifier is not random. The values below 0.5 show that the classification results are accidental. The F1-measure is calculated as follows:

$$F1 - measure = 2 \times \frac{recall \times precision}{recall + precision}$$

6.1 Performance of five classifiers

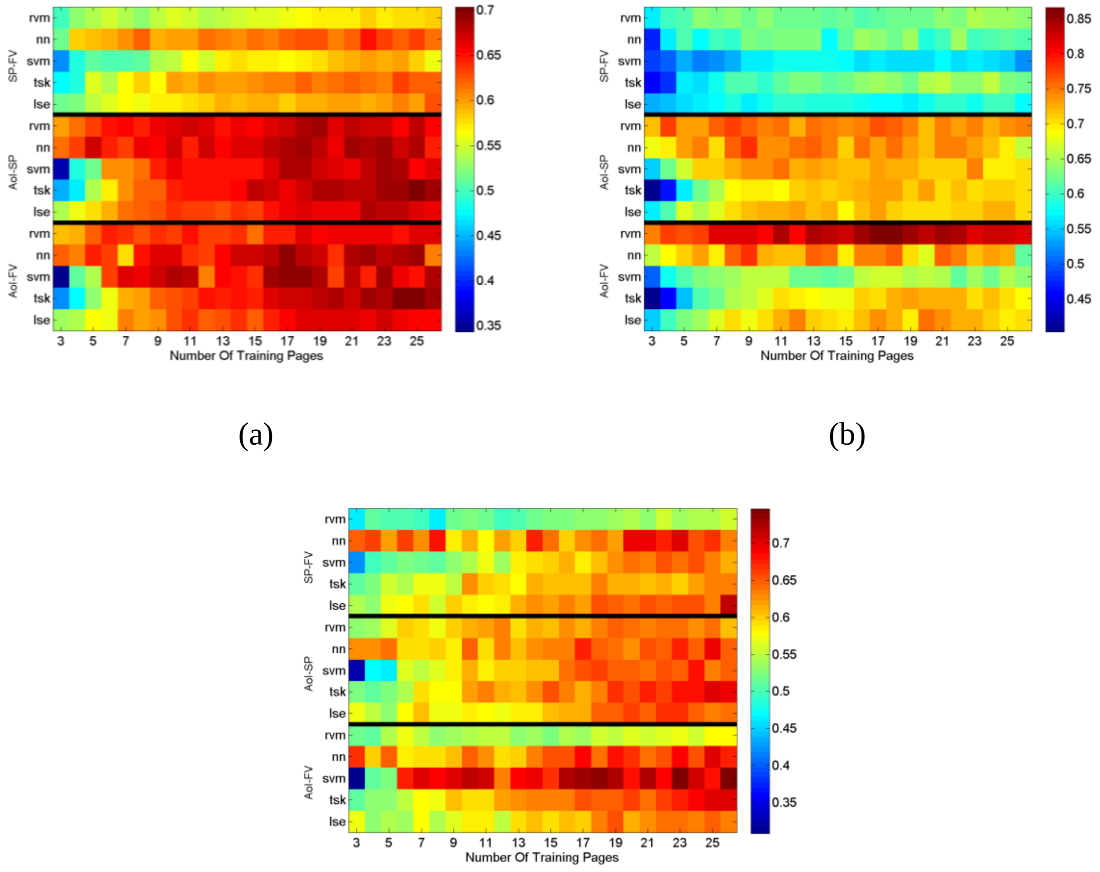
To classify the images into TC+ and TC- five different classifiers are used. These classifiers include

- TSK-FIS: A fuzzy inference classifier based on Takagi Sugeno Kang – Fuzzy Inference System
- NN: Feed Forward Back Propagation Neural Network classifier
- SVM: Support Vector Machine classifier with the RPF Kernel [**93**].
- RVM: Relevance Vector Machine classifier [**134**] with (Direct Kernel & RPF Kernel) [**135**].
- LSE: Least Square Estimation classifier

To produce the results in this section, each classifier was trained multiple times by increasing the number of training pages from 3 to 28 for the entire acquired data from all users separately. This was done by partitioning the recorded data from every user into two training and test sets based on the number of training pages. This form of training, removed the user dependency of the classification results. For example if the number of training pages is set to 28, the first 28 pages of the user’s records are used for training the classifier and the rest are used for evaluating the trained classifier. For every number of the training pages for every user the performance of the classifier is measured by the recall, precision and F1 performance measures. This is performed by calculating the average of the performance of the classifier for all users for a specific number of training pages. For example the average F1 performance of SVM classifier for 28 training pages is the average of all of the F1 measures found for all of the users with 28 pages of training. The entire mentioned computations are performed separately for the AoI-FV and the SP-FV. Finally three sets of results are produced:

- Performance based on the AoI-FV
- Performance based on the SP-FV
- Performance based on the average output of the classifier for both AoI-FV and SP-FV which will be denoted by AoI-SP.

Figure 15 illustrates the average performance of all of the mentioned classifiers for all users based on different number of training pages. The first point that we can see is that for all of the classifiers the value of F1 measure (Figure 15-a) exceed 0.5 for 10+ number of training pages. Knowing that the F1 values less than 0.5 show the random classification behaviour of a classifier, the graph shows that for all of the classifiers the F1 measure is significant enough for accepting the classification result.



(a)

(b)

(c)

Figure 15: Average performance of 5 classifiers for all users based on different number of training pages a) F1 Measure b) Precision c) Recall

The results show that by increasing the number of training pages from 3 to 17, for all of the classifiers the F1 measure performance keeps improving from under 0.5 to over 0.65. Then after 17 pages it stays relatively the same with slight fluctuations around 0.67. The reason for this fluctuation is that by increasing the number of training pages we are constantly removing the number of available TC+ images from the testing set which makes misclassifications more apparent. For example when there are 100 TC+ images in the testing set, 1 misclassification results in 1% change in recall however with only 10 TC+ images 1 misclassification ends up with 10% reduction in the recall value.

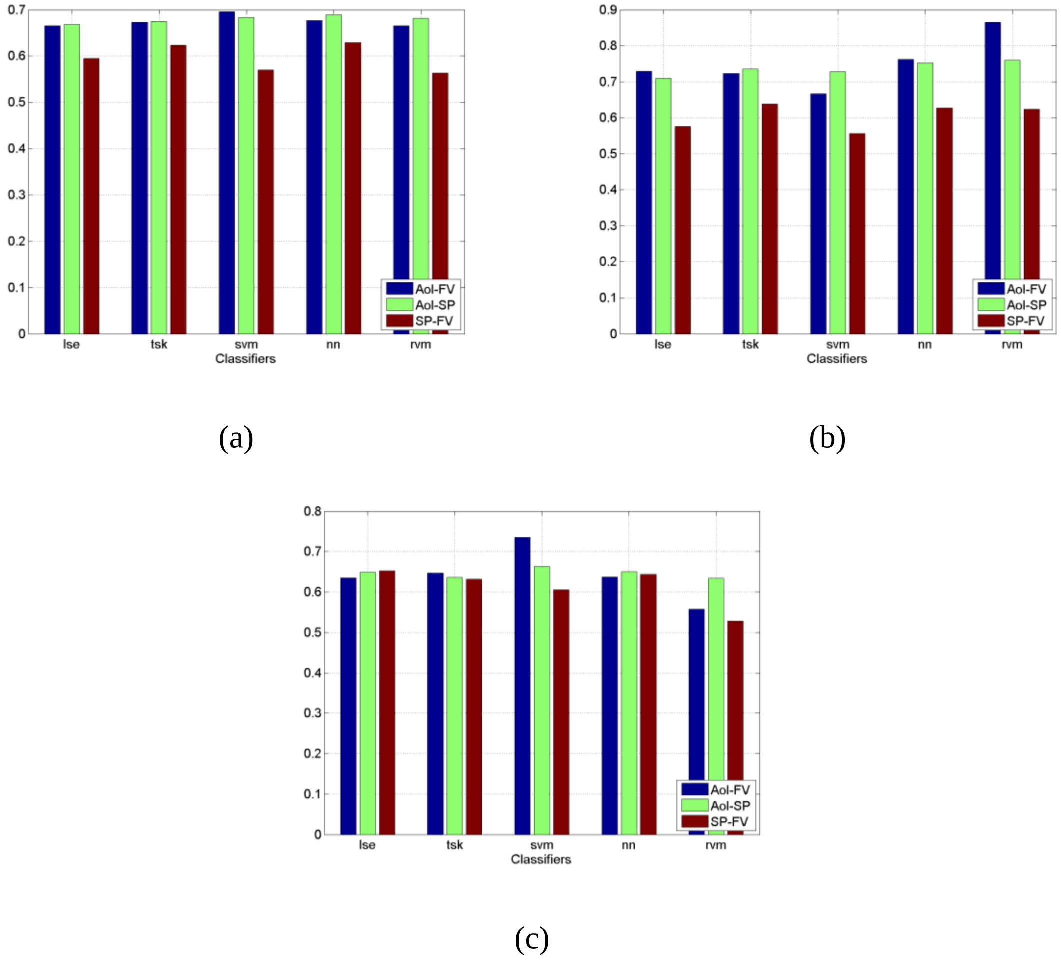
It is also clear from these graphs that the performance of the classifiers for the features in the AoI-FV is distinctively better than the performance of the classifiers for the features in the SP-FV by at least 5%. The main reason is that the values in the AoI-FV are the aftermath of all of the visits that a user paid to an image while the values in SP-FV are the instances of the users' interaction with the image.

We can see that in general the F1 performance for the average of AoI-FV and SP-FV (middle section in each graph) is equal or better than the performance for each of the feature vectors separately. This is because by calculating the average, they balance the output of each other. For example if an image is marginally considered as TC- for one feature vector and it is determined strongly as TC+ by another feature vector the final assigned class will be TC+. Now if the class variable is TC+ then this will improve the classification result otherwise this will worsen the classification performance. The depicted F1 results show that for the LSE, NN and RVM classifiers overall the number of times that an incorrect mild misclassification by one feature vector is corrected by a strong correct classification by another feature vector is equal or more than the inverse situation as a result we gave better values for AoI-SP compared to each of the feature vectors seperately. This is not the case for the 'TSK' and 'SVM' classifiers. However with a closer look we can see although integrating the classification results of the two vectors has little merit for the levels of performance, the resulting fluctuations in the F1 is reduced and there is a smoother change in these values.

In these graphs it is deceptive that the best precision belongs to the RVM starting from 0.72 for 3 training pages and constantly improving to over 0.85 for 26 training pages for the AoI-FV, 0.6 to 0.65 for the same training page numbers for the SP-FV and over 0.72 for the combination of the two vectors. However the RVM pays the price by recall performance with lowest values amongst all of the classifiers for both feature vectors. The lowest Precision belongs to the SVM which never exceeds 0.67 for both feature vectors however the combination of the two feature vectors results in precision values up to 0.7, but this combination ends up with lower recall values for SVM. We can see that the SVM has over 0.74 recall values for AoI-FV for 15+ number of training pages however after combining the vectors this reduces the overall recall performance for the SVM to 0.65. Another interesting point in the F1 performance is RVM and NN both start with a high value of F1, which means they can be well trained for fewer number of training pages however the rest of the classifiers need at least 7 training pages to be able to perform like RVM and NN. By continuously increasing the number of training pages we can see both SVM and TSK keep improving their performance until they exceed the performance of RVM and reach the NN performance.

Although RVM shows promising performance regarding precision and an acceptable F1 measure value that shows it is not behaving randomly, however the low level of recall values which affects its overall performance and the considerably slower training process compared to the SVM [**136**] [**137**] make SVM the preferred classifier for this task. We can also see that the LSE which is the least expensive classifier with regards to performance cost shows acceptable output but its F1 never exceeds 66% where with a slight compromise in performance cost we can reach up to 70% for the F1 measure by using other classifiers. As a result for the rest of the processing in this chapter we used TSK-FIS, NN and SVM.

Furthermore for the rest of this chapter, the number of training pages for the rest of processing is set to 18 pages. This is because the goal is to set the number of training pages as low as possible in order to have more test samples to evaluate the classifiers while the classifiers are trained well enough that adding more training pages does not improve their performance significantly. Figure 16 shows a closer view to the performance of the five classifiers for 18 training pages.



(a)

(b)

(c)

Figure 16: Performance of the five classifiers for 18 training pages with all features a) Average F1 Measure b) Average Precision c) Average Recall

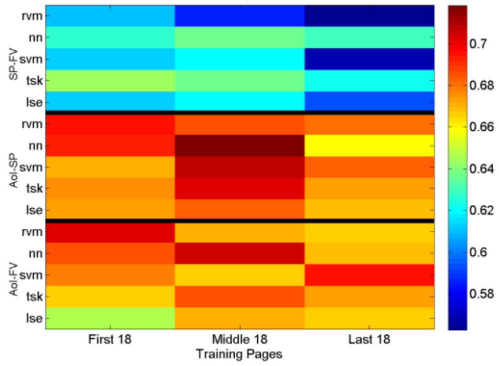


Figure 17: F1 Measure of the 5 classifiers after partitioning the data set in 3 formats with 18 training pages and 12 test pages

To investigate whether the classifiers produce acceptable results with different sets of training data after choosing 18 pages as the optimum number of pages for the rest of the processing, the data set was partitioned into training and test sets in three combinations and for every combination the classifiers were trained with the training set and the F1 performance was calculated. These three combinations include:

1- Choosing the first 18 pages of the experiment as the training pages and the remaining pages as the test pages
2- Choosing the middle 18 pages of the experiment as the training pages and the remaining pages as the test pages
3- Choosing the final 18 pages of the experiment as the training pages and the rest of the pages as the test pages

Figure 17 shows the F1 performance of all of the classifiers with these three combinations. As we can see for all of the classifiers for both feature vectors and their combination the F1 measure is more than 0.5 which shows that neither of the classifiers behaves randomly for any choice of partition. Also we can see that there is no dramatic fluctuation in the performance of any of the classifiers for any combination of partitioning. The biggest change is for the Neural Networks that by choosing the middle 18 pages as training pages we have F1 value with 0.71 and this value drops to 0.66 by choosing the final 18 pages of the experiment as the training pages for the combination of AoI-FV and SP-FV. It should also be mentioned that except the mentioned case for the Neural Networks we can see that the performance of the classifiers for the combination of the AoI-FV and SP-FV is less sensitive to the change in the training set as the change in F1 values is smaller. We can see that except for the Neural Networks the values in the F1 measure change between 0.67 and 0.67 for the AoI-SP combination compared to 0.57 to 0.65 for the SP-FV and 0.64 to 0.71 for the AoI-FV.

6.2 Feature selection

In previous section the gaze data were evaluated by 5 classifiers with all of the extracted features. However, there is possibility that some of the features are redundant or as much as their variance is affected by factors like similarity and saliency, it is not affected by the

Target Concept factor. This section mainly investigates these possibilities and tries to extract a subset of features that are most informative for the classification of TC+ images in order to improve the performance.

In this section two feature selection methods are studied:

i. Feature selection by Maximum dependency to Target Concept and minimum dependency to Saliency and similarity.
ii. Feature selection by backward elimination.

Both of these methods will be followed by further dimension reduction where we use the Principal Component Analysis (PCA) [**138**] to extract feature vectors with 98% stored variance of information from the original features.

6.2.1 Feature selection by Maximum dependency on Target Concept and minimum dependency on saliency and similarity

In the previous chapter the results of ANOVA showed that only few features in both SP-FV and AoI-FV feature vectors are not affected by all factors of saliency, similarity and Target Concept. When looking at the correlation coefficient graphs in the same chapter we can see that the results are in line with the ANOVA output. Consequently in this section it is investigated whether the features that are strongly affected by the saliency and similarity factors, compared to the Target Concept factor, affect the performance of the classifiers negatively. Consequently a Selection Score is assigned to every feature with the following equation:

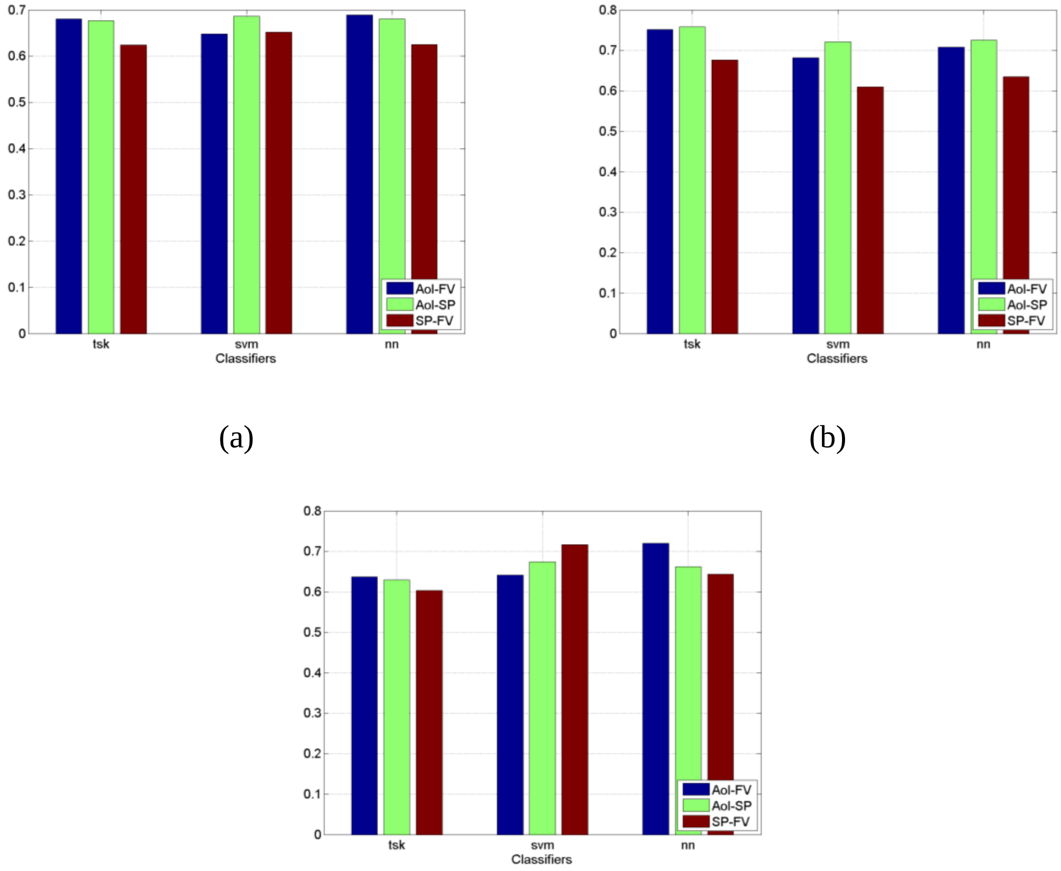
$$Selection\ Score = \frac{TC\ Score}{Simm\ Score + Simb\ Score + Sail\ Score}$$

where TC, Simm, Simb and Sail scores are correlation coefficients of the features with these vectors as stated in the previous chapter. This score shows the ratio of the correlation of a feature with TC over the rest of distracting factors considered in this thesis. If a feature is assigned a selection score below 1 it shows that the feature is more affected by the distracting factors of saliency and similarity rather than the Target Concept. We can see that in both feature vectors 6 features (highlighted by dark shading) have gained a score below 1. These features are removed from both AoI-FV and SP-FV vectors and the classifiers are trained again with the new subset of features. Table 14 shows the Selection Score assigned to each feature:

Table 14: Calculated selection score based on maximum dependency on Target Concept and minimum dependency on saliency and similarity

Area of Interest Feature Vector		Scan Path Feature Vector	
Features	**Score**	**Features**	**Scores**
iAd	1.2189	**tAngle**	0.2963
iAdp	1.2725	**tPnPSpd**	1.4141
iAdr	0.9757	**tPoSpd**	1.2592
iAv	1.1367	**tPrDist**	0.1316
iAvp	1.2096	**tPrPTime**	2.3868
iAvr	0.7760	**tPrSpd**	1.5176
iFt	1.1209	**tTime**	1.6229
iFtp	1.2145	**tPre_grad**	1.8404
iFtr	0.9318	**tMn_grad**	0.1850
iMx	1.1833	**t2reg**	0.3836
iMxp	1.2475	**tVn2reg**	0.4841
iMxr	0.9836	**lmn**	0.3028
iVisitNum	0.8096	**lmx**	2.4066
iSn	0.2064	-	-

Figure 18 shows the performance of the TSK, SVM and NN classifiers for 18 training pages after removing the features that gained a selection score below 1 and are regarded as the features that are affected by the distracting effects more than the TC effect. When comparing the results to the classifiers' performance in Figure 16 where no feature reduction took place, we can see that the F1 performance of the TSK is the least affected compared to the other two classifiers remaining at a value between 0.65 and 0.7. The F1 performance reduces for each feature vector for both SVM and NN however it stays the same for AoI-SP for all of the classifiers around 0.68. The reason is for the TSK and NN there is no significant difference in the F1 performance before and after feature reduction and for the SVM although the performance for the AoI-FV is reduced it is compensated by the increase in the SP-FV performance. It can be concluded that by removing the features affected by the distracting effects we are not making a significant difference to the overall outcome while reducing the dimension of both feature vectors to the half. As can be seen in the results the most significant change in one of the performance measures for all of the classifiers is the change in the Recall for the SVM. We can see that there is more than 10% drop in AoI-FV and almost a similar increase in for the SP-FV that eventually they cancel out each other.

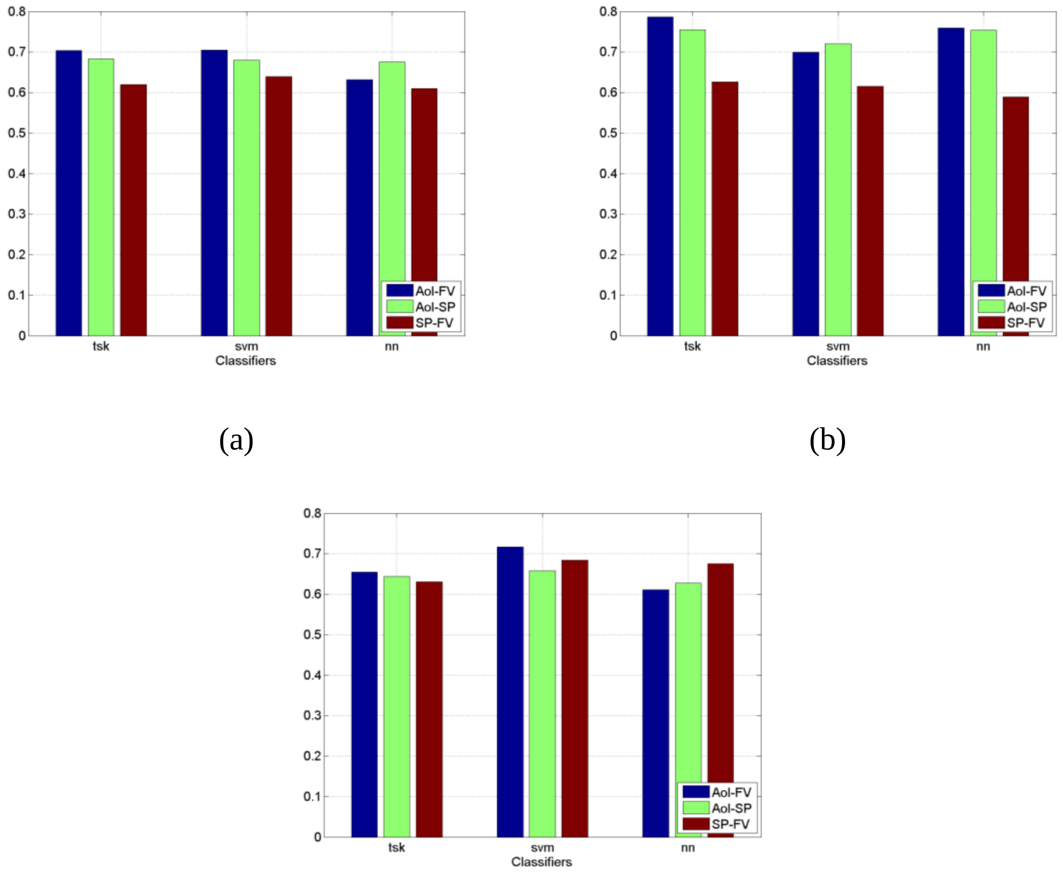


(a) (b)

(c)

Figure 18: Performance of the TSK, SVM and NN classifiers after choosing a subset of features based on Selection Score gained from correlation coefficient a) Average F1 Measure b) Average Precision c) Average Recall

After reducing the dimension of the feature vectors by removing the affected feature by distractors, the dimension in both feature vectors was reduced further by projecting the features to a lower dimension using PCA [**138**] and keeping the resulting vectors with accumulated 98% of variance for classification. The results of classification after this further reduction are shown in Figure 19. It is observable that there is a slight decrease in the F1 performance for all of the classifiers for both feature vectors separately and for the outcome of their combination. When looking at the precision and recall graphs it can be noticed that this decrease in F1 is more resulting from the reduction in the recall values and part of it is compensated by improvement in the precision values.



(a)

(b)

(c)

Figure 19: Performance of the TSK, SVM and NN classifiers after choosing a subset of features based on Selection Score gained from correlation coefficient and reducing the dimension of the feature vectors further by PCA a) Average F1 Measure b) Average Precision c) Average Recall

6.2.2 Feature selection by backward elimination method

To make sure that we are keeping the most informative features for the purpose of classification and removing the redundant features, another feature selection method was applied with the difference that except to focus on the distractor factors it was focused on the information of the features. In this method which is called backward elimination first the full set of features are selected and the classification performance is calculated, then with leave-one-out method one feature is set aside every time, the classifier is trained by the rest of the features and the performance of the classifier is calculated. Finally the performance values resulting from removal of every feature are compared to the performance of other classifiers and the feature that its removal results in the best performance is removed permanently from the set. This process continues until the further removal of the features results in a poorer classification performance.

Table 15 and Table 16 show the resulting maximum F1 measure gained after removing each corresponding feature for the three classifiers trained by AoI-FV and SP-FV respectively. For each classifier the dark shaded features are the ones that are removed from the feature vectors eventually. The following conclusions can be made from these tables:

1- For all three classifiers by reducing the number of features the performance of the classifier is improved where for the AoI-FV it has improved with 1.5+% for TSK, 2+% for SVM and 0.9+% for NN. Also for the SP-FV this has improved 1.2+% for TSK, 9%+ for SVM and 2.7+% for the NN.

2- For all three classifiers the improvement is more for the SP-FV than AoI-FV. This shows that there are more redundant and noisy features in the SP-FV compared to AoI-FV. AS stated earlier this stems from the fact that the features in SP-FV are calculated values from the instances of the gaze movement where the features in the AoI-FV are final outcome of a users' interaction with an image.

3- For all of the classifiers more than half of the features are removed in both feature vectors together.

4- tMnGrad, tPreGrad, tTime and iAdp features are not removed for the three classifiers. This shows these are the most informative features in both vectors independent from the classifier's type. When looking at the Table 12 and Table 13 it is noticeable that except tTime the rest of the features are children or grandchildren features derived from other features. This shows that calculating these features from other features not only did not result in redundancy of the features but also add more information to the feature vectors.

Table 15: Resulting F1 Measure for the TSK, SVM and NN classifiers after removing the features one by one for the AoI-FV

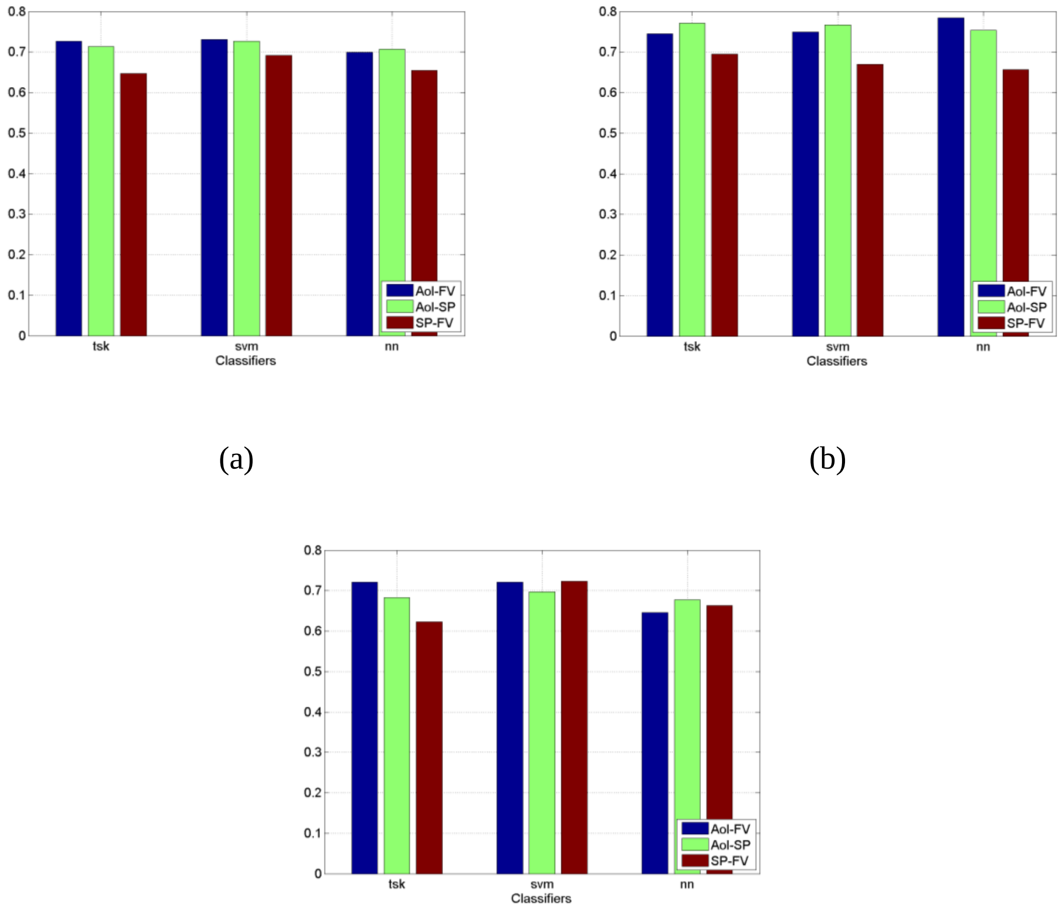
TSK		SVM		NN	
Removed Feature	Resulting F1	Removed Feature	Resulting F1	Removed Feature	Resulting F1
iAvr	0.6914	**iSn**	0.7096	**iAvr**	0.7060
iFtr	0.6971	**iMxr**	0.7092	**iFtp**	0.6984
iAd	0.7033	**iFtr**	0.7116	**iFtr**	0.7050
iMxp	0.7078	**iMx**	0.7138	**iSn**	0.6915
iFt	0.7163	**iFt**	0.6595	**iFt**	0.7053
iSn	0.7149	**iAv**	0.7148	**iAv**	0.7088
iMx	0.7269	**iFtp**	0.7215	**iAdr**	0.7065
iVisitNum	0.7222	**iAvp**	0.7229	**iMx**	0.7141
iFtp	0.7263	**iAvr**	0.7233	**iVisitNum**	0.6881
iMxr	0.7161	**iMxp**	0.7251	**iMxr**	0.7151
iAv	0.7229	**iAdr**	0.7307	**iAd**	0.7049
iAdr	0.7204	**iAd**	0.7282	**iMxp**	0.7004
iAvp	0.7074	**iVisitNum**	0.7258	**iAvp**	0.7150

Table 16: Resulting F1 Measure for the TSK, SVM and NN classifiers after removing the features one by one for the SP-FV

TSK		SVM		NN	
Removed Feature	Resulting F1	Removed Feature	Resulting F1	Removed Feature	Resulting F1
tAngle	0.6334	**tAngle**	0.6094	**tAngle**	0.6279
tPrPTime	0.6426	**tPrDist**	0.6310	**tPrDist**	0.6505
tNew_lmx	0.6445	**tPrPTime**	0.6514	**tPnPSpd**	0.6415
tNew_lmn	0.6451	**tNew_lmn**	0.6617	**tPrPTime**	0.6416
tPoSpd	0.6424	**tVn2reg**	0.6725	**tNew_lmx**	0.6492
tPrSpd	0.6466	**tPnPSpd**	0.6746	**tPrSpd**	0.6434
tPrDist	0.6463	**tPoSpd**	0.6773	**t2reg**	0.6543
tPnPSpd	0.6403	**tPrSpd**	0.6840	**tPoSpd**	0.6553
t2reg	0.6298	**tNew_lmx**	0.6917	**tVn2reg**	0.6559
tVn2reg	0.6283	**tMn_grad**	0.6891	**tMn_grad**	0.6468
tPre_grad	0.6332	**t2reg**	0.6869	**tNew_lmn**	0.6597
tMn_grad	0.6366	**tPre_grad**	0.6917	**tPre_grad**	0.6465

Figure 20 shows the performance of the three classifiers trained and tested by the new feature sets after removal of the shaded features in the above tables. When compared to the pre removal results and the results of the previous section (features subset selection based on selection score), the performance of all three classifiers has improved where for both previous feature sets the F1 measures hardly could reach 0.7 where after backward elimination they easily passed 0.7 for both AoI-FV and AoI-SP. Also It is observable that the classification results by SP-FV have improved. This improvement for TSK is mostly due to the improvement in the recall value while the precision stays the same and for both

NN and SVM it is due to the values of precision. Also when looking at the performance of the classifiers by further dimension reduction using PCA [**138**] shown in Figure 21, we can see that not only the results did not improve but also the F1 measure values decreased.



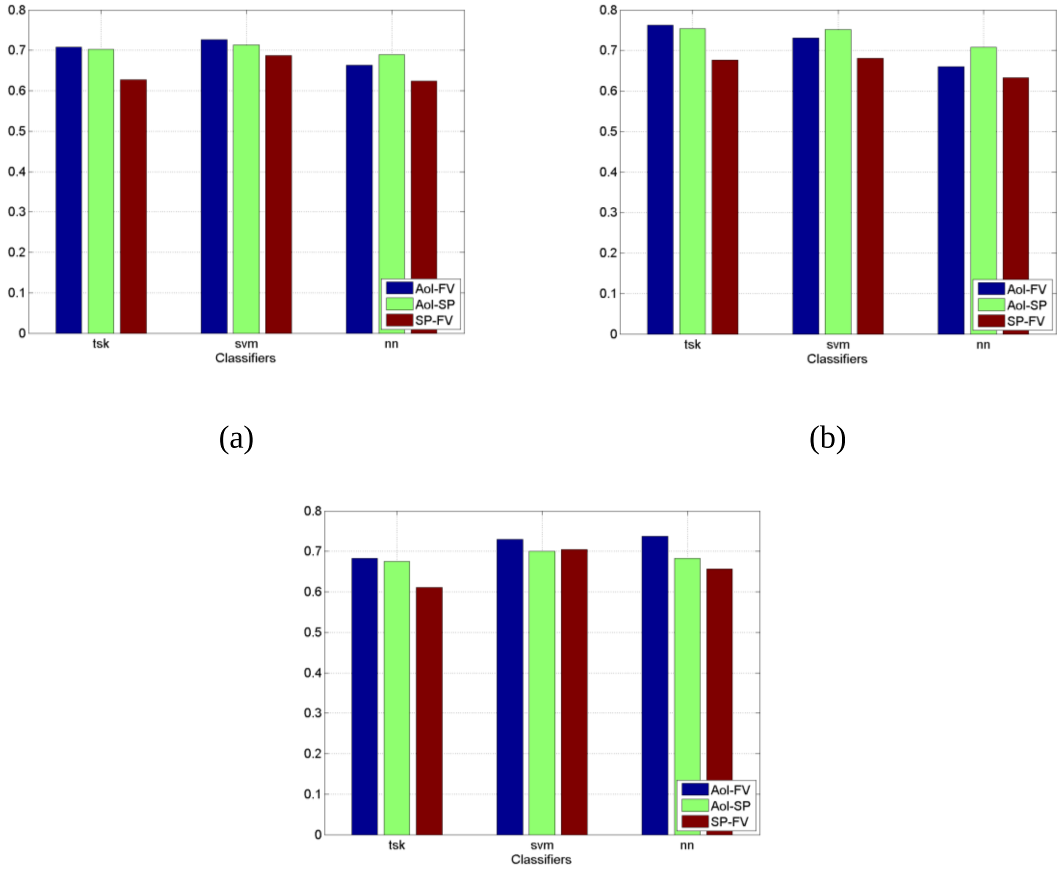
(a)

(b)

(c)

Figure 20: Performance of the TSK, SVM and NN classifiers after choosing a subset of features based on Backward Elimination method a) Average F1 Measure b) Average Precision c) Average Recall

In this section it was shown that that selecting a subset of the features based on the selection score, which removes the features that their values are affected by the distracting factors the most, helps to keep the same level of performance for the classifiers while reducing the dimension of the feature vectors hence the computation cost and time. However a more exhaustive search by testing the features one by one until there is no further improvement ends up with better results.



(a)

(b)

(c)

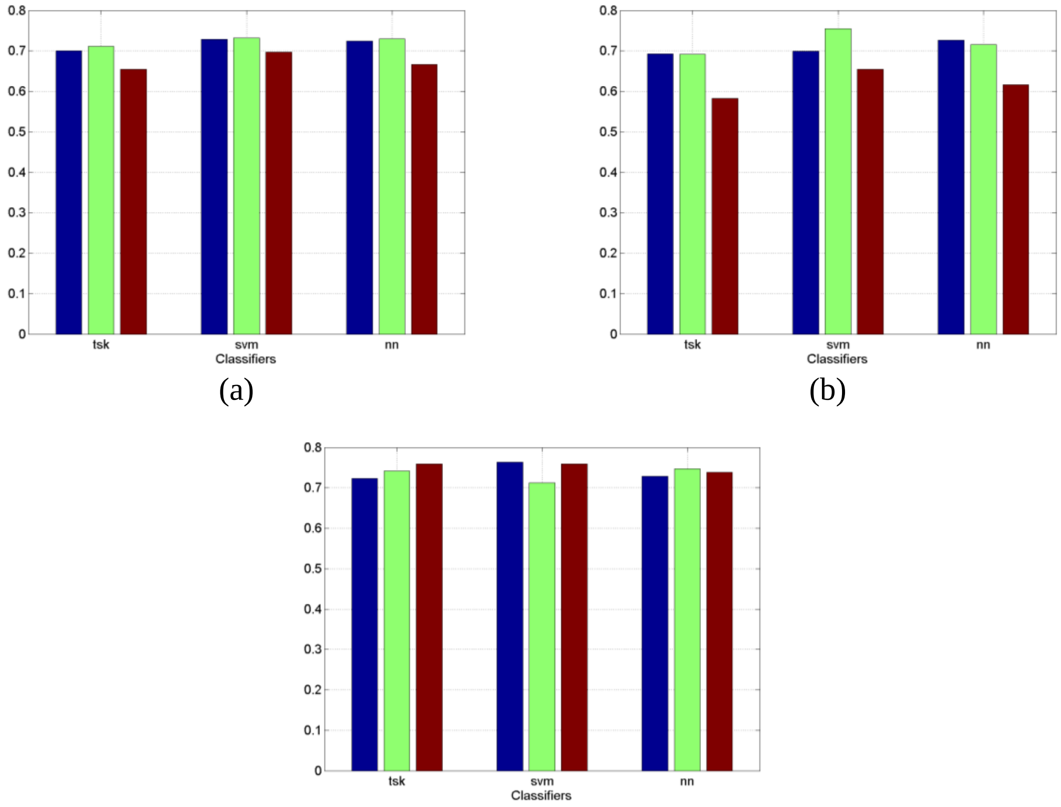
Figure 21: Performance of the TSK, SVM and NN classifiers after choosing a subset of features based on Backward Elimination Method and reducing the dimension of the feature vectors further by PCA a) Average F1 Measure b) Average Precision c) Average Recall

It was also illustrated that in both cases reducing the dimension further by using PCA had a very small negative impact on the overall performance of the classifiers. As a result it depends on the application to choose a slightly better classification performance or faster and less computation cost. Because the extra computation cost for training the classifiers with more number of features is trivial the author decided to perform the rest of the computations without reducing the dimension of the feature vectors by PCA.

6.3 Investigating the classifier performance for each user by using leave-one out method

Figure 22 shows the average performance of the TSK, SVM and NN networks using the leave-one-out method by classifying the visited images of every user when training the classifiers with the acquired data of the rest of the users. By looking at th results it can be seen that the performance of all three classifiers is in line with their performance when trained by the data of one user for themselves. The advantage of this method is that by training the classifiers with the data of other users there is no need to have a training phase for a user as a result more images can be classified and annotated. However there is a possible drawback which is depending on the chosen sets of user data for training the classifiers, they might not perform very well. This is because in the ANOVA results in APPENDIX B, the variance due to user factor is strong in the gaze features which means that the data in the features are strongly user dependant. In this thesis due to the background diversity of the participant it was not possible to draw a conclusion that whether this user dependency is due to cultural background.

The main difference between the output of this method and user adaptive method can be seen in their recall and precision performance where for the former there is almost 10% better precision performance in the classification results while for the latter there is 10% better recall performance. The reason is when the classifiers are trained by one's own gaze data they can separate the TC+ and TC- more accurately. As a result they will have a better qualitative performance. However because there is always a trade-off between precision and recall their recall performance decreases because by drawing a fine line between TC+ and TC- images they are overlooking all the TC+ images that fall in the grey area. On the other hand when the classifiers are trained by the data from for example 15 other users they cannot draw a fine line between the TC+ and TC- images but a very good approximation which results in choosing more TC+ images but it also ends up with misclassification of more TC- images which reduces the recall performance. As a result depending on the application one can choose between training the classifiers globally with better quantitative results or they can train user adaptive classifiers with better classification quality.



(a)

(b)

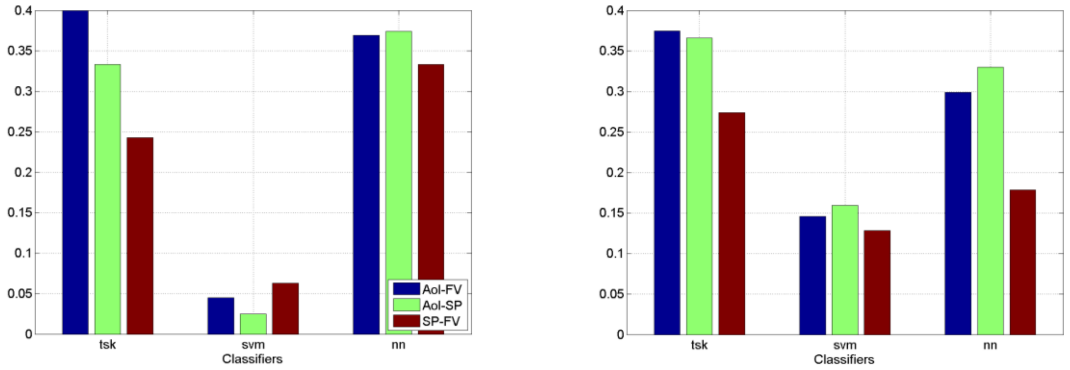
(c)

Figure 22: Performance of the TSK-FIS, NN and SVM by training the classifiers for every user using the leave one out approach a) Average F1 Measure b) Average Precision c) Average Recall

6.4 Performance for other datasets

To validate the performance of the proposed framework in this thesis, the used gaze data in [**91**] was acquired, the proposed features in this thesis were extracted and the images were classified based on the proposed ground truth data in [**91**].

In [**91**] the authors provided some random shots from different videos to the users and the users had to submit the images that they thought are the images that belong to a query topic. Overall 8 users, who were computer scientists, participated in this study where each user evaluated 4 different topics. Overall 32 user-topics where recorded throughout this experiment. The submitted images for a specific topic were considered as the ground truth in the experiment which is equivalent to TC+ images in this thesis. The arrangement of the images that appeared to the user was similar to the arrangement of the images that was used in this thesis with 24 (4 rows and 6 columns) images for each page.



(a) (b)

Figure 23: Performance of the proposed framework for Vrochidis et al. [91] dataset a) Trained from the first few pages of every user-topic b) Trained with the same method proposed in [91]

The authors in [**91**] used SVM with RPF kernel in order to classify the images in their experiment based on gaze data for a specific topic. Because there were 4 topics, it would make it 6 subsets of paired topics. The authors trained the SVM classifier 6 times each time they selected a new pair of topics as the testing data and the rest of the topics as the training data. With their method they could reach an F1- measure between 0.52 and 0.54 that shows the performance of their proposed method is acceptable because of F1 values greater than 0.5.

Figure 23 shows the F1-performance of the proposed framework in this thesis for the same dataset used in [**91**]. Every classifier was trained in two methods for the data.

1. For every user-topic the classifier was trained separately with the first 18 pages of the experiment and the rest of the data for that user-topic were considered as the testing data.
2. The same method that was used in [**91**] was used to train the classifiers in here.

As can be seen for neither of the methods the F1 performance of the classifiers exceeded 0.5 even for the SVM in Figure 23-b which is the same classifier and partitioning as mentioned in [**91**]. This shows that the output of the classifiers for this type of input is random as F1<0.5 indicates randomness in the performance of a classifier. The reason behind the randomness stems from the fact that the data that was used for this thesis was the raw data before any feature extraction and there was a different method to clean up the raw noisy data from [**91**]. The raw input data for each user-topic were stored in a *.txt file where every line in the text file indicated an entry. During the feature extraction on average for every user 44.92% of the input lines had to be removed. This includes all the data with missing tag values, all the corrupted lines with unrecognisable values and partially recorded lines. Removing this amount of data can result in missing important information in the data set and appearance of inconsistencies in the remaining data which make it unsuitable for feature extraction.

6.5 Overview of the results

The classification performance for all of the classifiers in Figure 15-a showed that given enough number of training pages, all of the classifiers are able to reliably classify the images into TC+ and TC- images with the proposed approach in this thesis. In this graph it was shown that all of the classifiers could reach F1 measures above 0.5 (meaning that the results are not accidental) for both feature vectors, SP-FV and AoI-FV, by training the classifiers with at least 10 training pages. When the number of training pages was increased to above 17 pages the F1 measure of some of the classifiers (i.e. TSK, NN and SVM) reached up to 0.7. These high values of F1 measure statistically show the high significance of the achieved recall and precision values.

In Figure 16-b it is demonstrated that when the classifiers are trained by the first 18 visited pages by the users, for all of the classifiers the classification is correct for over 70% of the images that are classified as TC+. This is accompanied by recall values (Figure 16-c) of over 60% which shows the proportion of the images that are classified as TC+ correctly over all of the images that really belong to the TC+ class. In Figure 17 it is illustrated that the results are consistent when the classifiers are trained by different sets of training pages. This is shown by choosing different sets of training pages with the same number of pages from the middle and last parts of the visited pages by the users.

In section 6.2 it was shown that although by removing the extra features that are more affected by the saliency and similarity factors we can improve the performance of the classifiers, the traditional approach of backward elimination ends with better results. Also it was shown that further dimensionality reduction by using PCA has a negative impact on the overall performance of the classifiers. Furthermore when the classifiers were trained for every user by the recorded data of other users, we could see that for all of the selected classifiers all three measures exceeded 0.7. The produced results show that with the proposed methods it is possible to keep a high performance of classification by relying on pre-recorded data which reduces the performance cost for training a new classifier for every user.

7 Conclusion

This research proposed advanced solutions for indexing the massive amount of visual information that is stored in the databases today. The goal was to discover innovative approaches to overcome the challenge of semantic gap by employing the power of truthfulness of human's judgement and the computational power of the machines.

In order to provide data for image annotation implicitly and purely based on human unconscious feedbacks, the valuable information inside human subject's eye movements were extracted when they were interacting with images on screen and the images that they visited was classified into two TC+ and TC- classes. The TC+ class is the class of the images that the user is hunting for and as a result can be annotated based on majority members of the class or the search query that a user has already given to the system.

In this thesis the research objectives were addressed as follows:

1) **To find an approach for empirically explanation of one's attention to an image on the screen in form of gaze features:**

For this purpose an interface is developed that records the coordinates of one's gaze intersection with the screen in front of them along with the time that this intersection has occurred while they are interacting with the appearing images on the screen. This interface provides the necessary tools for user interaction with the images on screen.

2) **To discover and extract visual features as a form of description of gaze movements that can be interpreted by machine for the task of classification:**

From the provided raw gaze movement data 27 gaze features are extracted. The ultimate goal was to study all of these 27 features and select the best subset. These 27 features are divided into two feature vectors namely AoI-FV and SP-FV; where the former holds the features that contain values that show visual attention to an Area of Interest (in here the area of an image on the screen) like average time and the number of times that the user looked at an area and the latter contains the features that represent the dynamics of gaze movements like speed, distance, etc. It was shown that not all of these features are independent and some of them are derived from other features or their variance partially contains the variance from other features. As a result the parent-child relationship between the features was depicted in a table. In addition the used features in other studies were reviewed and it was shown that out of the 27 features 17 feature have not been used prior to this thesis.

3) **To study the effect of different factors (i.e. saliency of an image on the screen and its similarity to other images) on one's visual attention that could affect the gaze features.**

Three factors of Saliency, Similarity and Target Concept class were defined that could affect the values of these features. The effect of these three factors on the features were studied by the ANOVA test and finding the correlation coefficient of the features with these factors. The observations from the ANOVA indicated that all of the three factors can influence shift of attention in all of the three scenarios. The magnitudes of the correlation coefficient and the F statistic values of the ANOVA proved that the Target Concept class variable factor was the strongest factor amongst all; however one's attention still can be diverted by the other two.

4) **To tailor different scenarios and choose the best scenario for the experiment that can best simulate the real life situation and provide the best gaze features for the purpose of classification**

Three scenarios were planned that each of them demanded the subjects a different quest for hunting a Target Concept in the appearing images on the screen. These scenarios included Best Choice scenario, First 100 scenario and All-in-Page scenario. It was tried to plan these scenarios so that they simulate different real life situations when someone is exploring the visual databases.

In the Best Choice scenario the users had to choose one image from similar images of a Target Concept which were mixed with other images containing different concepts. In the First 100 scenario the users had to identify the first 100 images with a predefined concept that were mixed with other images. In both of the previous two scenarios there was no pressure in finding all of the images with the Target Concept and the users were permitted to miss some Target Concept images. However in the All-in-Page scenario the users did not have this freedom and they had to locate every single Target Concept that appeared on the monitor screen.

By looking at the correlation coefficient and the two-way ANOVA test results it was concluded that the best scenario for the purpose of classification is the Best Choice scenario because of the higher dependency of the gaze features on the Target Concept factor and the design of the scenario which prevented the users to accidentally look at a TC+ image unintentionally.

5) **To classify the images based on gaze data into two different classes 1) the images that the user is searching for called TC+ class 2) The images that the user is not looking for them called TC- class**

In this thesis five different classifiers were tried for the purpose of classification of the images based on the gaze features. These include fuzzy based classifier (TSK-FIS), Feed Forward Back Propagation Neural Network (NN), Support Vector Machines (SVM), Relevance Vector Machines (RVM) and Least Square Estimation (LSE). The performance of all of these classifiers was compared when they were trained for each user with different number of training pages.

Out of these 5 classifiers the RVM and LSE were removed for further investigations because the first one provided similar results to other classifiers with a considerably more training time and the performance of the second one did not improve from a certain level, which was lower compared to the performance of other classifiers, by increasing the number of the training pages.

At the next stage 18 pages of training pages was chosen for the optimum number of training pages for partitioning the input data into the training sets and test sets. This number of training pages was chosen because naturally we intend to train our classifiers with minimum amount of training data while we are sure that we are not compromising the performance of the classifier.

6) To select the best feature set that describes the existence of one's interest to an image.

To investigate which gaze features are the most informative ones with two methods the dimension of the two gaze features were reduced. With the first method a dependency score was defined that described the maximum dependency of a gaze feature to the Target Concept and the minimum dependency of the feature to the distracting factors namely saliency and similarity. Based on this score the features that were assigned a score lower than one where removed. With the second method, called backward elimination, the gaze features were removed from the vectors one by one until there was no improvement in the classification performance. The results showed that although the dependency score helped remove some redundant features, the backward elimination helped keep the best features by improving the classification performance better than the other method.

Future works:

Although eye-tracking is an old topic with over 100 years of history, still it has a great potential for further advances. Due to limitations in the current technology there were shortcomings in the provided information for processing.

Throughout this research all of the calculations were solely based on the fixations. This is because with the provided technology it was not possible to monitor the saccade movements of the eyes as a result of the low working frequency (60Hz) of the eye trackers. There is a great potential for machines to understand one's thoughts better by processing saccade data like saccade speed and direction. There are various features that can be extracted from saccade information. These include:

- The angle of saccade: The angle by which the saccade starts/ends
- Duration of saccade: The total time that the saccade lasted
- Saccade travel distance: The distance that the saccade travelled (this is different from the tPrDist feature used in this thesis which is the distance between two visited images consecutively which can be formed of multiple saccades)
- Speed of saccade: The proportion of saccade duration and saccade distance.
- Saccade skip number: Number of saccades that passed over an image without any fixations
- Number of saccades in an eye movement: An eye movement can be formed of multiple saccades with very short fixations between them

At the current stage of technology the error that the eye-tracker producers claim for their eye-trackers in ideal conditions (User position, lighting of the room, etc.) is 0.5 degrees. This means 4.4 millimetres when the user is 50 cm far from the monitor screen (the experienced error in this research exceeded 1cm). This error prevents a reliable record of gaze movements within image boundaries. The accurate eye-movement data within an image can provide potentially useful information including:

- Number of fixations and saccades within an image
- Users' level of attention to an specific image
- Within image saliency versus within screen saliency
- The reason for attention to an image (whether an object inside the image attracted a user's attention or this was because of the texture of the image)

By removing the mentioned limitations it is possible to develop similar frameworks to the one proposed in this thesis with better performance as there are more data that can help the machines when investigating the gaze movements.

At the current stage the proposed framework is based on a limited database with standalone software. Further work is possible to integrate the proposed methods in this thesis in distributed environments such as search engines and study the outcomes in a real search environment.

List of Publications

S.N. Hajimirza, M.J. Proulx, and E. Izquierdo, "Reading Users' Minds From Their Eyes: A Method for Implicit Image Annotation," Multimedia, IEEE Transactions on, vol. 14, no. 3, pp. 805-815, June 2012.

S.N.H. Mirza and E. Izquierdo, "Examining visual attention: A method for revealing users' interest for images on screen," in Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on, sept. 2011, pp. 207-212.

S. Navid H., Michael Proulx, and Ebroul Izquierdo, "Gaze movement inference for user adapted image annotation and retrieval," in Proceedings of the 2011 ACM workshop on Social and behavioural networked media access, New York, NY, USA, 2011, pp. 27-32. [Online]. http://doi.acm.org/10.1145/2072627.2072636

Seyed Navid Haji and Ebroul Izquierdo, "Finding the user's interest level from their eyes," in Proceedings of the 2010 ACM workshop on Social, adaptive and personalized multimedia interaction and access, New York, NY, USA, 2010, pp. 25-28. [Online]. http://doi.acm.org/10.1145/1878061.1878070

S.N. Hajimirza and E. Izquierdo, "Gaze movement inference for implicit image annotation," in Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on, 2010, pp. 1-4.

References

[1] Catharine Smith. (2012, Apr.) www.huffingtonpost.com. [Online]. http://www.huffingtonpost.com/2012/04/23/facebook-s-1-amendment_n_1446853.html?ref=technology

[2] Yahoo! UK Ltd. (2012, Apr.) www.flickr.com. [Online]. http://www.flickr.com/photos/franckmichel/6855169886/

[3] K. Holmqvist et al., *Eye Tracking: A Comprehensive Guide to Methods and Measures*.: Oxford University Press, 2011. [Online]. http://books.google.co.uk/books?id=CjeGZwEACAAJ

[4] Andrew T. Duchowski, *Eye Tracking Methodology: Theory and Practice*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[5] Keith Rayner, "Eye Movements in Reading and Information Processing: 20 Years of Research," *Psychological Bulletin*, vol. 124, no. 3, pp. 372-422, 1998.

[6] Alfred L. Yarbus, *Eye Movements and Vision*. New York: Plenum Press, 1967, Translated from Russian by Basil Haigh. Original Russian edition published in Moscow in 1965.

[7] Seeing Machines Co., Facelab 5, Accessed on: 24/04/2012.

[8] Sheng-Wen Shih and Jin Liu, "A novel approach to 3-D gaze tracking using stereo cameras," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 1, pp. 234-245, feb. 2004.

[9] Michael F. Lan, "Vision, Eye Movements, and Natural Behavior," *Visual Neuroscienc*, vol. 26, no. 1, pp. 51-62, 2009.

[10] John D. Smith, Roel Vertegaal, and Changuk Sohn, "ViewPointer: Lightweight calibration-free eye tracking for ubiquitous handsfree deixis," in *Proceedings of UIST 2005*, 2005, pp. 53-61.

[11] David R. Hardoon and Kitsuchart Pasupa, "Image ranking with implicit feedback from eye movements," in *Proceedings of the 2010 Symposium on Eye-Tracking Research Applications*, New York, NY, USA, 2010, pp. 291-298. [Online]. http://doi.acm.org/10.1145/1743666.1743734

[12] Laszlo Kozma, Arto Klami, and Samuel Kaski, "GaZIR: gaze-based zooming interface for image retrieval," in *Proceedings of the 2009 international conference on Multimodal interfaces*, New York, NY, USA, 2009, pp. 305-312. [Online]. http://doi.acm.org/10.1145/1647314.1647379

[13] K. Pasupa et al., "Learning to rank images from eye movements," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, 27 2009-oct. 4 2009, pp. 2009-2016.

[14] Zhiwei Zhu, Kikuo Fujimura, and Qiang Ji, "Real-Time Eye Detection and Tracking under Various Light Conditions," in *In Proceedings of ETRA: Eye Tracking Research and Applications Symposium*, 2002, pp. 139-144.

[15] Y. Yang et al., "Web and Personal Image Annotation by Mining Label Correlation With Relaxed Visual Graph Embedding," *Image Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 1339-1351, march 2012.

[16] Chong Wang, D. Blei, and Fei F. Li, "Simultaneous image classification and annotation," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, Los Alamitos, CA, USA, 2009, pp. 1903-1910.

[17] Ilaria Bartolini and Paolo Ciaccia, "Multi-dimensional keyword-based image annotation and search," in *Proceedings of the 2nd International Workshop on Keyword Search on Structured Data*, New York, NY, USA, 2010, pp. 5:1--5:6.

[18] T. Sumathi and M. Hemalatha, "A combined hierarchical model for automatic image annotation and retrieval," in *Advanced Computing (ICoAC), 2011 Third International Conference on*, dec. 2011, pp. 135-139.

[19] Xiang Sean Zhou and Thomas S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems*, vol. 8, pp. 536-544, 2003.

[20] Divna Djordjevic, "User Relevance Feedback, Search and Retrieval of Visual Content," Queen Mary University of London, London, UK, Ph.D. dissertation 2006.

[21] Bryan Russell, Antonio Torralba, and William T. Freeman, Labelme the Open Annotation Tool, Accessed on: 09/04/2011.

[22] Yu Tang Guo and Bin Luo, "An automatic image annotation method based on the mutual K-nearest neighbor graph," in *Natural Computation (ICNC), 2010 Sixth International Conference on*, vol. 7, aug. 2010, pp. 3562-3566.

[23] Dongjian He, Yu Zheng, Shirui Pan, and Jinglei Tang, "Ensemble of multiple descriptors for automatic image annotation," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 4, oct. 2010, pp. 1642-1646.

[24] Liu Wenyin et al., "Semi-automatic image annotation," in *In Proc. of Interact 2001: Conference on Human-Computer Interaction*, 2001, pp. 326-333.

[25] A. Dorado and E. Izquierdo, "Semi-Automatic Image Annotation Using Frequent Keyword Mining," in *Information Visualisation, International Conference on*, Los

Alamitos, CA, USA, 2003, p. 532.

[26] Rui Shi, Huamin Feng, Tat-Seng Chua, and Chin-Hui Lee,.: Springer Berlin Heidelberg, 2004, vol. 3115, pp. 1951-1951.

[27] O.O. Karadag and F.T.Y. Vural, "HANOLISTIC: A Hierarchical Automatic Image Annotation System Using Holistic Approach," in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, june 2009, pp. 16-21.

[28] Linsen Yu, Yongmei Liu, and Tianwen Zhang, "Using Example-Based Machine Translation Method For Automatic Image Annotation," in *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*, vol. 2, 2006, pp. 9809-9812.

[29] M. Paradowski and A. Sluzek, "Automatic image annotation by image fragment matching," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, sept. 2011, pp. 83-89.

[30] S.A. Manaf and M.J. Nordin, "Review on statistical approaches for automatic image annotation," in *Electrical Engineering and Informatics, 2009. ICEEI '09. International Conference on*, vol. 01, aug. 2009, pp. 56-61.

[31] Lixing Jiang, Jin Hou, Zeng Chen, and Dengsheng Zhang, "Automatic image annotation based on decision tree machine learning," in *Cyber-Enabled Distributed Computing and Knowledge Discovery, 2009. CyberC '09. International Conference on*, oct. 2009, pp. 170-175.

[32] Hao Ma, Jianke Zhu, M.R.-T. Lyu, and I. King, "Bridging the Semantic Gap Between Image Contents and Tags," *Multimedia, IEEE Transactions on*, vol. 12, no. 5, pp. 462-473, 2010.

[33] Guiguang Ding and Na Xu, "Automatic semantic annotation of images based on Web data," in *Information Assurance and Security (IAS), 2010 Sixth International*

Conference on, aug. 2010, pp. 317-322.

[34] Tianxia Gong, Shimiao Li, and Chew Lim Tan, "A Semantic Similarity Language Model to Improve Automatic Image Annotation," in *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on*, vol. 1, oct. 2010, pp. 197-203.

[35] Bongwon Suh and Benjamin B. Bederson, "Semi-Automatic Image Annotation Using Event and Torso Identification," Computer Science Department, University of Maryland, College Park, MD, Tech. rep. 2004.

[36] Theodora Tsikrika, Christos Diou, Arjen P. de, and Anastasios Delopoulos, "Image annotation using clickthrough data," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, New York, NY, USA, 2009, pp. 14:1--14:8.

[37] Klimis Ntalianis, Nicolas Tsapatsoulis, Anastasios Doulamis, and Nikolaos Matsatsinis, "Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution," *Multimedia Tools and Applications*, pp. 1-25, 2012, 10.1007/s11042-012-0995-2. [Online]. http://dx.doi.org/10.1007/s11042-012-0995-2

[38] Jun Jiao and Maja Pantic, "Implicit image tagging via facial information," in *Proceedings of the 2nd international workshop on Social signal processing*, Firenze, Italy, 2010, pp. 59-64.

[39] R. Jesus, D. Goncalves, A.J. Abrantes, and N. Correia, "Playing games as a way to improve automatic image annotation," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, june 2008, pp. 1-8.

[40] L. Seneviratne and E. Izquierdo, "Image annotation through gaming (TAG4FUN)," in *Digital Signal Processing, 2009 16th International Conference on*, 2009, pp. 1-6.

[41] Ashkan Yazdani, Jong-Seok Lee, and Touradj Ebrahimi, "Implicit emotional tagging of multimedia using EEG signals and brain computer interface," in *Proceedings of the first SIGMM workshop on Social media*, New York, NY, USA, 2009, pp. 81-88.

[42] S.N. Hajimirza and E. Izquierdo, "Gaze movement inference for implicit image annotation," in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, 2010, pp. 1-4.

[43] Seyed Navid Haji and Ebroul Izquierdo, "Finding the user's interest level from their eyes," in *Proceedings of the 2010 ACM workshop on Social, adaptive and personalized multimedia interaction and access*, New York, NY, USA, 2010, pp. 25-28. [Online]. http://doi.acm.org/10.1145/1878061.1878070

[44] S. Navid H., Michael Proulx, and Ebroul Izquierdo, "Gaze movement inference for user adapted image annotation and retrieval," in *Proceedings of the 2011 ACM workshop on Social and behavioural networked media access*, New York, NY, USA, 2011, pp. 27-32. [Online]. http://doi.acm.org/10.1145/2072627.2072636

[45] S.N.H. Mirza and E. Izquierdo, "Examining visual attention: A method for revealing users' interest for images on screen," in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, sept. 2011, pp. 207-212.

[46] Klimis Ntalianis, Anastasios Doulamis, and Nicolas Tsapatsoulis, "Implicit visual concept modeling in image / video annotation," in *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*, Firenze, Italy, 2010, pp. 33--38.

[47] C. Koch and S Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, pp. 219-227, 1985.

[48] Velia Cardin, Karl J Friston, and Semir Zeki, "Top-down modulations in the visual form pathway revealed with dynamic causal modeling," *Cereb Cortex*, vol. 21, no. 3, pp. 550-562, 2011.

[49] Michael J. Proulx, "Bottom-Up Guidance in Visual Search for Conjunctions," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, no. 1, pp. 48-56, 2007.

[50] S.N. Hajimirza, M.J. Proulx, and E. Izquierdo, "Reading Users' Minds From Their Eyes: A Method for Implicit Image Annotation," *Multimedia, IEEE Transactions on*, vol. 14, no. 3, pp. 805-815, june 2012.

[51] Peter Auer et al., "Pinview: Implicit Feedback in Content-Based Image Retrieval," in *Workshop on Applications of Pattern Analysis*, 2010, pp. 21-57.

[52] Georg Buscher, Andreas Dengel, and Ludger van Elst, "Eye movements as implicit relevance feedback," in *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, Florence, Italy, 2008, pp. 2991--2996.

[53] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V. Elst, "Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond," *ACM Trans. Interact. Intell. Syst.*, vol. 1, no. 2, pp. 9:1--9:30, Jan 2012.

[54] A. Faro, D. Giordano, C. Pino, and C. Spampinato, "Visual attention for implicit relevance feedback in a content based image retrieval," in *Proceedings of the 2010 Symposium on Eye-Tracking Research Applications*, Austin, Texas, 2010, pp. 73--76.

[55] He Zhang, Teemu Ruokolainen, Jorma Laaksonen, Christina Hochleitner, and Rudolf Traunmüller, *Gaze- and Speech-Enhanced Content-Based Image Retrieval in Image Tagging*, Timo Honkela et al., Eds.: Springer Berlin Heidelberg, 2011.

[56] Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, and Julian Mennenoh, "Increasing the user's attention on the web: using implicit interaction based on gaze behavior to tailor content," in *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, Copenhagen, Denmark, 2012, pp. 544-553.

[57] Arto Klami, Craig Saunders, Teo filo E. de, and Samuel Kaski, "Can relevance of images be inferred from eye movements?," in *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, New York, NY, USA, 2008, pp. 134-140. [Online]. http://doi.acm.org/10.1145/1460096.1460120

[58] M. Pantic and A. Vinciarelli, "Implicit human-centered tagging [Social Sciences]," *Signal Processing Magazine, IEEE*, vol. 26, no. 6, pp. 173-180, November 2009.

[59] Mohammad Soleymani and Maja Pantic, "Human-centered implicit tagging: Overview and perspectives.," in *IEEE International Conference on Systems, Man and Cybernetics*, Oct 2012, pp. 3304-3309.

[60] Sander Koelstra and Ioannis Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, 2013. [Online]. http://www.sciencedirect.com/science/article/pii/S0262885612001825

[61] N. Wade and B.W. Tatler, *The Moving Tablet of the Eye: The Origins of Modern Eye Movement Research*.: Oxford University Press, 2005. [Online]. http://books.google.co.uk/books?id=hJg2xhz7XKUC

[62] A. Plotkin, O. Shafrir, E. Paperno, and D.M. Kaplan, "Magnetic Eye Tracking: A New Approach Employing a Planar Transmitter," *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 5, pp. 1209-1215, may 2010.

[63] F. Träisk, R. Bolzani, and J. Ygge, "A comparison between the magnetic scleral search coil and infrared reflection methods for saccadic eye movement analysis," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 243, pp. 791-797, 2005, 10.1007/s00417-005-1148-3. [Online]. http://dx.doi.org/10.1007/s00417-005-1148-3

[64] R. M. da and A. Gonzaga, "Dynamic Features for Iris Recognition," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. PP, no. 99, pp. 1-11, 2012.

[65] L.H. Yu and M. Eizenman, "A new methodology for determining point-of-gaze in head-mounted eye tracking systems," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 10, pp. 1765-1773, oct. 2004.

[66] Zhiwei Zhu and Qiang Ji, "Novel Eye Gaze Tracking Techniques Under Natural Head Movement," *Biomedical Engineering, IEEE Transactions on*, vol. 54, no. 12, pp. 2246-2260, dec. 2007.

[67] G. Iannizzotto and F. La Rosa, "Competitive Combination of Multiple Eye Detection and Tracking Techniques," *Industrial Electronics, IEEE Transactions on*, vol. 58, no. 8, pp. 3151-3159, aug. 2011.

[68] J.J. Magee, M. Betke, J. Gips, M.R. Scott, and B.N. Waber, "A Human ;Computer Interface Using Symmetry Between Eyes to Detect Gaze Direction," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 38, no. 6, pp. 1248-1261, nov. 2008.

[69] C. Hennessey and P. Lawrence, "Noncontact Binocular Eye-Gaze Tracking for Point-of-Gaze Estimation in Three Dimensions," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 3, pp. 790-799, march 2009.

[70] V. Rantanen et al., "A Wearable, Wireless Gaze Tracker with Integrated Selection Command Source for Human;Computer Interaction," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 15, no. 5, pp. 795-801, sept. 2011.

[71] A. Villanueva and R. Cabeza, "A Novel Gaze Estimation System With One Calibration Point," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 38, no. 4, pp. 1123-1138, aug. 2008.

[72] Dongsoo Kim and Gunhee Han, "A 200 s Processing Time Smart Image Sensor for an Eye Tracker Using Pixel-Level Analog Image Processing," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 9, pp. 2581-2590, sept. 2009.

[73] Michal Jacob and Shaul Hochstein, "Graded recognition as a function of the

number of target fixations.," *Vision research*, vol. 50, pp. 107-17, 2010 Jan 2010.

[74] Melissa M. Kibbe and Eileen Kowler, "Visual search for category sets: Tradeoffs between exploration and memory," *Journal of Vision March 18*, vol. 11, no. 3, p. article 14, March 2011.

[75] Helga Mazyar, Ronald van den, and Wei Ji Ma, "Does precision decrease with set size?," *Journal of Vision*, vol. 12, no. 6, pp. 1-16, 2012. [Online]. http://www.journalofvision.org/content/12/6/10.abstract

[76] Paul V. McGraw, Neil W. Roach, David R. Badcock, and David Whitaker, "Size-induced distortions in perceptual maps of visual space," *Journal of Vision*, vol. 12, no. 4, pp. 1-14, 2012. [Online]. http://www.journalofvision.org/content/12/4/8.abstract

[77] Marnix Naber, Maximilian Hilger, and Wolfgang Einhauser, "Animal detection and identification in natural scenes: Image statistics and emotional valence," *Journal of Vision*, vol. 12, no. 1, pp. 1-24, 2012. [Online]. http://www.journalofvision.org/content/12/1/25.abstract

[78] Supriya Ray, Neha Bhutani, and Aditya Murthy, "Mutual inhibition and capacity sharing during parallel preparation of serial eye movements," *Journal of Vision*, vol. 12, no. 3, pp. 1-22, 2012. [Online]. http://www.journalofvision.org/content/12/3/17.abstract

[79] Ruth Rosenholtz, Jie Huang, Alvin Raj, Benjamin J. Balas, and Livia Ilie, "A summary statistic representation in peripheral vision explains visual search," *Journal of Vision*, vol. 12, no. 4, pp. 1-17, 2012. [Online]. http://www.journalofvision.org/content/12/4/14.abstract

[80] Gheorghita Ghinea and Gabriel-Miro Muntean, "An eye-tracking-based adaptive multimedia streaming scheme," in *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, Piscataway, NJ, USA, 2009, pp. 962-965. [Online]. http://dl.acm.org/citation.cfm?id=1698924.1699160

[81] S.R. Gulliver and G. Ghinea, "Stars in their eyes: what eye-tracking reveals about multimedia perceptual quality," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 34, no. 4, pp. 472-482, july 2004.

[82] Hantao Liu and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 7, pp. 971-982, july 2011.

[83] O. Oyekoya and F.W.M. Stentiford, "A performance comparison of eye tracking and mouse interfaces in a target image identification task," *IEE Seminar Digests*, vol. 2005, no. 11099, pp. 139-144, 2005.

[84] U. Rajashekar, I. van der, A.C. Bovik, and L.K. Cormack, "GAFFE: A Gaze-Attentive Fixation Finding Engine," *Image Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 564-573, april 2008.

[85] Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jurgen Ziegler, "A cognitive cost model of annotations based on eye-tracking data," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1158-1167. [Online]. http://dl.acm.org/citation.cfm?id=1858681.1858799

[86] Yun Zhang, Hong Fu, Zhen Liang, Zheru Chi, and Dagan Feng, "Eye movement as an interaction mechanism for relevance feedback in a content-based image retrieval system," in *Proceedings of the 2010 Symposium on Eye-Tracking Research \& Applications*, New York, NY, USA, 2010, pp. 37-40. [Online]. http://doi.acm.org/10.1145/1743666.1743674

[87] Antti Ajanki, David R. Hardoon, Samuel Kaski, Kai Puolamaki, and John Shawe-Taylor, "Can eyes reveal interest? Implicit queries from gaze patterns," *User Modeling and User-Adapted Interaction*, vol. 19, no. 4, pp. 307-339, #oct# 2009. [Online]. http://dx.doi.org/10.1007/s11257-009-9066-4

[88] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective*

Computing, vol. 3, no. 1, pp. 42-55, April 2012.

[89] Ramanathan Subramanian, Victoria Yanulevskaya, and Nicu Sebe, "Can computers learn from humans to see better?: inferring scene semantics," in *Proceedings of the 19th ACM international conference on Multimedia*, Scottsdale, Arizona, USA, 2011, pp. 33-42. [Online]. http://doi.acm.org/10.1145/2072298.2072305

[90] Tina Walber, Ansgar Scherp, and Steffen Staab, "Identifying objects in images from analyzing the users' gaze movements," in *Proceedings of the 18th international conference on Advances in Multimedia*, Klagenfurt, Austria, 2012, pp. 138--148.

[91] S. Vrochidis, I. Patras, and I. Kompatsiaris, "Exploiting gaze movements for automatic video annotation," in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2012*, Dublin, 2012, pp. 1--4.

[92] Åke Björck, *Numerical Methods for Least Squares Problems*. 3600 Market Street, 6th Floor, Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1996.

[93] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[94] Chris Rorres Howard, *Elementary Linear Algebra (9th ed.)*.: John Wiley and Sons, 2005, ISBN 978-0-471-66959-3.

[95] Tilo Strutz, *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond*.: Vieweg and Teubner, 2010, ISBN 3834810223.

[96] John Wolberg, *Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments*.: Springer, 2006, ISBN 978-3540256748.

[97] Stephen L. Chiu, "Fuzzy Model Estimation Based on Cluster Esimation," *Journal of Intelligent and Fuzzy Systems*, vol. 2, no. 3, pp. 267-278, September 1994.

[98] C. Wagner and H. Hagras, "Toward General Type-2 Fuzzy Logic Systems Based on zSlices," *Fuzzy Systems, IEEE Transactions on*, vol. 18, no. 4, pp. 637-660, 2010.

[99] Mohammad Biglarbegian, William W. Melek, and Jerry M. Mendel, "On the stability of interval type-2 TSK fuzzy logic control systems," *Trans. Sys. Man Cyber. Part B*, vol. 40, no. 3, pp. 798-818, Jun 2010. [Online]. http://dx.doi.org/10.1109/TSMCB.2009.2029986

[100] W.A. Lodwick and J. Kacprzyk, *Fuzzy Optimization: Recent Advances and Applications*.: Springer, 2010. [Online]. http://books.google.co.uk/books?id=PTynZpHIer8C

[101] Jyh-Shing Roger Jang, "Neuro-Fuzzy Modeling :Architectures, Analyses, and Applications," University of California, Berkeley, CA 94720, Ph.D. dissertation 1992.

[102] F.A. Marquez, A. Peregrin, and F. Herrera, "Cooperative Evolutionary Learning of Linguistic Fuzzy Rules and Parametric Aggregation Connectors for Mamdani Fuzzy Systems," *Fuzzy Systems, IEEE Transactions on*, vol. 15, no. 6, pp. 1162-1178, 2007.

[103] A. Ferrero, A. Federici, and S. Salicone, "Instrumental Uncertainty and Model Uncertainty Unified in a Modified Fuzzy Inference System," *Instrumentation and Measurement, IEEE Transactions on*, vol. 59, no. 5, pp. 1149-1157, may 2010.

[104] Dan Li, Chongquan Zhong, and Liyong Zhang, "Fuzzy c-means clustering of partially missing data sets based on statistical representation," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, vol. 1, 2010, pp. 460-464.

[105] N.K. Verma, P. Gupta, P. Agrawal, and Yan Cui, "MRI brain image segmentation for spotting tumors using improved mountain clustering approach," in *Applied Imagery Pattern Recognition Workshop (AIPRW), 2009 IEEE*, 2009, pp. 1-8.

[106] A. Celikyilmaz and I. Burhan Turksen, "Enhanced Fuzzy System Models With Improved Fuzzy Clustering Algorithm," *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 3, pp. 779-794, 2008.

[107] Shang-Ming Zhou and John Q. Gan, "Extracting Takagi-Sugeno Fuzzy Rules with Interpretable Submodels via Regularization of Linguistic Modifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1191-1204, 2009.

[108] S.N. Sivanandam, S. Sumathi, and S. N. Deepa, *Introduction to Fuzzy Logic using MATLAB*. Berlin Heidelberg: Springer, 2007.

[109] H. W. Sorenson, "Least-squares estimation: from Gauss to Kalman," *Spectrum, IEEE*, vol. 7, pp. 63-68, 2009.

[110] I. Steinwart and A. Christmann, *Support Vector Machines*.: Springer-Verlag, New York, 2008.

[111] V. N. Vapnik, *The Nature of Statistical Learning Theory (Information Science and Statistics)*, 2nd ed. New York: Springer-Verlag, 2000.

[112] S. R. Gunn, "Support Vector Machines for Classification and Regression," University of Southhampton, Technical Report 1998.

[113] P. W. Goldberg, "Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parameterized by Real Numbers," *Machine Learning*, vol. 18, no. 1, pp. 131-148, 1995.

[114] N. Chowdhury and M.A. Kashem, "A comparative analysis of Feed-forward neural network 26; Recurrent Neural network to detect intrusion," in *Electrical and Computer Engineering, 2008. ICECE 2008. International Conference on*, 2008, pp. 488-492.

[115] D. C. Montgomery, *Statistical Quality Control*, 6th ed.: NJ: Wiley, 2009.

[116] Joseph L. Rodgers and Alan W. Nicewander, "Thirteen Ways to Look at the

Correlation Coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59-66, 1988. [Online]. http://dx.doi.org/10.2307/2685263

[117] Rozic N., F. Chiaraluce, and Radic J., "Analysis of the Correlation Coefficient Between Component Noise Squared Norms for OFDM Systems," *Signal Processing Letters, IEEE*, vol. 18, no. 5, pp. 311-314, may 2011.

[118] Jean-Baptiste Durand, Damien Camors, Yves Trotter, and Simona Celebrini, "Privileged visual processing of the straight-ahead direction in humans," *Journal of Vision*, vol. 12, no. 6, p. article 34, June 2012.

[119] Maoguo Gong Jingjing, Yan Liang, Jiao Shi, and Wenping Ma, "Fuzzy C-Means Clustering With Local Information and Kernel Metric for Image Segmentation," *Image Processing, IEEE Transactions on*, vol. 22, pp. 573 -584, Feb 2013.

[120] Dacheng Tao. (2013, Mar.) The COREL Database. [Online]. https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval

[121] Michelle R. Greene, Tommy Liu, and Jeremy M. Wolfe, "Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns," *Vision Research*, vol. 62, no. 0, pp. 1-8, 2012. [Online]. http://www.sciencedirect.com/science/article/pii/S0042698912000922

[122] Vincent and Fua, Pascal Lepetit, "Keypoint recognition using randomized trees," *Pattern Analysis and Machine Intelligence (PAMI), IEEE*, vol. 28, no. 9, pp. 1465--1479, 2006.

[123] Zhang Jing, Zhuo Li, Gao Jingjing, and Liu Zhixing, "A Study of Top-Down Visual Attention Model Based on Similarity Distance," in *2nd International Congress Image and Signal Processing (CISP)*, 2009, pp. 1-5.

[124] Zhiyong Xiong and Ke Chen, "An algorithm of image retrieval based on content similarity," in *Computer Science and Automation Engineering (CSAE), 2011 IEEE International on*, 2011.

[125] E. Chalom, E. Asa, and E. Biton, "Measuring image similarity: an overview of some useful applications," *Instrumentation Measurement Magazine, IEEE*, vol. 16, no. 1, pp. 24-28, 2013.

[126] JeremyM. Wolfe, "Guided Search 2.0 A revised model of visual search," *Psychonomic Bulletin Review*, vol. 1, no. 2, pp. 202-238, 1994.

[127] Jeffrey R.W Mounts and Brandon E Gavett, "The role of salience in localized attentional interference," *Vision Research*, vol. 44, no. 13, pp. 1575 - 1588, 2004.

[128] Jeremy Wolfe, "Guided Search 2.0 A revised model of visual search," *Psychonomic Bulletin \& Review*, vol. 1, pp. 202-238, 1994, 10.3758/BF03200774. [Online]. http://dx.doi.org/10.3758/BF03200774

[129] L Itti and C Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194-203, 2001.

[130] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*, 2007, pp. 545-552.

[131] J. Harel, A Saliency Implementation in MATLAB, Accessed on: 20/06/2012.

[132] César Ferri, José Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification.," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27-38, 2009.

[133] David L. Olson and Dursun Delen, *Advanced Data Mining Techniques*, 1st ed.: Springer Publishing Company, Incorporated, 2008.

[134] Michael E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. 34, pp. 211--244, sep 2001.

[135] Peter Torrione Morton, Sam Keene, and Kenneth, "The Pattern Recognition Toolbox for MATLAB," 2011. [Online]. http://newfolderconsulting.com/prt

[136] Michael E and others Tipping, "The relevance vector machine," *Advances in neural information processing systems*, vol. 12, no. 1, pp. 652--658, 2000.

[137] Chao and Tian, Lianfang Dong, "Accelerating Relevance-Vector-Machine-Based Classification of Hyperspectral Image with Parallel Computing," *Mathematical Problems in Engineering*, vol. 2012, 2012.

[138] Ian T Jolliffe, *Principal component analysis*.: Springer verlag, 2002.

[139] Jeremy M. Wolfe, Todd S. Horowitz, Naomi Kenner, Megan Hyle, and Nina Vasan, "How fast can you change your mind? The speed of top-down guidance in visual search," *Vision Research*, vol. 44, no. 12, pp. 1411-1426, 2004.

[140] J. M. Wolfe, G.A. Alvarez, and T. S. Horowitz, "Attention is fast but volition is slow," *Nature*, vol. 406, p. 691, 2000.

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

[142] James E Hoffman, , HEditor Pashler, Ed.: Psychology Press, 1998, vol. 31, pp. 119-153.

[143] Casimir Ludwig and Iain Gilchrist, "Goal-driven modulation of oculomotor capture," *Attention, Perception, and Psychophysics*, vol. 65, pp. 1243-1251, 2003.

[144] Harish Katti and Mohan Kankanhalli, "Eye-tracking methodology and applications to images and video," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 641-642.

[145] H. J. Bierens, "Introduction to Hilbert Spaces," Pennsylvania State University, 2007.

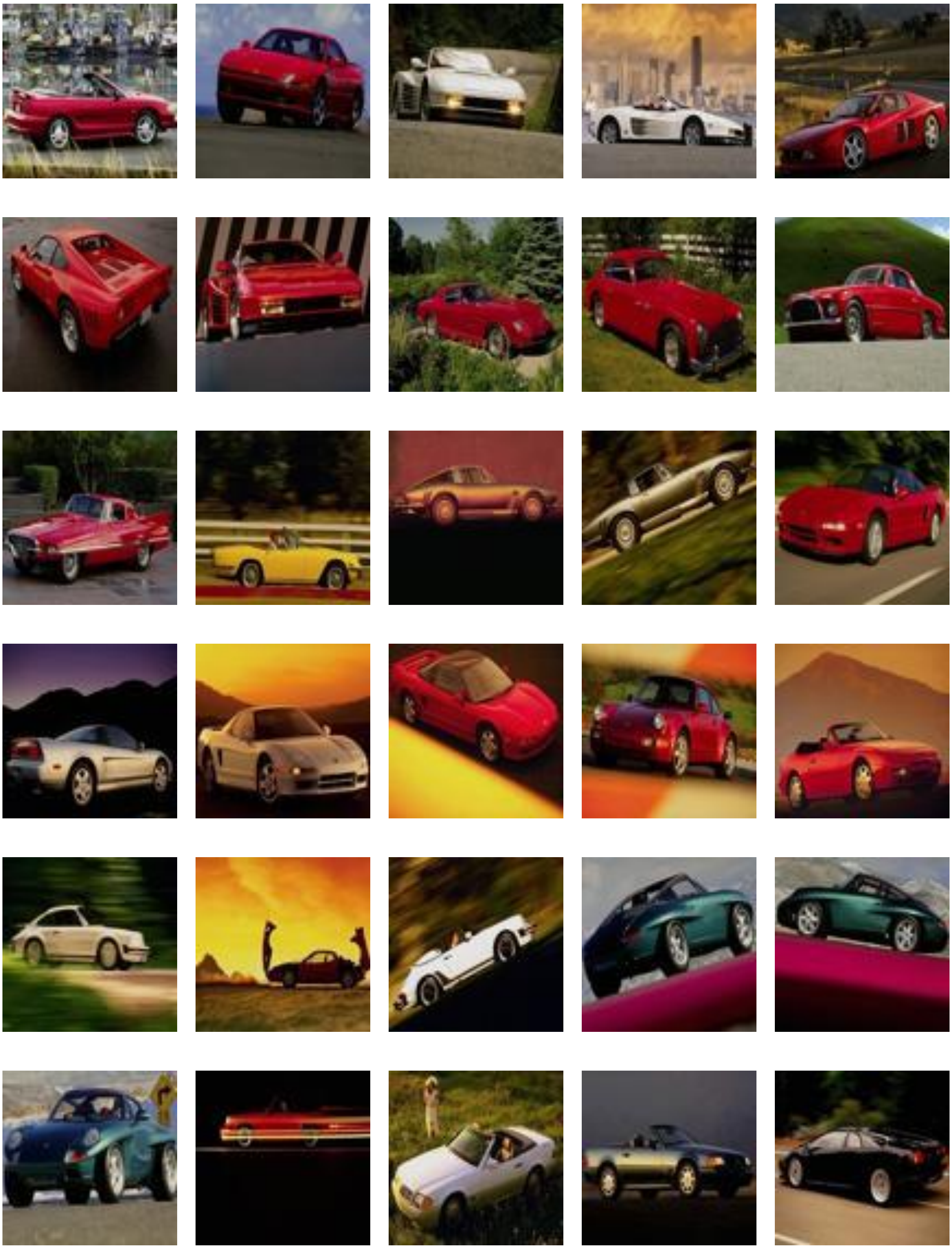
[146] M. I. Jordan and R. Thibaux, "The Kernel Trick," Berkeley, University of California, Technical Report 2004.

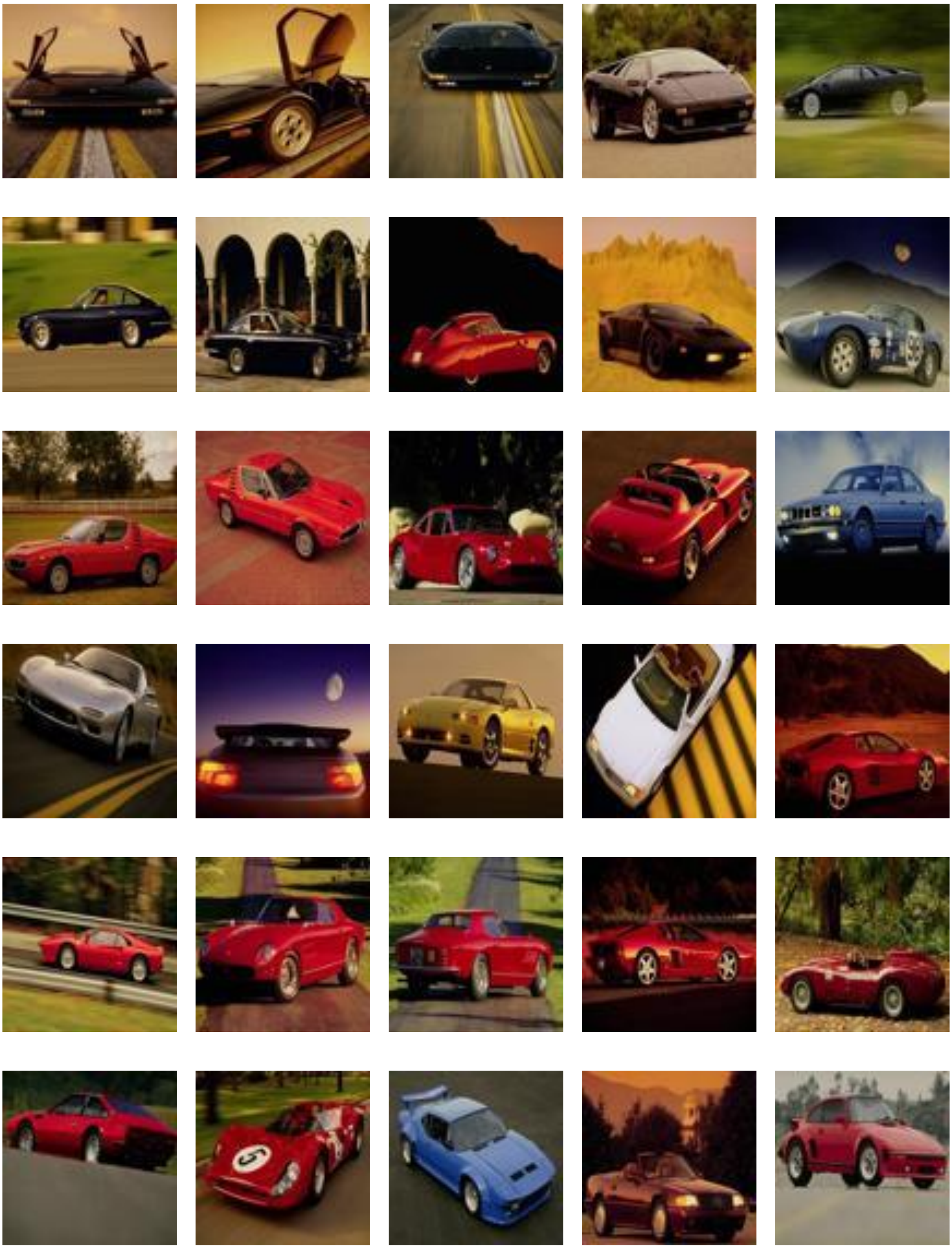
[147] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Bell Laboratories, Lucent Technologies, Kluwer Academic Publishers 1998.

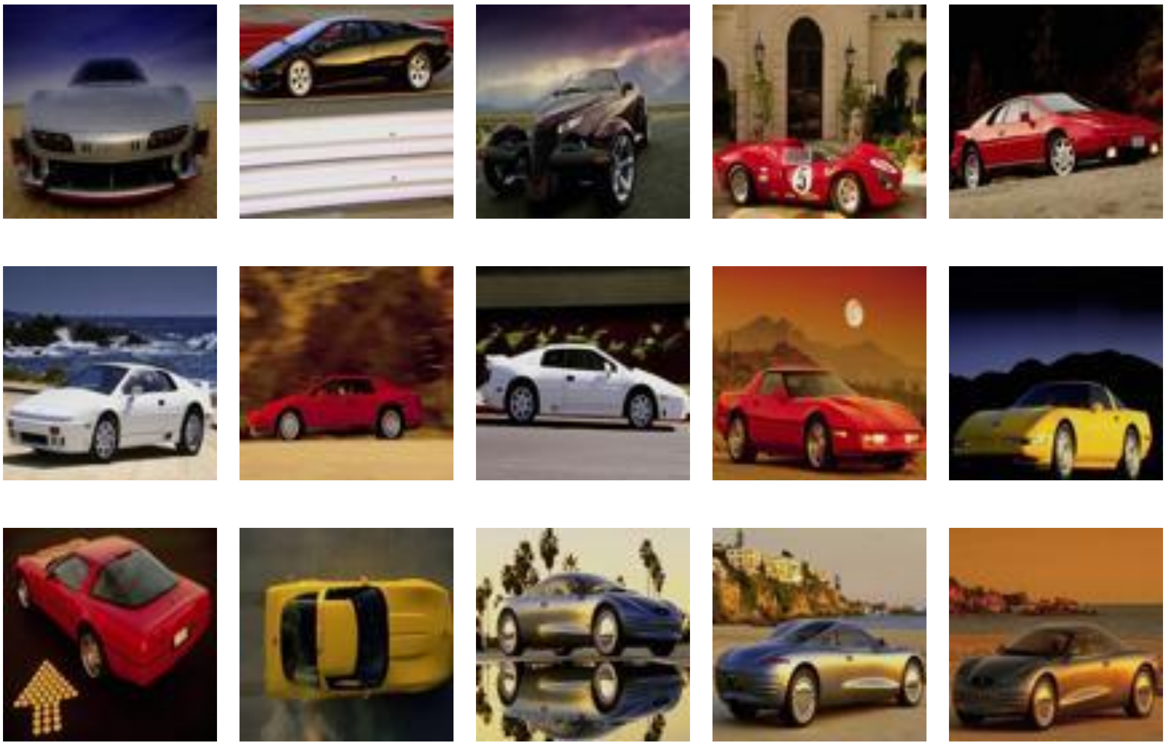
APPENDIX A GROUND TRUTH IMAGE DATASET

This APPENDIX illustrates the subset of 700 images taken from the Corel database [**120**].

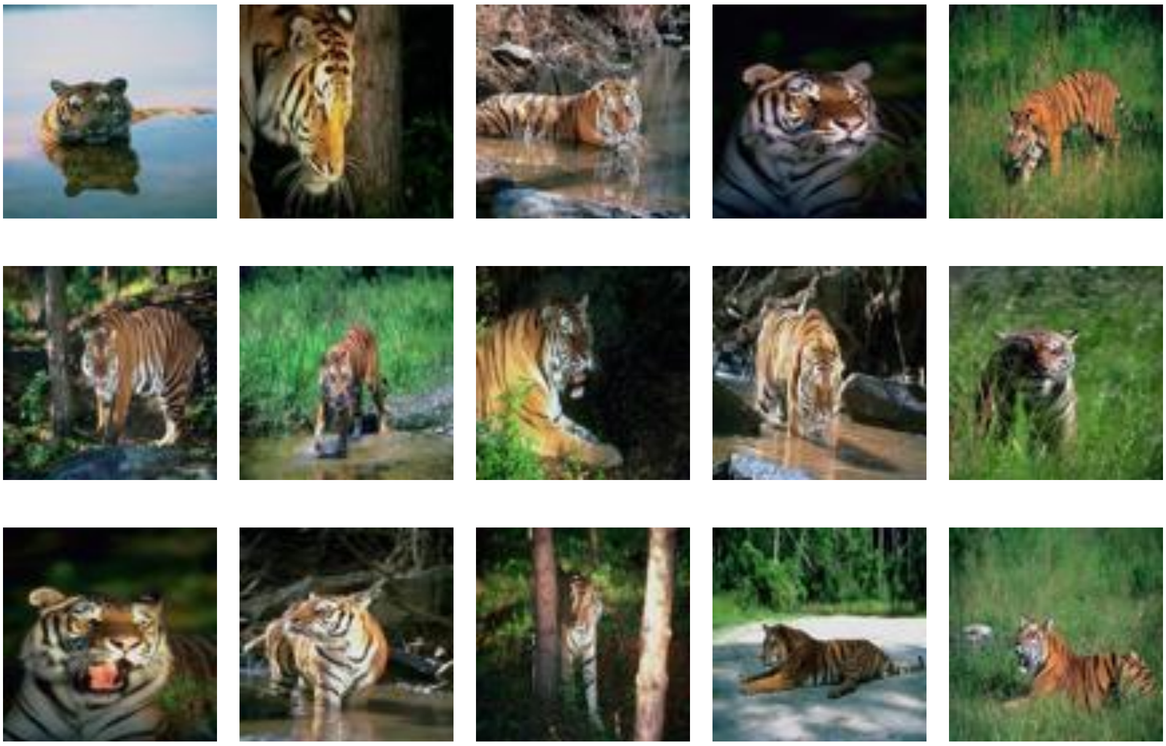
The following images are annotated as Car:



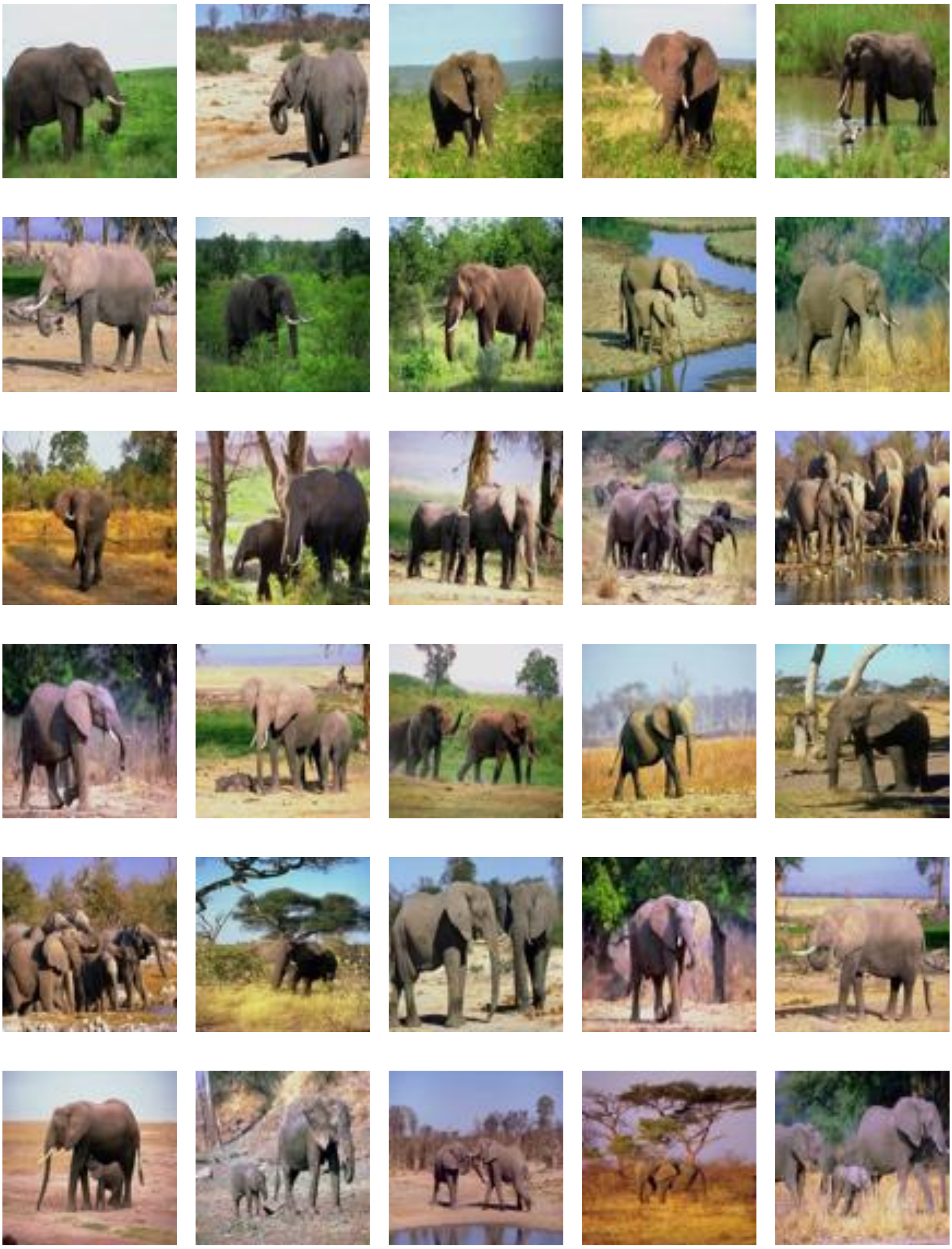


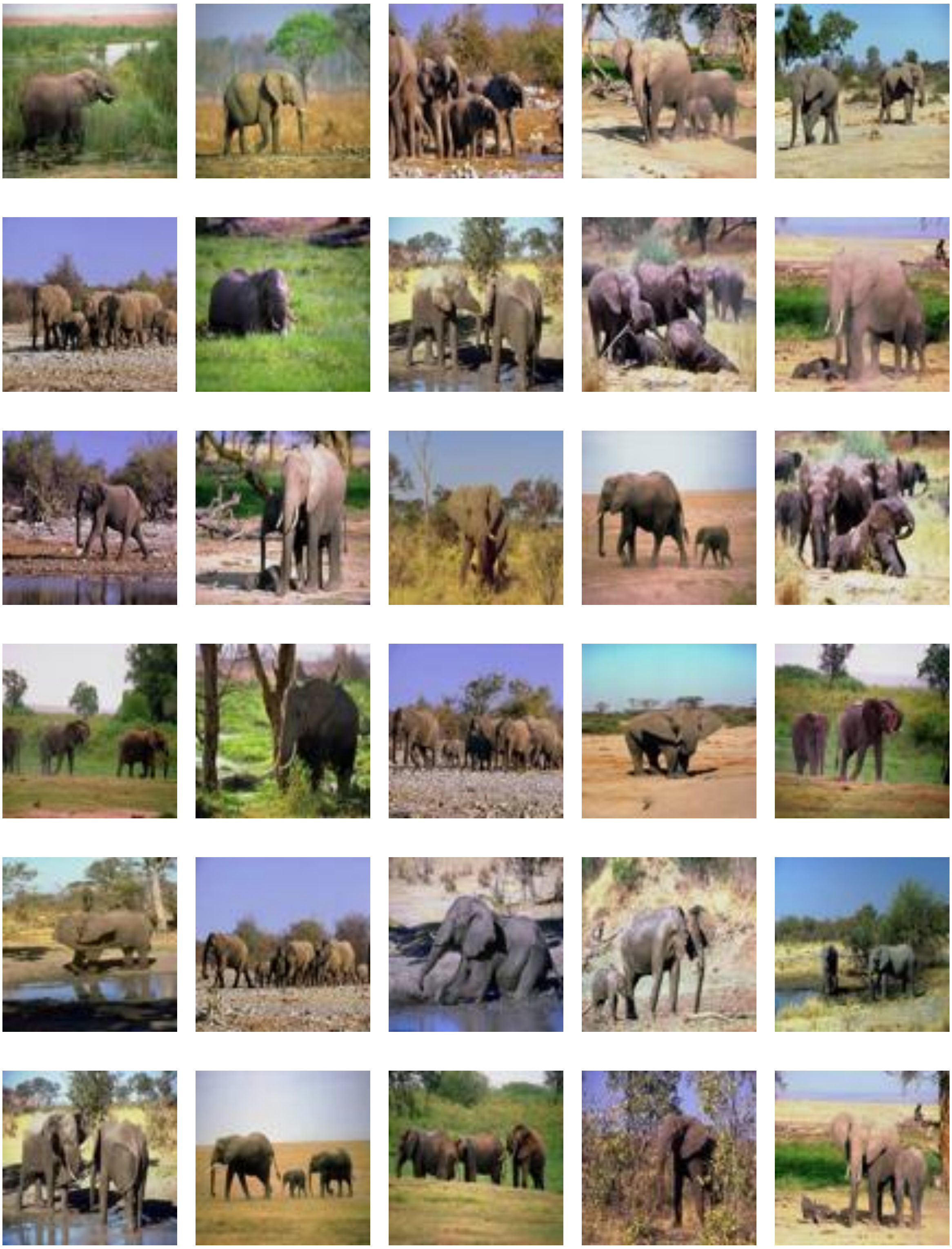


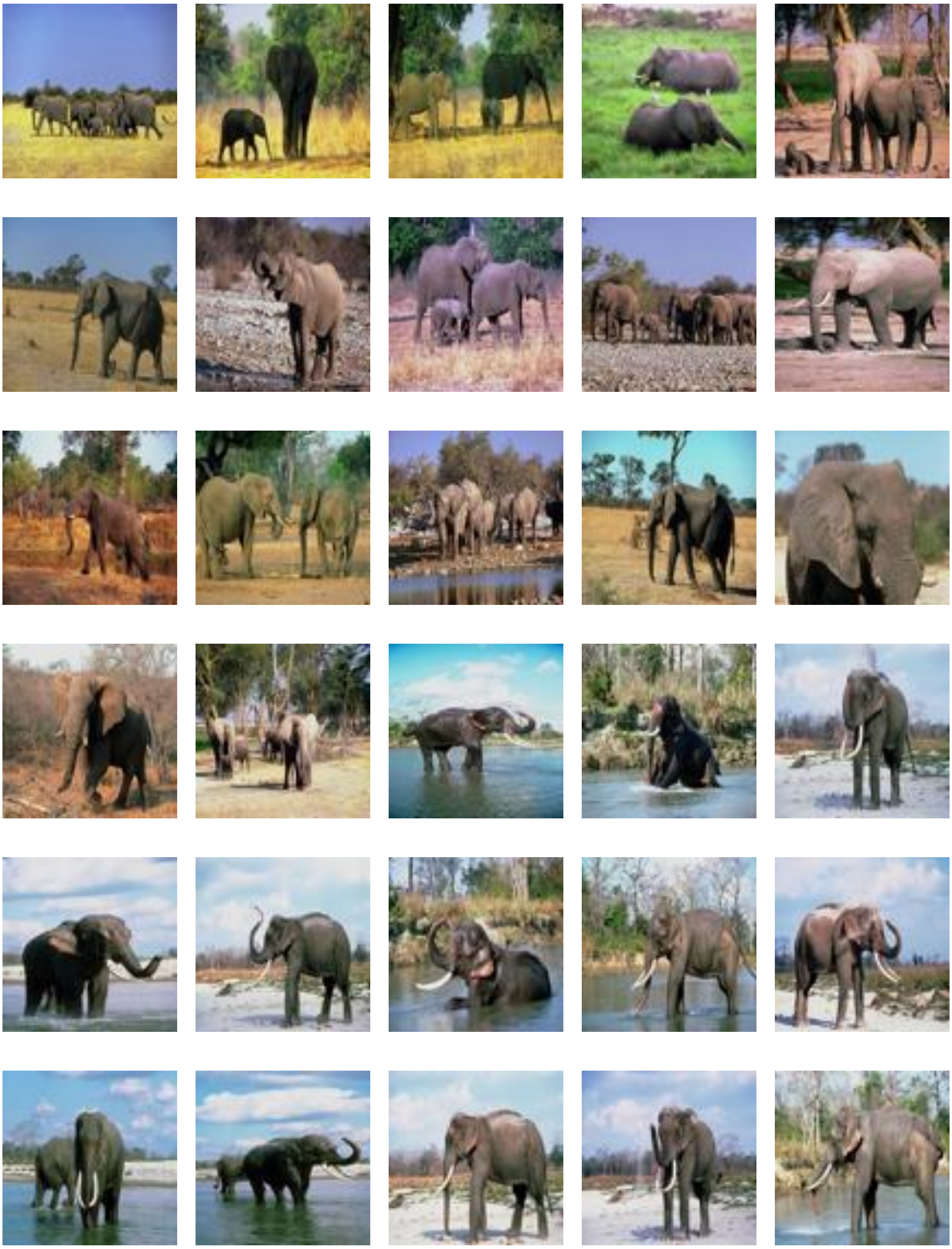
The following images are annotated as Tiger:



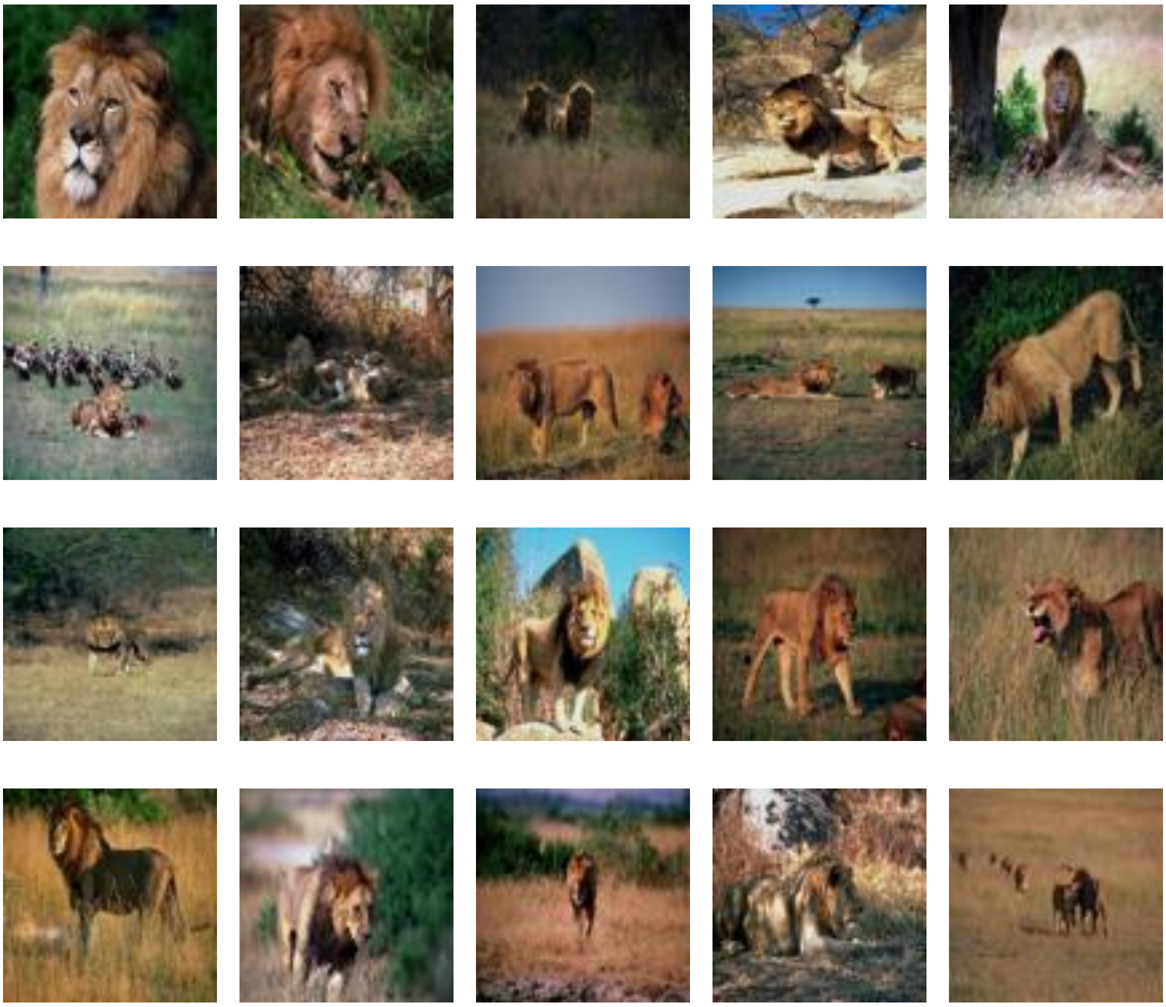
The following images are annotated as Elephant:

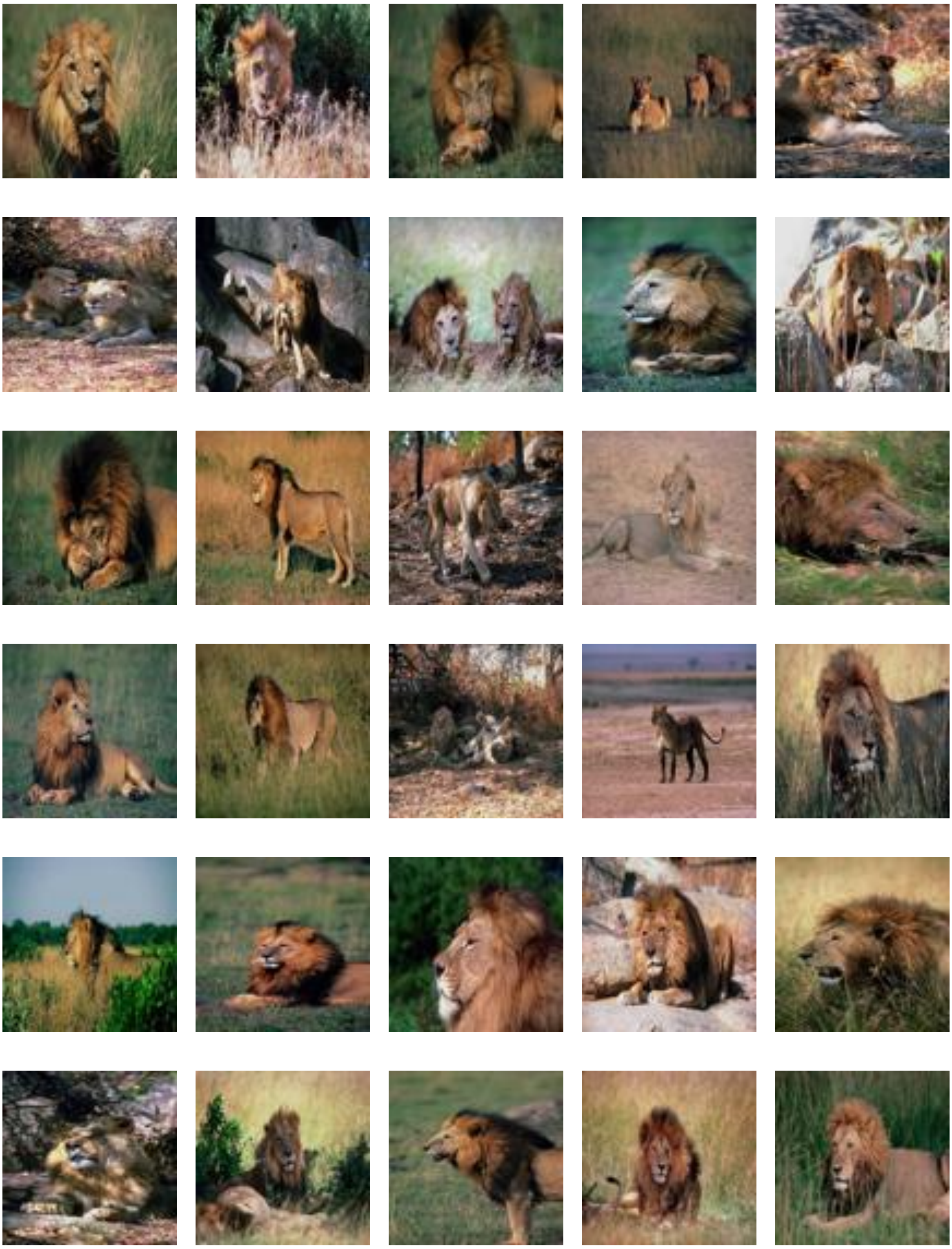


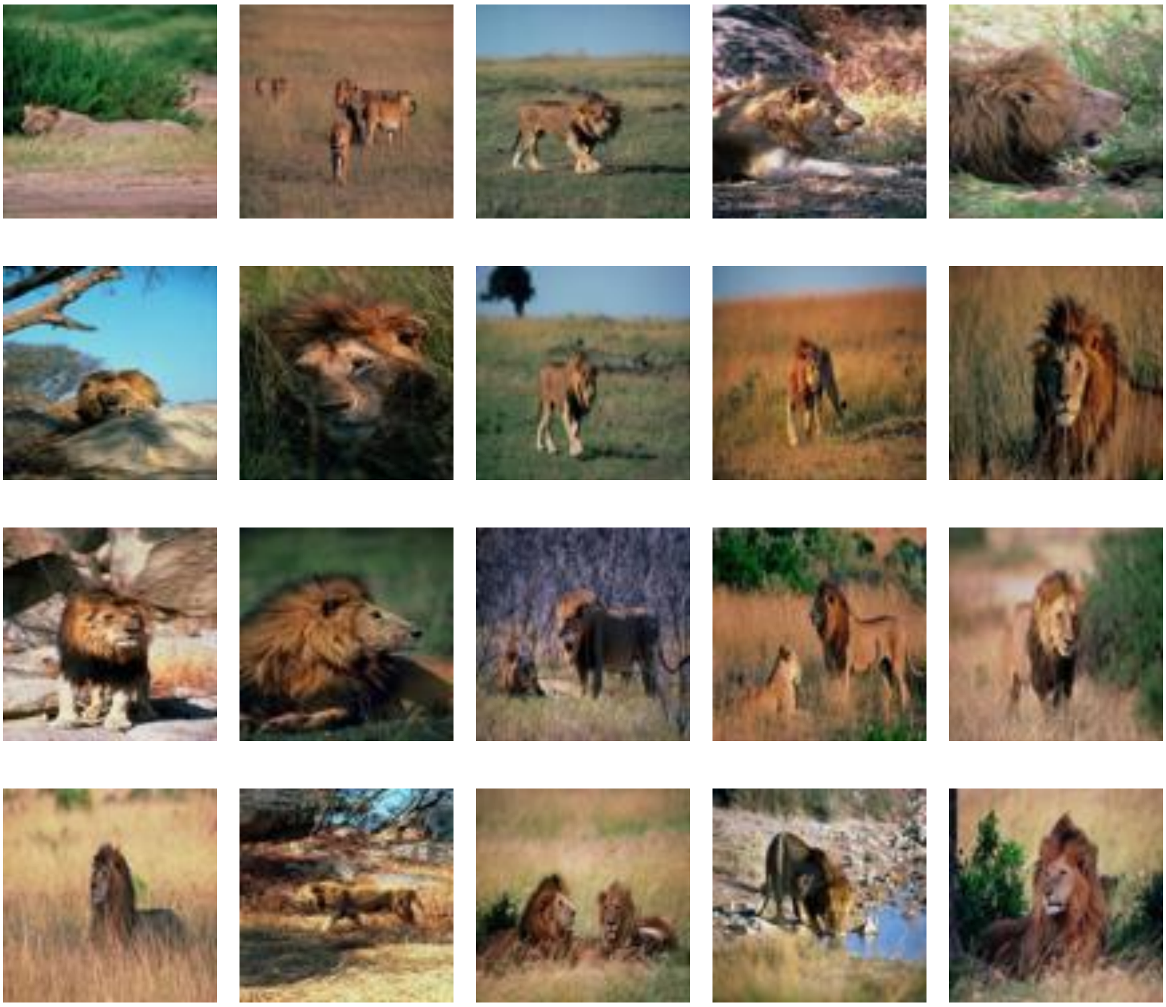




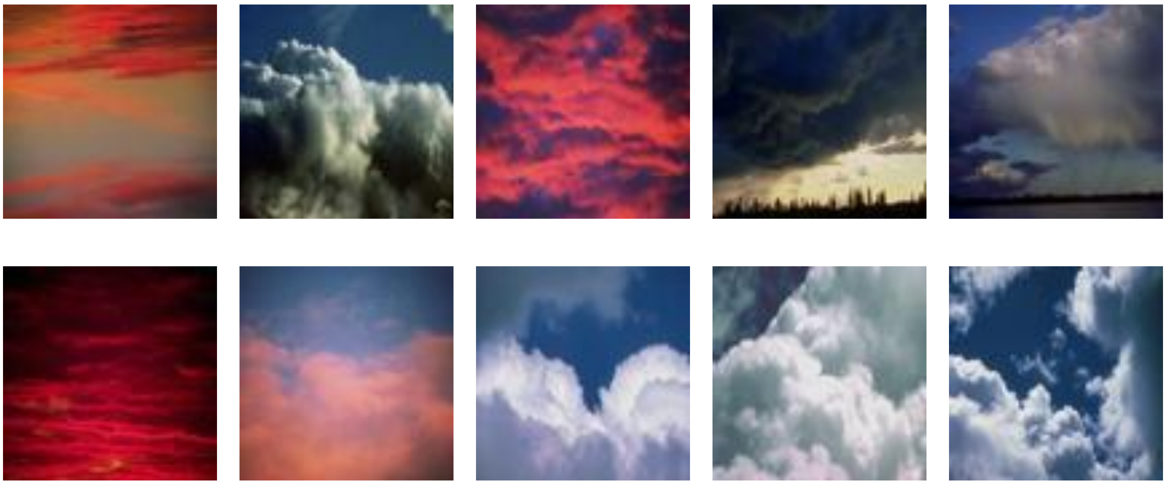
The following images are annotated as Lion:

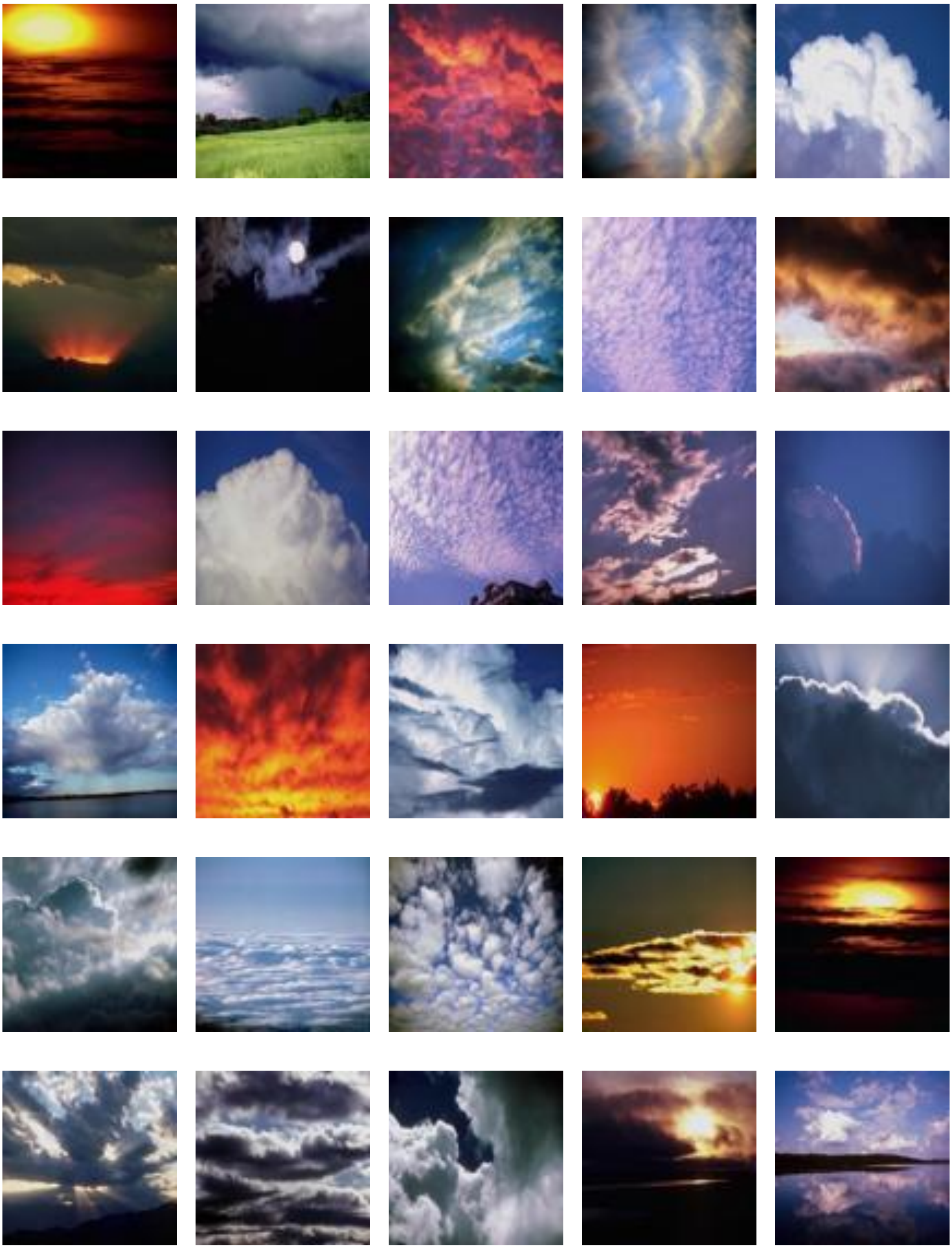


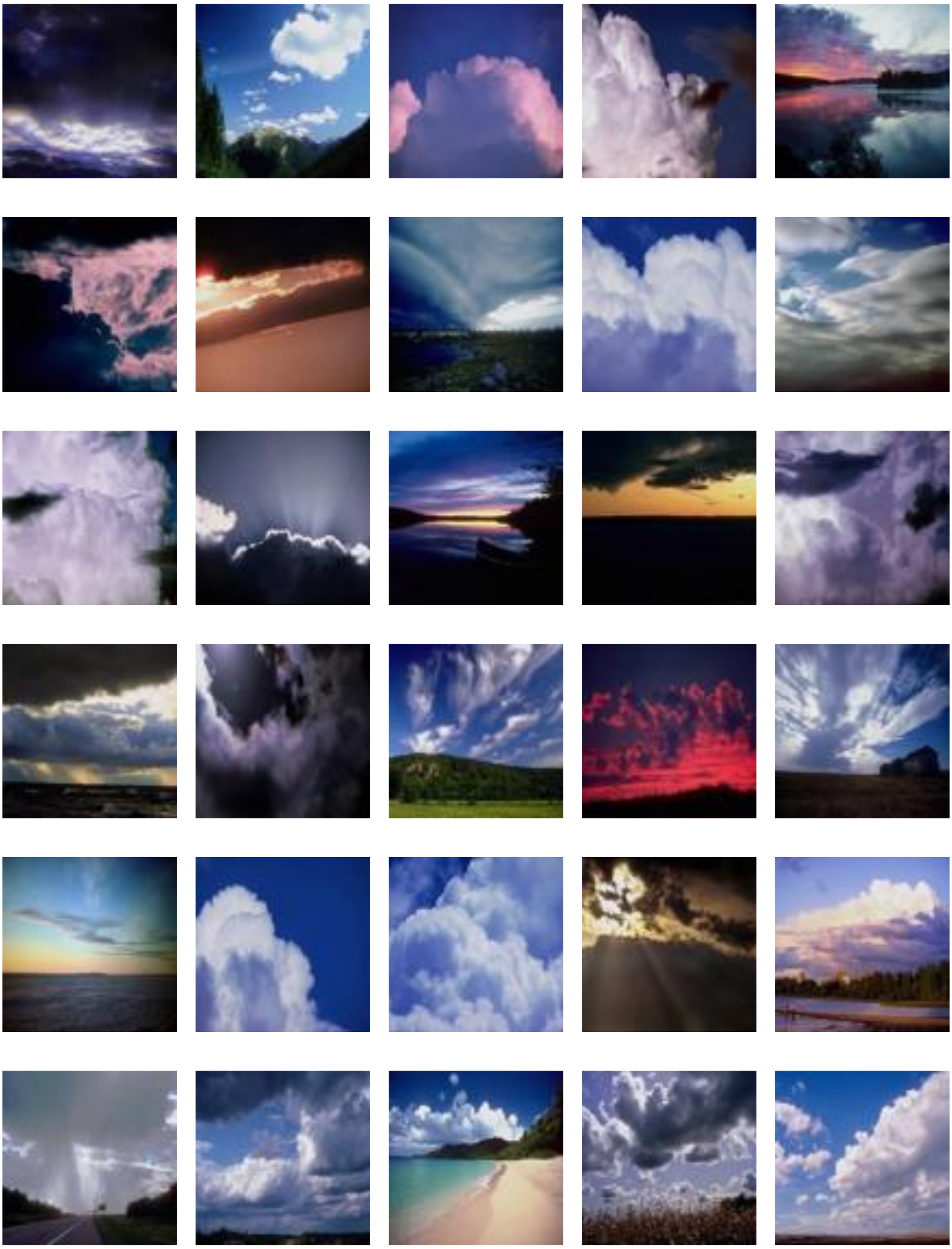


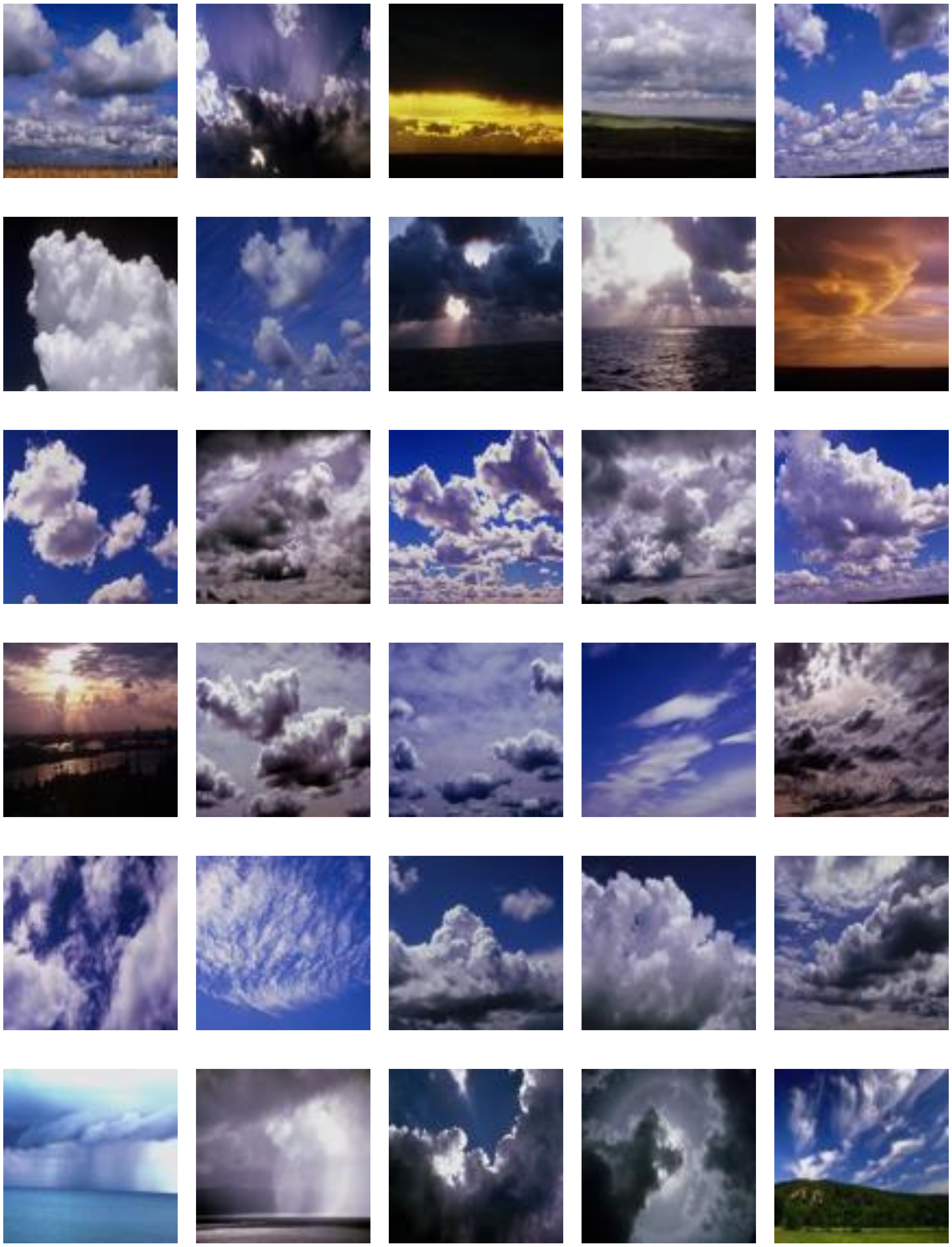


The following images are annotated as Sky:

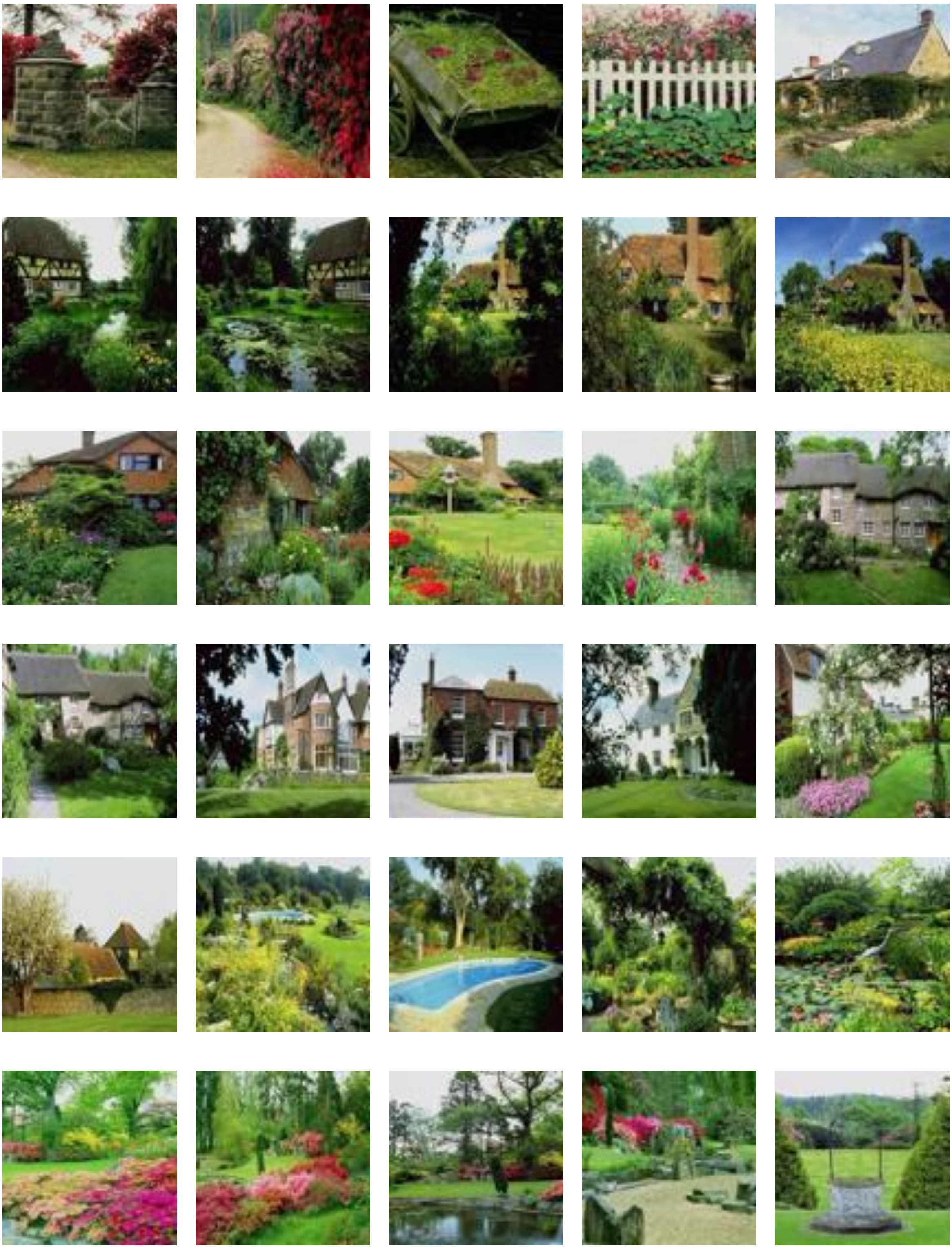




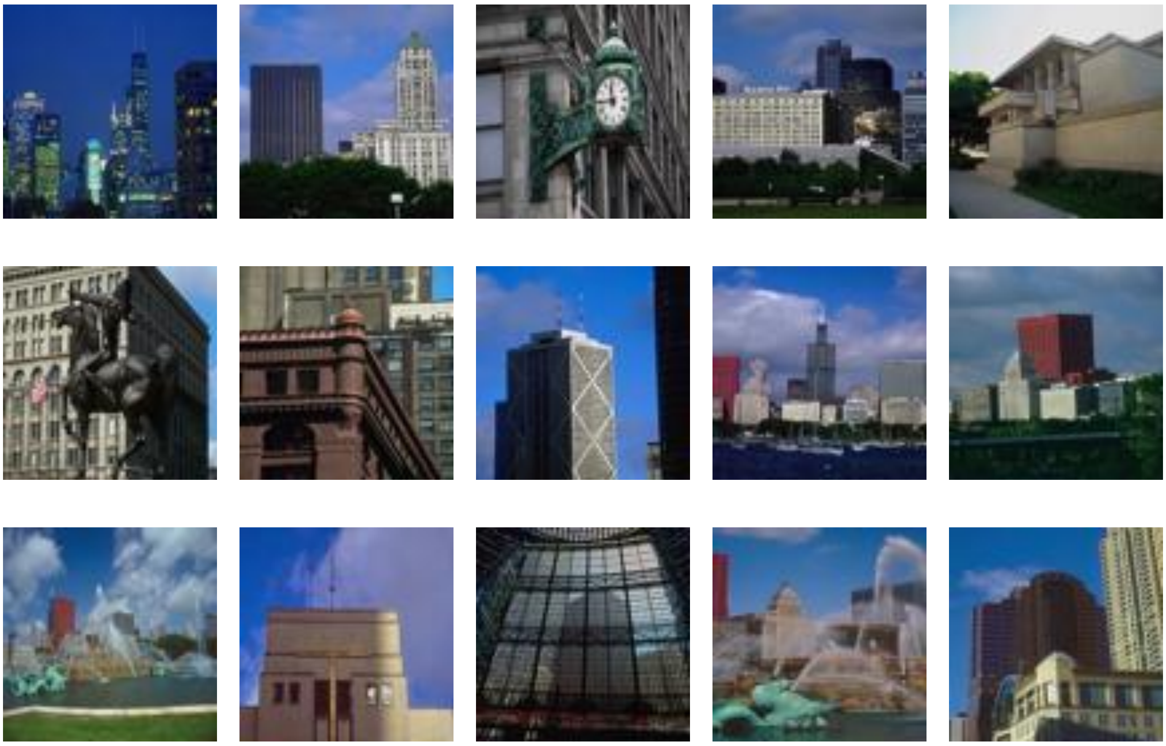


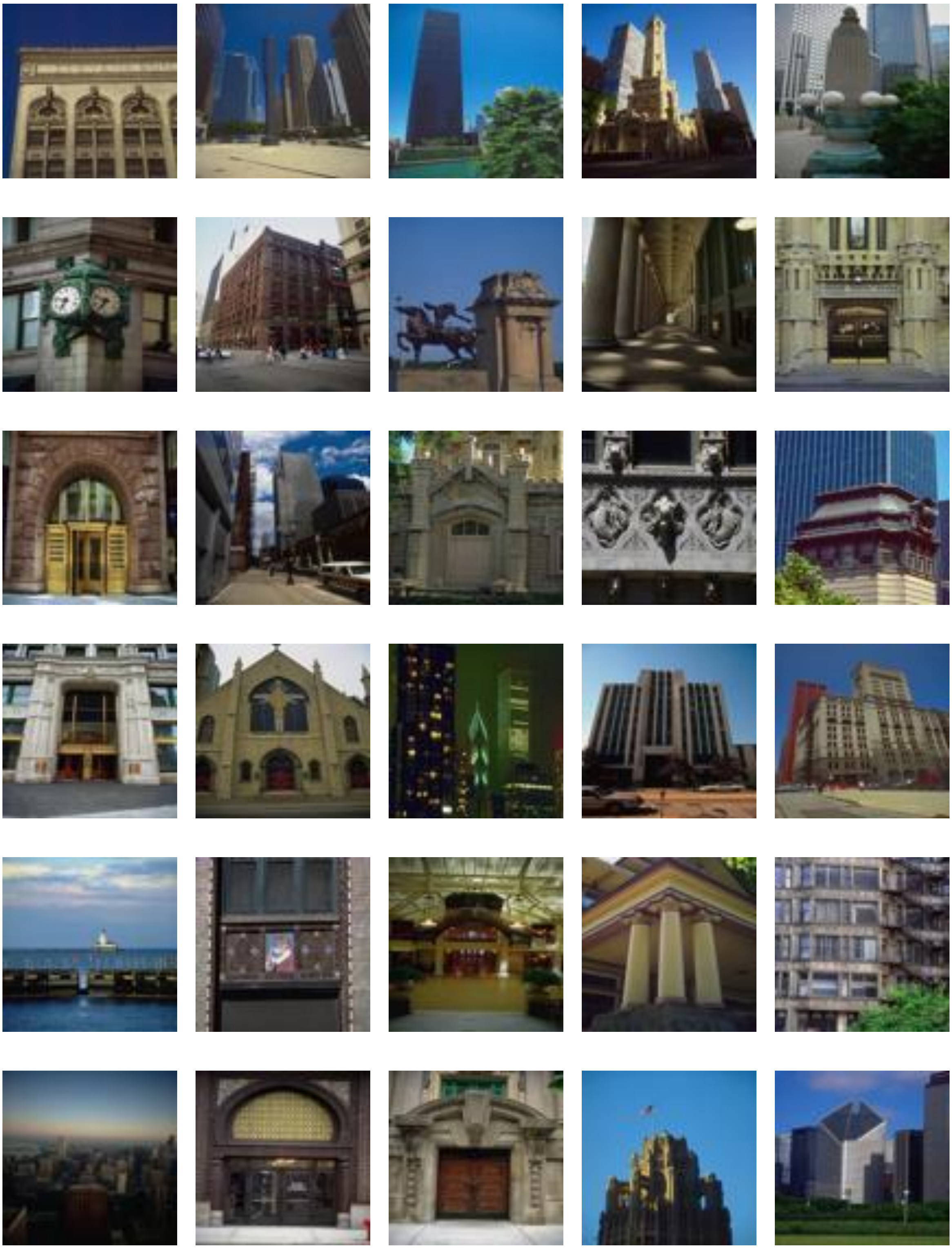


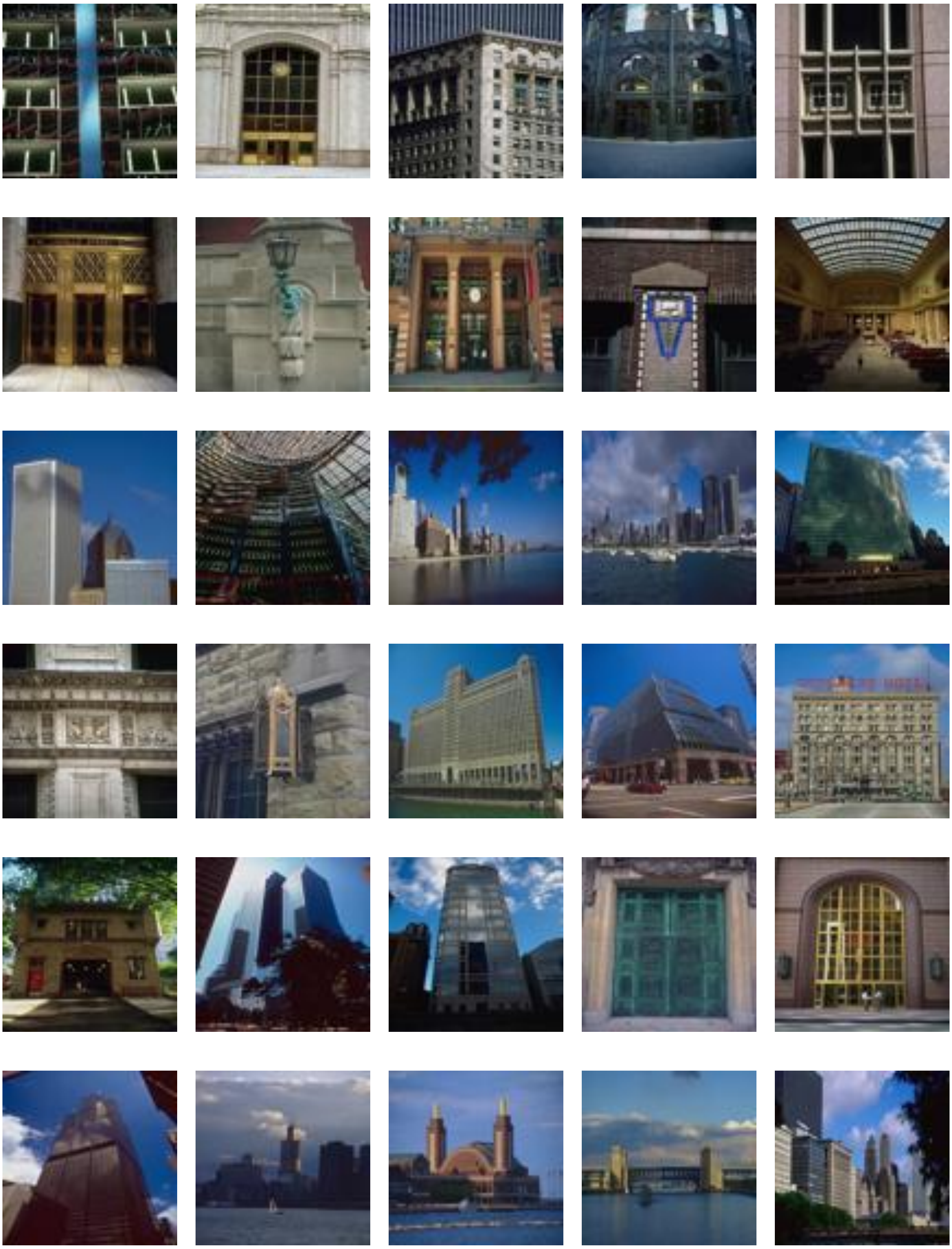
The following images are annotated as Vegetation:

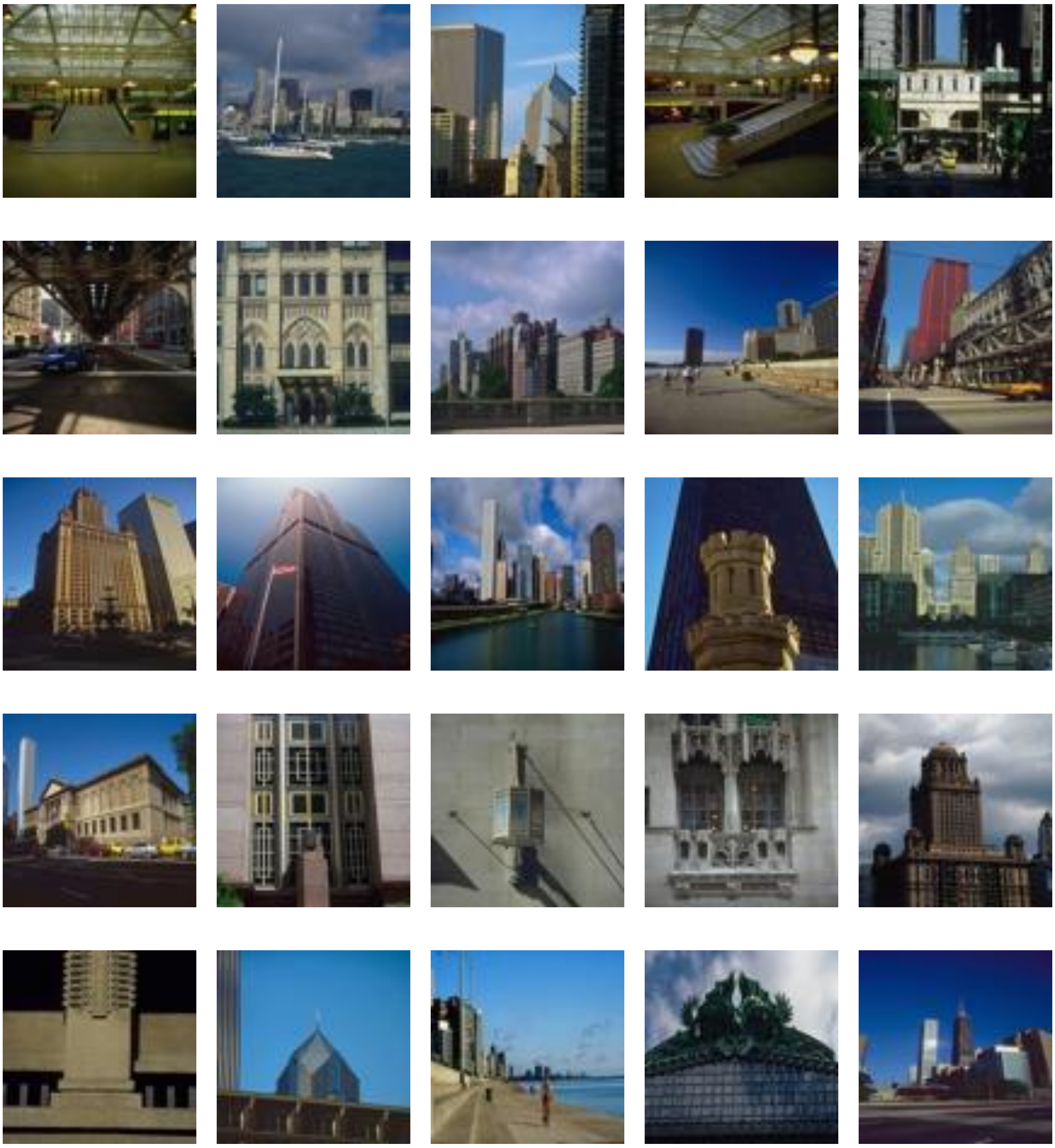


The following images are annotated as Building:









APPENDIX B Two-Way ANOVA RESULTS

The Following tables show the results of two-way ANOVA test based on the saliency, Main-concept Similarity, Background Similarity, and Target Concept factors when considering the effect of the users and the mutual effect of users and saliency on the gaze features in the AoI-FV and SP-FV; where Users-F, Users-P show the F statistic and p-value of the users' effect also U*G-F and U*G-P show the F statistic and p-value of the mutual effect of the users and the four factors respectively.

a. Saliency factor

Table 17: Two-way ANOVA results based on Saliency factor for Best Choice scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	10.6130	0.0000	1.1770	0.2816
iAdp	8.5245	0.0000	1.0757	0.3734
iAdr	27.1352	0.0000	1.3160	0.1825
iAv	9.9618	0.0000	1.6150	0.0617
iAvp	29.7073	0.0000	0.8637	0.6058
iAvr	11.9504	0.0000	0.9227	0.5377
iFt	6.9971	0.0000	0.9972	0.4546
iFtp	23.8166	0.0000	0.7997	0.6793
iFtr	14.4705	0.0000	0.7091	0.7778
iMx	7.7248	0.0000	1.1341	0.3185
iMxp	20.6802	0.0000	0.8680	0.6008
iMxr	21.0288	0.0000	1.2994	0.1927
iVn	32.6770	0.0000	2.3863	0.0019
iSn	24.4050	0.0000	1.2781	0.2066
tAngle	4.3982	0.0000	1.0865	0.3626
tPnPSpd	10.2287	0.0000	0.7420	0.7432
tPoSpd	9.9183	0.0000	0.7721	0.7104
tPrDist	32.6765	0.0000	0.8141	0.6630
tPrPT	3.8500	0.0000	1.9094	0.0180
tPrSpd	9.0877	0.0000	0.7673	0.7157
tTime	14.0726	0.0000	2.1252	0.0067
tPreGrad	0.0688	1.0000	1.7474	0.0360
tMnGrad	0.0291	1.0000	1.1235	0.3279
t2reg	36.4503	0.0000	3.3034	0.0000
tVn2reg	24.5665	0.0000	2.9234	0.0001
Lmn	6.5053	0.0000	1.3263	0.1769
Lmx	14.3569	0.0000	1.9611	0.0144

Table 18: Two-way ANOVA results based on Saliency factor for First 100 scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	40.9453	0.0000	2.6738	0.0005
iAdp	41.0921	0.0000	1.6811	0.0475
iAdr	24.3736	0.0000	2.3586	0.0022
iAv	18.0459	0.0000	3.0879	0.0000
iAvp	92.0943	0.0000	1.5555	0.0778
iAvr	16.4586	0.0000	1.0323	0.4174
iFt	14.8100	0.0000	1.7367	0.0378
iFtp	81.3891	0.0000	1.1997	0.2633
iFtr	18.7870	0.0000	1.9097	0.0181
iMx	28.6650	0.0000	2.8194	0.0002
iMxp	69.3050	0.0000	1.3381	0.1697
iMxr	20.5791	0.0000	2.1121	0.0072
iVn	29.3504	0.0000	1.6248	0.0594
iSn	26.6330	0.0000	1.1016	0.3486
tAngle	3.9750	0.0000	1.8079	0.0279
tPnPSpd	11.5419	0.0000	1.1977	0.2648
tPoSpd	9.5203	0.0000	0.7434	0.7417
tPrDist	21.9603	0.0000	1.4393	0.1193
tPrPT	2.5616	0.0008	0.9512	0.5054
tPrSpd	11.5273	0.0000	1.4752	0.1048
tTime	26.6556	0.0000	2.9677	0.0001
tPreGrad	0.0298	1.0000	1.5320	0.0848
tMnGrad	0.0205	1.0000	0.4523	0.9633
t2reg	24.1410	0.0000	0.6827	0.8041
tVn2reg	15.0808	0.0000	0.6690	0.8173
lmn	3.9339	0.0000	0.9224	0.5382
lmx	28.5430	0.0000	2.5605	0.0008

Table 19: Two-way ANOVA results based on Saliency factor for All-in-Page scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	21.7909	0.0000	2.2854	0.0032
iAdp	11.1556	0.0000	4.7656	0.0000
iAdr	14.2364	0.0000	2.8415	0.0002
iAv	29.9686	0.0000	2.5463	0.0009
iAvp	20.0566	0.0000	3.4758	0.0000
iAvr	15.3555	0.0000	1.6811	0.0475
iFt	20.2136	0.0000	2.4718	0.0013
iFtp	19.9386	0.0000	3.1633	0.0000
iFtr	13.6854	0.0000	1.9364	0.0160
iMx	31.7807	0.0000	2.6371	0.0005
iMxp	16.4028	0.0000	4.6996	0.0000
iMxr	14.5175	0.0000	2.0780	0.0084
iVn	29.7700	0.0000	5.9786	0.0000
iSn	41.5909	0.0000	5.5813	0.0000
tAngle	5.8449	0.0000	1.6578	0.0519
tPnPSpd	26.3400	0.0000	0.8964	0.5679
tPoSpd	24.6837	0.0000	0.7826	0.6986
tPrDist	46.3224	0.0000	1.7228	0.0398
tPrPT	1.7341	0.0380	1.2568	0.2207
tPrSpd	22.8476	0.0000	1.0144	0.4361
tTime	40.5732	0.0000	3.6664	0.0000
tPreGrad	0.0590	1.0000	1.1984	0.2640
tMnGrad	0.0341	1.0000	0.9375	0.5208
t2reg	51.9077	0.0000	2.6411	0.0005
tVn2reg	7.2715	0.0000	1.8841	0.0201
lmn	25.6589	0.0000	1.2484	0.2270
lmx	33.6895	0.0000	3.2118	0.0000

b. Main-concept Similarity factor

Table 20: Two-way ANOVA results based on Main-concept Similarity factor for Best Choice scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	10.9649	0.0000	19.9549	0.0000
iAdp	9.0418	0.0000	17.9013	0.0000
iAdr	29.4729	0.0000	28.5735	0.0000
iAv	10.8547	0.0000	20.8363	0.0000
iAvp	29.7429	0.0000	11.7737	0.0000
iAvr	13.0595	0.0000	13.5223	0.0000
iFt	7.1618	0.0000	13.2204	0.0000
iFtp	23.5099	0.0000	7.5983	0.0000
iFtr	16.0833	0.0000	12.7702	0.0000
iMx	8.2052	0.0000	21.1224	0.0000
iMxp	20.8455	0.0000	15.4865	0.0000
iMxr	22.7527	0.0000	26.4652	0.0000
iVn	32.7472	0.0000	5.2205	0.0000
iSn	27.3618	0.0000	1.2129	0.2531
tAngle	3.3499	0.0000	1.3345	0.1715
tPnPSpd	6.8084	0.0000	1.8353	0.0248
tPoSpd	6.4785	0.0000	1.9271	0.0166
tPrDist	31.1312	0.0000	1.7902	0.0301
tPrPT	2.8604	0.0002	1.0129	0.4377
tPrSpd	6.5134	0.0000	1.5862	0.0689
tTime	11.8351	0.0000	1.4370	0.1202
tPreGrad	0.2882	0.9965	1.1785	0.2802
tMnGrad	0.2359	0.9989	1.5232	0.0876
t2reg	24.7007	0.0000	2.0999	0.0076
tVn2reg	16.2793	0.0000	2.0863	0.0081
Lmn	6.5859	0.0000	2.6437	0.0005
Lmx	11.0946	0.0000	1.5514	0.0790

Table 21: Two-way ANOVA results based on Main-concept Similarity factor for First 100 scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	34.7322	0.0000	33.4726	0.0000
iAdp	43.2820	0.0000	20.6481	0.0000
iAdr	27.5364	0.0000	21.5459	0.0000
iAv	15.0866	0.0000	22.7208	0.0000
iAvp	92.6359	0.0000	12.1583	0.0000
iAvr	19.0940	0.0000	11.8603	0.0000
iFt	12.0830	0.0000	17.7203	0.0000
iFtp	82.8850	0.0000	10.2122	0.0000
iFtr	21.2750	0.0000	13.0096	0.0000
iMx	24.2984	0.0000	30.0668	0.0000
iMxp	72.0188	0.0000	15.8375	0.0000
iMxr	23.8124	0.0000	20.4128	0.0000
iVn	28.0593	0.0000	4.8684	0.0000
iSn	27.8884	0.0000	0.7045	0.7824
tAngle	3.8815	0.0000	1.8876	0.0198
tPnPSpd	7.5045	0.0000	0.7530	0.7313
tPoSpd	6.1460	0.0000	0.7701	0.7125
tPrDist	12.8175	0.0000	0.8034	0.6752
tPrPT	1.9273	0.0167	1.1436	0.3100
tPrSpd	7.8243	0.0000	0.8948	0.5697
tTime	23.7305	0.0000	2.6419	0.0005
tPreGrad	0.1966	0.9996	1.6036	0.0644
tMnGrad	0.3864	0.9829	1.9573	0.0146
t2reg	17.9781	0.0000	1.1808	0.2787
tVn2reg	11.0988	0.0000	0.9764	0.4774
lmn	4.3052	0.0000	2.3428	0.0025
lmx	26.4356	0.0000	2.5313	0.0010

Table 22: Two-way ANOVA results based on Main-concept Similarity factor for All-in-Page scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	22.8790	0.0000	20.8982	0.0000
iAdp	8.7258	0.0000	34.3795	0.0000
iAdr	18.4366	0.0000	26.2272	0.0000
iAv	32.9809	0.0000	20.0920	0.0000
iAvp	17.9083	0.0000	20.7488	0.0000
iAvr	18.1627	0.0000	10.3454	0.0000
iFt	21.9666	0.0000	15.3266	0.0000
iFtp	17.5801	0.0000	17.5203	0.0000
iFtr	16.4232	0.0000	13.1970	0.0000
iMx	34.9989	0.0000	25.7450	0.0000
iMxp	14.1991	0.0000	29.4821	0.0000
iMxr	18.2987	0.0000	25.6719	0.0000
iVn	30.2087	0.0000	3.5697	0.0000
iSn	47.2157	0.0000	1.6011	0.0653
tAngle	5.1404	0.0000	0.7932	0.6867
tPnPSpd	18.1923	0.0000	2.9390	0.0001
tPoSpd	16.5710	0.0000	2.8083	0.0002
tPrDist	36.7530	0.0000	0.9740	0.4799
tPrPT	1.3286	0.1749	1.6346	0.0570
tPrSpd	16.3331	0.0000	2.8715	0.0002
tTime	28.0580	0.0000	7.1401	0.0000
tPreGrad	0.5041	0.9401	2.6300	0.0006
tMnGrad	0.3996	0.9798	2.8325	0.0002
t2reg	37.1292	0.0000	0.9632	0.4920
tVn2reg	5.9821	0.0000	0.8260	0.6492
lmn	16.0871	0.0000	2.9997	0.0001
lmx	24.6323	0.0000	5.1088	0.0000

c. Background Similarity factor

Table 23: Two-way ANOVA results based on Background Similarity factor for Best Choice scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	9.8224	0.0000	34.2915	0.0000
iAdp	8.2310	0.0000	29.2703	0.0000
iAdr	29.4975	0.0000	34.6576	0.0000
iAv	9.8345	0.0000	27.7809	0.0000
iAvp	29.2808	0.0000	14.6473	0.0000
iAvr	12.6498	0.0000	11.3070	0.0000
iFt	6.4393	0.0000	17.0464	0.0000
iFtp	23.3822	0.0000	9.6911	0.0000
iFtr	15.5121	0.0000	13.2883	0.0000
iMx	7.0815	0.0000	33.7222	0.0000
iMxp	20.3386	0.0000	22.5576	0.0000
iMxr	22.6702	0.0000	32.9586	0.0000
iVn	32.4795	0.0000	11.1103	0.0000
iSn	27.5301	0.0000	1.5729	0.0728
tAngle	3.5245	0.0000	1.3491	0.1632
tPnPSpd	7.7202	0.0000	2.1293	0.0066
tPoSpd	7.3707	0.0000	2.0626	0.0090
tPrDist	32.1939	0.0000	2.0341	0.0103
tPrPT	3.0081	0.0001	0.9536	0.5027
tPrSpd	7.4287	0.0000	2.1503	0.0060
tTime	11.7556	0.0000	1.7699	0.0328
tPreGrad	0.2474	0.9985	1.6291	0.0582
tMnGrad	0.1526	0.9999	1.9339	0.0161
t2reg	29.3904	0.0000	2.6466	0.0005
tVn2reg	19.0408	0.0000	2.1378	0.0064
lmn	6.5795	0.0000	2.4394	0.0015
lmx	11.5291	0.0000	1.8908	0.0196

Table 24: Two-way ANOVA results based on Background Similarity factor for First 100 scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	37.1196	0.0000	32.0211	0.0000
iAdp	37.0204	0.0000	27.4532	0.0000
iAdr	27.6666	0.0000	25.5759	0.0000
iAv	16.0693	0.0000	25.2965	0.0000
iAvp	81.5452	0.0000	20.5532	0.0000
iAvr	19.4838	0.0000	17.0358	0.0000
iFt	12.7649	0.0000	21.5080	0.0000
iFtp	74.1513	0.0000	17.0945	0.0000
iFtr	21.0524	0.0000	16.5031	0.0000
iMx	26.1699	0.0000	30.4368	0.0000
iMxp	63.1105	0.0000	23.6473	0.0000
iMxr	23.8460	0.0000	25.1420	0.0000
iVn	28.3077	0.0000	4.9956	0.0000
iSn	28.2220	0.0000	1.3342	0.1722
tAngle	4.5459	0.0000	2.0506	0.0095
tPnPSpd	8.4494	0.0000	0.7235	0.7629
tPoSpd	6.9569	0.0000	0.8468	0.6253
tPrDist	14.9611	0.0000	0.5208	0.9310
tPrPT	2.1923	0.0049	1.2394	0.2332
tPrSpd	8.7985	0.0000	0.7104	0.7765
tTime	23.8191	0.0000	4.7724	0.0000
tPreGrad	0.1128	1.0000	1.8190	0.0266
tMnGrad	0.3200	0.9937	2.5236	0.0010
t2reg	21.4252	0.0000	1.4627	0.1101
tVn2reg	13.1814	0.0000	1.2438	0.2303
lmn	3.9820	0.0000	3.0568	0.0001
lmx	27.0134	0.0000	4.5874	0.0000

Table 25: Two-way ANOVA results based on Background Similarity factor for All-in-Page scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	23.2273	0.0000	28.2961	0.0000
iAdp	8.3867	0.0000	47.1720	0.0000
iAdr	18.4940	0.0000	29.3168	0.0000
iAv	32.8610	0.0000	24.6365	0.0000
iAvp	17.5751	0.0000	25.9695	0.0000
iAvr	17.9339	0.0000	10.8590	0.0000
iFt	21.8957	0.0000	19.1210	0.0000
iFtp	17.0772	0.0000	23.8868	0.0000
iFtr	16.6005	0.0000	17.2878	0.0000
iMx	35.2330	0.0000	33.2924	0.0000
iMxp	13.7114	0.0000	39.0152	0.0000
iMxr	18.3406	0.0000	31.1186	0.0000
iVn	30.4998	0.0000	5.6178	0.0000
iSn	47.2009	0.0000	1.2837	0.2030
tAngle	5.2608	0.0000	0.6850	0.8020
tPnPSpd	21.2940	0.0000	3.4344	0.0000
tPoSpd	19.6858	0.0000	3.2927	0.0000
tPrDist	41.2928	0.0000	1.5516	0.0786
tPrPT	1.5497	0.0792	1.6622	0.0510
tPrSpd	18.7598	0.0000	3.4021	0.0000
tTime	33.3558	0.0000	9.0658	0.0000
tPreGrad	0.4395	0.9680	3.6640	0.0000
tMnGrad	0.3451	0.9905	3.6427	0.0000
t2reg	40.9368	0.0000	1.0589	0.3900
tVn2reg	6.3792	0.0000	0.9889	0.4636
lmn	18.9408	0.0000	3.2036	0.0000
lmx	28.4954	0.0000	7.0356	0.0000

d. Target Concept factor

Table 26: Two-way ANOVA results based on Target Concept factor for Best Choice scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	37.7998	0.0000	27.4838	0.0000
iAdp	8.8098	0.0000	5.6862	0.0000
iAdr	59.2872	0.0000	13.5936	0.0000
iAv	22.4834	0.0000	14.4488	0.0000
iAvp	38.6757	0.0000	10.8810	0.0000
iAvr	18.0271	0.0000	5.2029	0.0000
iFt	15.7111	0.0000	10.7666	0.0000
iFtp	30.0236	0.0000	8.8938	0.0000
iFtr	22.0048	0.0000	5.2365	0.0000
iMx	25.6329	0.0000	18.7609	0.0000
iMxp	27.5604	0.0000	9.4212	0.0000
iMxr	47.8687	0.0000	12.6182	0.0000
iVn	44.0564	0.0000	7.6211	0.0000
iSn	17.6529	0.0000	1.2265	0.2428
tAngle	4.1315	0.0000	2.8490	0.0002
tPnPSpd	10.8640	0.0000	4.4449	0.0000
tPoSpd	10.0366	0.0000	4.1254	0.0000
tPrDist	32.4849	0.0000	1.6048	0.0640
tPrPT	3.8491	0.0000	5.1816	0.0000
tPrSpd	9.8987	0.0000	4.4943	0.0000
tTime	22.5374	0.0000	14.3677	0.0000
tPreGrad	0.8204	0.6558	10.8983	0.0000
tMnGrad	0.0635	1.0000	0.7969	0.6825
t2reg	31.4350	0.0000	4.2303	0.0000
tVn2reg	20.2072	0.0000	3.6071	0.0000
lmn	6.8074	0.0000	3.0487	0.0001
lmx	13.3866	0.0000	7.3547	0.0000

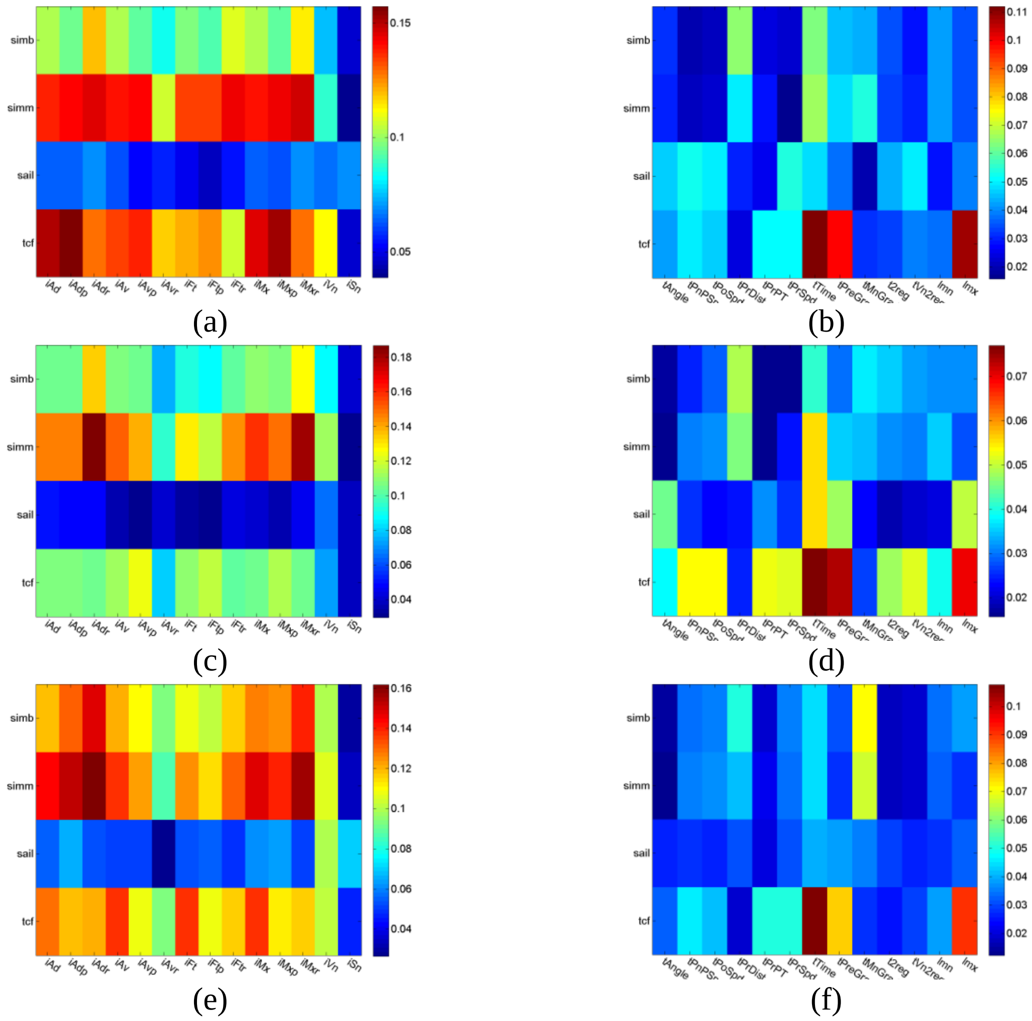
Table 27: Two-way ANOVA results based on Target Concept factor for First 100 scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	117.8981	0.0000	76.4125	0.0000
iAdp	48.4807	0.0000	21.2104	0.0000
iAdr	52.2400	0.0000	15.0180	0.0000
iAv	44.2406	0.0000	31.3608	0.0000
iAvp	101.8925	0.0000	13.8501	0.0000
iAvr	30.9063	0.0000	8.3619	0.0000
iFt	37.2097	0.0000	27.3382	0.0000
iFtp	88.1540	0.0000	10.4249	0.0000
iFtr	33.5661	0.0000	8.6648	0.0000
iMx	82.4444	0.0000	56.2098	0.0000
iMxp	77.3680	0.0000	13.8558	0.0000
iMxr	46.6664	0.0000	14.8987	0.0000
iVn	35.8460	0.0000	7.4025	0.0000
iSn	20.3805	0.0000	0.7126	0.7741
tAngle	2.9674	0.0001	2.4834	0.0012
tPnPSpd	9.0026	0.0000	1.7276	0.0391
tPoSpd	7.3841	0.0000	1.7263	0.0394
tPrDist	17.3491	0.0000	0.9014	0.5621
tPrPT	2.5282	0.0009	3.2261	0.0000
tPrSpd	9.1642	0.0000	1.8428	0.0241
tTime	38.6352	0.0000	24.1447	0.0000
tPreGrad	1.9805	0.0131	16.5473	0.0000
tMnGrad	0.1462	0.9999	0.6224	0.8592
t2reg	22.1405	0.0000	4.3578	0.0000
tVn2reg	13.9818	0.0000	2.3544	0.0023
lmn	5.0827	0.0000	2.9833	0.0001
lmx	25.5112	0.0000	12.2564	0.0000

Table 28: Two-way ANOVA results based on Target Concept factor for All-in-Page scenario

Features	Users-F	Users-P	U*G-F	U*G-P
iAd	38.9222	0.0000	24.4970	0.0000
iAdp	19.8877	0.0000	22.2794	0.0000
iAdr	25.5524	0.0000	12.4384	0.0000
iAv	37.7222	0.0000	20.2511	0.0000
iAvp	25.7332	0.0000	12.2865	0.0000
iAvr	18.2219	0.0000	5.6962	0.0000
iFt	27.5627	0.0000	16.2793	0.0000
iFtp	25.4073	0.0000	11.3207	0.0000
iFtr	19.0140	0.0000	7.1136	0.0000
iMx	52.8677	0.0000	28.8987	0.0000
iMxp	25.2766	0.0000	18.6945	0.0000
iMxr	29.9028	0.0000	13.8664	0.0000
iVn	28.2859	0.0000	3.9288	0.0000
iSn	28.6584	0.0000	0.2954	0.9959
tAngle	4.9006	0.0000	2.1058	0.0074
tPnPSpd	19.6258	0.0000	3.7817	0.0000
tPoSpd	18.4027	0.0000	3.4191	0.0000
tPrDist	38.0530	0.0000	0.6566	0.8291
tPrPT	3.4807	0.0000	4.3175	0.0000
tPrSpd	17.0601	0.0000	3.9071	0.0000
tTime	51.2028	0.0000	20.0053	0.0000
tPreGrad	2.5295	0.0009	11.1992	0.0000
tMnGrad	0.2114	0.9994	1.2037	0.2599
t2reg	46.3631	0.0000	2.7724	0.0003
tVn2reg	7.5163	0.0000	2.9702	0.0001
lmn	13.6702	0.0000	1.7089	0.0425
lmx	36.3886	0.0000	11.2262	0.0000

APPENDIX C Standard Deviation of Correlation Coefficient



(a) (b)

(c) (d)

(e) (f)

Figure 24: Standard deviation of the Correlation Coefficient between the features and Background Similarity (simb), Main-concept Similarity (simm), Saliency Score (sail) and Target Concept (tcf) vectors for all Users. The colours show the magnitude of the standard deviation of the correlation coefficient a) First 100 scenario, AoI-FV b) First 100 scenario, SP-FV c) Best Choice scenario, AoI-FV d) Best Choice scenario, SP-FV e) All-in-Page scenario, AoI-FV f) All-in-Page scenario, SP-FV