FINDING INTERESTING PHOTOS IN ALBUMS USING VISUAL ATTENTION

KARTHIKEYAN VAIAPURY

NATIONAL UNIVERSITY OF SINGAPORE

2007

FINDING INTERESTING PHOTOS IN ALBUMS USING VISUAL ATTENTION

KARTHIKEYAN VAIAPURY

B.Tech.(Information Technology), Bharathidasan University, India

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2007

Acknowledgements

I would like to express my sincere and heartful thanks to my advisor Prof. Mohan S Kankanhalli of the Department of Computer Science at School of Computing, National University of Singapore, for his immense advice, enduring patience, kind support and cordial encouragement.

I thank almighty for his divine guidance, grace and strength to finish this research work on time.

I would like to render my sincere thanks to School of Computing for giving me excellent opportunity to pursue higher studies at NUS.

I would also like to thank my seniors Pradeep, Vivek, Shiva and William for their constant support.

I am highly indebtful to my parents, sisters and friends for their support and affection.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and background to e-Chronicles

A chronicle is an extended account, in prose or verse, of historical events, sometimes including legendary material, presented in a chronological order and without authorial interpretation or comment. Due to advances in sensor processing and storage technologies, there has been a transition from human-reported alphanumeric records to electronic records. e-Chronicles are the chronicles which are mediated either in the form of videos or photos and sometimes even documents produced by the authors or others. The term "e-Chronicles" refers to the collection of all significant media events digitally recorded in various phases during the lifetime of organizations or individuals or groups. e-Chronicles can be broadly categorized into *personal* e-Chronicles which are the mediated collections of personal lives and *organizational* e-Chronicles which are the mediated archival records of institutions and corporations. Due to people's interest in remembering or gaining from the past experiences and memories, personal e-Chronicles play a significant role in our life as well as in many aspects of society. Personal e-Chronicles,

also known as family e-Chronicles are the media collections concerning family members capturing their life and events such as wedding, birthday, convocation *etc.* The reasons behind the sustained dynamic growth of family media collections include-

1. Availability of technology that is needed for capturing and storing the experiences,
2. Affordable cost of cheaper digital cameras,
3. Minimal photographic or videographic skills needed by users.

In general, as stated by Kim *et al.* an e-Chronicle system includes following aspects [32]

1. Recording data using multiple sensors,
2. Supporting rich tags for access and presentation of appropriate information,
3. Providing access to data at multiple levels of granularity and abstraction.

In e-Chronicles, recording data using multiple sensors is an ongoing process in most families. Indeed, researchers also have shown keen interest in recording an individual's whole life and then mining the important events [12, 14, 13]. Thus the family e-Chronicle needs to include support for rich tags for access, categorization of media collections into meaningful groupings and providing access to these groupings via web at multiple levels of granularity and abstractions using appropriate mechanisms. The technical issues in family e-Chronicles include -

1. *Media capture and storage:* With digital cameras, one can capture media of good quality (e.g. 10 megapixels resolution) which are subsequently stored in its onboard memory or memory card reader and then downloaded to a digital media album that possibly resides on a personal computer. Once transferred to a computer, the images may be viewed, processed, or exported to multiple image file formats. In olden days, using analog cameras, photos were captured with film rolls and then processed in a batch. The advantages of media captured with digital cameras over analog cameras include better resolution and easy replication of media. With the media captured by analog camera, the following difficulties are incurred.

    - it does not support easy making of multiple copies of media

    - it does not have the ability to provide instant feedback of capture quality

2. *Media annotation and representation:* The stored digital images need to be annotated which is the extension of interpretation. Annotation is a descriptive form of metadata that assists users in the reuse and composition of media. It helps to identify structured information of the resource and makes data more manageable to identify and explore the available resources. Metadata makes media archives more accessible and facilitates flexible searching.

3. *Media querying:* Posing the search constraints that could help in efficient searching is a difficult problem for large collections.

4. *Media retrieval:* Media retrieval is the process that enables users to access media resources. Content-based retrieval is an entire area of

study devoted to this problem. How to engineer a system which allows multiple users to access from the same collection is challenging in terms of efficiency.

5. *Media presentation:* Customizing the presentation through intuitive interfaces thereby facilitating access of *appropriate information* to all types of users. The key idea is to minimize information overload and present only relevant information.

6. *Media sharing:* Sharing of media to multiple users across different locations via email, ftp etc is considered a necessary feature nowadays.

7. *Media security:* Providing media at different access levels for different people at multiple levels of granularities is necessary. This is because different people have different trust levels established with the owner of the echronicles.

As Gray [15] has elucidated, we are on the verge of realizing Bush, Babbage and Turing visions to develop a system that automatically organizes indexes, digests, evaluates and summarizes information. It has been argued that organization is a basic human need - "Even if improved search means we can always find the information we need, we may continue to organize it for other reasons including to support serendipitous browsing and provide the satisfaction of putting our things in order" [63].

Indeed, when there is large pile of photos, people prefer to see images that seem interesting to them than to tediously search and retrieve images based on content based retrieval. For example, when people remember collections, they look into entire temporal cluster of images, say a marriage event at some time period, and at a step one level further, they prefer to see "*interesting*"

images in that folder. However, the term interesting mentioned above varies from person to person due to their individual interests, context, experiences and preferences.

1.2 Interestingness

Interestingness is the power of attracting or holding one's attention. Based on one's intention, one will pay attention. Intention is an objective or goal a person is willing to accomplish. Usually, attention is driven by intention which in turn is driven by interestingness. Assuming that *intention is related to interestingness*, we define interestingness as an entity that arises from

1. interpretation and experience,
2. surprise,
3. beauty,
4. aesthetics and
5. desirability.

The items 2,3,4 and 5 are based on how one interprets and his/her accumulation of experience as embodied in the human cognition system. It can be noticed that not all types of attention are associated with interestingness (refer figure 1.1).

As stated in [28], attention can be categorized into following six types:

- *Sensorial Attention:* It refers to the attention towards objects that makes reasonable sense to a person.
- *Intellectual Attention:* The attention towards represented objects that is known to a person is called as Intellectual attention.

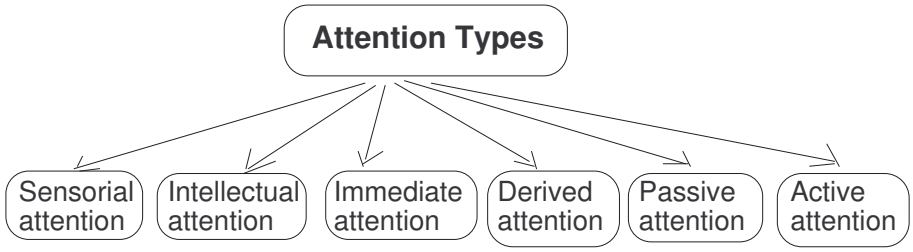

Figure 1.1: *Types of attention (Source: James [28])*

- *Immediate Attention:* It is the attention that is drawn within short time period.
- *Derived Attention:* When the topic or stimulus is interesting in itself, its interest to associate with some other immediately interesting thing is known as derived attention.
- *Passive attention:* It is the attention that arises from non-voluntary, reflex, effortless action.
- *Active attention:* The attention that arises from voluntary action is called as active attention. Intention is central to concept of voluntary action [35].

In fact, this work is inspired by the fact that conscious control of actions would be possible if attention to intention is combined as a single mechanism [35]. Attention is the process of selectively concentrating on certain parts of the environment exclusive to others. Intention is the underlying specific purpose that may lead to an action in order to accomplish the goal. In fact, one way of accomplishing intention could be probably attending towards objects of relevant type. The relationship between interestingness and attention can be seen in figure 1.2. As one can see from figure 1.2, there exists relationship

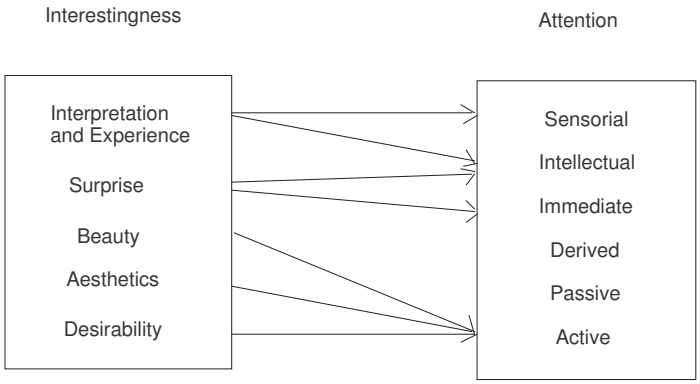

Figure 1.2: *Relationship between Interestingness and Attention*

between desirability, aesthetics, beauty and active attention since attention is paid voluntarily. Similarly, since surprise has the ability to make us attend immediately, it is related to immediate attention. Also, since sensorial and intellectual attention arise from one's interpretation and thinking that depends on one's background and experience, interpretation and experience is related to sensorial and intellectual attention.

Since interesting objects are often attended by human beings [28], now we would provide a brief introduction to the phenomenon of attention in section 1.2.1.

1.2.1 Attention

Attention is the cognitive process in which a person concentrates on some features of the environment to the relative exclusion of the others. It can also be explained as the neurobiological conception (*i.e* concentration of mental powers) upon an object by close or careful observing or listening [39, 40]. According to [28], *"it is viewed as the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible ob-*

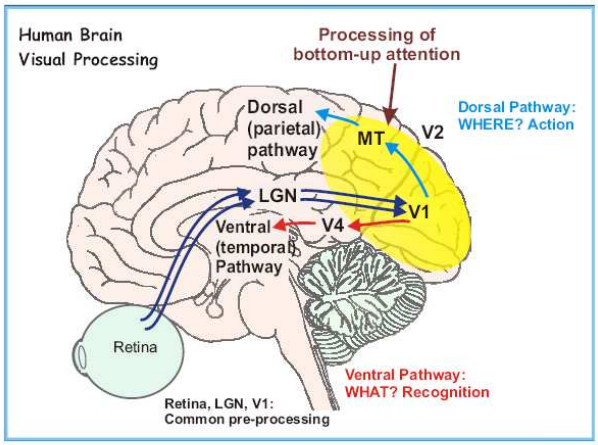

Figure 1.3: *Human Brain (Source: Edgington et al. [9])*

jects or trains of thought...It implies withdrawal from some things in order to deal effectively with others". The small part of incoming visual information reaches short term memory and visual awareness providing the ability to investigate closely [30]. It has survival value to keep an eye on every thing that is happening even if accuracy is lost. This trade-off is a part of the phenomenon of attention [61]. Attention in images is based on the fact that real images often contain vast areas of insignificant data from the perspective of cognition. Hence, if we can attend to the relevant parts, the image can be interpreted more quickly using less resources. The process by which people attend objects based on their own interest is called attention. In general, ***what*** object is attended to and ***where*** the attention is likely to be is controlled by the ventral and dorsal pathway respectively [23]. The dorsal and ventral pathway of the human brain depicting its role in attention processing is shown in figure 1.3.

Essentially, to find what object is attended to and where the attention likely to be, filtering and prioritizing the information is vital. This is analagous to the nature to our *human fovea*, which acts corresponding to a particular stimulus [23]. The stimulus mentioned above refers to the action that accelerates a physiological or psychological activity or response.

In general, the two most commonly used computational models of attention are: -

- **Bottom-up attention:** It is based on the combinations of low level features which include both oriented as well as non-oriented features such as colour, contrast and orientation.

- **Top-down attention:** It involves task dependent processing which generally requires some prior knowledge about the scene. In effect, the user attention is guided by what he sees. These two models are explained with specific examples in section 2.1.1 and section 2.1.2 respectively.

Though interpretation varies from person to person (as discussed earlier), a *reasonable generic interpretation* can be arrived at by studying how images are perceived or interpreted in human brain. It can be used to calculate the saliency regions from which most attended regions can be found out. *Attended regions* are the regions (parts) of an image which attract human attention. The attended region needs to be determined on the basis of either surprise/interestingness factor or the task at hand (refer section 2.2). If we know the attended region before hand, then we can adopt a top-down approach.

It is also a challenging problem to build visually appealing interfaces

which not only enable browsing and searching, but also give better information for the purpose of visualization, authoring, story telling and annotation [7] based on user attention/interestingness.

1.3 Motivation

As collections of images are growing even larger, tools are needed to efficiently manage, organize, and navigate through them. Recently, there has been some pioneering research in showing "interesting" images. For *example*, Yahoo's Flickr [74] addresses this problem based on social network analysis. They make use of available web-based information as follows:

1. the number of times image has been viewed for the last 7 days and
2. the number of times it has been marked as favourite *etc.*

The main motivation behind our research is to find "*interesting*" images based on the image content rather than just social network analysis. The related CBIR (Content Based Information Retrieval) on one hand, in the top down perspective are based on low level features. In other words, their focus is efficiently utilizing the low-level visual content information. On the other hand, in the bottom up perspective, visual attention of images is computed by a saliency based approach. The saliency based approach is based on the fact that human brain selectively attends to the information available to human eyes due to its limited capacity to process all the data perceived up by 125 million photoreceptors in each eye. This is also evident from cocktail effect. The cocktail effect is the process by which human brain selectively attends to a person's talk in a crowd of conversations.

Also, neuroscientists and psychologists have determined that attention is driven by intention. This is based on the fact that interestingness varies

from person to person in terms of age, gender, cultural background, context, interpretation and experience. Hence it would be better to find the user's need by directly asking the user itself *i.e* top-down approach. Since there is a gap between low level visual feature information and high level semantic information, it is still challenging to build a system which attempts to capture ***user interest*** that is *dynamically evolving over time*. This reveals that when only the bottom-up approach is used, the real intent of user interest might be lost since only relative saliency is considered (refer to section 2.1). When only top-down approach is considered, information from the human cognition system might be lost. These issues establish the need for a framework which potentially utilizes both bottom-up and top-down methodology. This work is targeted towards finding interesting images based on user interest and attention models. Thus, there is a strong case for building a system that:

1. needs to be flexible enough to adaptively learn the environment with respect to the context

2. utilizes attention information in order to capture user's interest.

An important aspect in visual attention is the computation of *attended regions* in an image [23]. As said earlier, attended regions are the regions (parts) of an image which attract human attention. To define what are all the attended regions utilising both bottom up and top down approach is a research issue nowadays. For example, consider the problem of displaying the most attended regions in a particular mobile device display [6]. Here, the key issue is showing the most attended regions in a higher resolution on the provided display area. As said earlier, the real intent of user interest should not be lost. The past experience include the images that are viewed

by user. Thus, the motivation is to show images that seem interesting to user based on both bottom up approach (visual cognition system) and top down approach (goal-oriented) while considering *past experiences* as well. Also, the result of this set varies dynamically along with time and user's interest.

1.4 Problem Statement and Scope

We assume that a user has an e-Chronicle system which has a collection of photos. Our problem is to find "***interesting***" photos in the system using the visual attention model such that user interest (top-down) and visual attention models (bottom-up) are considered. The **user interest** can be known by asking the user himself/herself what seems interesting to user via a relevance feedback mechanism. The **visual attention models** are chosen based on the objective function which in our case is finding saliency/attended regions ***according to a specific goal***. The goal is to find images that seem interesting to user. Though attention models are applicable to multimedia types as video, image and audio, our scope is limited to visual image attention models in this research work. Our problem is to form a framework to ***identify*** the common attention samples that *evolve dynamically* based on user interest.

The issues are:

1. how to define interestingness for images
2. how to get the common attentive features of attended regions
3. how to capture interestingness that changes dynamically
4. how to combine the top down and bottom up approaches

We provide a ***novel framework*** that integrates both top down and bottom up approach while **considering interestingness** via relevance feedback. The key functionality of the system is its ability to adapt by learning from the past experience while preserving user interestingness depending on the context of how user select images in the environment. The key issue is choosing the attention models such that interestingness attributes are satisfied. In a summary, this is done by:

1. Feature extraction from each of the selected images by attention model,
2. The term weights are revised accordingly based on the relevance feedback provided by the user and then displayed images.

1.5 Overview of thesis

This thesis is organized as follows:
In chapter 1, we provided a brief introduction to echronicles and *interestingness* of images with respect to attention. We also provided a broader view of attention and its categories in section 1.2. We also explained the motivation behind this research work in section 1.3. The problem statement and scope of this research work has been stated in section 1.4.
In chapter 2, we provide the literature survey made on four key areas related to our work such as visual attention models, interestingness, relevance feedback and non identical duplicate detection. We provide discussion at the end of survey on each of the aformentioned key areas. Finally, we give a general discussion of the inferences made through this study. The state of the art made on attention under three sub categories such as bottom-up model, top-down model and hybrid model is presented in section 2.1. The various applications of attention models have been summarized in section

2.1.4 followed by a discussion. The survey made on interestingness is presented in section 2.2. We give a brief introduction to Itti-Koch saliency method and Bayesian theory in section 2.6 and section 2.7. We present the state of the art made on relevance feedback models in section 2.3.

In chapter 3, we explain the framework - attention based interestingness. The overview of our framework is summarized in section 3.1. We explain saliency extraction process, non identical duplicate detection and relevance feedback in section 3.2, section 3.3 and section 3.4 respectively.

The experimental details and evaluation results (subjective study and analysis) are discussed in chapter 4. In section 4.1, we outline the software structure and implementation platform followed by a brief description about system interface in section 4.2. The illustration of calculated saliency attention values is given in section 4.3. In section 4.4, we report the performance speed of various methods involved in the system. In section 4.5, we present the user study results as a system evaluation process. In section 4.6, we explain how SIFT aids in detecting non identical duplicates with sample image pairs. The future work and conclusions are given in chapter 5. In section 5.1, we give a summary of the work. Recommendations for future work are suggested in section 5.2.

Chapter 2

Related Works

This thesis describes the use of RFB (Relevance FeedBack) based visual attention model to show the user-specific interesting photos in a multimedia echronicle. The key areas, that were reviewed before the start of this work *include:*

1. **Visual Attention models**- To get a comprehensive knowledge about the existing visual attention models.

2. **Interestingness**- After getting adequate knowledge from the existing visual attention models, we studied interestingness which is primarily based on Bayesian Surprise theory [21].

3. **Relevance Feedback**- We studied relevance feedback from the perspective of information retrieval which is an efficient control mechanism to elicit users interest in order to customize the set of "interesting" photos.

4. **Non Identical Duplicate Detection**- We intend to increase the surprisingness by removing non identical duplicates while retaining the intention of user.

2.1 Visual Attention models

Attention is typically based on two major facts:

- Human beings do not perceive all things as equally important,
- Some objects have "pop - out" effect from the environment.

The exact location of the attentional bottleneck is an issue due to limited capacity information processing capability of human brain [55].

Navalpakkam *et al.* have stated that the number of objects attended in a human brain varies as follows [46]:

- only one spatio-temporal structure can be represented at a time (according to coherence theory),
- three or four objects in visual short term memory,
- many number of attended objects in visual short term and long term memory if attended objects are previously attended before.

Generally, the interesting part in an image is referred to as ROI (Region of Interest). It can be determined either by using its low level feature information or saliency information according to human cognition system. Saliency are of following two types [38]:

1. ***self saliency***: It refers to what determines how conspicuous a region on its own with respect to color, saturation, brightness and size,
2. ***relative saliency*:** It refers to how distinctive the region appears when there are regions of competing distinctiveness in the neighbour-

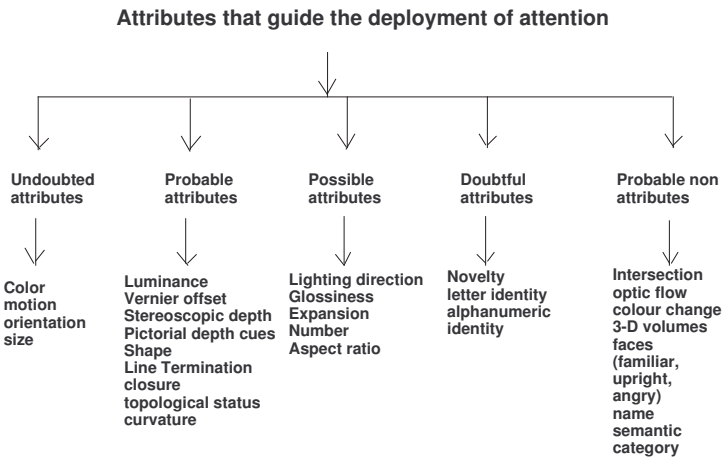

Figure 2.1: *Attention Guidance Attributes (Source: Wolfe et al. [70])*

hood. In other words, how salient a region is relative to its surrounding.

The information obtained from low level features of an image and cognition based saliency map represents self saliency and relative saliency respectively.

Wolfe *et al.* [70] have summarized the attributes that could guide attention as shown in figure 2.1.

Now, we present existing attention models under three main categories namely, a) bottom-up , b) top-down and c) hybrid (bottom-up + top-down) & undertake a detailed discussion about the most commonly used model known as the Itti-Koch attention model.

2.1.1 Bottom-Up model

The ***saliency map*** is constructed through a bottom-up approach which is based on the combinations of the low level features (which include both oriented as well as non-oriented) such as colour, contrast and orientation of

the image itself. It can then be used in many applications such as detecting surprising events/irregularities in video [2, 21]. The term "saliency" refers to the process of laying emphasis on the areas of images that attract high visual attention. Saliency map exhibits the following properties:

1. It topographically encodes for conspicuity (or "saliency") at every location in the visual input.

2. It predicts subjective human performance on a number of psychophysical tasks.

It is also found by Berg *et al.* that humans and monkeys follow same type of bottom-up attentional mechanism [1]. We note that the computation models such as [2, 40, 68] are based on this purely bottom-up approach methodology. While the saliency map of Ma *et al.* is based on the contrast alone, all the other models rely on multiple features such as Contrast, Colour and Orientation etc [41].

By combining multiple image features into a single topographical saliency map, the attended locations are found. The authors define attention at three levels namely, attended view, attended areas and attended points. The attended area is compared and correlated with the early selection of the human perception. The location of the attended area is found using fuzzy growing algorithm based on the assumption that the attention is usually directed towards the center of the image. The static saliency attention model proposed by Ma *et al.* is based on the number of attended regions and their position, size and brightness in saliency map [40]. The brightness information of the luminance component has significant impact on the image than that of the other two color components since eye has fairly little color sensitivity. This is one of the reasons it is perceived as attention model.

More details about the model can be found in section 3.2.1.

The authors further proposed an attention model which is based on the intuition that humans tend to pay more attention to the region near to the center of frame. A normalized Gaussian template is used to assign a weight to the position of the saliency regions. Ma *et al.* [39] have also proposed a more generic user attention model which covers static, motion, face, camera and linguistic attention models. Their approach is based on how viewers attentions are attracted by motion, object, audio and language while viewing a video program. Since they consider multifarious streams, a linear combination has been adopted as the fusion scheme.

However, in this research work, we limit our scope to visual attention models only. Chen *et al.* have adopted a similiar visual attention model for adapting images on small displays [6]. The MPEG7 attention model has been proposed by Wolf *et al.* that can be used for ranking images [69]. Once the salient region is found, global interest value can be calculated that can be used to organize image collections and to prioritize data for further processing. The existing bottom-up based attention models are generic since they rely on the saliency map ***without a specified goal***. The disadvantages of the bottom-up models include the following:

1. They use the method first and then exploit the solution. This means that the approach may not be well suited for specific goal oriented tasks.

2. Implicitly or explicitly, these models tend to adopt the ***low level human attention phenomenon*** without taking the semantic aspects into account.

2.1.2 Top-Down Model

Top-down attention, also called task dependent processing model generally requires some prior knowledge about the scene, for *example*, detecting and classifying the animals in the underwater video [9]. We note that only [18] has adopted this approach purely. In fact, there are research works which fall under both top-down as well as bottom-up categories such as [9, 11, 10, 20, 29, 45]. However, we will briefly look into the work of Navalpakkam *et al* [45] in the perspective of top-down approach as an example.

Navalpakkam *et al.* have aimed for a goal oriented attention guidance model [45]. Their approach is based on the task dependence graph such as large and small objects in which one of the aims is to prune the search area. For example, if a person is searching for a pen then it seems intuitively reasonable to look for a category of small objects rather than large objects. In fact, Navalpakkam *et al.* state that there is lot of evidence that our human brain may adopt a need-based approach. A need-based approach is one where only desired objects are quickly detected in the scene, identified and represented [46]. Given that user needs to find the interesting images and the need based approach is adopted by human brain, we note that top-down approach also plays vital role along with bottom-up approach.

The shared attention model which is called so because of group has been proposed by Matthew *et al.* for gaze imitation [18]. The gaze imitation is *for example*, infants as young as one year of age can follow the gaze of an adult to determine the object the adult is focusing on. Their methodology consists of finding the gaze vectors with bottom-up saliency maps of visual scenes to produce estimates (maximum a posteriori) of objects being looked at by an observed instructor. This is used for meeting indexing only and typically applicable where a group of people gazing at a particular object.

Table 2.1: *A comparative table depicting the state of the art*

Reference	Attention Stream	Methodology	Application
Ma *et al.* [40]	V_s,M_d,A,C,L	B_U	Video summarization
Edgington *et al.* [9, 11, 10, 65]	V_s,M_d	B_U,T_D	Detection, classification in underwater video
Wang *et al.* [68]	V_s,M_d	B_U	Surveillence
Hoffman *et al.* [18]	V_s	T_D	Shared Imitation
Boiman *et al.* [2]	V_s,M_d	B_U	Irregularity in images, video
Navalpakkam *et al.* [45]	V_s	B_U,T_D	Goal oriented model
Hu *et al.* [20]	V_s	B_U,T_D	Image Transmission
Kankanhalli *et al.* [29]	V_s,M_d	B_U,T_D	Sampling multimedia streams

Note: In the table 2.1, ***V,M,A,C,L*** represent the Visual, Motion, Audio, Camera and Linguistic attention model respectively. B_U , T_D represent Bottom-Up and Top-Down methodology. The suffix ***s*** and ***d*** denote nature of the attention stream whether it is static or dynamic.

2.1.3 Hybrid: Bottom-Up + Top-Down Model

Hu *et al.* have stated that the visual attention is not only affected by low level features but also guided by high level information. Hence, it is likely to consider both the bottom-up as well as top-down methodology while forming the attention [20].

Visual experience depends on convolution of bottom-up salience and top-down modulation specified by behavioral goals. The existing attention based on both top-down and bottom-up attention models include [9, 11, 10, 20, 29, 45]. Kankanhalli *et al.* have developed the experiential sampling technique, which is a goal oriented dynamic attention model for multimedia streams. This framework has been earlier applied to the problems of traffic monitoring, face detection and monologue detection [29, 30].

This has significant advantages due to:

1. its ability to use the prior experiences
2. its dynamic nature.

Table 2.2: *A comparative table depicting the state of the art - interestingness*

Reference	Stream	Methodology	Application
Butterfield *et al.* [3, 73]	visual	Social network analysis	Flickr Interestingness
Wolf *et al.* [69]	visual	B_U	Ranking of images
Dubinko *et al.* [8]	visual	T_D	Interesting tags
Chen *et al.* [6]	visual	B_U	Display Image in mobile devices

2.1.4 Applications of Attention models

It is found from the existing literature that visual attention models have been deployed for a variety of applications such as:

1. *Video summarization*: Finding the significant video frames using an attention model and summarizing accordingly [39],
2. *Detect Surprise events/irregularities in video*: Finding the surprise events [2, 24],
3. *Real time surveillance video display with salience*: Using saliency map for surveillence [68],
4. *Image transmission*: Sending the coarse version of the image first [20],
5. *Meeting Indexing*: Estimating head pose gazing [18] and
6. *Detecting Visual Events in Underwater Video*: Finding visual events using an bottom-up attention model [9, 11, 10].

Discussion: As it can be seen, in all of the above mentioned/cited applications, the bottom-up attention model has been used. This is also evident from the table 2.1. In table 2.1, we provided a comparative table depicting state of the art which summarizes the research work, type of attention stream, the adopted methodology and its application.

A significant amount of work [2, 9, 11, 10, 20, 45, 67] have been done on

computational models based on bottom-up methodology, in particular Itti-Koch saliency method. The brief details about Itti-Koch saliency method has been explained in section 2.6. We did a survey and found that attention is the better aid to semantic image understanding that can be used in adaptive content delivery and region based image retrieval. Navalpakkam *et al.* suggested the task dependency graph methodology for goal oriented task which requires high level semantic understanding [45]. Edgington and Walther *et al.* proposed a neuromorphic saliency based model which is based on both bottom-up and top-down [65]. However, their goal is just detecting the moving objects in underwater video. The work which is close to our intention of finding interesting images is [69]. However, their methodology is only based on bottom-up approach and does not consider user specific interest.

2.2 Interestingness

The patent by Butterfield *et al.* [3] describes the use of ranking media objects in determining interestingness through social network analysis. In Flickr [73], the notion of interestingness has been introduced to show the pictures that are seen by the people at that instant depending on a score based on the ideas in [3]. This score is based on the social network which is a measure of some combination of how many times a picture has been viewed, how many comments it has and how many times it has been tagged or marked as a favorite. In particular, Flickr interestingness is based on tags, comments, annotations or favorites. It is noted that no attention based modeling or any content based analysis has been done in Flickr's interestingness. MPEG-7 attention model which is based on a bottom-up methodology approach has been adopted by Wolf *et al.* to rank images [69]. In [8], Dubinko *et al.* have

attempted the problem of identifying most interesting tags over time. Their definition of interestingness has the following properties [8]:

- A more interesting object during a particular interval will occur more frequently within the interval, and less frequently outside the interval.
- A highly infrequent object need not be the most interesting object for that time interval.

The most attended region of an image is found using bottom up methodology and further used for display in mobile devices [6]. Based on the fact that people are interested in images which contains faces at the centre, Ma *et al.* have proposed face attention model [40]. Actually, we have discussed this in earlier section also since it is viewed as a attention model. It is also true that people would also be interested in scenes and hence we used the SIFT method which is better for detecting textured scenes Mikolajczyk *et al.* [42]. The more details can be found in section 2.9.

The reason for choosing Ma group's Itti-Koch attention model is of two fold as stated below:

1. Firstly, it makes use of neuromorphic based saliency information that is inline with the human cognition system and
2. Secondly, the model uses a fact that human eye has fairly little color sensitivity than luminance component.

As it can be seen in equation 3.1, we use brightness information of the saliency regions obtained from Itti-Koch saliency method described in section 3.2.1.

Also, the reasons for using face based attention model as mentioned in section 3.2.2 is based on the fact that human tend to concentrate at the centre portion of the image than towards the edge.

A notion of "***boring***" video frames is developed by detecting whether or not there is an interesting candidate object for an animal present in a particular sequence of underwater video [9]. This is determined by comparing each scanned location of the saliency map with the events that are already being tracked. If it does not belong to any of these events, a new tracker for the detected object is initiated.

The existing research work based on interestingness is summarized in table 2.2. Though interestingness is based on many attributes, we would describe surprise in a detailed manner since we intend to perform non identical duplicate removal such that surprise can be increased while maintaining user intention. Surprise is one of the attributes that triggers interest in human beings. According to Itti *et al.*, the key factor to our survival is ***surprise*** which is our ability to rapidly attend to, identify and learn from surprising events, to decide our present and future courses of action [21]. They state that there would be usually no surprise from the data that does not change prior beliefs. More details about prior beliefs is provided at the end of this section. Now, we present three examples for surprise.

1. Even the most liked TV programs, when telecasted for long time, become boring.

2. The other example could be, let us assume that there is more or less some continuous movement on a busy freeway. Here, one is surprised when there is no such movement detected at some time instant or during some specific interval of time.

3. In our photo album, consider the following scenario where and how surprise can be increased. Suppose if a person is interested in some category of photos in his mind, but in spite of many rounds of relevance

feedback process (refer section 2.3), if the system still shows totally unrelated photos that is of not of the user's interest, then it is surprising. For example, if a user is selecting images related to scene based photos, but if the system responds with totally unrelated images, then it is surprising. However, assume that the user poses same query and the system returns retrieved images. Now, within those retrieved images, surprise can be increased by removing non identical duplicates. In this case, surprise can be increased *while the user intention is still retained.*

Finding the attended regions is primarily based on "wow" factor *i.e.*, unit of Bayesian surprise (refer to section 2.7). The attentive level where to look is based on the intermediate level that analyzes the content of the fovea and the associative level that integrates the information in time (temporal). This is similiar to people's tendency to look back into past experiences, at temporal as well as attentive level. The fovea is a part of the eye which has high concentration of cone cells that are responsible for color vision in human beings. Also, it is noted that since fovea does not have rod cells, it is not sensitive to dim lights. As seen in figure 1.3 earlier, neural activity within the area V4 of the human brain system also indexes the degree to which a stimulus within the neuron's RF expresses a target-defining feature reflecting attentional modulations influenced by the prior knowledge of target identity. The two elements that are essential for a formal definition for surprise [21] are:

1. Surprise exists in the presence of uncertainty and
2. Surprise is related to the expectations of the observer. (single synapse, neuronal circuit, organism or computer device)

Discussion: It is understood from the literature survey that there still exists a need to develop a framework for defining interestingness such that both bottom-up and top-down approaches are adopted. To be specific, a user's individual interest that varies from person to person need to be considered.

2.3 Relevance Feedback:

In the context of image retrieval, the process of selecting those images that appear to be relevant to the query image is known as relevance feedback. A comprehensive review about relevance feedback has been given in [77].

In general, it is better to use the relevance feedback mechanism because of the following reasons: [77]

1. More ambiguity arises when interpreting images than words
2. Judging a document takes time while an image reveals its content almost similar to a human observer.

It is found that relevance feedback generally improves retrieval performance by improving search criteria in context of retrieval through user interaction [62]. The authors inferred that by using the relevance feedback, the accuracy is significantly increased about 6% to 30% in retrieval precision. Pseudo relevance feedback (PRF), also known as blind feedback refers to a methodology where instead of relying on the user to choose the top k relevant documents, the system simply *assumes that its top-ranked documents are relevant*, and uses these documents to augment the query with a relevance feedback ranking algorithm. Yan *et al.* describes PRF as the process of identifying potential positive and negative class labels of unlabeled images in the collection with aid of hints from the initial search results [72].

Yu *et al.* have used pseudo relevance feedback for information retrieval [75]. The authors have explored the advantage of partitioning the web pages into segments so that better expansion terms could be selected which in turn boost the retrieval performance. In spite of many works done using relevance feedback, we are interested in this approach because we assume our attended regions are analogous to segments in web pages.

Since we are interested in showing photos by capturing user interest via relevance feedback in particular pseudo relevance feedback, we would discuss it in section 2.8.

The feedback is in the form of questions by the user to obtain results. The system learns from the training examples to achieve improved performance next round, iteratively if the user desires [77]. To achieve this, the authors [75] have used VIPS (Vision based Page Segmentation) algorithm which is used for the purpose of selection of query expansion terms in pseudo relevance feedback. Here, the query expansion is used to bring some relevant documents missed in the initial round that can then be retrieved to increase the overall performance. The VIPS method depends on vision based cues and DOM (Document Object Model). This is done for grouping semantically related content of webpage into a single segment for web processing.

Discussion: Relevance Feedback is widely used in content based image retrieval. It has many advantages including facilitation of top down methodology or goal based approach. It is found that most widely used formula for query refinement is Rochio formula [33].

2.4 Non-Identical Duplicate Detection

Duplicate media content can exist because of two reasons -

- first, for transcoding purposes or for illegal copying of potential content;
- second, the consumers often shoot multiple photos and videos of the same scene.

The problem of duplicate detection in the first case is the problem of matching exactly two similar media contents, the solutions for which have been proposed using various digital signature / watermarking based methods [17]. In the second case, the duplicate detection is performed by matching two media contents which are not exactly identical but almost similar (such media are called "non-identical duplicates."[26])

Discussion: We have earlier applied this technique to identify non identical duplicates in video [64]. The video is a sequence of frames that have a high degree of temporal correlation among them. Each frame is an image in the two-dimensional spatial plane. Since image is analagous to video frames, indeed it can be applied to our data set. Since we assume that interestingness is related to surprise and surprise originates from uncertainty of data, our idea is to increase in surprise as well as Shannon entropy information locally by removing non identical duplicates.

As said earlier, we would provide some background information about each of the methodology in upcoming sections.

2.5 General Discussion

From our survey, to the best of our knowledge, it appears that so far no attention modeling system has been done for the purpose of *interestingness* in particular for home based chronicles which might be centered around people, events, locations *etc.* The interestingness needs to be defined precisely

which potentially should consider the past experiences as well. Since relevance feedback can capture the dynamic attention and use past experiences, a computational model based on the relevance feedback technique coupled with interestingness would be useful. Also, it is noted that most of the visual attention-based research works such as [2, 9, 11, 10, 20, 45, 67] are centered around this model.

In a work by Ma *et al.* [39], the authors have used the above Itti-Koch model to find the static saliency value of images which can be used for suitable applications. Further, the face centric attention model has also been proposed by them which assumes that people are interested in images having people at the centre. We have used and explained this model in section 3.2.2. Since people would also be interested in scenes, we used SIFT method which is better for detecting textured scenes Mikolajczyk *et al.* [42]. We have provided a brief introduction to the SIFT method in section 2.9.

We also noted from the Bayesian surprise theory that by removing redundant information, surprise can be increased. We intend to make use of this idea and increase the surprise by removing the non identical duplicates.

2.6 The Itti-Koch Visual attention model

The pictorial representation of Itti-Koch saliency model is as shown in figure 2.2. The input image is subsampled into a dyadic Gaussian pyramid by convolution with a Gaussian filter [67]. This can be clearly understood from the schematic diagram of Gaussian pyramid construction as shown in the figure 2.3.

Each level of the image pyramid is then decomposed into maps for Red-Green (RG), Blue-Yellow (BY), Intensity (I) and local orientation (O). The

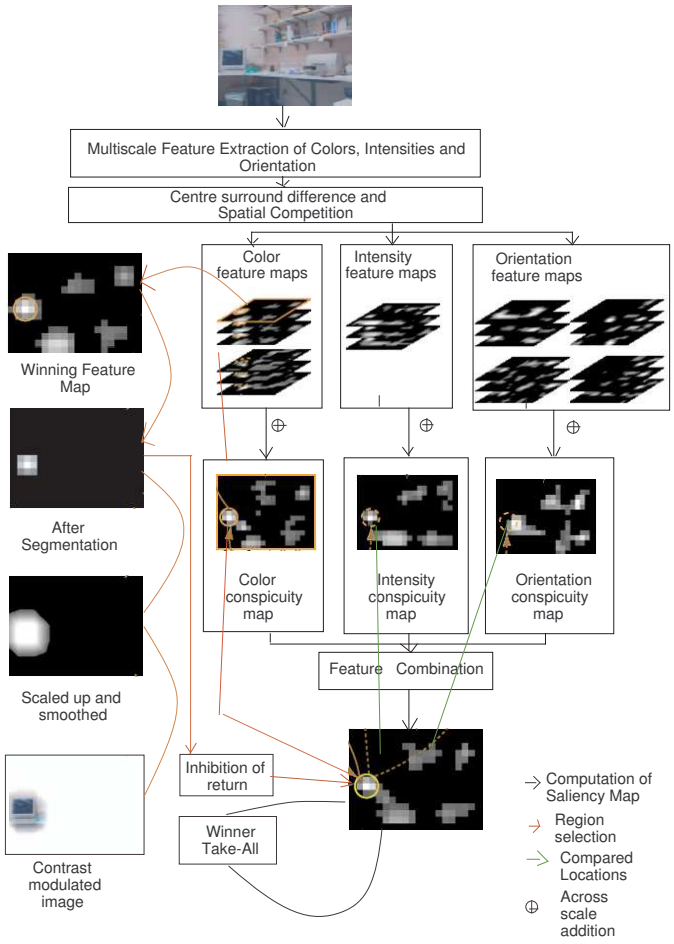

Figure 2.2: *Itti-Koch Saliency Model [Source: Walther [66, 67]]*

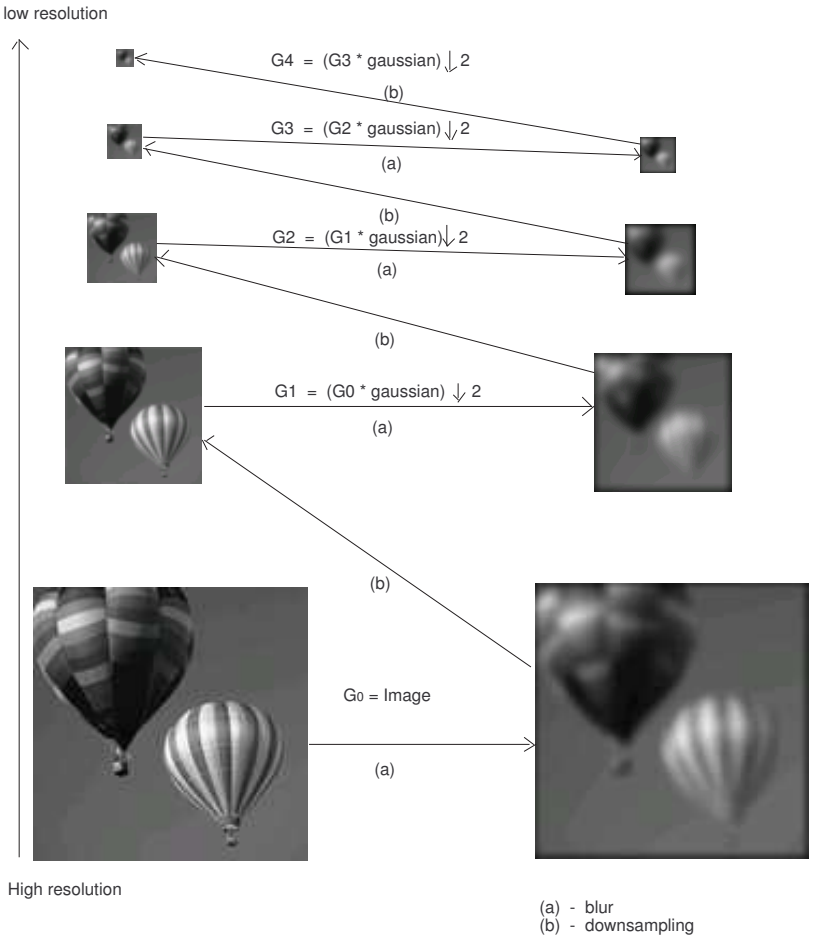

Figure 2.3: *Gaussian Pyramid*

intensity map is calculated as

$$M_I = \frac{(r+g+b)}{3} \tag{2.1}$$

Here dyadic denotes the pyramids with downsampling by a factor of 2 (default) and r, g, b denote the red, green and blue values of the color image respectively.

$$M_{RG} = \frac{(r-g)}{max(r,g,b)} \tag{2.2}$$

$$M_{BY} = \frac{b - min(r,g)}{max(r,g,b)} \tag{2.3}$$

The center surround feature maps obtained from all the above mentioned channels are summed using across scale addition (point to point addition) and the sums are normalized again.

$$F_l = N(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} F_{l,c,s}) \forall l \in L_I \cup L_C \cup L_O \tag{2.4}$$

where $L_I = \{I\}$, $L_C = \{RG, BY\}$ and $L_O = \{0^o, 45^o, 90^o, 135^o.\}$ The across scale addition denoted by $\oplus$ represents the point to point addition followed by reduction of map to scale 4. c and s represent the center and surrounding levels of respective feature pyramids [66]. N denotes the iterative non linear normalization operator. Equation 2.4, 2.5 and 2.6 are also known as conspicuity map for intensity(I), color(C) and orientation(O) respectively.

$$C_C = N(\sum_{l \in L_C} F_l) \tag{2.5}$$

$$C_O = N(\sum_{l \in L_O} F_l) \tag{2.6}$$

It can be seen that $C_I = F_l$ (refer to equation 2.4).

Here C_I, C_C and C_O represent the the conspicuity map for intensity, color and orientation respectively. This is followed by the *combination of all conspicuity maps into a single saliency map* (S) as follows:

$$S = 1/3 \sum_{k \in \{I,C,O\}} C_k \tag{2.7}$$

The locations in the saliency map compete for the highest saliency value by winner−take−all network to integrate and fire neurons. The winning location (x_w, y_w) of this process is attended to and the saliency map is inhibited within a given radius of (x_w, y_w). Itti-Koch model is successful in identifying the salient location. The extent to which the location is salient is attempted by Rutishauser *et al.* [56]. The authors have provided an extension to the formal framework of Itti-Koch model, aiming object recognition.

The overall brief description of Itti-Koch model is summarized below:

1. *Feature Extraction*: The visual features from the entire visual scene are extracted parallely in several multiscale feature maps. Feature extraction is achieved through linear filtering for a given feature type (*example*: intensity, color, orientation *etc*).

2. *Center Surround Operation*: The linear filtering is followed by a center-surround operation which extracts local spatial discontinuities for each feature type. The center-surround difference is determined by parameter basis (*example*: dyadic).

3. *Supervised Learning*: The importance of color discontinuity over orientation or intensity discontinuity or vice-versa is found by involving *supervised learning* using manually defined target regions ("binary

target mask"). As stated in [22], the supervised learning procedure is used when there is a need to detect specific targets. Each feature map is globally multiplied by a weighting factor. The final input to the saliency map is the point-wise sum of all feature maps. The learning procedure for the weight $w(\mu)$ of a feature map μ consists of the following:

(a) Compute the global maximum max_{glob} and minimum min_{glob} of the feature map μ.

(b) Compute its maximum max_{in} inside the manually outlined target regions [22] and its maximum max_{out} outside the target regions.

(c) Update the weight following an additive learning rule independent of the map's dynamic range: $w(\mu) \longleftarrow w(\mu) + \eta \frac{max_{in} - max_{out}}{max_{glob} - min_{glob}}$ here $\eta > 0$ determines the learning speed. As said earlier, this method is adopted if there is a needed to find specific targets. Only positive or zero weights are allowed. Here, max_{in} and max_{out} are the maximum values inside and outside the manually outlined target regions respectively. Also, max_{glob} and min_{glob} are the global maximum and global minimum values of the feature map respectively.

(d) The learning procedure promotes through an increase in weights, the participation to the saliency map of the feature maps which show higher peak activity inside the target regions than outside. This means that the priority is given preferably inside the target regions and then try to find the saliency object.

As inhibition of return can be seen in the figure 2.2, it basically represents a brief (about 300 milliseconds) period of facilitating the pro-

cessing at a location where attention is directed at.

4. *Maximum Detector/WTA (Winner-Take-All Rule)*: After such combination is computed, a maximum detector selects the most salient location in the saliency map and shifts attention towards it. The saliency map is sequentially scanned in the order of decreasing saliency by focus of attention which is achieved by the Winner-Take-All Rule (WTA) that *selects the most salient location* at any given time. Mozer *et al.* describes about WTA that the saliency units compete with each other and the unit that is most active will inhibit others [43]. Also, only one attentional unit is active at a time - corresponding to the selection of a particular location. Then this location is inhibited (suppressed) to allow the system to focus on the next most salient location. Commonly, the time period during which attention is inhibited from returning to the previously attended location is called as Inhibition of Return (IOR). IOR usually lies in between 300 milliseconds and 3 seconds. Whenever a portion of an image is attended, saliency region is in the short term memory for the time mentioned as above.

Also, it is noted that Itti *et al.* have made a comparison of the following feature combination strategies for saliency based system [22] such as

1. simple normalized summation,
2. linear combination with learned weights,
3. global non linear normalization followed by summation and
4. local non linear competition between salient locations.

It is stated that the above mentioned 3^{rd} and 4^{th} method yielded significant performance whereas 2^{nd} one yielded poor generalization.

2.7 Bayesian Surprise theory

We provide background information about Bayesian surprise theory, proposed by Itti *et al.* [21] in order to show how surprise is primarily based on the prior information a person or model possess. Based on the prior probability, background of an information is known, say $P(M)_{M\in\Re}$ over the hypotheses or models M in a model space $\Re$. The fundamental effect of a new data distribution D on the observer is to change the probability distribution into the posterior distribution via Bayes theorem

$$\forall M \in \Re, P(M/D) = \frac{(P(D/M) \times P(M))}{P(D)} \tag{2.8}$$

New data observation D carries no surprise if it leaves the observer beliefs unaffected, *i.e* if the posterior is identical to the prior; conversely, D is surprising if the posterior distribution resulting from observing D significantly differs from the prior distribution.

Surprise is defined by the average of log odd ratio

$$S(D,M) = KL((P(M)|D), P(M)) = \int_M P(M|D) log \frac{P(M|D) \times dM}{P(M)} \tag{2.9}$$

In the above equation, KL denotes (Kullback-Leibler) divergence which is the distance measure between prior and posterior distribution. Itti *et al.* have formulated a unit of surprise as "wow". The definition of wow is provided as follows: wow is defined for a single model M as the amount of surprise corresponding to a two-fold variation between $P(M|D)$ and $P(M)$, i.e., as $logP(M|D)/P(M)$ (with log taken in base 2).

The integration over model class denotes the total number of "wows" (refer section 2.2) experienced considering all models.

In the next section, we present a pseudo relevance feedback algorithm used by Yu *et al.* [75].

2.8 Pseudo Relevance Feedback Algorithm

The steps in VIPS (Visual based Page Segmentation) are as given below:

- **Initial Retrieval: -** An initial list of ranked web pages is obtained by using any traditional information retrieval methods.
- **Page Segmentation: -** In this step, the VIPS algorithm is applied to divide retrieved web pages into segments. After the vision-based content structure is obtained, all the leaf nodes are extracted as segments. Since it is very expensive to process all retrieved web pages, a few top pages are selected for segmentation. The candidate segment set is made up of these resulting segments.
- **Segment Selection: -** This step chooses most relevant segments from the candidate segment set. Some ranking methods [53] are used to sort the candidate segments and the top (*eg*:20) segments are selected for expansion term selection in the next step.
- **Expansion Term Selection: -** This approach is used to select expansion terms. The difference is that expansion terms are selected from the selected segments instead of from the whole web pages. All terms except the original query terms in the selected segments are weighted according to the following Term Selection Value TSV:

$$TSV = w^{(1)} * \frac{r}{R} \tag{2.10}$$

where $w(1)$ is the Robertson/Sparck Jones weight; R is the number of

selected segments; and r is the number of segments which contain this term. They have considered top 10 terms are selected to expand the original query.

- **Final Retrieval: -** The term weights for the expanded query are set as the following: For original terms, the new weight is tf * 2 where tf is its term frequency in the query; For expansion terms, the new weight is $1 - \frac{(n-1)}{10}$ if the current term ranks n^{th} in TSV rank. 10 terms are selected to expand the query. The expanded query is used to retrieve the data set again for the final results.

Now, we discuss SIFT method in the next section.

2.9 SIFT method

SIFT (Scale Invariant Feature Technique) has been proposed by Lowe [37] for finding distinctive feature points in an image. SIFT features which have been widely used in many object recognition applications, possess the following properties:

- they are scale invariant
- they are resistant to translation, rotation and scaling.

It maps an image data into scale-invariant coordinates relative to local features. In a paper by Mikolajczyk *et al.*, it is found that scenes are of two types: structured scene and textured scene. The authors have inferred that SIFT is best for textured scene [42]. Also, indoor environments often have large homogeneous textured objects, such as walls and furniture [36]. Our idea is that by capturing such homogeneous textured images from images of the whole dataset, interestingness might be captured. Though it has been

used widely in many object recognition oriented applications, we propose one more application for finding user interest. The major steps in SIFT method include the following [37]:

1. Detection and Localization: The detection and localization of keypoints are done as follows:
    - *Scale-space extrema detection:* The first stage identify potential interest points by using a Difference of Gaussian (DOG) function in the image scale space.
    - *Keypoint localisation:* The location and scale of each candidate point is determined and keypoints are selected based on measures of stability.
2. Orientation assignment: This is based on the major gradient direction around each keypoint at the selected scale.
3. Keypoint descriptor: A descriptor is generated for each keypoint from local image gradients information at the scale found in step 2.

An important aspect of the algorithm is that it generates a large number of features over a broad range of scales and locations. The number of features generated is dependent on image size and content.

Thus, in this Chapter 2, we have explained the state of the art made on

- Attention models,
- Interestingness, and
- Non Identical Duplicate detection.

Now, we would provide our framework / algorithm / technique for attention based interestingness next in Chapter 3.

Chapter 3

Attention-based Interestingness

This thesis proposes a novel framework for browsing interesting images in the eChronicles Attention system. We shall start the discussion by describing the typical user interaction flow with the system. This is followed by detailed description of the framework and system architecture. In order to provide better understanding of the system, we present the attention and relevance feedback algorithms in detail which takes advantage of both bottom-up attention as well as top-down (goal-based) methodologies.

Now, we would provide our overview of the framework for attention based interestingness.

3.1 Overview of the Framework

Initially, the user selects images from the collections that are displayed randomly (refer figure 3.1,3.2). The images which closely match the feature values of the selected images are displayed as the result set. From the result

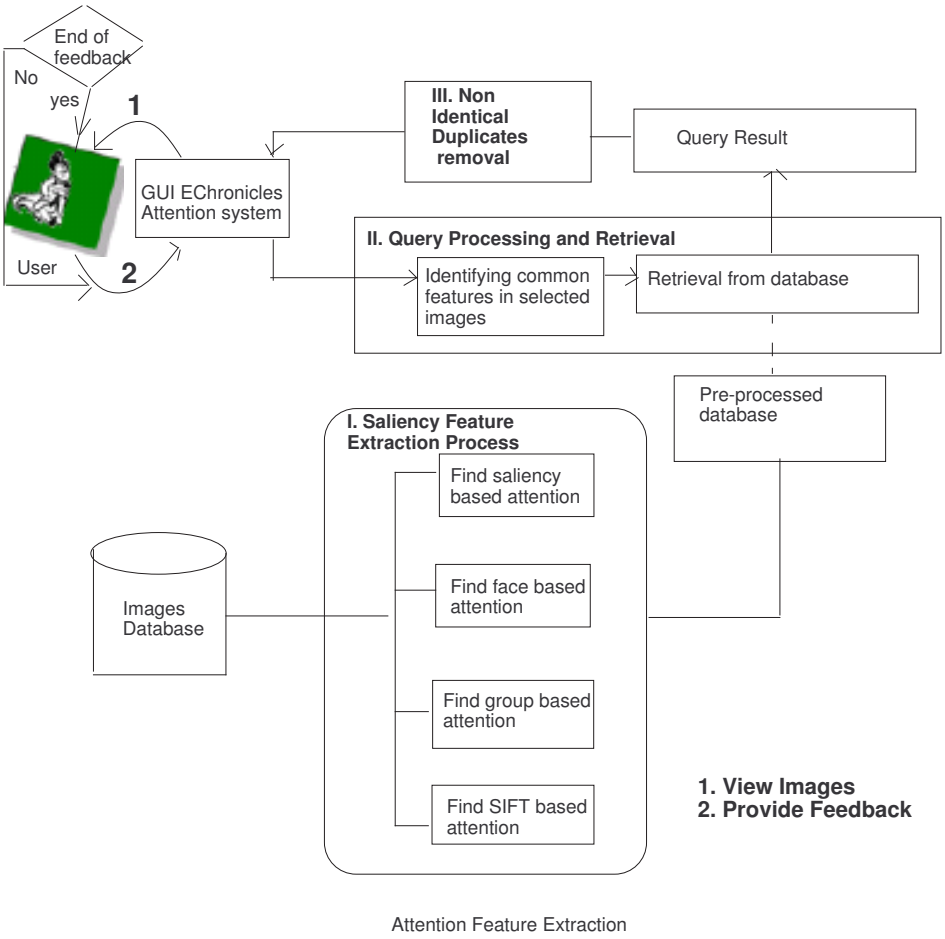

Figure 3.1: *Proposed Framework for EChronicles Attention System: finding interesting Images*

set of displayed images, the user is asked once again to select images that are interesting within this displayed set. The key issue is that user is allowed to select different images of her own interest repeatedly until the precise interest of user is succinctly captured and she is able to view all the images which are of interest to her. To have a better understanding of the system, graphical user interface of the system is shown in figure 3.2. Our framework consists of three important components

1. *Saliency Feature Extraction process:* This part studies how people provide attention to images. The saliency attention values are arrived at from four attention models, namely *Itti-Koch based attention, Face centric based attention, SIFT based attention and Group based attention.* (more details provided in section 3.2.1 to 3.2.4). The reason for choosing Ma group's Itti-Koch attention model is of two fold as stated below:

   (a) Firstly, it makes use of neuromorphic based saliency information that is inline with the human cognition system and

   (b) Secondly, the model uses a fact that human eye has fairly little color sensitivity than luminance component.

   As it can be seen in equation 3.1, we use brightness information of the saliency regions obtained from Itti-Koch saliency method described in section 3.2.1.

   Also, the reasons for using face based attention model as mentioned in section 3.2.2 is based on the fact that human tend to concentrate at the centre portion of the image than towards the edge. The group based attention is based on the fact that people are interested in images that contain group of people.

2. *Query processing and retrieval results*: The query is formed from the set of relevant images selected by the user. The query is formed by the feature vector obtained from the selected images. Based on the query constraints, the similiarity measure is done between the query vector (feature obtained from user selection) against the entire database feature vector and the result is displayed. (more details provided in section 3.4)

3. *Non Identical Duplicates Removal:* This is a process of removing non identical duplicate images in the retrieved results. (for details refer to section 3.3)

The saliency feature extraction undertakes the offline preprocessing on the image database to prepare it to provide the relevant information to the relevance feedback components. The component is the online component which identifies common features across selected images and displays images which are most likely to be interesting to the user. We shall look at each of these three components in further detail in the following sections.

3.2 Saliency Feature Extraction process:

Saliency Feature extraction is the important component of the framework which aids the system in identifying the important features of the various images present in the database. This process comprises four attention methodologies.

3.2.1 Itti-Koch Based Attention

Each image can be represented in YUV model where Y stands for the luminance component (brightness) and U and V are the chrominance (color)

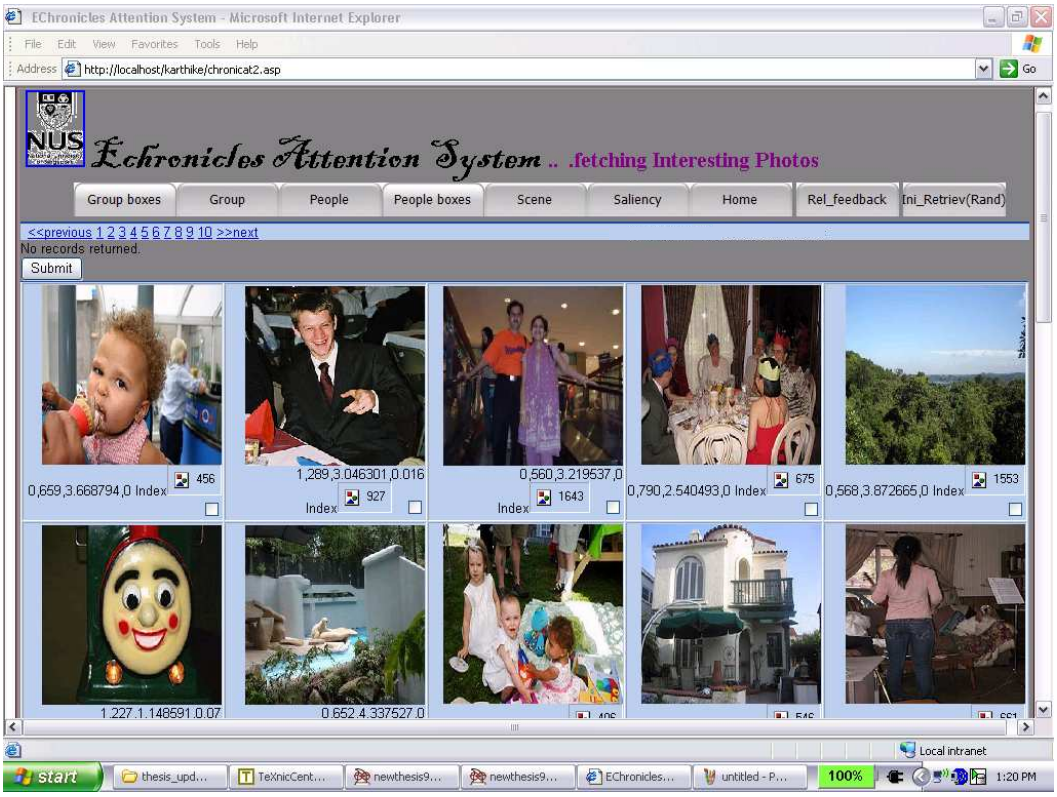

Figure 3.2: *GUI of the system*

components. $B_{i,j}$ represents the luminance component at pixel $(i,j)^{th}$ position. The brightness information of the luminance component has more impact on the image than that of the other two color components since the eye has less color sensitivity. This is one of the reasons it is perceived as attention model.

The centre of Itti-Koch saliency region is found and a Gaussian template is centered around mean of the frame and the weightage is given accordingly as shown in equation 2.7. As explained in section 2.6, a number of visual cues are extracted from the scene by computing the cue maps F_l. The use of cues is motivated by a study on primate visual systems. It uses two chromatic channels that are inspired from human vision namely Red/Green (RG) and Blue/Yellow (BY). Each map F_l is transformed in its conspicuity map C_l. Each conspicuity map highlights the parts of the scene that potentially differs according to a specific cue, from their surroundings. Then, the conspicuity maps are integrated together to form a saliency map S with respect to the earlier mentioned equation 2.7. The detail steps to calculate saliency value is given in algorithm given below.

Algorithm : Static saliency

Input: Images Database

Output: Static Saliency attention values

Method:

1. for all images in the database, obtain the saliency attention value ϕ as from equation 3.1.

$$\phi = \frac{1}{A_{frame}} \sum_{k=1}^{N} \sum_{i,j \in R_k} B_{i,j} \times w_{pos}^{i,j} \tag{3.1}$$

A_{frame} is the area of the frame

k is an integer that denotes the number of saliency regions where $k : 1 -> N$

N denotes number of saliency regions in an image

i, j denotes the pixel position in saliency regions

$B_{i,j}$ denotes brightness of the pixels in saliency regions

R_k denotes the saliency region

$w_{pos}^{i,j}$ is a normalized Gaussian template with the mean located at center of the frame.

2. end

Now, we would discuss the face based attention model in section 3.2.2.

3.2.2 Face Based Attention

To find face attention value, we adopt an approach as described in [40]. The face is the salient region in an image and to be specific, the size and position of a face usually reflect the importance of the face. We used the OpenCV face detector and obtained the face information in each image of the entire dataset such as number of faces, sizes and positions. A face detected on sample image of our dataset with position weight is shown in figure 3.3. As it can be clearly seen from the formula and the figure 3.3 that if the face is detected at the centre, then a full weightage of 1 is given (since weightage of the centre block 8 is divided by 8). Based on where the centre of the face overlaps with the index of the block, it is mutiplied with the corresponding

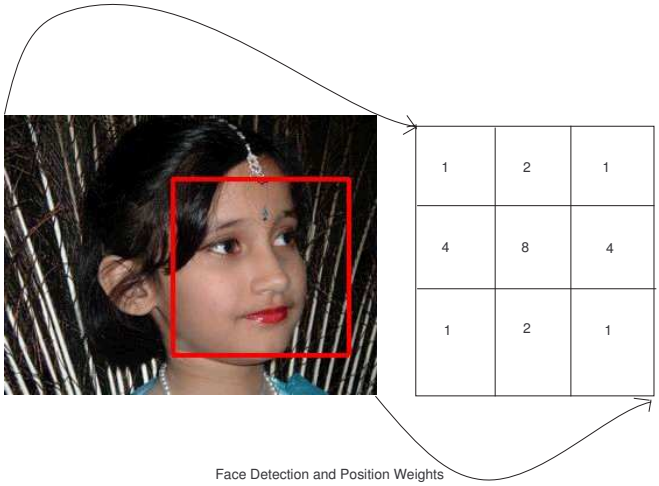

Figure 3.3: *Face detection and position weights*

index weight, If the detected face centre is say within the 5^{th} block, then the full weightage 1 is given. The intuitive idea is that if the face is detected at the centre position, then more weightage is given comparitively to faces detected at other positions.

The detailed steps to calculate face attention value is given in algorithm.

Algorithm : Face Attention model

Input: Images Database

Output: Face Attention values

Method:

1. for all images in the database, obtain face attention value κ as from equation 3.2.

$$\kappa = \sum_{k=1}^{N} \frac{A_k}{A_{frame}} \times \frac{wpos^i}{8} \tag{3.2}$$

where A_k denotes the size of k^{th} face in a frame

A_{frame} denotes the area of frame

$wpos^i$ is the weight of position defined in figure 3.3

$i..[0, 8]$ is the index of position.

2. end

Now, we provide SIFT based attention algorithm in section 3.2.3.

3.2.3 SIFT based attention

As discussed in section 2.9, SIFT is considered to be useful in finding textured scene such as walls and furniture [36]. Our idea is that by capturing such homogeneous textured images from images of the whole dataset, interestingness might be captured. In other words, we can find images similiar to those selected by the user and be able to display images which are interesting to him. We define SIFT-based attention as the number of scale invariant feature keypoints in an image.

$$\delta = \frac{\#S}{A_{frame}} \tag{3.3}$$

where A_{frame} denotes area of the frame (image).
S is the number of SIFT points in an image. Now, we let us discuss group based attention algorithm in section 3.2.4.

3.2.4 Group based attention

We define user's group-based attention η as the number of faces in an image. The reason behind why we call it as attention is as follows :

- face is the natural candidate that can guide features.

- face is considered as a probable non - attribute that guide attention (refer figure 2.1).

$$\eta = (nf) \tag{3.4}$$

where nf represents the number of faces in image. Thus as a whole, we denote the obtained Itti-Koch attention value, face attention value, number of faces and number of SIFT points in an image as $\phi, \kappa, \eta, \delta$ respectively. Thus as a whole, each image is represented by the attention feature vector as

$[\phi, \kappa, \eta, \delta]$

where ϕ represents Itti-Koch attention value, κ represents face based attention value, η represents number of faces in an image and δ represents number of SIFT points in an image. Now we would discuss about Non Identical Duplicate Detection in section 3.3.

3.3 Non Identical Duplicate Detection

This is yet another component of the framework which helps in increasing the surprise while retaining the intention of the user. Now, we shall discuss how non-identical duplicates are detected using the SIFT method [37]. Let m_{ij} be the number of match points between the images i and j, P_i and P_j be the number of key points found using SIFT method for the image i and image j, respectively. Then, the similarity matches between all the Image pairs (I_i, I_j), $1 \leq i \leq n_1$, $1 \leq j \leq n_2$ are obtained using the SIFT method [37]. As a result of this, we obtain a matrix M_{ij}, $1 \leq i \leq n_1$, $1 \leq j \leq n_2$.

The match score M_{ij} is computed using the following equation -

$$M_{ij} = 2 \times \frac{m_{ij}}{P_i + P_j} \tag{3.5}$$

Algorithm : Non Identical Duplicate Detection

1. for all image pairs (I_i, I_j), $1 \leq i \leq n_1, 1 \leq j \leq n_2$,
2. calculate similiarity match (NID metric) as follows: obtain a matrix M_{ij}, $1 \leq i \leq n_1, 1 \leq j \leq n_2$ $M_{ij} = 2 \times \frac{m_{ij}}{P_i + P_j}$ where m_{ij} is the number of match points between the image i and image j, P_i and P_j are the number of key points in image i and j respectively.
3. end

Now, we shall discuss relevance feedback mechanism used in our echronicle attention system in section 3.4.

3.4 Query Analysis and Retrieval (Relevance Feedback)

This is an vital component which helps to capture the user interest dynamically that evolves over time. We make a set of assumptions as given below:

- $A1$: That user selects images that have some common attributes among them.
- $A2$: That people might not select all images that seem interesting to them.

Attention Feature Extraction: Initially, we processed all of our images in the dataset and the set of attention features extracted from each

image according to the attention models is stored in feature vector format as

$$\begin{bmatrix} \phi \\ \kappa \\ \eta \\ \delta \end{bmatrix}_{4\times 1} \tag{3.6}$$

where ϕ represents the Itti-Koch attention value, κ represents the face based attention value, η represents the number of faces in an image and δ represents the number of SIFT points in an image.

The steps in our pseudo feedback algorithm include:

1. **Initial retrieval:** Initially, a list of ranked photos can be obtained by using any of the reasonable method. This, we have done by randomly selecting photos from the database. Actually, from this set, the user needs to select the images that appear interesting to him.

2. **User selection:** This step selects the most relevant attended images from candidate image set selected by user in the initial set as discussed in step 1. Let i represents the index of the images in the database where $i = 1 \cdots n$. Here $n = 2023$ where n is the total number of images in the dataset. Let C be the number of images shown in the display window. Here $C = 30$. Let R represent the relevant images (images selected by the user) and NR represent the non-relevant images that are not selected by the user. Then, $|R| \leq |C| < |D|$
   where $|R|$, $|NR|$ and $|D|$ represents the cardinality of relevant images, non relevant images and whole image database respectively. Then $|NR| = 30 - |R|$ holds true since we consider 30 images for displaying images in photo album interface. Let us assume that query is posed by the user at time instant t_1, t_2, t_3, $t_4..t_n$ respectively.

3. **Query Formulation:** The first results before any query are thirty images, randomly selected from the database of 2023 images. The query is formulated as follows:

   First retrieval: Let us assume that first retrieval f_1 is obtained from q_1 which is the set of feature vectors obtained from user selected images. We represent q_1 as the combination of the query terms such as

$$q_1 = q_{t1} \wedge q_{t2} \wedge q_{t3} \wedge q_{t4} \tag{3.7}$$

   Here $\wedge$ represents the AND operator. So, we can denote q_1 as the combination of query terms qt_1, qt_2, qt_3 and qt_4.

$$\begin{bmatrix} qt_1 \\ qt_2 \\ qt_3 \\ qt_4 \end{bmatrix}_t = \begin{bmatrix} min < \phi < max \\ min < \kappa < max \\ \eta \geq max \\ min < \delta < max \end{bmatrix} \tag{3.8}$$

   min and max represent the minimum and maximum number of images selected by the user.

   The first retrieval that occurs at time instant 1 is denoted by f_1. For the first retrieval $f_1 = q_1$. We can conceptually view the generalized query q_1 as a compound query which is the composition of 4 atomic queries (qt$_1$,qt_2,qt_3,qt_4) .

   Second retrieval: Let the second retrieval be denoted by f_2 that happens at time instant 2. This is calculated as follows:

$$f_2 = \alpha \times q_2 + (1 - \alpha) \times f_1 \tag{3.9}$$

Now, q_2 represents the query formed from the set of images selected from f_1. The user selects set of images (feedback) and the feature vector values obtained corresponding to those selected images at this particular time instant t_2 be represented by q_2.

Third retrieval: The third retrieval f_3 at time instant t_3 is formed from the set of displayed images f_2.

Now,

$$f_3 = \alpha \times q_3 + (1 - \alpha) \times f_2 \tag{3.10}$$

General feedback retrieval: The general feedback query for final retrieval at time instant t_n can be represented as The query for n^{th} retrieval denoted by q_n is the query formed from the displayed images of f_{n-1}. The final retrieval is

$$f_n = \alpha \times q_n + (1 - \alpha) \times f_{n-1} \tag{3.11}$$

where

$$n \geq 2 \tag{3.12}$$

In the case where $n = 1, f_1 = q_1$

α represents the weightage given to the query term values at that time instant. The number of images displayed on the interface depends on the the number of images which meet the query constraints. The *reason* why we call it as *pseudo relevance feedback* is that initial retrieval is based on images that are selected *randomly* where *surprise is high.* This is in accordance with Bayesian theory. This is explained as follows: There is no redundant information in the initial retrieval which means that it is highly unlikely that there would be similar type of

images in the first round. However, as the relevance feedback process goes on, the images are grouped together based on user selection of images. Since we know from the Bayesian surprise theory that redundant information carries no surprise, the surprise is high in the initial retrieval.

The key idea is to fix a bound from the initial user selection and narrow it down to capture the user's interest. The query refinement is based on giving mutual weightage to the minimum and maximum values of the current and previous feedback values. In this way, one can find the interesting images in minimal time. Since we assume that people selects images that have common attributes, we use the min and max values of each of the individual features in the images selected by user and use it for our further term reweighting. Based on the images selected, an initial query is formed from which it is then refined based on further selection. The query matching is done with the database where the value meets the query constraints and results are displayed. One limitation of this approach is that user interest is dependent on the initial set of random number of images displayed from which he selects the images. However, one has the option to choose different types of images until he is willing to attend and then followed by the selection such that user interest can be narrowed down.

Chapter 4

Implementation and Results

In the preceding chapter, we discussed our framework. In this chapter, we shall look into the implementation details and results. Preliminary experiments have been performed on the dataset using the developed system to analyze the quality, accuracy and efficiency of our framework in finding the interesting images. This is done by evaluation of the results of our system by quantitative performance of each adopted method as well as for the whole system and qualitative subjective analysis through an user study. This analysis aims at investigating the user's perception on whether the system is able to meet quality attributes or not.

Now we shall explain the implementation details with respect to software structure of the system.

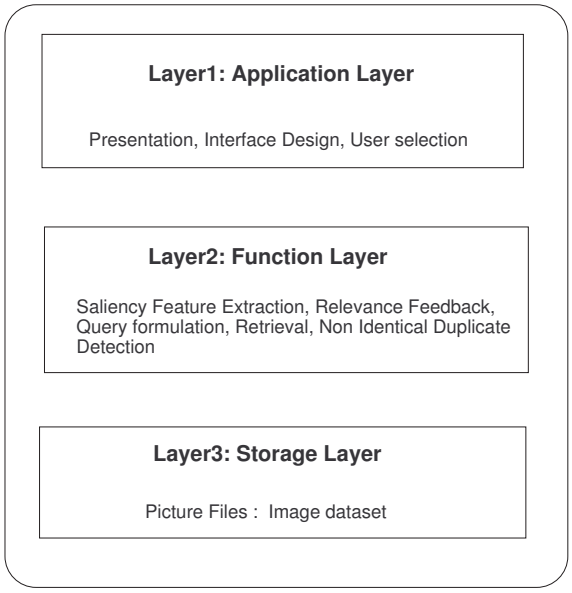

Figure 4.1: *Software Structure of Echronicles Attention System*

4.1 Software structure

The software structure of the system constitutes three layers such as application layer, function layer and storage layer as given in the diagram 4.1.

(a) *Application layer:* It is concerned with interface design issues, presentation, results viewing and result selection etc.

(b) *Function layer:* It constitutes three main functional modules such as saliency feature extraction, Relevance feedback mechanism and Non Identical Duplicate Detection.

(c) *Storage layer:* It comprises the images related information (2023 images dataset).

4.1.1 Implementation Platform

The application layer which includes GUI interface design, presentation, results viewing and result selection is implemented using ASP (Active Server Pages) and a VB Script environment that acts as front-end. As a back-end tool for storage layer, we used MS Access. The platform is Windows, ASP (Active Server Pages) and VB Script environment as front-end and MS access as back end. We used Intel Open CV (Visual C++ environment) for face detection, Matlab for calculation of SIFT and saliency points of images.

Data-set: We collected a data-set of 2023 images (a combination of personal collections and downloaded pictures from Flickr) and we used Pentium-IV 2.4 GHz with 512 MB RAM for our experiments.

4.2 Description about interface

We used ASP and designed a user interface which has a check button for each of the images that are displayed initially as shown in figure 4.2. The checkbutton enables the user to select any number of images ranging from 1 to 30. The relevance feedback will be initiated once the user confirm the details returned by the system such as how many images that user has selected etc. The relevance feedback modules take care of query analysis and retrieval and return the set of images. Then, the user can once again select any number of images using which the system will extract common attentive information, then process and display the new set of images based on the information extracted from the selected images. This process is accomplished in an iterative

manner until the user is satisfied with the returned results. The aim of the system is to grasp the relevant information need from the user and process the saliency features in the database on user interest-basis.
To have a clear picture about our implementation interface in the initial round and further subsequent rounds, we provide the screenshot of our system at various time instants.

A. Fig 4.2 represents the screenshot of the system in the time step 0 where 30 images are displayed initially from which user is allowed to select images of his interest. Here, the user has selected images 1 and 4. Then the system performs query analysis and retrieval after which the results are displayed at the second stage.

B. The displayed results are shown in the figure 4.3. At this time step 1, it can be seen that user has selected the images 2 and 4.

C. After the completion of query analysis and retrieval part, the results are displayed by the system as shown in screen shot refer figure 4.4. Though the system GUI can display 30 images, only 10 images are displayed in the screenshot for visual clarity purpose.

4.3 Experimental Results and Discussion

In this section, we present our experimental results followed by a detailed discussion how visual attention model along with relevance feedback aids in improving the process of capturing user interest.

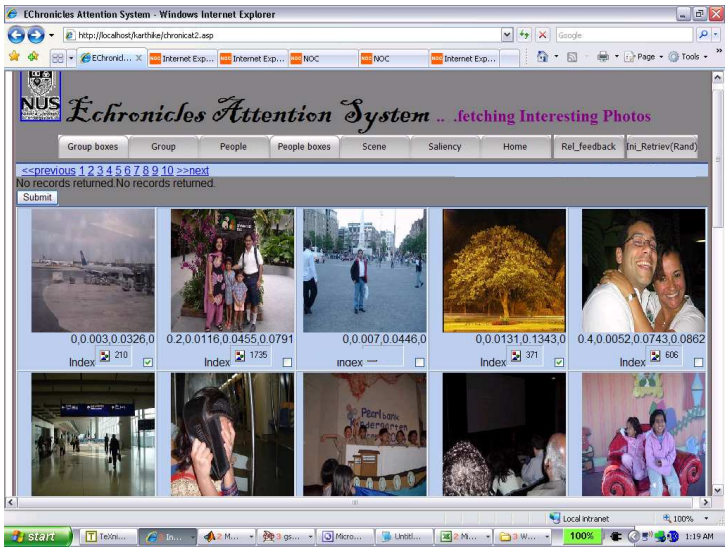

Figure 4.2: *Echronicles Attention system : Initial Round*

Figure 4.3: *Echronicles Attention system : Second Round*

Figure 4.4: *Echronicles Attention system : Third Round*

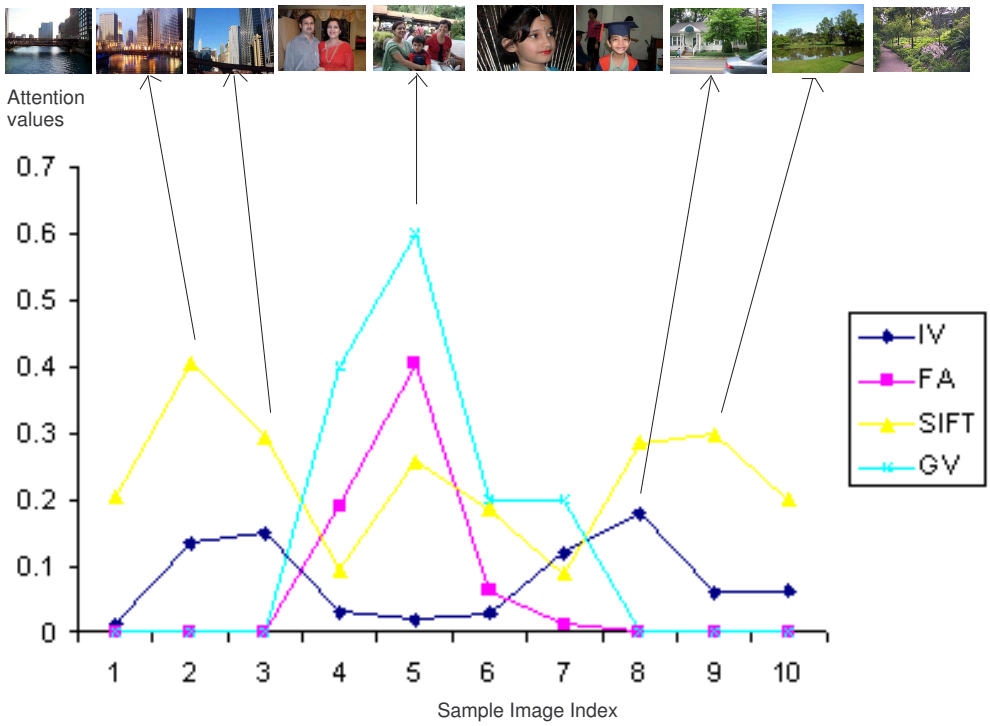

Figure 4.5: *Attention value graph*

Table 4.1: *Attention values for a sample of 10 images from our dataset*

SampleImageIndex	Image	*IV*	*FA*	*SIFT*	*GV*
(1)	*Building1.JPG*	0.014029	0	0.0092	0
(2)	*Building2.JPG*	0.134954	0	0.0183	0
(3)	*Building3.JPG*	0.149079	0	0.0132	0
(4)	*Pradeepfamily.JPG*	0.030523	0.190644	0.0042	0.4
(5)	*Pranjal.JPG*	0.018136	0.403447	0.0116	0.6
(6)	*Akansha.JPG*	0.028184	0.062493	0.0083	0.2
(7)	*Pranjalfriend.JPG*	0.1184	0.012428	0.004	0.2
(8)	*Scene1.JPG*	0.177858	0	0.0129	0
(9)	*Scene2.JPG*	0.060937	0	0.0134	0
(10)	*Scene3.JPG*	0.063202	0	0.009	0

4.3.1 Illustration of calculated saliency attention values

Now, we shall have a look at the calculated attention values for sample images of the dataset (refer table 4.1) and have a discussion pertaining to how each of the attention features are useful in capturing user interest. The attention values and images corresponding to peak attention values can be seen in the graph (refer figure 4.5).

In the figure, IV represents Itti-Koch Static Attention value, FA represents Face Attention value, SIFT represents Scale Invariant Feature value and GV represents Group Attention value. The following observations are made from the attention value graph.

The attention value graph is explained using 10 images that belong to four groups. The purpose of the graph is to show the utility of each of the attention value features in identifying each of these groups. The SIFT based attention value as shown in yellow color arrow represents the images which has textured scenes such as buildings. The face based attention value which is shown in pink color arrow represents the images where the face is in center or more number of faces in the center. The group based attention value which is shown by green color arrow represents the images with more number of faces ignoring

whether the face is at the center or not. The Itti-Koch attention value which is shown in blue color represents the images which have more brightness information in the saliency regions obtained through Itti-Koch static saliency attention model underlying human cognition system.

It can be clearly seen that SIFT attention value has its higher values centered around scene based images (such as buildings). Indeed, it works well for identifying user interest images related to textured scenes (refer figure 4.5, table 4.1).

The face user attention model for Ma *et al.* works well to identify the images where face is at the centre underlying the hypothesis that people often attend images where the face is at centre. This is evident from the images index 4,5,6 and 7. The combination of SIFT and Itti-Koch helps to find the scene images (refer sample images index 8,9,10). The combination of face attention value and group attention value help together to capture group images with people at centre (refer image index 5). The SIFT points alone helps to capture textured images such as buildings (refer sample image index 2). The good and bad examples of saliency map is shown in the figure 4.6. It can be noticed that the saliency map of images b, j, q seems not to be reasonable though it is computed based on Itti-Koch attention model.

Group based attention value is good enough to identify the group images. (refer sample images index 5,4). So, as a whole it can be clearly seen that each of these features help together to capture interest via a relevance feedback. In this way we use both bottom up as well as top down methodology. Now let us provide the illustration of each of attention feature extraction with an example.

Figure 4.6: *Saliency plotted on Sample Images Dataset*

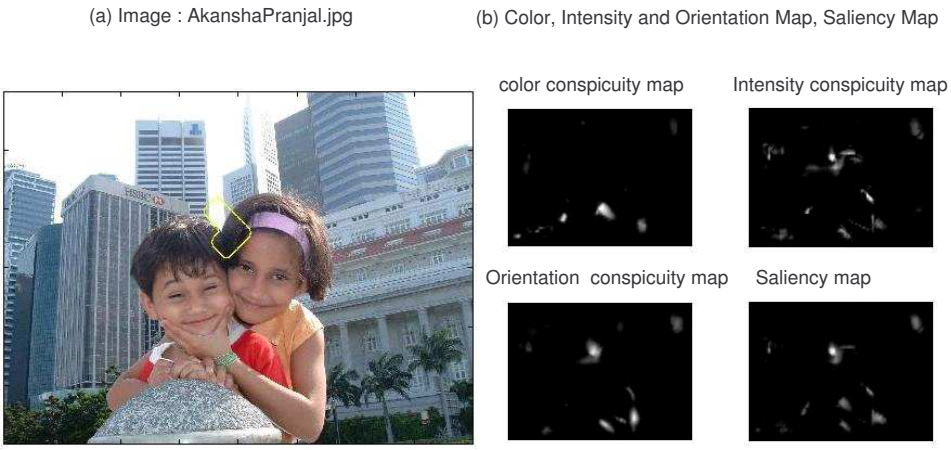

Figure 4.7: *Example Saliency map*

4.3.2 Illustration of attention feature extraction with examples

Here we would take an example image as shown in figure 4.7. We used static saliency method to find the saliency map (refer figure 4.7(a)) which is the combination of color, contrast and orientation map (refer figure 4.7(b)). The area of the saliency region is calculated and brightness information for the saliency region is found. So ultimately this method will help to find the images which has the saliency region with more brightness information. This is based on the fact that people will attend more towards the region which has brightness information according to human cognition system and Ma attention model.[39]

The saliency map can be seen as shown in the figure 4.7(b). For the above example, the obtained Itti-Koch static attention value is 0.030158388. The face attention value for the sample figure 4.10 is 0.062492672. The centre of the detected face is found and multiplied with the corresponding index of position (0:8) as seen in earlier figure

3.3. The idea is that if the face is at the centre, then the weightage is given more. This is actually a normalized Gaussian template with mean centered around the area of the frame. If the number of faces is more than 2, then the individual weights are considered. The SIFT attention value of the image shown in figure 4.12 is 0.0081. The group attention value of the image shown in figure 4.11 is 0.6. The attention value takes the maximum value of 1 to 5 faces in our database and 0 to zero faces.

To show how attention values are useful in identifying the image clusters, we picked 100 images with 25 images for each category such as buildings, portraits, group images and scenaries. The sample set images from each of those categories is shown in figure 4.9 A, B, C and D respectively. In the figure, IV represents Itti-Koch Static Attention value, FA represents Face Attention value, SIFT represents Scale Invariant Feature value and GV represents Group Attention value. The attention values for each of those categories are shown (refer to figure 4.8). It is noted that SIFT value is useful in identifying building images, GV value in finding the group based images, FA in identifying portraits and IV in finding the saliency regions based on cognition system. Also, it is noted that face attention value is higher than group based attention value though they are based on facial information. This is due to factors such as size of the face detected, number of faces and position of the face whether it is at center or at corner ends of the image. Though it is seen that IV does not have peak values significantly, it aids the system in retaining cognition based visual information such as brightness of the pixels associated with saliency regions.

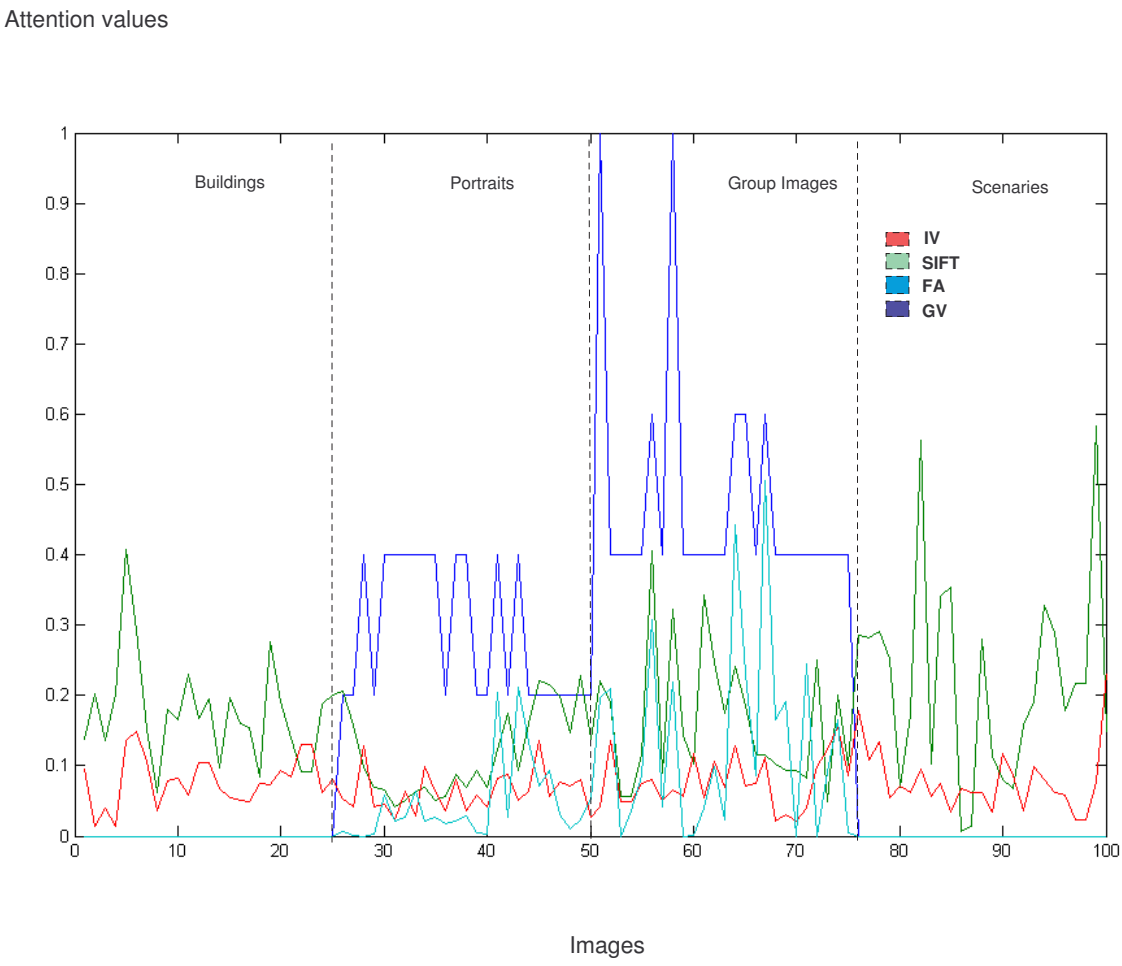

Figure 4.8: *Attention values for different image categories*

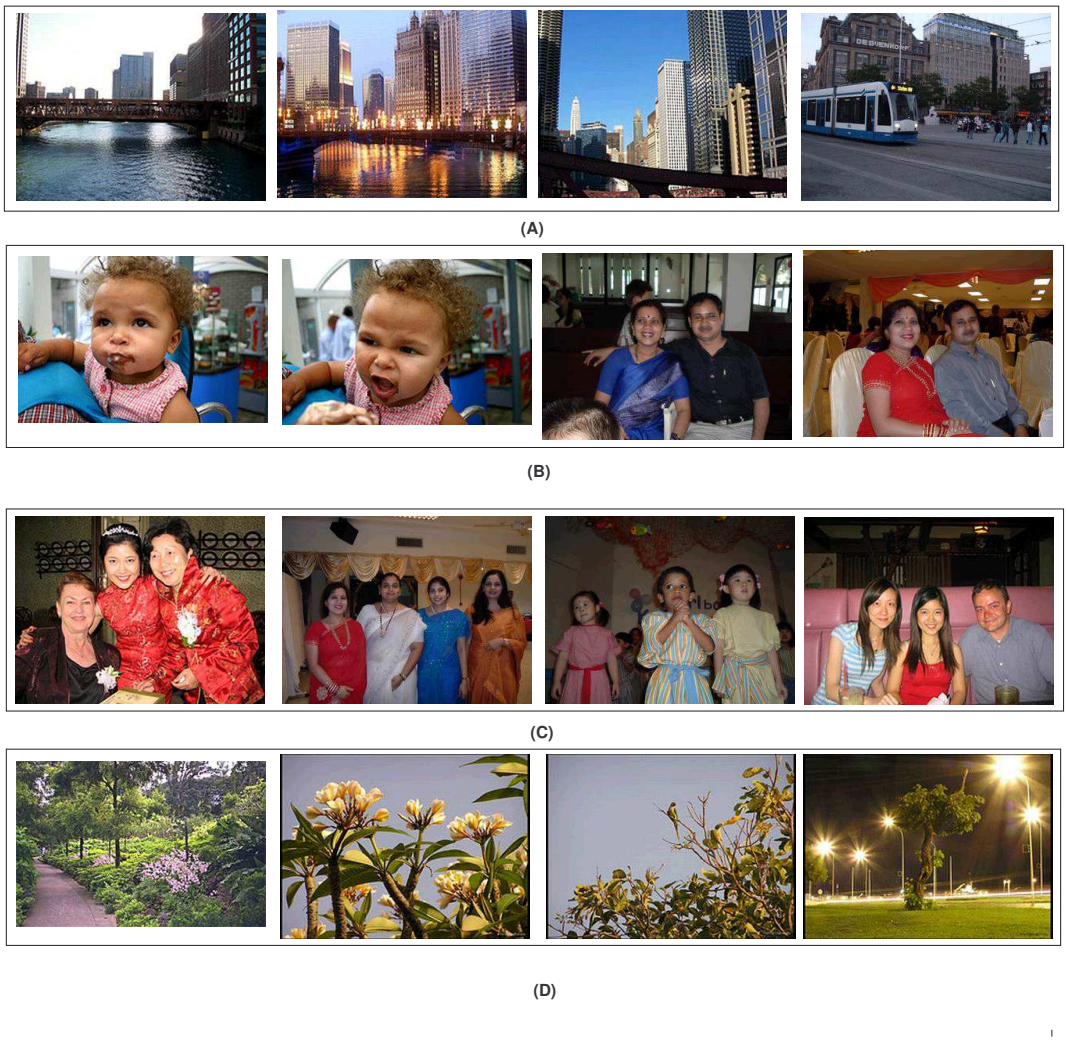

Figure 4.9: *Sample image set from different image categories*

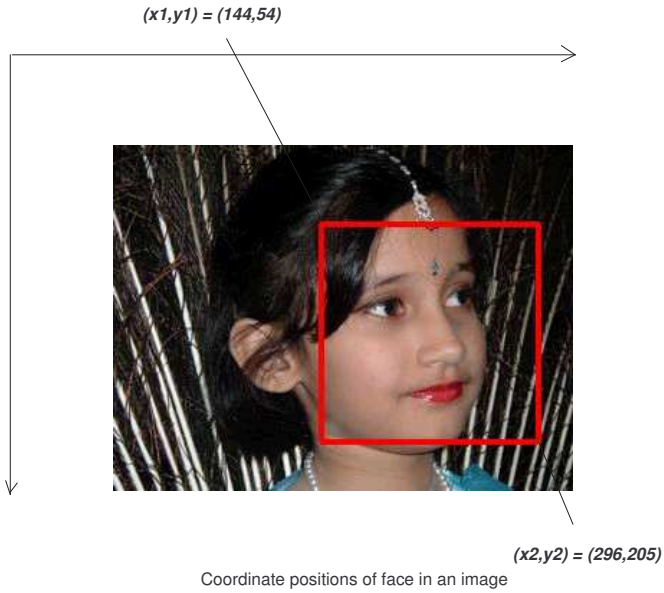

Figure 4.10: *Face coordinate position*

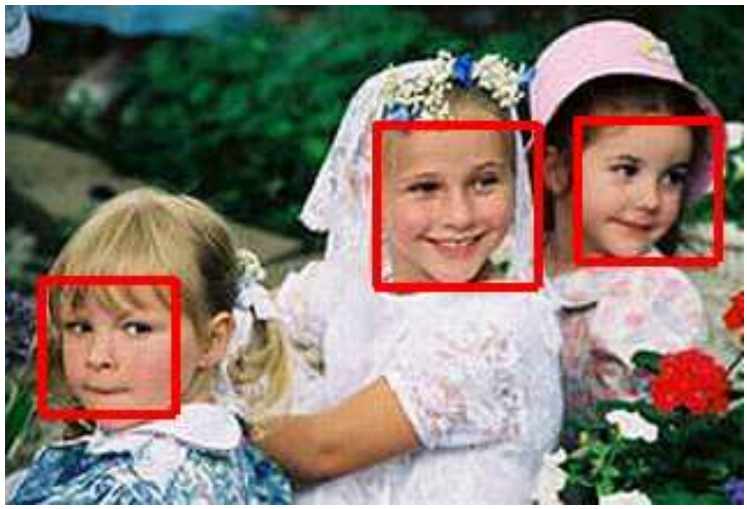

Figure 4.11: *Group photo sample image*

Figure 4.12: *SIFT points of an image*

4.4 Performance

Average time for computation:

The graph is plotted for each of attention method vs average computation time in figure 4.13. The time taken for calculation of Itti-Koch region, SIFT points, FA (face attention) and number of faces are 2355 *sec*, 16588.6 *sec*, 2200 *sec* and 5249 *sec* respectively.

Average time for computation for RFB Query: The time taken for 1R (First Round), 2R (Second Round), 3R (Third Round) are given as follows: We conducted 4 trials and the average run time for 1 R is 2.25 sec, 2R is 1.45 sec and 3R is .8 sec. The average run time is plotted as shown in the figure 4.14.

Average time for computation of NIDs: For the sample 20 nid images, the computation time for finding non identical duplicates is 1

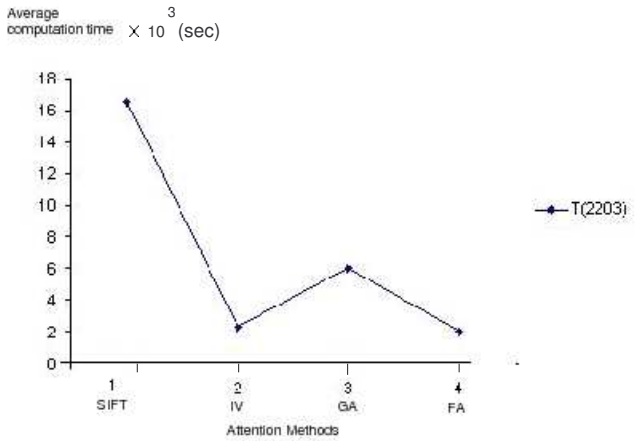

Figure 4.13: *Average time for computation Vs. Attention methods*

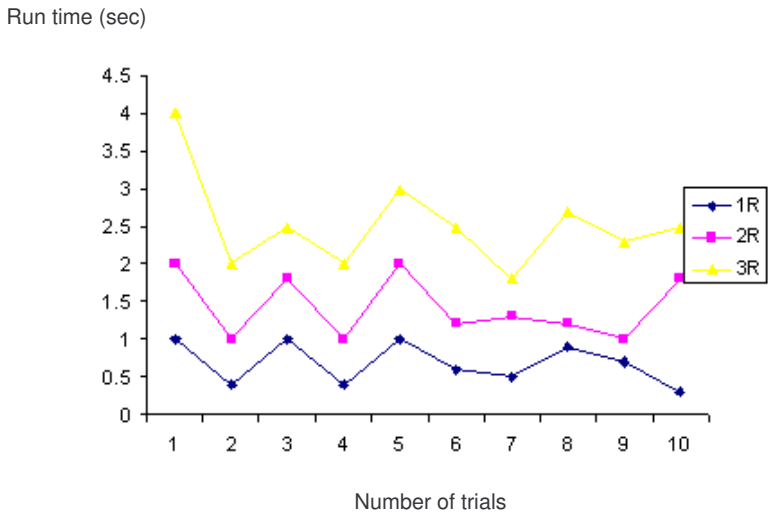

Figure 4.14: *Runtime for computation of RFB*

hrs 01 *min i.e* 3660 sec.

Conclusion: Thus, the following observations are made from the performance study.

- The feature computation is time-consuming and hence needs to be done offline,
- Pre-computation leads to acceptable runtime performance and
- The computation for NID is also time-consuming.

4.5 User Study

We have collected a data-set of 2023 images (a combination of personal collections and downloaded pictures from Flickr). To assess our system, we performed subjective analysis of our system through a user study. The aim of the user study is to understand the synergy between people expectations, need and the real time experience that the system gives. We prepared a questionaire that aims at judging the image quality attribute via subjective scores ranging from 1 which is minimum to 7 which is maximum. The questionaire is shown in Appendices.

Experiment 1: In the first user study, twenty three human subjects participated and are asked seven questions as shown in the questionaire pertaining to quality attributes of our system as well as Flickr. The users are asked to give IAQ (Image Attribute Quality) score ranging from 1 (lowest) to 7 (highest) purely based on the attributes such as a) enjoyability b) surprise c) aesthetics d) desirability e) RFB quality f) RFB usefulness and g) ease of use *etc.* Each user is given a gift as

Table 4.2: *User Study Results: Part I*

System	$IQA1$	$IQA2$	$IQA3$	$IQA4$	$IQA5$	$IQA6$	$IQA7$
Flickr	4.5	4.4	4.9	4.3	N/A	5.5	5.4
EChronicles	4.7	3.9	5.2	4.8	4.8	5.2	6.3

a token of appreciation for spending his/her time in the survey. The user study I results are provided in table 4.2.

$IQA1$ - Enjoyability $IQA2$ - Surprise $IQA3$ - Aesthetics $IQA4$ - Desirability $IQA5$ - RFB Quality $IQA6$ - RFB usefulness $IQA7$ - ease of use.

Experiment 2: In the second user study, ten subjects participated and are asked same set of 7 questions. The difference is that the user is shown random number of images without relevance feedback (query analysis and retrieval) component. To be fair, we informed the user that relevance feedback is available but it actually was non-functional in the system and the user study is subsequently made to find image attribute quality. The user study results for our system with and without relevance feedback are as seen in table 4.3.

Discussion: Social network analysis (as used in Flickr) is useful in finding interesting photos for most people in a large group. However, it does not imply a personalized interest. It is interesting to note that both of the results for Flickr as well as EChronicle Attention system are comparable. It can be noted that interest can be at a personal level or a generic level. The former strategy has been used in EChronicle Attention system while later in Flickr. The study results from table 4.2 revealed that surprise is higher for Flickr than our system. This we understand to be correct since Flickr has wide variety of collections of unique images marked as interesting. Also this is true since it is

not based on content based processing and hence likely of having less redundant information (similiar kind of pictures *i.e* obtained via relevance feedback in our echronicle attention system). A content based methodology combined with RFB (as used by our system) can provide comparable performance to social network analysis. People valued the utility of RFB. RFB Quality (score scale 4.8 out of 7). Aesthetics and enjoyability in EChronicle attention system is high due to

(a) Attention features

(b) RFB (Relevance feedback mechanism)

People valued the utility of RFB. Personalized interest can be achieved via a Relevance Feedback mechanism. Bottom up approach + top down (pseudo-relevance feedback) methodology is deployed for estimating user's interestingness. Social network analysis is better for surprise. Use of multiple attention features can increase the variety of user's interestingness which can be captured. Social network analysis + attention features would be better. But the enjoyability and aesthetics are slightly higher for our attention system than Flickr. *However, it is to be noted that about half of our dataset comes from Flickr.* Thus, the potential reasons we believe for this slightly higher value is due to relevance feedback mechanism and saliency features. This is also confirmed with the RFB usefulness score. However, people ranked the quality of RFB as 4.8 in the scale out of 7. This is basically the user intention score. Overall our system is able to capture user intention as indicated by the score for RFB quality.

The desirability is higher with RFB than without RFB. This is also higher when compared with Flickr as it can be seen in the first user

Table 4.3: *User Study Results: Part II*

System	$IQA1$	$IQA2$	$IQA3$	$IQA4$	$IQA5$	$IQA6$	$IQA7$
EChronicle (without RFB)	2.9	4.3	3.8	3.1	1.9	2.2	6.1
EChronicle (with RFB)	4.7	3.9	5.2	4.8	4.8	5.2	6.3

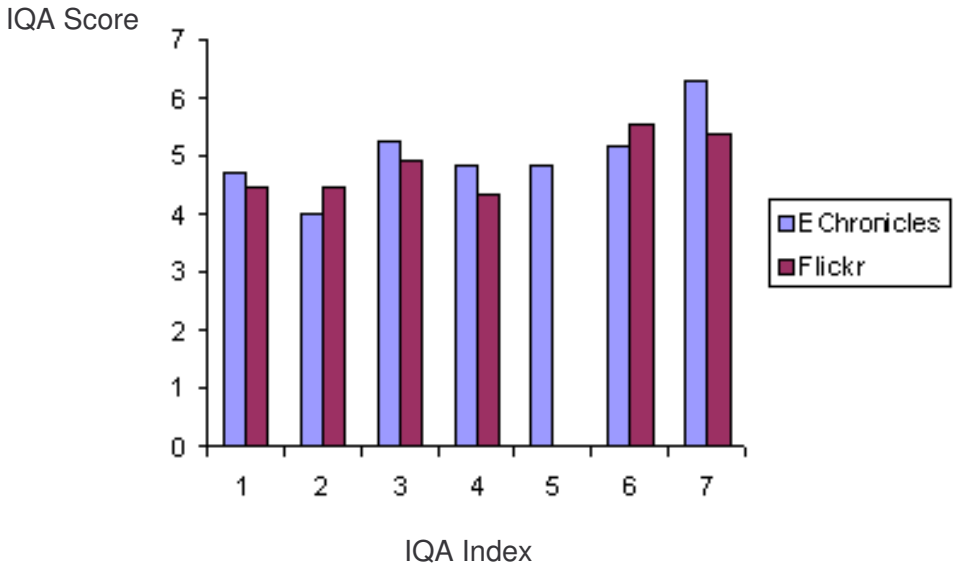

Figure 4.15: *Results: User Study EChronicle Attention System Vs. Flickr*

study. This shows that people do like to have images based on attention features and relevance feedback mechanism.

The study reveals the importance of RFB usefulness as 5.5 out of scale 7 in Flickr. The users are asked to give RFB usefulness score based on how useful the system would be if RFB is used in the system. RFB usefulness refers to the extent to which RFB would be useful in the Flickr. Surprise, which is defined as unexpectedness in terms of quality, may not arise when there is redundancy of information. The interestingness in the sense of surprise can be increased by removing non identical duplicates while still retaining the intention of the user.

Figure 4.16: *Results: User Study EChronicle Attention with and without RFB*

4.6 Illustration of Non Identical Duplicate Detection

We discussed how surprise can be increased by removing Non-Identical Duplicates (refer to example 3 of section 2.2). Though Relevance feedback and NIDs (Non-Identical Duplicates) are at odds with each other, we try to remove the NIDs such that surprise is increased while user intention is maintained. For example, consider the scenario as mentioned in third example of section 2.2 . As the relevance feedback process goes on, the images fall under one of the mentioned attention categories and given the assumption people select images either belonging to scene, people or group based category, the images move towards the cluster. Now, we would show the utility of SIFT method in detecting non

identical duplicates. As introduced in section 2.4, non identical images are not exactly similiar but almost similiar, (for example refer figure 4.17). Here, one can notice that the number of key points in image a) *AkanshaPranjal1.jpg* and image b) *AkanshaPranjal2.jpg* are 14468 and 10079 keypoints and the match between them is 490 since they are NIDs. However, if one consider the image *Pradeep.jpg* which has 18785 keypoints and the image *AkanshaPranjal2.jpg*, number of matches between them is as less as 5. Thus it can be inferred that the number of matches between NID pairs is high whereas for non-NID pairs, it is significantly less. We have examined the efficacy of SIFT method in detecting the non identical duplicates of videos [64]. An sample image with scale invariant points has been shown in figure 4.10.

Now, we we will show how NID metric varies for NID images and non-NID images.

Though there are many NID images available in our dataset, we have considered 10 pairs of images to demonstrate the NID utility (refer figure 4.20). The NID metric is obtained by the equation 3.5 as discussed in section 3.3. The NID metric has been plotted for our sample dataset as shown in figure 4.18 and 4.19. The peak value represents the highest match score between the images. The diagonal represents the NID metric obtained between the image pairs $(1, 1)$, $(2, 2)$.. up to $(20, 20)$ respectively. The key point values and the matches obtained between corresponding key points are given in the table 4.4.

In the table 4.4, the pairs $(1, 2)$, $(3, 4)$, $(5, 6)$, $(7, 8)$ and $(9, 10)$ are NID pairs.

Now, we would provide the conclusions and futurework in the last

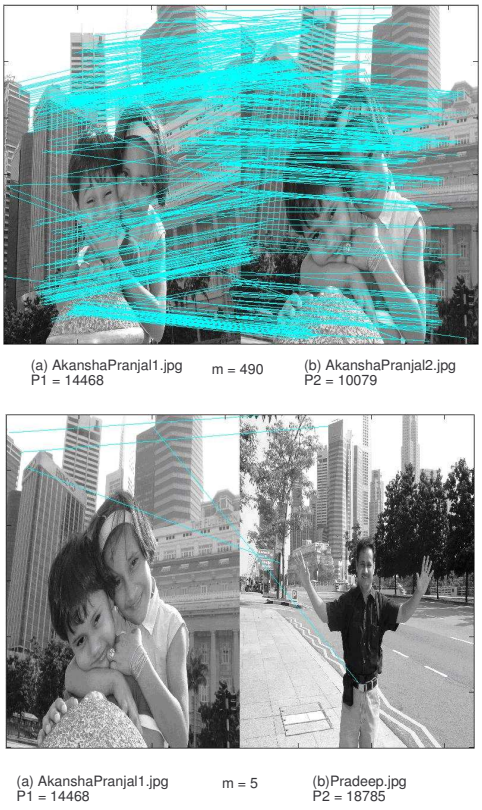

Figure 4.17: *Matching NIDs and Non-NIDs using SIFT*

Table 4.4: *SIFT matches for NIDS : Sample Images*

NID image index pair	*Keypoints*1	*Keypoints*2	*Matches*
1,1	420	420	420
1,2	420	406	10
1,3	420	21	1
1,4	420	21	0
1,5	420	300	4
2,2	406	406	406
2,3	406	21	1
2,4	406	21	0
3,3	21	21	21
3,4	21	21	4
3,5	21	300	1
3,6	21	311	0
3,7	21	194	0
3,8	21	225	0
4,4	21	21	21
4,5	21	300	0
5,5	300	300	300
5,6	300	311	117
6,6	311	311	311
6,7	311	194	1
7,7	194	194	194
7,8	194	225	9
8,8	225	225	225
8,9	225	394	1
9,9	394	394	394
9,10	394	286	92
9,11	394	555	0
10,10	286	286	286
10,11	286	555	0
10,12	286	682	0

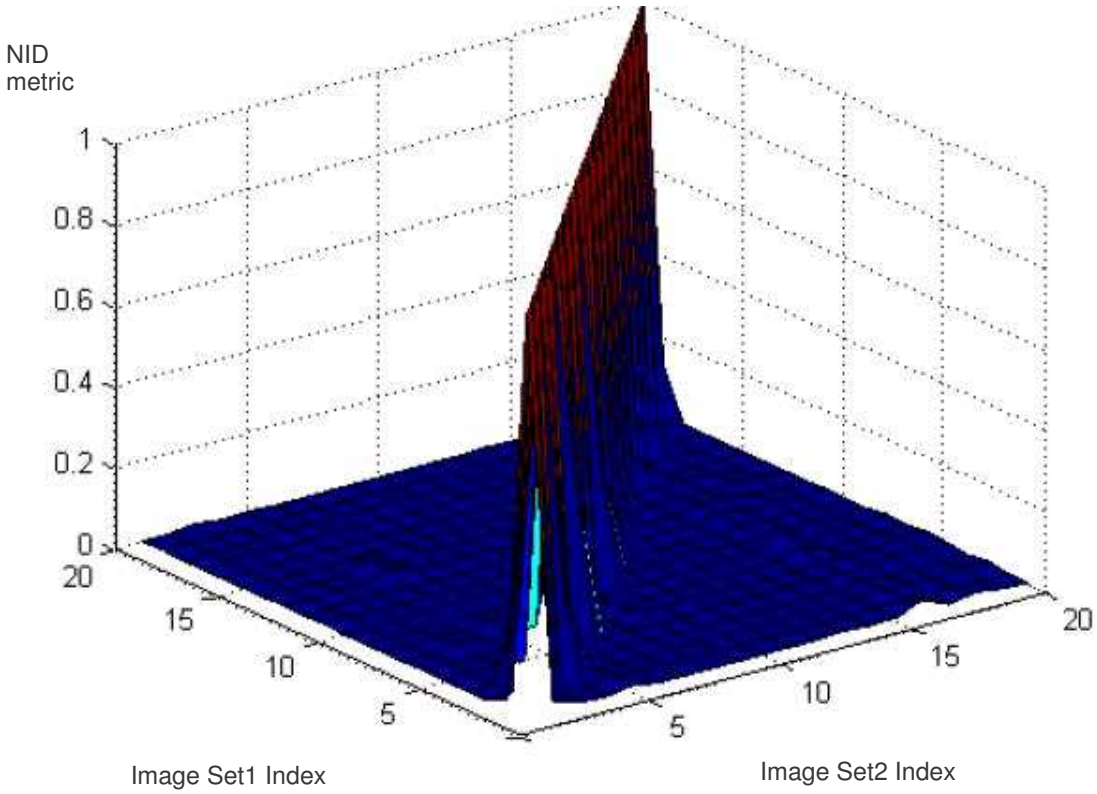

Figure 4.18: *NID metric for sample image set*

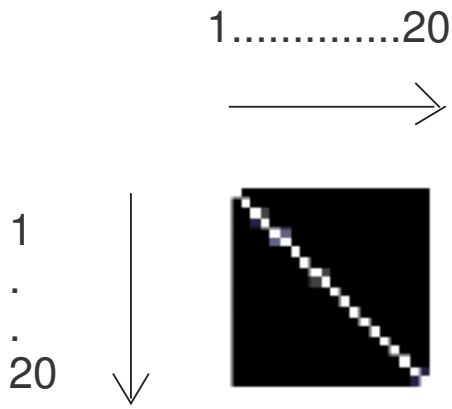

Figure 4.19: *NID metric matrix*

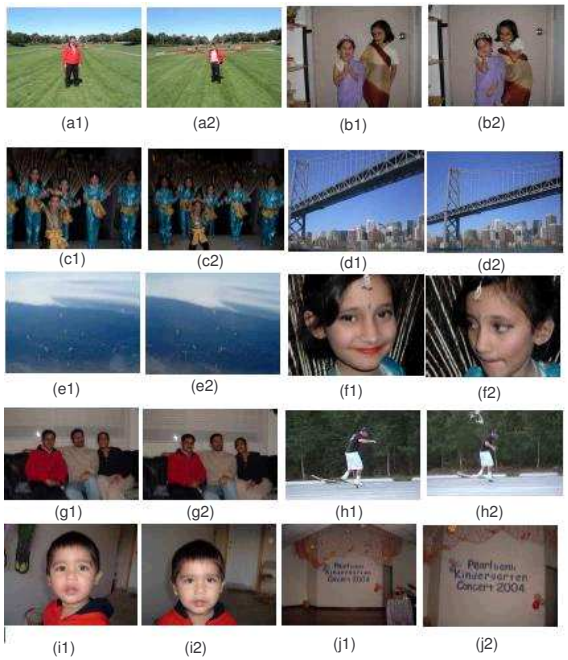

Figure 4.20: *NID Images in our dataset*

chapter5.

Chapter 5

Conclusion and Future Work

5.1 Summary

In this work, we have proposed a framework for finding "interesting" images by deploying visual attention and feedback mechanism. We implemented a system known as "EChronicles Attention system" based on the proposed framework. EChronicles Attention system is an integration of attention models and a pseudo relevance feedback mechanism.

- Since ***interesting objects*** are often attended to by human beings, we built a system based on visual attention features.
- Since interestingness also depends on his/her own interpretation and his/her accumulation of experience, we obtained the necessary information by asking user itself what he needs by a relevance feedback.

The attention models deployed in the system are such as Itti-Koch saliency attention model for finding bottom up saliency attention value, face centric attention model for finding whether face is at centre, group based attention model for finding whether user is interested in more number of faces, and SIFT based attention model for finding images that have scale invariant feature points.

The limitations of this work include:

(a) limited number of visual attention cues, for example current model only considers scene based, group based, face based and itti-koch based attention values. As discussed earlier, the system is fixed with attention features but dynamically changes based on how user selects the images.

(b) simple query analysis and retrieval part to identify how top down approach can be used. The idea is to use user's information to identify user's interest information.

(c) a relatively small database of 2023 images was used.

A relevance feedback method is adopted based on the feature weights obtained from selected images. For evaluation purposes, we conducted an user study and it is compared with Flickr (which actually introduced the notion of interestingness) in terms of our own interestingness quality attributes. The attributes that we consider for defining interestingness notion are a) interpretation and experience, b) surprise, c) beauty, d) aesthetics and e) desirability.

Our comparison with Flickr is centered around the above mentioned interestingness attributes. From the user study results, it is inferred that combination of attention features and relevance feedback mech-

anism is better for showing "interesting" images rather than using any adhoc methodology. We propose a new application using a novel combination of attention models and relevance feedback mechanism to identify "user interest" of images.

The social network analysis would be better choice if higher value of surprise attribute needs to be obtained. Also, interestingness in the notion of surprise can be captured by removing non identical duplicates according to Bayesian surprise theory. The idea is that when there is no information redundancy, surprise might be high. Thus as a whole, in this research work, we investigated whether content based interestingness coupled with relevance feedback mechanism would be useful or not. Typically, this is achieved by an user study on our system with and with out having relevance feedback. We finally conclude that by using both attention features and relevance feedback mechanism, user interest can be identified in an even better manner.

5.2 Recommendations for Future work

One possible direction of future work could be extending our framework to handle combination of both social network analysis and our methodology. In terms of framework itself, the system is fixed with predefined attention features while it can adapt and learn the environment. The other potential attention features which could cover other aspects of user interest can be explored. Face recognition can be done to find the exact person and then search for interestingness with in that cluster or groupings. A more comprehensive user study can be performed on a realistically sized database.

Bibliography

[1] D.Berg, S. Boehnke, R. Marino, P. Baldi, D. Munoz and L. Itti, Characterizing Surprise in Humans and Monkeys, *In: HFSP (Human Frontier Science Program) 6th Annual Meeting*, Paris, France, 2006

[2] O.Boiman and M.Irani, Detecting Irregularities in Images and in Video, *10th IEEE International Conference on Computer Vision ICCV*, pp 462-469, ISBN 0-7695-2334-X, 2005

[3] S.ButterField, E.Costello, C.Fake, H.Begg, C.James, M.Sergeui, and E.Schachter, Interestingness ranking for media objects, *United states Patent Application No 0060242139*, 2006

[4] S.ButterField, E.Costello, C.Fake, H.Begg, C.James, M.Sergeui, and E.Schachter, Media Object metadata association and ranking, *United states Patent Application No 0060242178*, 2006

[5] E.Chang, C.Li and J.Wang, P.Mork and G.Wiederhold, Searching Near-Replicas of Images via Clustering, *SPIE Multimedia Storage and Archiving Systems IV*, pp 281-292, Vol 3846, 1999

[6] L.Chen, X.Xie, X.Fan, W.Ma, H.Zhang and H.Zhou, A visual attention model for adopting images on small displays, *Multimedia Systems, Springer-Verlag*, 2003

[7] M.Cooper, J.Foote, A.Girgensohn and L.Wilcox, Temporal Event clustering for digital photo collections, *ACM Transactions on Multimedia Computing, Communications, and Applications*, pp 269-288, 2005

[8] M.Dubinko, R.Kumar, J.Magnani, P.Raghavan and A.Tomkins, Visualizing Tags over Time, *In Proceedings of the 15th International Conference on World Wide Web WWW*, pp 193-202, 2006

[9] D.R.Edgington, D.Walther, D.E.Cline, R.Sherlock, K.A.Salamy, A.Wilson and C.Koch, Detecting Visual Events in Underwater Video using a Neuromorphic Saliency-based Attention System, *Eos Trans. AGU, 84(46), Fall Meet. Suppl.,* Abstract H11F-0912, 2003

[10] D.R.Edgington, I.Kerkez, D.E.Cline, D.Oliver, M.A.Ranzato and P.Perona, Detecting, Tracking and Classifying Animals in Underwater Video, *IEEE International Conference on Computer Vision and Pattern Recognition CVPR*, 2004

[11] D.R.Edgington and D.Walther, Detection of visual events in underwater video using a neuromorphic saliency-based attention system, *Workshop on Neuromorphic Engineering*, 2002

[12] J.Gemmell, G.Bell, R.Leuder, S.Drucker and C.Wong, My LifeBits: fulfilling the Memex vision, *ACM Multimedia MM*, pp 235-238, 2002

[13] J.Gemmell, A.Aris and R.Leuder, Telling Stories with my lifebits, *IEEE International Conference on Multimedia and Expo ICME*, 2005

[14] J.Gemmell, R.Leuder and G.Bell, The MyLifeBits Lifetime Store, *ACM SIGMM 2003 Workshop on Experiential Telepresence ETP*, 2003

[15] J.Gray, What next?: A Dozen Information Technology Research goals, *Microsoft Technical Report*, MS-TR-99-50, 1999

[16] M.Guironnet, N.Guyader, D.Pellerin and P.Ladret, Static and Dynamic Feature-based Visual Attention Model: Compared to Human Judgement, *Communications of ACM* , Vol 49, 1, 2006

[17] A.Hampapur and R.Bolle, Comparison of distance measures for video copy detection, *IEEE International Conference on Multimedia and Expo ICME*, 2001

[18] M.Hoffman, D.Grimes, A.Shon and R.Rao, A probabilistic model of gaze imitation and shared attention, *Neural Networks*, Vol 19, 3, pp 299-310, 2006

[19] S.C.H.Hoi, A Unified Log-Based Relevance Feedback Scheme for Image Retrieval, *IEEE Transactions on Knowledge and Data Engineering*, Vol 18, No 4, 2006

[20] Y.Hu, X.Xie, Z.Chen Wei, Y.Ma, Attention Model Based Progressive Image Transmission, *Proceedings of the 33rd Hawaii International Conference on System Sciences*, Vol 3, pp 3037, 2000

[21] L.Itti and P.Baldi, Bayesian Surprise Attracts Human Attention, *Advances in Neural Information Processing Systems NIPS*, Vol 19, pp 1-8, 2005

[22] L.Itti and C.Koch, A Comparison of feature Combination Strategies for Saliency based Visual Attention Systems, *In SPIE Conference on Human Vision and Electronic Imaging IV. SPIE*, Vol 3644, pp 373-382, 1999

[23] L.Itti and C.Koch, Computational modeling of visual attention, *Nature Reviews: Neuroscience, Macmillan Magazines Limited*, Vol 2, Issue 3, pp 194-203, 2001

[24] L.Itti and P.Baldi, A Principled Approach to Detecting Surprising Events in Video, *Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR 05*, pp 631-637, 2005

[25] L. Itti, N. Dhavale and F. Pighin, Photorealistic Attention-Based Gaze Animation, *In Proceedings of IEEE International Conference on Multimedia and Expo ICME*, pp 1-4, Jul 2006

[26] A.Jaimes, S.F.Chang and A.C.Loui, Detection of non-identical duplicate consumer photographs, *Pacific Rim International Conference on Multimedia*, vol 1, pp 16-20, 2003

[27] A.Jaimes, S.F.Chang and A.C. Loui, Duplicate detection in consumer photography and news video, *Proceedings of the tenth ACM international conference on Multimedia MM*, pp 423 - 424, 2002

[28] W.James, Classics in the History of Psychology, Chapter 11, http://psychclassics.yorku.ca/James/Principles/prin11.htm

[29] M.S.Kankanhalli, J. Wang and R. Jain, Experiential Sampling on Multimedia Data Streams, *IEEE Transactions on Multimedia*, Vol 9, 1, 2006

[30] M.S.Kankanhalli, J.Wang and R.Jain, Experiential based sampling on Multiple Data Streams, *Proceedings of the ACM International Conference on Multimedia MM*, 2003

[31] Y.Ke, R. Suthankar and L. Huston, Efficient Near-duplicate Detection and Sub-Image Retrieval, *ACM International Conference on Multimedia MM*, 2004

[32] P.Kim, U.Gargi and R.Jain, Event based Multimedia Chronicling systems, *Proceedings of the Second ACM Workshop on Continuous Archival and Retrieval of Personal Experiences CARPE*, pp 1-12, 2005

[33] K.Kishida, Experiment on Pseudo Relevance Feedback Method using Taylor Formula at NTCIR-3 Patent Retrieval Task, *Proceedings of the Third NTCIR Workshop*, 2003

[34] O.Komogortsev and J.Khan, Predictive Perceptual Compression for Real Time Video Communication, *Proceedings of the 12th annual ACM international conference on Multimedia MM*, pp 220-227, 2004

[35] H.C.Lau, R.D.Rogers, P.Haggard and R.E.Passingham, Attention to Intention, *Science*, 303, 1208, DOI: 10: 1126/science. 1090973, 2004

[36] L.Ledwich and S.Williams, Reduced SIFT Features For Image Retrieval and Indoor Localisation, *Australian Conference on Robotics and Automation*, 2004

[37] D.G.Lowe, Distinctive image features from scale invariant key points, *International Journal of Computer Vision*, pp 91-110, 2004

[38] J.Luo and A.Singhal, On Measuring Low-Level Saliency in Photographic Images, *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition CVPR*, Vol 1, pp 84-89, 2000

[39] Y.Ma, X.Sheng Hua, L.Lu and H.Zhang, A generic framework of user Attention Model and Its Application in Video Summarization, *Journal of the IEEE Transactions on Multimedia*, Vol 7, No 5, pp 907-919, 2005

[40] Y.Ma and L.Lu, H.Zhang and M.Li, A User Attention model for video summarization, *Proceedings of the ACM Multimedia MM*, pp 533-542, 2002

[41] Y.Ma, H.Zhang, Contrast-based Image Attention Analysis by Using Fuzzy Growing, *Proceedings of the Eleventh ACM International Conference on Multimedia MM*, Vol 49, Issue 1, 2006

[42] K.Mikolajczyk and C.Schmid, A performance evaluation of local detector and descriptors, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 27, No 10, pp 1615-1630, 2005

[43] M.C.Mozer and M.Sitton, Computational Modeling of Spatial Attention, *In H. Pashler (Ed.), Attention London UCL Press 1996*, pp 341-393, 1996

[44] H.Muller, W.Muller, D.M.Squire, S.M.Maillet and T.Pun, Strategies for positive and negative relevance feedback in image retrieval, *Technical Report*, University of Geneva, Jan 30, 2001

[45] V.Navalpakkam and L.Itti, A Goal Oriented Attention Guidance Model, *Proceedings of the Second International Workshop on Biologically Motivated Computer Vision: Lecture Notes In Computer Science*, pp 453-461, 2525, Dec 2002

[46] V.Navalpakkam and L.Itti, Modeling the influence of task on attention, *Vision Research*, Vol 49, Issue 1, 2006

[47] A.Parker, Seven Deadly Colors, *Free Press*, 2005

[48] Shared attention detection system and method, United States Patent 7106204, http://www.freepatentsonline.com/7106204.html

[49] K.Porkaew and K.Chakrabarti, Query refinement for multimedia similarity retrieval in MARS, *Proceedings of the seventh ACM international conference on Multimedia MM*, pp 235 - 238, 1999

[50] P.Quelhas, F.Monay, J.M.Odobez, D.Gatica Perez, T.Tuytelaars and L.Van Gool, Modeling Scenes with Local Descriptors and Latent Aspects, *Proceedings of the Tenth IEEE International Conference on Computer Vision ICCV*, Vol 1, pp 883 - 890, 2005

[51] N.Ouerhani and H.Hugli, Robot Self-Localization using Visual Attention, Proceedings 2005, *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, June 27 - 30, 2005

[52] O.Ramstrom and H.Christensen, Distributed Control of Attention, Book: Attention and Performance in Computational Vision, *Lecture Notes in Computer Science*, Vol 3540, pp 639-648, 2005

[53] S.Robertson and K.Sparck-Jones, Relevance weighting of search terms, *Journal of the American Society of Information Science*, Vol 27, pp 129-146, 1976

[54] K.Rodden and K.R.Wood, How do People Manage their Digital Photographs?, *Proceedings of ACM Conference on Human Factors in Computing Systems ACM CHI*, pp 409-416, 2003

[55] M.Rugg, Cognitive Neuroscience, *First MIT Press edition*, pp 221, 1997

[56] U.Rutishauser, D.Walther, C.Koch and P.Perona, Is bottom-up attention useful for object recognition, *Vision Research*, Vol 49, Issue 1, 2006

[57] T.Sakai, T.Manabe and M.Koyame, Flexible Pseudo Relevance Feedback via Selective Sampling, *ACM Transactions on Asian Language Information Processing*, Vol 4, No 2, pp 111-135, 2005

[58] A.Salah, E.Alpaydin and L.Akuran, A Selective Attention Based Method for Visual Pattern Recognition, *In Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp 96-100, 2001

[59] J.Satinover, The Quantum Brain, *Wiley Publications*, 2002.

[60] J.Sivic and A.Zisserman, Video Google: A Text Retrieval Approach to Object Matching in Videos, *International Conference on Computer Vision*, pp 1470-1477, 2003

[61] Taylor, The Natural History of the Mind, Chapter 12, pp 182-184, *Martin Secker and Warburg Limited*, 1979

[62] Z.Stejic, Y.Takama and K.Hirota, Relevance feedback based image retrieval interface incorporating region and feature saliency patterns as visualizable image similiarity criteria, *IEEE Transactions on Industrial Electronics*, pp 839- 852, Vol 50, Issue 5, 2003

[63] J.Teevan, W.Jones and B.B.Bederson, Personal Information Management: Introduction, *Communications of ACM*, Vol 49, Issue 1, 2006

[64] K.Vaiapury, P.K.Atrey, M.S.Kankanhalli and K.R.Ramakrishnan, Non Identical Video Duplicate Detection using SIFT method, *International*

Conference on Visual Information in Engineering, VIE, pp 537-542, 2006

[65] D.Walther, D.R. Edgington and C.Koch, Detection and Tracking of Objects in Underwater Video, *IEEE International Conference on Computer Vision and Pattern Recognition CVPR*, 2004

[66] D.Walther, U.Rutishauser, C.Koch and P.Perona, Selective visual attention enables learning and recognition of multiple objects in cluttered scenes, *Computer Vision and Image Understanding*, pp 41-63, 2005

[67] D.Walther, Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics. *PhD thesis, California Institute of Technology, Pasadena, CA*, Chapter 2, 23rd February 2006

[68] G.Wang, T.T.Wong and P.A.Heng, Real-Time Surveillance Video Display with Salience, *Proceedings of the third ACM international workshop on Video surveillance and sensor networks VSSN*, pp 37-44, ISBN:1-59593-242-9, 2005

[69] H.Wolf and D.Deng, How interesting is this? Finding interest hotspots and ranking images using an MPEG-7 visual attention model, *Annual Colloqium of Spatial Research Centre, SIRC*, 2005

[70] J.M.Wolfe and T.S.Horowitz, What attributes guide the deployment of visual attention and how do they do it?, *Nature Reviews Neuroscience*, pp 1-7, 2004

[71] J.M.Wolfe, How might the rules that govern visual search constrain the design of visual displays, Brigham and Women's Hospital Harvard Medical School.

[72] R.Yan, A.Hauptmann and R.Jin, Multimedia search with pseudo relevance feedback, *International Conference on Image and Video Retrieval CIVR*, 2003

[73] Yahoo's Flickr, http://www.flickr.com

[74] Yahoo's Flickr Interestingness http://www.flickr.com/explore/interesting/7days/

[75] S.Yu, D.Cai, J.Wen and W.Ma, Improving Pseudo Relevance feedback in Web Information Retrieval Using Web page segmentation, *Proceedings of The Twelfth International World Wide Web Conference WWW*, 2003

[76] H.Zhang, R.Rahmani, S.R.Cholleti and S.A.Goldman, Local image representations using pruned salient points with applications to CBIR, *Proceedings of the 14th annual ACM international conference on Multimedia MM*, pp 287-296, 2006

[77] X.Zhou and T.Huang, Relevance Feedback in Image Retrieval: A Comprehensive Review, *Proceedings of CVPR Content based Access of Image and Video Libraries CBAIVL*, 2003

[78] ASP tutorial http://www.w3schools.com/asp/default.asp

[79] VBScript tutorial http://www.w3schools.com/vbscript/default.asp

[80] Opencv http://www.intel.com/technology/computing/opencv/

Appendix

Part 1: User Study for EChronicles Attention System:

Please circle the scale 1-7 for the following questions

1. How will you rate the system in terms of enjoyability? Echronicles Attention System

   1 2 3 4 5 6 7

   where 1 is less enjoyable and 7 is highly enjoyable

2. Can you rate the system in terms of surprise (unexpectedness in terms of quality)?

   Echronicles Attention System

   1 2 3 4 5 6 7

   where 1 is low surprise and 7 is highly surprise

3. How will you rate the system in terms of beauty/aesthetics (*sensory emotional values) while browsing the system?

   Echronicles Attention System

   1 2 3 4 5 6 7

   where 1 is low on aesthetics and 7 is high on aesthetics

4. How will you rate the system in terms of desirability ?

Echronicles Attention System

1 2 3 4 5 6 7

where 1 is low desirability and 7 is high desirability

5. How well does the relevance feedback process work?

   Echronicles Attention System

   1 2 3 4 5 6 7

   where 1 is very bad and 7 is extremely well

6. How useful is the Relevance feedback in the system?

   Echronicles Attention System

   1 2 3 4 5 6 7

   where 1 is useless and 7 is very useful

7. How easy is the system to use?

   Echronicles Attention System

   1 2 3 4 5 6 7

   where 1 is very difficult and 7 is very easy

* - modern aesthetics attribute (refer: http://en.wikipedia.org/wiki/Aesthetics)

Part 2: User study for Flickr:

1. How will you rate the system in terms of enjoyability?

   Flickr 1 2 3 4 5 6 7

   where 1 is less enjoyable and 7 is highly enjoyable

2. Can you rate the system in terms of surprise (unexpectedness in terms of quality)?

   Flickr 1 2 3 4 5 6 7

   where 1 is low surprise and 7 is highly surprise

3. How will you rate the system in terms of beauty/aesthetics(sensory emotional values) while browsing the system?

   Flickr 1 2 3 4 5 6 7

   where 1 is low on aesthetics and 7 is high on aesthetics

4. How will you rate the system in terms of desirability ?

   Flickr 1 2 3 4 5 6 7

   where 1 is low desirability and 7 is high desirability

5. Do you think RFB will be useful to have in Flickr?

   Flickr 1 2 3 4 5 6 7

   where 1 is useless and 7 is highly useful

6. How easy is the system to use?

   Flickr 1 2 3 4 5 6 7

   where 1 is not easy and 7 is extremely easy