Exploiting Human Visual Attention for Automatic Image Selection and Annotation

Tina Walber

July 2014

Vom Promotionsausschuss des Fachbereichs 4: Informatik der Universität Koblenz-Landau zur Verleihung des akademischen Grades **Doktor der Naturwissenschaften (Dr. rer. nat.)** genehmigte Dissertation.

Vorsitzender des	
Promotionsausschusses:	Prof. Dr. Ralf Lämmel
Promotionsvorsitz:	Prof. Dr. Susan P. Williams
Berichterstatter:	Prof. Dr. Steffen Staab
	Prof. Dr. Ansgar Scherp

Datum der wissenschaftlichen Aussprache: 18.07.2014

Institute for Web Science and Technologies
University of Koblenz-Landau
Germany

Veröffentlicht als Dissertation an der Universität Koblenz-Landau.

Abstract

The availability of digital cameras and the possibility to take photos at no cost lead to an increasing amount of digital photos online and on private computers. The pure amount of data makes approaches that support users in the administration of the photo necessary. As the automatic understanding of photo content is still an unsolved task, metadata is needed for supporting administrative tasks like search or photo work such as the generation of photo books. Meta-information textually describes the depicted scene or consists of information on how good or interesting a photo is.

In this thesis, an approach for creating meta-information without additional effort for the user is investigated. Eye tracking data is used to measure the human visual attention. This attention is analyzed with the objective of information creation in the form of metadata. The gaze paths of users working with photos are recorded, for example, while they are searching for photos or while they are just viewing photo collections.

Eye tracking hardware is developing fast within the last years. Because of falling prices for sensor hardware such as cameras and more competition on the eye tracker market, the prices are falling, and the usability is increasing. It can be assumed that eye tracking technology can soon be used in everyday devices such as laptops or mobile phones. The exploitation of data, recorded in the background while the user is performing daily tasks with photos, has great potential to generate information without additional effort for the users.

The first part of this work deals with the labeling of image region by means of gaze data for describing the depicted scenes in detail. Labeling takes place by assigning object names to specific photo regions. In total, three experiments were conducted for investigating the quality of these assignments in different contexts. In the first experiment, users decided whether a given object can be seen on a photo by pressing a button. In the second study, participants searched for specific photos in an image search application. In the third experiment, gaze data was collected from users playing a game with the task to classify photos regarding given categories. The results of the experiments showed that gaze-based region labeling outperforms baseline approaches in various contexts. In the second part, most important photos in a collection of photos are identified by means of visual attention for the creation of individual photo selections. Users freely viewed photos of a collection without any specific instruction on what to fixate, while their gaze paths were recorded. By comparing gaze-based and baseline photo selections to manually created selections, the worth of eye tracking data in the identification of important photos is shown. In the analysis of the data, the characteristics of gaze data has to be considered, for example, inaccurate and ambiguous data. The aggregation of gaze data, collected from several users, is one suggested approach for dealing with this kind of data.

The results of the performed experiments show the value of gaze data as source of information. It allows to benefit from human abilities where algorithms still have problems to perform satisfyingly.

Zusammenfassung

Mit der zunehmenden Verbreitung digitaler Kameras nimmt die Anzahl der aufgenommen Fotos drastisch zu. Fotos werden sowohl für den privaten Gebrauch aufgenommen und auf eigenen Festplatten gespeichert, als auch im Internet verbreitet. Die Verwaltung dieser großen Datenmengen stellt eine Herausforderung dar, bei der Benutzer zunehmend unterstützt werden müssen. Die automatische Analyse von Bildinhalten anhand von Algorithmen ist ein ungelöstes Problem und kann kaum die Bedürfnisse menschlicher Nutzer erfüllen. Daher werden häufig Metainformationen genutzt, um z.B. abgebildete Szenen textuell zu beschreiben oder Bewertungen zu Fotos zu speichern. Im Rahmen dieser Arbeit wird untersucht, wie diese Metainformationen ohne zusätzlichen Aufwand für Benutzer generiert werden können. Dazu werden Augenbewegungen von Benutzern mit einem Eyetrackinggerät erfasst und die daraus abgeleitete visuelle Aufmerksamkeit als Informationsquelle genutzt.

Aufgrund von fallenden Hardwarepreisen bei gleichzeitig zunehmender Konkurrenz sind die Preise für Eyetracker in den letzten Jahren stark gefallen und ihre Bedienbarkeit wurde vereinfacht. Es wird angenommen, dass die Erfassung von Blickdaten bald mit alltäglichen Geräten wie Laptops möglich sein wird, während Benutzer z.B. verschiedenen Beschäftigungen mit digitalen Bildern nachgehen. Die Auswertung dieser Blickinformationen erlaubt es, Informationen ohne zusätzlichen Aufwand für den Menschen bereitzustellen.

Im ersten Teil dieser Arbeit wird untersucht, ob durch die Auswertung von Blickinformationen, Schlagworte Bildregionen zugewiesen werden können, mit dem Ziel abgebildete Szenen zu beschreiben. Insgesamt wurden drei Experimente durchgeführt um die Qualität der Beschreibungen zu untersuchen. Im ersten Experiment entschieden Teilnehmer durch das Drücken bestimmter Tasten, ob ein gegebenes Objekt auf einem Foto zu sehen war. In der zweiten Studie suchten Benutzer mit einer simulierten Bildersuche nach Fotos von bestimmten Objekten. Im dritten Experiment klassifizierten Benutzer Fotos bezüglich gegebener Objektnamen in einem eyetracking-gesteuerten Spiel. In jedem Experiment wurden die Augenbewegungen aufgezeichnet und die Objektnamen bzw. Suchbegriffe entsprechenden Bildregionen zugeordnet. Die Ergebnisse zeigen, dass in den verschiedenen Anwendungen Bildinhalte durch Blickpfadanalysen sinnvoll beschrieben werden können. Im zweiten Teil wird die Identifizierung von interessanten Fotos in einer Sammlung von Fotos anhand von Blickbewegungen erforscht, mit dem Ziel, Benutzern individuelle Fotoauswahlen anzubieten, nachdem sie Fotos in einer Sammlung betrachtet haben. Durch den Vergleich der unter Einbeziehung der visuellen Aufmerksamkeit automatisch erstellten Auswahlen mit manuell von den Benutzern erstellten Auswahlen, wird das Potential von Blickinformation in der Erkennung wichtiger Fotos deutlich.

Die Ergebnisse dieser Arbeit zeigen das große und bisher ungenutzte Potential der impliziten Nutzung von Blickdaten. Es kann von menschlichen Fähigkeiten profitiert werden, besonders dort, wo Algorithmen die menschliche Wahrnehmung noch lange nicht simulieren können.

Acknowledgments

I thank Prof. Steffen Staab for giving me the opportunity to obtain a PhD and the scientific education. I thank Prof. Ansgar Scherp for supervising and supporting me during the last years. I wish to acknowledge the help provided by the members of the topic group Human Computer Interaction. Thank you to all WeST colleagues for participating in my experiments and giving feedback in several talks and discussions; particularly Leon Kastler, Christoph Kling, Jérôme Kunegis, Julia Perl, and, Christoph Schaefer. Thank you to Prof. Ramesh Jain and the members of his research group for the fruitful research stay. I thank the student assistants Chantal Neuhaus and Annika Wießgügel for their dedicated work. I would like to express my great appreciation to all volunteers for participating in the experiments performed in the scope of this thesis. I thank Tanja and Thorsten Prinz, Lena Gieseke, Hendrik Ziezold, and Katrin Heinen for their feedback on this dissertation. I would like to offer my special thanks to my family and friends for supporting me during the last years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The interaction of users with digital data can provide information about the data. The viewing behavior can be recorded with an eye tracking device and analyzed with the goal to gain information on the viewed stimuli. This thesis deals with the viewing of photos and with the creation of meta-information on these photos to improve the understanding of the underlying semantics.

In this section, two big challenges in the creation of photo meta-information are discussed, and the approach for creating them by means of gaze data, suggested in this work, is introduced. Subsequently, the research questions and an overview of the contributions of this work are given. The outline of this thesis completes this section.

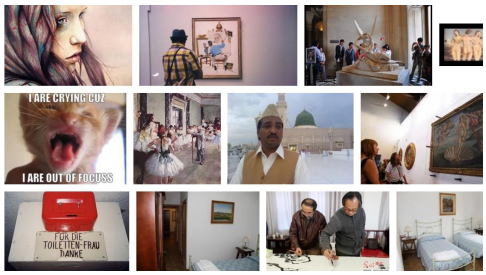
1.1 Motivation

The amount of digital data is increasing in consequence of technical achievement in the last decades. One area of growth is digital photography and the spread of photos on the Web and on hard drives or other personal devices. The first digital cameras for the mass market were introduced in the 1990s, and until the mid-2000s, most users switched from analog to digital cameras. Today, even most mobile phones and tablet PCs have integrated cameras, and a majority of the digital photos are taken with these devices, as shown by a statistic published by the photo sharing web portal Flickr. Here, several Apple iPhone devices are listed as the most popular camera models.[1]

Digital photography leaded to a huge amount of photos because digital photos are easy to take and do not cost anything. On newswiretoday.com, an estimate of 3.5 billion cameras and camera phones in use worldwide and over 1 trillion personal digital photos stored on computers, on mobile devices and on external web servers was published[2]. The pure amount of data makes it hard to keep track of the photos, for example, when searching for a specific photo or trying to gain an overview of a photo collection.

[1] http://www.flickr.com/cameras, last visited September 27, 2013
[2] http://www.newswiretoday.com/news/84943/, last visited December 20, 2013

(a) Input for *Search by image*. Visitors taking photos of the Mona Lisa with their mobile devices.



(b) The first results of the search results list for input image (a).

Figure 1.1: Example for Google's *Search by image* functionality.

1.1.1 Search for Photos and Labeling of Photos

The search for specific photos is not trivial, and different approaches are available for online and file system search. Photos that are available online are often accessed directly by image search engines such as Google Images[3] or Yahoo! Image Search[4] or by search functionalities provided by specific photo storage platforms such as Flickr.[5] Photos that are stored in file systems can be managed by on-board functionality such as browsing the folder structure or search for photos by their file names. Applications supporting the user in these search tasks on personal computers are available; for example, Google Picasa[6] and iPhoto[7] offer additional search functionalities.

The search can be performed based on the pixel information of the photos, such as color schema or structure. These approaches are called *content-based*. For example, the Google Images tool *Search by image* belongs to this category of search applications. An example photo has to be uploaded to the web page, and the *Search by image* functionality delivers a search results list with photos that are visually similar to the input photo. In Figure 1.1, an example search is depicted. Although the photos in the search results list show some similarity concerning color schema and structure, it is obvious that the visually similar photos often not show a semantically similar content.

The search for photos that were taken within a certain period or at a concrete location can be performed based on photo context information, if this information is available, for example, as EXIF information attached to the photo. The approaches based on the context information are called *context-based*. A search function, using the capture time of a photo, can be performed, for example, on Flickr. Geo information are used, for example, by *iPhoto*, a photo management tool offered by Apple, where the photos can be displayed on a map. These

[3] http://images.google.com
[4] http://de.images.search.yahoo.com
[5] http://www.flickr.com
[6] http://picasa.google.com
[7] https://www.apple.com/mac/iphoto/

approaches depend on the availability of the context information, which has to be stored during the photo capturing or added by hand.

Despite the aforementioned possibilities, most users intend to search for photos based on their semantic content, that is, the depicted scene and objects. The process of understanding visual scenes is complex. The image, projected on the retina during the visual perception or on the chip of a digital camera, is ambiguous because the same image can be built by different objects and scenes during the projection of the 3-D world on a 2-D representation. The so-called *inverse projection problem* has to be solved to derive real-world objects from depictions [Gol13]. The depicted objects can be only partly depicted, covered by other objects, or just blurred. In addition, the perspective from which an object is depicted strongly influenced the photo. The so-called *view point invariance* describes the different perspectives of one object depending on the viewing direction [Gol13], whereby unusual view points can complicate the identification of an object.

Thus, the complex process of human visual reception and image understanding is based on factors such as a high level of abstraction, background knowledge, and emotions. It is described in more detail in Section 2.1. This cognitive process solves the aforementioned problems but can hardly be reproduced by computer algorithms. The semantic gap [SWS+00] characterizes the differences between this human image understanding and description with a high level of abstraction, and the digital image representation that can be performed by algorithms, which mainly uses low-level pixel data and the results of image processing.

Fully automatic approaches that deal with the understanding of photo content are far from delivering results that are at the level of human understanding of visual content [VJ01]. This is why often metadata is used for describing what is depicted on a photo. Tjondronegoro and Spink [TS08] showed in their survey that the majority of search engines for multimedia contents such as photos are still based on keywords. Google image search extracts the keywords from the information surrounding an image on a web page and other context information such as the image name. Other applications provide the possibility to manually add keywords as tags, for example, Flickr and Google Picasa. This manual creation of tags can be very tedious, especially when considering the huge amount of photos. The need of high-quality metadata is obvious but the creation is a challenge.

The search for photos is highly influenced by the semantic gap, as the understanding of photo contents is needed for supporting users (e.g., in terms of high-quality tags) but can hardly be delivered fully automatically. This results in the following:

Challenge 1 *The automatic understanding of photo semantics and the creation of photo descriptions as metadata are challenging tasks for computer algorithms.*

1.1.2 Interestingness of Photos

The large amount of personal digital photos makes the management of photo collections an increasingly challenging task. Users easily take hundreds of photos

during vacations or personal events such as weddings or birthday parties. Often, selections of "good" photos have to be created to reduce the amount of photos stored or shared with others [FKP+02, KSRW06, NF09, RW03]. While users enjoy certain photo activities such as the creation of collages for special occasions such as anniversaries or weddings, these tasks are seen as "complex and time consuming" for normal photo collections [FKP+02]. Algorithms are needed for support users in the creation of selections.

In the automatic selection of photos from large collections, choosing photos only based on pixel and context information is often not sufficient. Even knowledge of the semantic content of photos cannot perform the task of creating a satisfying subset because more complex and abstract criteria come into play. The selection process itself can be very individual, and the decision on which photos should be part of a selection is based on several factors such as the depicted scene, the quality of the photo, and the depicted persons but also on factors such as interestingness and personal preferences, which can hardly be determined by algorithms. The selection criteria can be very diverse, and even objective evaluations of which photos represent a photo set are difficult. Figure 1.2 shows two photos of the Taj Mahal; the decision which photo is "better" can barely be made without context information (e.g., is the depicted person a family member of the person who should make the decision?) and the knowledge of personal preferences (e.g., does a person like funny photos?). Thus, photo selections are very individual and accordingly hard to be created automatically.

These problems can be summarized in the following statement:

Challenge 2 *The automatic identification of photos interesting or important to users is a challenging task.*

Figure 1.2: Two photos of the Taj Mahal — which photo is the better one? The decision is very individual and depends, for example, on a possible relation to the depicted girl on the right photo.

1.1.3 Analysis of User Behavior

Online image repositories are highly frequented. The website flickr.com, an image and video hosting community, has about 13 million users per day.[8] The users search for images, view images, upload images, and tag images. Besides the analysis of the photo content at pixel level and the context information, user behavior can be analyzed for benefiting from the intuitive photo understanding of humans and for understanding their preferences. A lot of research has been done in the area of collecting implicit user feedback for improving retrieval quality in online search. The relevance of web pages is derived from the browsing behavior of search engine users. Implicit feedback systems have the benefit that the cost of explicit user ratings or feedback can be removed. Every user interaction with the system can contribute to an implicit rating. In the work of Claypool et al. [CLWB01], implicit rating methods for recommender systems were presented. Joachims [JGP+05] and Jung et al. [JHW07] used click-through data of search engine users as implicit source of information to determine the importance of search results. Other information such as how long a document was displayed were investigated, for example, by Agichtein et al. [ABD06]. Zhang et al. [ZGS+10] identified attention durations, click-through rates, and mouse movements as implicit feedback measures. Cursor movements were used in the detection of important and characterizing sentences of an article, which then can be used for improved summarizations by Lagun et al. [LAGA14]. Yao et al. [YMNL13] presented an approach of video tagging based on click-through data. From this data, relations between videos are explored and used for annotating online videos, for example, by assigning tags from similar videos. Comparing the results, for example, with those from feature-based similarity measures showed that the approach is promising.

The mentioned approaches cannot consider which contents a user perceives, for example, on a web page. It is limited to the analysis of performed actions such as clicking, while the main part of the behavior is unknown, such as the visual scanning of web pages, the reading of texts, or the viewing of photos. Knowing more about this kind of user behavior can extend the existing approaches by learning more about objects such as images from the user behavior. This information can be gained from analyzing the user's eye movements. Eye tracking devices analyze the position of the users' eyes relative to a monitor as well as the viewing direction to compute the fixated points on a computer screen.

Photos on private computers are usually viewed or skimmed at least once. Most users like the viewing of photos, for remembering an event, or while sharing experiences with others. Frohlich et al. [FKP+02] showed that the work with photos has a big emotional payoff for the users. Rodden and Wood [RW03] showed that photo collections are browsed frequently. However, the frequency of browsing decreases over time. The viewing of photos usually happens shortly after the capturing or downloading to a computer. Even in this scenario, information can be extracted when knowing the viewing behavior. Photos that obtain more visual attention can be assumed as being more interesting or important.

[8] http://websiteworths.com/flickr.com, statistics July 27, 2010

1.1.4 Eye Tracking Approach for Obtaining Implicit Information from Users

In the past, professional eye tracking devices were expensive and were solely used in laboratory experiments. But the interest in eye tracking technology is growing, and the technological progress is impressing. It can be assumed that eye tracking technology will be wider spread in the near future, for example, in common devices such as laptops or mobile phones. One reason for this assumption is the rapid development of sensors in IT hardware. Nowadays, eye trackers can be developed using low-cost hardware, and more and more open-source eye tracking projects appear. Another reason is that the prices for professional system are falling, and the hardware becomes easier to use. For more details on the development of eye tracking hardware, see Section 2.2. Based on this assumption, the usage of gaze information in the context of everyday task is reasonable.

The aim of this work is to implicitly gain meta-information about images from the users' viewing behavior during the work with photos. While the user is performing image-related tasks like viewing photos or searching for images, the gaze data is recorded by an eye tracker in the background. The gaze data is analyzed to enrich the image meta-information, without any additional effort from the users. In the experiments, the participants were never told to consciously control their gaze and to fixate, for example, on specific photos or specific regions of the photos. Biedert et al. [BBD10] consider that "a high promising approach is just to observe eye movements of the user during his or her daily work in front of the computer, to infer user intentions based on eye movement behavior." The goal is to benefit from human perception skills for annotating images and to analyze the eye movement to get to know individual preferences.

1.1.5 Challenges Faced in the Use of Eye Tracking Technology

The use of gaze data as source of information involves some limitations and challenges this work has to deal with.

First of all, today's eye tracking technology brings some technical limitations, such as a limited freedom of head movements for the users. Professional devices such as the Tobii X60 eye tracker allow, for example, head movements in dimensions of $44 \times 22 \times 30$ cm. If the user's head leaves this area of head movements, no or only low-quality gaze data can be recoded by the device. This can be restrictive for some users, depending on their behavior. During the eye tracking experiments, it can be necessary to bring the sitting position of the participants to mind. For a small number of users, eye tracking technology has serious problems and the eyes cannot be clearly detected. This can be caused by anatomical characteristics (like distinct strabismus) or corrective eyesight devices such as some contact lenses. These kinds of problems rarely occurred during the experiments and were not further analyzed because of the low impact. In addition, the quality of gaze data is limited in its accuracy. Modern devices has an accuracy of $0.5°$, which corresponds to about 5 mm on the screen.

All aforementioned problems can negatively influence the gaze data quality, and one has to deal with noisy data. The gained data is not comparable with mouse data such as clicks or the pressing of buttons on the keyboard, which is much more exact and controlled. That means, when dealing with gaze data, inaccurate data has always to be considered.

Furthermore, there are biological and/or psychological factors that complicate the gaze analysis. Humans are, for example, capable to recognize visual information in the corner of their eyes with the so-called *covert attention* [Gol13]. Even large objects are usually not scanned completely with the eyes but only parts of them are fixated and the rest are perceived with the extra-foveal vision. Thus, parts of, for example, a photo can be perceived even without directly fixating on it. This problem can also occur vice versa; fixations can be on areas that are not consciously perceived.

When analyzing a human gaze trajectory, it is also not possible to distinguish between fixations that are part of a scanning process and fixations on an object of importance. In addition, the level of concentration on the perceived visual content cannot be measured. Vetegaal [Ver02] said, "Although eye fixations provide some of the best measures of visual interest, they do not provide a measure of cognitive interest." From the information on which area was observed on a photo, it cannot be derived if the information has been processed and was perceived. In addition, it is also not clear why an object caught the visual attention of a human. An area can be fixated because of a given task or because something else caught the observer's attention. This can be a familiar, a weird or a funny object.

The information gained from eye tracking devices can be ambiguous and inaccurate. The challenge is to examine how suitable the information, gained from gaze data, is and if it is reliable enough to be used in annotation tasks.

1.2 Research Questions

In the forgoing discussion on problems in the management of digital photos, two big challenges were identified. In this thesis, the potential of generating information on photos from gaze data is analyzed for these two challenges presented above — the generation of labels describing the depicted scene and the identification of interesting photos. Thus, on the one hand, it is investigated if the human capacity of identifying objects in photos can be exploited to obtain information about objects in images. This information is used for the labeling of image region with the aim to describe the content of photos in more detail. On the other hand, the capabilities of identifying interesting photos in a collection of photos for creating an individual photo selection is examined.

As an exploitative approach, the human behavior is recorded in the form of eye movements and analyzed without additional effort for the users.

Understanding Image Semantics — Region Labeling

Earlier in this section, the importance of labels describing the content of photos was discussed. These labels can be defined at photo level, assigning tags to

images as a whole. A more detailed description of photo contents at pixel level can be worthwhile, for example, for improving image search. Several works showed that a more detailed description of photo contents by region labeling can improve the search [KTS01, KY08, YLZ07]. Region labels can also serve as ground truth data for computer vision algorithms [RTMF08]. Chum and Zisserman [CZ07] found that regions of interest improve the classification and localization in object detection. Region labels are also helpful in recognizing depicted scenes, for example, in photos of indoor scenes [QT09]. The manual assignment of tags to images can already be tedious [Rod99] but the manual region labeling is even more burdensome.

In the first part of this thesis (Sections 4 to 7), the human capacity to intuitively identify objects in photos is exploited with the aim of labeling image regions. Users' gaze paths are analyzed to assign tags to regions with the aim to benefit from this intuitive identification of objects. The potential of this idea is investigated in three experiments. The examples in Figure 1.3 illustrate the performance of the human visual perception. Although the photos show a chair from an unusual perspective (left), only parts of a chair (center), or a designer chair with an unusual form (right), the identification of these objects is not difficult to a human observer but can be difficult for computer algorithms.

Figure 1.3: Photos of chairs — Easy to identify for human viewers but can cause problems for computer vision algorithms because of perspective and cropping, unusual form of appearance, and depicted scene.

The following research questions concerning region labeling are investigated in this thesis:

RQ 1 ***Can information about depicted scenes on photos be gained from gaze data analysis?***

The gaze data shows the visual interest of users. In this work, the question if the data is precise enough to give evidence on the depicted scene is investigated.

This research question is split up into the following sub question:

RQ 1.1 *Is it possible to identify an object, from a given set of objects, by means of gaze data from users who had decided if they can see that specific object on a photo?*

RQ 1.2 *Can the identification be improved when considering inaccurate data?*

RQ 1.3 *Does the aggregation of gaze data gained from several users improve the region identification results?*

RQ 1.4 *Can objects on photos be identified from gaze analysis when no high-quality object regions are given?*

RQ 1.5 *Can the region labeling approach be applied to daily routine tasks such as online image search, with users searching for photos in a simulated web search application?*

RQ 1.6 *Does the approach perform well in a distracting situation?*

Identification of Interesting Photos — Creation of Selections

In the second part of this work (Section 8), the gaze-based approach is applied to the creation of photo selections. Interesting photos from a collection of photos are selected by means of gaze analysis with the goal to create a representative, high-quality selection. Image collections can be used, for instance, for the automatic creation of photo books [SB11] or for creating presentation for friends or family. The visual attention on specific photos in a collections is measured by analyzing the gaze data. This data is interpreted as interest, and the photos with highest results are assumed to be more interesting to the user than the photos with lower results. Photo selections are then automatically created based on the level of interestingness.

The following research questions are investigated in this thesis:

RQ 2 ***Can important photos in a collection be identified from gaze analysis, and is this information worthwhile in the creation of individual photo selections?***

RQ 2.1 *Does a gaze-based selection outperform objective selections based on content and context analysis when comparing the selections with those created manually by the users?*

RQ 2.2 *Does the personal interest in a viewed photo set have an impact on the obtained selection results?*

1.3 Contributions

This thesis is an experiment-driven work. In total, four user experiments were performed to determine the problem of using gaze information in the creation of photo metadata. Three experiments were conducted in the region labeling part, one for the photo selection part. An overview of the research questions and the conducted experiments can be found in Figure 1.4. In total, 141 test execution of four different experiments were performed with 122 unique volunteers.

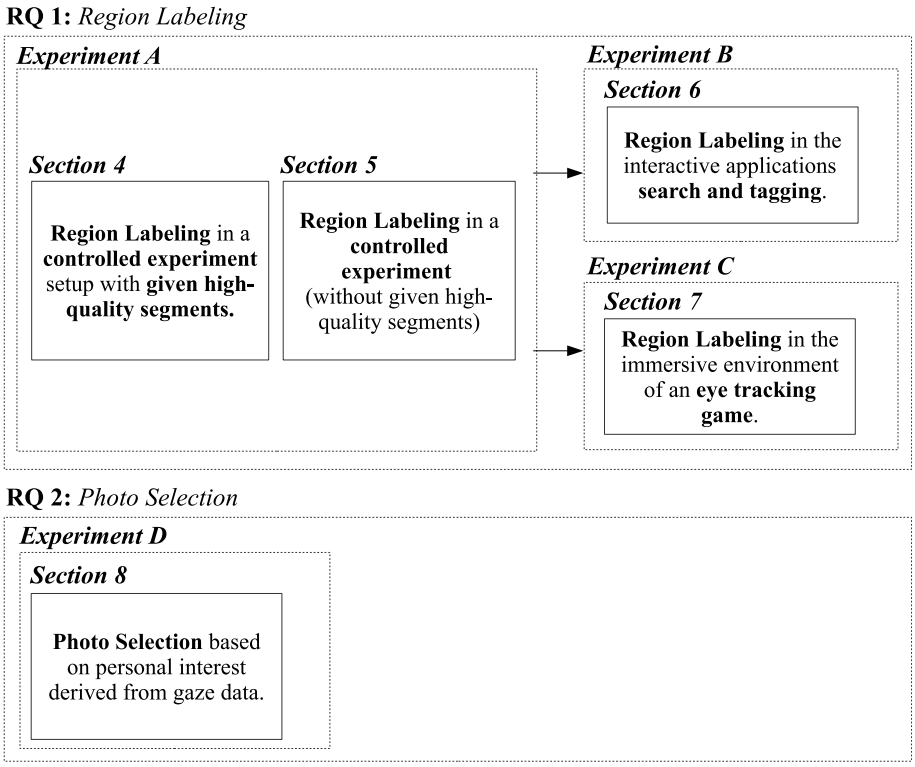


Figure 1.4: Structure of this work structured by the performed experiments. The sections 2 (Background) and 3 (Related work) provide the basis for all experiments and sections.

1.3.1 Contributions to Research Question 1: Photo Region Labeling

Research question **RQ 1** and the sub questions **RQ 1.1** to **RQ 1.6** are dealing with the problem of describing the content of a photo by assigning tags to image regions. Three consecutive experiments ***A*** to ***C*** with in total 100 unique participants were conducted concerning these questions. The assignment of labels to image regions was performed by analyzing the gaze paths of users completing different image tasks. In several steps, it was investigated if the eye movements provide reliable information for performing these assignments.

The main contribution to the first research question 1 is that it can be shown that:

C 1 ***The labeling of image regions is possible by means of gaze data for describing the photos' semantics and it outperforms baseline approaches for region labeling.***

More details on the region labeling approach and its potential are given in the following contributions concerning the more detailed research questions that led to contribution **C 1**.

In the first experiment, ***Experiment A***, gaze data was collected in a controlled experiment. The users had to decide whether they can see a given object on given photos. First, a tag was presented to the participants. Then a photo was displayed on the screen in full-screen mode, and he/she had to decide whether an object described by the given tag could be seen on the photo. The decision was made by pressing a key on the keyboard. For all photos, high-quality segmentations in the form of manually marked objects were given. In this first step in answering **RQ 1.1**, 13 different eye tracking measures were investigated for selecting one of the given objects by means of the recorded gaze data. As a result, it could be shown that 63% of the selected objects were in fact described by the given tag. The results significantly outperformed three baseline approaches on performing photo region labeling based on the given object regions. It can be concluded that

C 1.1 *The identification of an object region for a specific tag (from a given set of manually created object regions) in photos presented in a controlled experiment, outperforms baseline approaches not using gaze data.*

In the gaze analysis performed in ***Experiment A***, two approaches for dealing with the specific characteristics of gaze data were introduced. The region extension considers the possible inaccuracy of gaze data and assumes that fixations that are positioned close to a region could have in fact been on this region. An extension of 13 pixels led to a maximum improvement of 9% compared with the results without region extension. The region extension was thus applied to all following analysis. The weighting of small regions was introduced to compensate the fact that it is more likely that a big region is fixated by chance or during a scanning process than a small region. Thus, the idea is that fixations on small region should have a higher validity. The results for the weighting approach were diverse, the weighting can improve the results but it can also worsen the results. A definition of good parameters for the weighting was difficult. Consequently, the weighting is not applied to following analysis. The findings on research question **RQ 1.2** are concluded as follows:

C 1.2 *The extension of image regions improves the identification of relevant image regions, while the weighting of small image regions can improve the results but the identification of good parameters is difficult.*

In ***Experiment A***, some participants viewed the same photos and decided about the same objects. In the analysis, it was investigated if the aggregation of gaze paths, thus the usage of all fixations from all users in the analysis, improves the results. It pointed out that the results improved with an increasing number of aggregated gaze path. The improvement was 109% when comparing the results for single gaze path analysis with the results for 10 aggregated gaze paths. Related to research question **RQ 1.3**, this result shows that:

C 1.3 *The aggregation of gaze paths of different users improves the labeling results compared to single gaze paths.*

The contributions described before were published in [WSS12] and [WSS13a].

In the previous gaze analysis, manually created image regions that described the depicted objects in the form of polygons were used. In a next step, the approach was extended by using segments gained from automatic image segmentation obtained from a state-of-the-art segmentation algorithm instead of the high-quality polygons. The region labeling task was more challenging as the whole image was segmented and the segmentation was less exact as a consequence of under- and over-segmentation. An additional measure was introduced in this work. It is not based on a segmented image but uses information extracted from so-called heat maps, which represent the areas which obtained the highest visual attention. The data obtained from ***Experiment A*** was used in the analysis. It could be shown that despite the additional severity, the labeling of image regions was possible with an average precision of 56% at pixel level over all photos in the data set. The results significantly outperformed baseline approaches. This work was published in [WSS13b] and answered research question **RQ 1.4** by showing that:

C 1.4 *Region labeling is possible without the availability of high-level object regions.*

In the second experiment, ***Experiment B***, participants were asked to search for photos using a simulated online image search interface. The experiment application was designed to resemble a common online images search consisting of a search query page and a search results list. In the analysis of the data, the denoted search terms were assigned to specific regions in the images of the search results list. Therefore, the previously introduced measures were used. It became apparent that the gaze-based measures significantly outperformed baseline measures. The results were published in [WNS14] and answer research question **RQ 1.5** with contribution:

C 1.5 *Region labeling can be performed in image search scenarios by assigning search terms to image regions by means of gaze analysis.*

In the third experiment, ***Experiment C***, it was investigated if the region labeling approach also performs well in a very different scenario, while the user is playing a gaze-controlled game. As in the first experiment, the task was to decide whether a specific object can be seen on a photo. The photos were classified concerning these object categories by fixating indicated areas on the gaming screen. An additional rating was performed during this classification. While in the first studies, the user had no time constraints and the photos were displayed full screen resp. static in a search results list, the gaze-controlled game *EyeGrab* was developed to demand fast decision making from the participants and to break up the full concentration on the photo viewing. The user was brought into the immersive situation of a game with distractions from the game setup, the gaze control, and the emotional pressure of success and failure. It can be shown that, however, the region labeling can be performed at a precision at pixel level of 61%, outperforming baseline approaches. The game was presented in [WNS12]. The results of the region labeling were published in [WSS14a] and answer research question **RQ 1.6**:

C 1.6 *Region labeling by means of gaze analysis can be performed in the immersive scenario of a gaze-controlled computer game.*

1.3.2 Contributions to Research Question 2: Creation of Photo Selections

The selection of photos based on the analysis of gaze paths of users who viewed photos in a collection was investigated in ***Experiment D***. During the viewing of the photos, gaze paths were recorded and subsequently used in the creation of an individual photo selection. These selections were compared with selections manually created by the user. In the creation process, gaze data and other measures obtained from the photo's content or context information were combined. This work answered research question **RQ 2** and showed that:

C 2 ***Individual photo selections created based on eye tracking information significantly outperforms photo selections not using gaze information.***

This contribution is concluded from finding presented subsequently. The results were published in [WNS+13] and [WSS14b].

In the creation of photo selections, eye tracking measures, content-based measures, and context-based measures were combined by means of machine learning. It turned out that the selections based on all measures and selections based on eye tracking measures significantly outperformed the baseline selections, based only on content and context information with an improvement of up to 22%. Research question **RQ 2.1** can thus be answered by contribution:

C 2.1 *Photo selections based on gaze data significantly outperformed objective selections based on content and context analysis alone.*

The photo collection used in ***Experiment C*** was designed in such a way that it contained photo sets that were of interest to the participants and photo sets that were of less interest. The photos of interest showed the participant itself, its colleagues, or it depicts situation of an event the participant participated in. The capability of the gaze-based photo selection approach was separately investigated for these photo sets for answering research question **RQ 2.2**. The analysis showed that:

C 2.2 *Higher personal interest in the viewed photo sets has a positive influence on the photo selection results.*

1.4 Publications

This thesis provides a number of contributions to the literature about the usage of gaze data in annotating photos and in the identification of interesting photos. Parts of this work were presented in the following main publications:

- Tina Walber, Ansgar Scherp, and Steffen Staab. Smart photo selection: Interpret gaze as personal interest. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems,* CHI'14, pages 2065–2074, New York, NY, USA, 2014. ACM.

- Tina Walber, Chantal Neuhaus, and Ansgar Scherp. Tagging-by-search: automatic image region labeling using gaze information obtained from image search. In *Proceedings of the 19th international conference on Intelligent User Interfaces,* pages 257–266. ACM, 2014.
- Tina Walber, Ansgar Scherp, and Steffen Staab. Exploitation of gaze data for photo region labeling in an immersive environment. In *MultiMedia Modeling*, pages 424–435. Springer, 2014.
- Tina Walber, Ansgar Scherp, and Steffen Staab. Can you see it? two novel eye-tracking-based measures for assigning tags to image regions. In *Advances in Multimedia Modeling*, pages 36–46. Springer, 2013.
- Tina Walber, C. Neuhaus, Steffen Staab, Ansgar Scherp, and Ramesh Jain. Creation of individual photo selections: read preferences from the users' eyes. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 629—632. ACM, 2013.
- Tina Walber, Ansgar Scherp, and Steffen Staab. Benefiting from users' gaze: selection of image regions from eye tracking information for provided tags. *Multimedia Tools and Applications*, pages 1–28, 2013.
- Tina Walber, Ansgar Scherp, and Steffen Staab. Identifying objects in images from analyzing the users' gaze movements for provided tags. In *Advances in Multimedia Modeling*, pages 138–148. Springer, 2012.

Published Data Sets

The data sets on which the publications [WSS13a] and [WSS14a] are based were made available on `http://west.uni-koblenz.de/Research/DataSets/gaze`. The data sets contain the photos used in the experiment, the labeled ground truth image regions, as well as the recorded gaze data.

1.5 Outline

An overview of the research questions and conducted experiments were shown in Figure 1.4. In this overview, also the sections dealing with the indicated topics are given.

First in this thesis, background information on the human visual perception process and on eye tracking technology is given in Section 2 (Background), as far it is needed as foundation for the work. The subsequent Section 3 (Related Work) presents important research approaches and results.

In the Sections 4 to 8, the performed research is presented. The four Sections 4 to 7 are related to the region labeling approach (Research Question **RQ 1**); Section 8 refers to the Photo Selection Topic (Research Question **RQ 2**).

The chapters are related the four experiments performed in this thesis. Sections 4 and 5 are based on ***Experiment A***, which was conducted as a strongly controlled experiment. The analysis was based on manually created regions in Section 4 and on automatic photo segmentation in Section 5. Sections 6 and 7

apply the introduced region labeling approach to different usage contexts. Section 5 describes the research on performing region tagging in search and tagging scenarios investigated in ***Experiment B***, while Section 7 describes the usage in the gaming scenario of ***Experiment C***. In Section 8 the research on photo selection creation based on ***Experiment D*** is presented. Finally, in Section 9, the results of this work are concluded and future work is outlined.

Chapter 2

Background on Eye Tracking and Gaze Analysis

First, a short introduction into the human visual perception is given. Then the techniques for measuring and analyzing gaze data are presented. Considering the focus of this work, this section concentrates on the perception of photos displayed on a computer screen.

2.1 Human Visual Perception

The visual perception is one of the most important elements of the human sensory system. It includes the incidence of reflected light on the retina, the control of the eyes by muscles, and the signal processing in the human brain with the recognition process that includes knowledge and emotions. The eye is mainly an input medium; it receives visual information. In the communication between humans, the eyes can become a source of information, for example, by fixation on objects for directing the attention of other humans to it or by expressing emotions.

While the biggest part of the perception process cannot be measured or predicted as it takes place in the human brain, the movement of the eyes can be observed. The positions of the eyes and the focused points, for example, on a computer screen, deliver unique information on the human visual attention, as the input of the perception process can be derived from it. Only a short overview on the visual perception can be given in this section; for more details, see Goldstein's book *Sensation and Perception* [Gol13].

2.1.1 The Human Visual System

Light, reflected by the surrounding environment or emitted by a computer screen, falls on the back of the human eye, which is covered by the retina.

The retina contains two kinds of light receptors: cones and rods. Rods are highly light sensitive but they do no deliver color information. Cones are less light sensitive but highly color sensitive. The receptors are not equally distributed on the retina. The cones are concentrated in a small center region of the retina, while the rods build an outer ring. The outer ring corresponds to the so-called peripheral vision surrounding a center area of high resolution, called fovea. Parts of an image that are depicted in the center of the retina are thus perceived with a higher resolution than the parts depicted in the outer areas.

Parameters and dimension in the visual perception are usually given as degrees of the visual angle. The center of the visual angle is built by the fovea. The visual angle A is calculated as follows:

$$A = 2 \arctan \frac{O}{2D} \tag{2.1}$$

where O is the size of the scene object and D is the distance between the eye and the object.

The human field of view, thus, the area in which visual information is perceived, has a size of about 180°. The area of highest acuity, where the image is depicted on the fovea, is a circular region of about 2° [Duc07]. For a distance of 60 cm between the eye and the monitor, a circular area on the computer screen with a diameter of about 4 cm can be perceived at highest resolution. This circle is surrounded by a ring called the para-foveal area, from about 2° to about 5°. At 5°, the acuity has decreased to only 50% compared with the foveal vision. The area in which detailed information can be perceived is limited to about 30°, which corresponds to about 30 cm, also given a distance of 60 cm to the computer screen. Outside this area, mainly movement can be perceived.

In Figure 2.1, the different areas of the visual field are visualized on an example photo. The depicted areas were calculated for a photo displayed on a computer screen with 20 cm of height, again for an assumed distance of 60 cm.

Because of the relatively small area of highest resolution, the eyes have to be moved to scan a complete scene. Each eye is controlled by three pairs of muscles [HNA+11]. Seven kinds of eye movements occur during the scanning of a scene [Gol13]. Most of them are not part of the analysis in this work because they are not part of the controlled perception process. An example for those not considered movements are microsaccades, which avoid the visual receptors to be continually simulated by the same input signal and ensure continuously small, jerk-like movements. Another example are tremors, involuntary and rhythmic contraction of the muscle, whose function is not clear but could be caused by imprecise muscle control. The important phases of the eye movements that are considered in this work are fixations and saccades. These phases take place independently from the involuntary, small eye movements mentioned previously. Fixations are periods when the eyes steadily gaze at one point for at least 80-100 milliseconds and in an area of 1 to 2 minutes of arc in amplitude. The normal dwell time of fixations lies between 200 and 400 ms [GSL+02]. Saccades are fast movements between these fixations, when the focal area is relocated. The maximum of the visual information is perceived during the fixations. Thus, the identification of fixations in gaze trajectories is of big importance, and the

Figure 2.1: Illustration of the visual field with (fa) foveal area, (pfa) para-foveal area, and (pa) peripheral area for a photo displayed with a height of 20 cm.

analysis of the fixated points in these periods are mainly used in the analysis of gaze data.

Scan paths are often visualized by depicting fixations and saccades. Fixations are displayed as circles, with the diameter encoding the duration of a fixation. The saccades are displayed as lines, linking the fixations. An example for a visualization can be found in Figure 2.2.

Pupil Size

Besides the information on which point was fixated, the human eye provides additional insights. Psychological studies have revealed that there are correlations between the pupil behavior and emotional states. For example, the results of Partala and Surakka [PS03] show that the pupil size is significantly larger during emotional stimuli than during neutral stimuli. Larger pupil size changes were recognized when participants viewed emotionally arousing photos, compared with neutral photos, by Bradley at al. [BMEL08]. However, it cannot be distinguished between pleasant or unpleasant stimuli.

2.1.2 Perception Process and Information Interpretation

The challenge in the visual perception process is to derive a 3-D scene from a 2-D presentation, when a scene is depicted on the retina of the human eye. The reduction of the dimension comes along with ambiguity — the same image can

(a) (b) (c)

Figure 2.2: Examples for scan paths recoded with an eye tracking device. Fixations are depicted as circles, the radius encodes the duration of the fixations. The lines between the fixations are saccades.

be built by different objects and scenes. That makes the conclusive identification of, for example, 3-D objects in 2-D presentation impossible. The "inverse projection problem" describes this problem. Effects such as blurry images, partly depicted objects, or unusual perspectives aggravate the problem.

The automatic understanding of visual contents is a challenging task for computer algorithms, while humans usually have no problems in perceiving a scene. For humans, the derivation of semantic information from the pure visual stimuli is intuitive. This impressing performance of the human brain is caused by several concepts, investigated in psychological research.

The identification of the gist of a scene can take place within less than a second, even without identifying objects. This is performed based on global image features such as "degree of naturalness" (is the depicted scene natural or artificial?), "degree of expansion," or "color." From this information, inferences on the presented scene can be rapidly drawn. It can be estimated, for example, where a situation takes place and what are the areas containing information. Thus, already the very first fixations on a scene are meaningful, as they are not placed randomly.

In the advanced perception process, the perceived data has to be structured and interpreted for obtaining information on the depicted scene and objects. The main perceptual organization strategies in processing the perceived color information are grouping and segmentation principles. In the grouping step, the visual information is organized based on organization principles. Examples are the "principle of similarity," which describes that similar things are grouped together or the "principle of good continuation," which means that points are assumed as belonging together when they are linked by a straight or only smoothly curbing line. More principles are known and can be found, for example, in [Gol13]. In the figure-ground segmentation, the distinction between

figure and ground takes place. The segmentation follows principles, for example, specific "properties of figure and ground." They are automatically considered and comprise, for example, the occurrence of borders, which separate the figure from the ground, or unformed material, without a specific shape, which is characteristic for the ground. "Image-based factors" (e.g., an object is usually depicted lower in the image than the background) or "subjective factors" (e.g., past experiences or instructions) also play a role.

In the more advanced perception of a visual scene, which is mainly the interpretation of the perceived forms and figures, prior knowledge and experiences are of importance. Besides the aforementioned principles, the human visual system is adapted to the physical characteristics of our environment. One example is the light-from-above assumption, which is grounded on the experience that usually light falls from above on a scene and the shadows are projected accordingly. This knowledge serves the orientation in a scene and the perception of the geometric characteristics. Furthermore, semantics has an influence on how scenes are perceived, as humans have a learned knowledge on which objects occur in which scenes. For example, Torralba et al. [TOCH06] showed the mandatory role of scene context in an experiment with participants searching for objects in real-world images. Considering all these principles, the difficulties of algorithms to simulate human perceptions get clearer.

2.1.3 Visual Attention

Because of the small area in which the perceived image is of high resolution, the eyes have to be moved to perceive a complete scene. Two models build the basis for understanding how a scene is scanned.

On the one hand, **bottom-up models** explain eye movements by saliency in the image itself. Experiments showed that persons who freely view images tend to fixate salient regions [PIKI05, IK00, IK01]. The bottom-up approach considers mainly the incoming data and is thus data-driven. The low-level image features are, for example, brightness, color, or contrasts [PIKI05, IK00, IK01].

On the other hand, **top-down models** consider the viewing process as actively controlled attention. Research has shown that eye movements can be controlled consciously. A very fast visual subsystem delivers a first overview of the viewed scene, based on gist and coarse layout, as shown by Itti [Itt03] and described in the previous section. This can be performed within a fraction of a second. In addition, experience and assumptions on where specific elements are positioned in a scene are automatically considered. After this first impression of the gist of scene, the visual attention follows the individual interest. When a task is given, the interest strongly depends on this task, as shown, for example, by Yarbus [Yar67] in 1967. In his experiment, the gaze paths of one user viewing one image with different tasks were compared. The visualized gaze paths can be found in Figure 2.3. Newer work by Henderson et al. [HBCM07] showed that users with a strong task can even ignore low-level saliency. Saliency can already be ignored during the very first fixations, as shown by Einhäuser et al. [ERK08].

For free-viewing situations, when no task is given to the users, Goldstein summarizes that "we attend to what interests us" [Gol13]. Calvo and Lang have shown that emotional scenes [CL04] attract attention earlier. In addition,

the tendency of humans to fixate faces in images is well-known, and even the identification of parts of the faces from gaze paths can be performed [SBL+09].

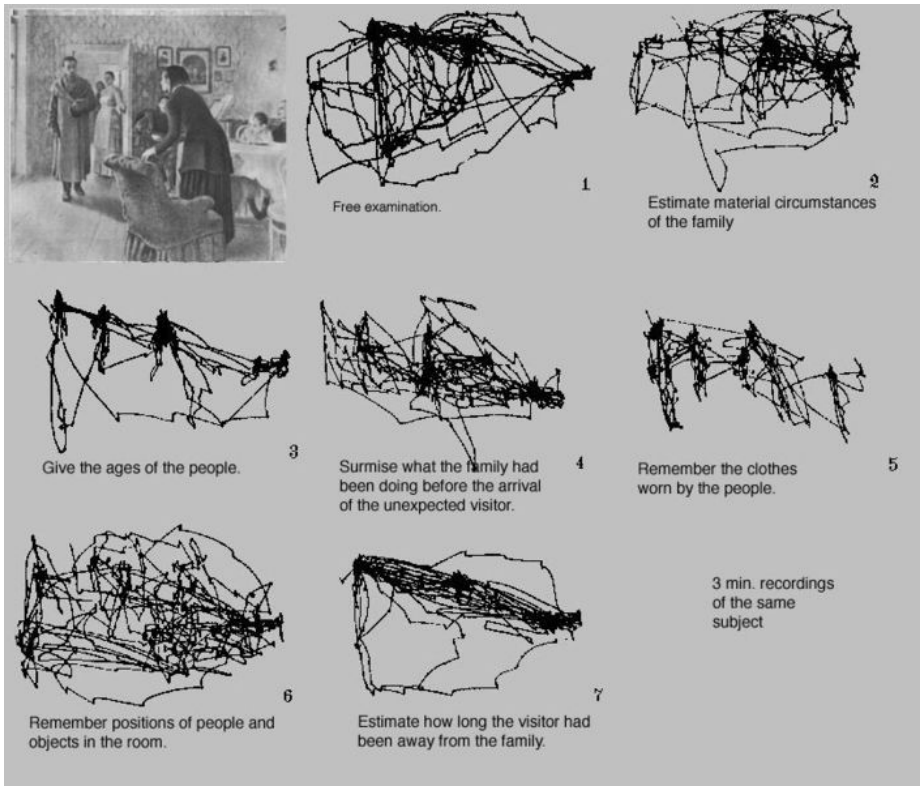


Figure 2.3: Yarbus experiment from 1967 — Visualizations of gaze paths show the strong influence of the given task.

Perception of Objects

Work of Triesch et al. [TBHS03] showed that the perception is a need-based approach and that desired objects are quickly detected in visual scenes. Ramanathan et al. [RKS+10] declared that "visual attention is not subjective but is directed towards salient objects." Selective visual attention is the mechanism by which we can rapidly direct our gaze towards objects of interest in our visual environment [TG80]. The results of Nutmann and Henderson [NH10] and Foulsham and Kingstone [FK13] pointed out that the preferred viewing location for objects in complex scenes is the center of the object. This finding holds even for the very first fixations, the so-called landing points on an image.

Scan patterns

The scan paths are also influenced by where an information is expected. In natural scenes, the sky is expected to be on the top, and a car on a road. Besides

these expectations, learned patterns for information display can influence gaze paths in specific situations. One effect is called center bias and it was described amongst others by Judd et al. [JEDT09] and Zhao and Koch [ZK11]. It describes an concentration of fixations in the center of an image. The appearance of such a bias is based on different factors like the experience that photographers place the most important objects in the center of an image or simply the straight-ahead position in front of the screen. Also other inherent photographic bias such as image compositions following the golden section can occur [Fre07]. On web pages, the logo can often be found in the left upper corner and the navigation on top of the page [BDC10]. The reading direction influences the scanning of list [Duc07]. In user experiments, a button for navigating through different pages of an experiment application is expected at a specific position on the screen when it appeared at that position in the previous experiment steps [PHG+04].

Covert Attention and Inattentional Blindness

Fixations in gaze paths show the areas of the highest visual perception and they are an indicator for the users' attention. However, it cannot be measured if the perceived visual information is important to the human, why it is important, or even if the information is indeed perceived and remembered. A well-known effect is the *inattentional blindness* [MR98]. It describes the lack of conscious perception of a scene or specific objects, even though it was fixated. This can happen when the "thoughts are elsewhere" or one is concentrated on a specific task, fading down information not important for this task. A well-known work is the Gorillas Experiment by Simons and Chabris [SC99], where even a gorilla in a group of persons can be overseen if the concentration lies on the given task to count the persons.

The opposite effect of the *covert attention* can also appear. It describes the effect that parts of a scene are perceived even without directly fixating it. Although this covert attention can happen, it usually takes place when a whole scene has to be overseen while the main attention is directed to one point. For example, in a baseball game, the players have to track the ball but also the whole scene on the playing field. However, it is unusual not to fixate an object of interest directly in other situations without this special need of keeping an overview.

2.2 Eye Tracking Hardware

Eye tracking devices measure the position and the orientation of the eyes in space and calculate the viewing direction from this data. By calculating the intersection of this viewing direction ray and an object in the real-world, for example, a computer monitor, the look-at positions can be determined. The fixated position is called point of regard (POR) [Duc07]. Additional information such as the user's pupil size are usually available as well. Other data such as blink rates can be extracted from the raw eye monitoring data.

In this section, a short overview of recent eye tracking hardware is given as well as an assumption on the future development. Limitations of the systems and specific characteristics of gaze data are discussed. Finally, the equipment used in experiments presented in this thesis and its parameters are described.

2.2.1 State of the Art

An eye tracking experiment setup consists of different components that are usually an eye tracking module, a monitor, and a computer running the eye tracking software.

Different approaches for tracking the eyes were invented for measuring the human's viewing behavior. The early devices were based on measuring skin's electric potential differences by placing electrodes on the skin around the eye. Contact lens-based methods were also developed. They allow very sensitive measurement but are also coupled with the need of physical contact with the eyeball and a distinct discomfort for experiment participants.

Nowadays, the most common eye tracking devices are based on measuring the pupil and the corneal reflection in a video sequence to identify the viewing direction. They are non invasive and thus more comfortable to be used. The hardware components of these systems are cameras and optionally additional sources of light. Most often, they make use of infrared light, as it is invisible for the user and less sensitive to daylight and other surrounding light sources. Depending on the position of the source of light and the human eyes, the pupils appear dark or bright in the recorded video sequence. In Figure 2.4, pictures of a bright pupil (A) and a dark pupil (B) are shown. For the bright pupil tracking, the illumination has to be placed close to the viewing direction of the camera. The light is then reflected off the back of the retina, which causes a light pupil, as also known from the red eyes effects in photographies. When the camera is offset from the viewing direction and the retro reflection from the retina is directed away from the camera, the pupil appears black. Most systems use the dark pupil approach while in some cases, depending on, for example, the race, the bright pupil approach delivers better results. The two methods can be combined in advanced eye tracking systems.

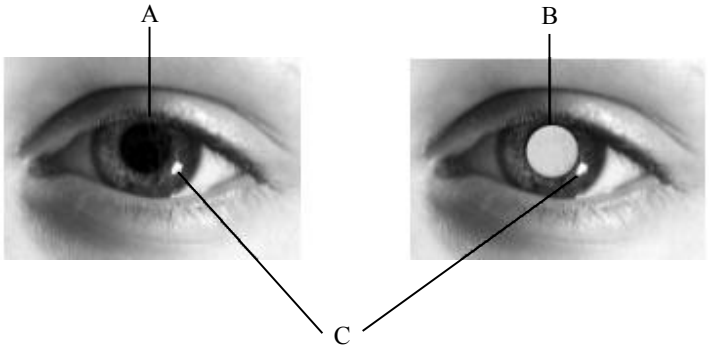


Figure 2.4: Pictures of the human eye, showing the pupils (A and B) and the corneal reflection C [Mil03].

Computer vision techniques are used to identify a person's head, eyes, pupils, and corneal reflection. The corneal reflection is the reflection of light sources on the surface of the eyeball, also called Purkinje reflection. See Figure 2.4 for an example of corneal reflection labeled with C. Usually, at least four reflections of the surrounding can be seen on one human eyeball but most systems rely on the first Purkinje reflection, the corneal reflection. From the position of the pupil and the corneal reflection and the changing offset between them, the viewing direction can be calculated. Taking the geometry of the complete scene with the position of a computer screen relative to the user's eye allows the trigonometric calculation of PORs. For increasing the robustness of an eye tracking system, several cameras and light sources can be deployed. Both eyes are detected, and consequently, two PORs are calculated which can be not congruent. The two points are usually combined by building the average, or if the differences are big, only the data from the dominant eye is used.

The eye tracking systems can be head-mounted or table-mounted, which describes the position of the camera in the experiment setup. It can be placed on the human head, for example, by wearing a helmet with an attached camera. This approach avoids the need of considering head movements in the computer vision part of the eye tracking process. But as for the invasive techniques, head-mounted systems bring disadvantages in use, for example, discomfort associated with wearing a helmet. In table-mounted systems, the camera is placed statically in front of the research participants. These systems are used when eye movements on a computer screen are being recorded. The alignment between eye tracking device (camera) and the computer monitor has to be given manually.

Video-based eye tracking devices work at specific sampling rates, usually between 30 Hz and 1250 Hz. Very high rates are necessary for high-quality eye tracking, for example, for analyzing human gaze during the reading process or for neurological experiments. Most devices have a sampling rate of 50 to 60 Hz. Vendors of professional eye tracking systems offer solutions for various experiment setups, from systems recording the human gaze while driving a car with several cameras to devices integrated in computer monitors. Varieties of sampling rates and tracking accuracy are available. Most systems provide pupil diameter and POR.

Calibration

For calculating the POR on a computer screen, the eye tracking system needs an internal model of the experiment scene, including the setup of the scene, the user's anatomic characteristics, and the imaging properties of the cameras. For the calculation of this model, a calibration period is performed during which reference data for known fixated positions on the screen is collected. During the calibration phase, the user is asked to follow a point on the screen with his/her eyes. The point stops for short periods at several calibration points. The number of points can vary; it is usually 5 or 7. The positions of this specific calibration dots on the screen are known, and a model of the scene can be calculated by the eye tracking system.

The quality of the calculated model can be tested against the collected calibration data. Figure 2.5 shows a visualization of the quality of the model by showing the differences between the known positions of the calibration points and the calculated fixation points. Depending on the quality of the calibration, which could be visualized by the calculated offset, the investigator can decide whether the calibration has to be repeated or if it is accepted.

Figure 2.5: Visualization of the calibration results in Tobii Studio. The lines depict the offset between the calibration points on the screen and the fixations points calculated by the internal model.

Accuracy and Freedom of Head Movement

Accuracy is defined as the average difference between the real stimuli position and the measured gaze position. *Precision* is defined as the ability of the eye tracker to reliably reproduce the POR when the same point is fixated. Accuracy and precision need to be high when reliable eye tracking data should be recorded. A number of factors complicate the gaze analysis and can lead to low accuracy and precision results, for example, noise from small eye movements such as microsaccades, limitations of the resolution of the used camera system, an incorrect calibration or strong position changes of the users that cannot be compensated.

Professional devices reach an accuracy of around 0.5° (see Table 2.1 for some examples for professional eye tracking systems), which roughly corresponds to 5 mm on the screen at a distance of about 60 cm between the eyes and the monitor.

The area in which the eye movements can be recorded reliably with high accuracy is limited to an area described by the freedom of head movement. Eye tracker vendors provide information about the freedom of head movements for their devices. They are usually described by an operating distance, describing the supported distances between the user and the eye tracker and a head box, giving the range of possible head movements vertical to the visual axis. Examples are given in Table 2.1. The values show that head movements are supported in a range that allows natural behavior. However, the change of position in front of the computer can be too strong to be supported, for example, when a user moves back and forth.

Company	Product	Operating distance	Head Box	Accuracy
Tobii	X2-60[1]	40 - 90 cm	50 × 36 cm	0.5°
SMI	RED[2]	60 - 80 cm	40 × 20 cm	0.4°
myGaze	Eye Tracker[3]	50 - 75 cm	32 × 21 cm	0.5°

Table 2.1: Specifications for operating distance, freedom of head movements, and accuracy for a selection of professional eye tracking device.

Eye tracking software usually supports the users in finding the ideal position in front of the eye tracking device. Figure 2.6 shows the diagnosis tool of the Text 2.0 framework [BBS+10b]. This framework realizes gaze interaction for web applications, including the correct positioning of the users, the calibration of the eye tracking system, and the extension of web elements by gaze interaction. Here, the distance between the user and the screen as well as the position of the eyes are displayed for supporting the correct alignment before starting the calibration.

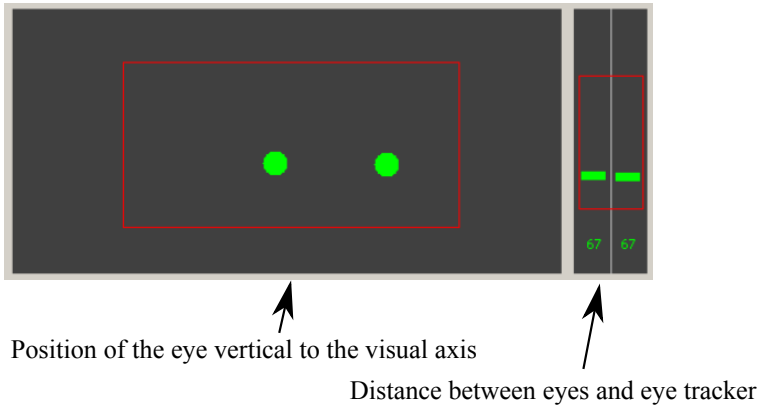


Figure 2.6: Text 2.0 framework: the diagnosis tool helps the participants to get into right tracking position in front of the eye tracking device by offering a visualization of the eye position and the distance to the eye tracking device.

2.2.2 Recent Development of Eye Tracking Hardware

While in the past, high-quality eye trackers were deployed mainly in research or usability laboratory, it can be assumed that eye tracking will be available to the average user in the near future. The interest in eye tracking technology

[1] http://www.tobii.com/en/eye-tracking-research/global/products/hardware/tobii-x2-60-eye-tracker, last visited February 15, 2013

[2] http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/red-red250-red-500.html, last visited February 15, 2013

[3] http://www.mygaze.com/fileadmin/download/Tech_Specs/130613_mygaze_techspecs.pdf, last visited December 20, 2013

in general is growing because of its unique possibilities in enriching the human interaction with computers and the information provided by analyzing visual attention.

Costs for professional eye tracking hardware have significantly dropped in the past. While the cost for eye tracking systems was about USD 30,000 just three years ago, embedded and low-cost solutions are available now for less than USD 100.[4] With the low-cost devices, Tobii tries to enter the end user market. The integration of eye tracking units into everyday devices such as laptops was realized by Tobii, although these devices are still prototypes (see Figure 2.7).

Figure 2.7: Tobii prototype of a laptop with eye tracking unit.

Systems that can detect the eyes and can calculate the viewing direction using cameras integrated in common devices such as tablet PCs are already on the market (e. g., Natural User Interface Technology, OKAO Vision[5]). Furthermore, eye trackers can be developed using low-cost hardware. Several approaches using webcams or low-cost equipment at open-source base were published within the last year. San Agustin et al. [SASHH09] compared a commercial eye tracker and a webcam system. The results for the webcam system are satisfactory and comparable with the commercial system, although still with limitations concerning the comfort of use because the webcam has to be very close in front of the eye and the user has very limited freedom for head movements. Lin et al. [LLLL12] presented an eye tracking system using a webcam that even works in real-time. Another open-source solution has been published by the ITU GazeGroup.[6] Promising results on using eye tracking data from unmodified, common webcams were also presented by Sewell and Komogortsev [SK10]. Lukander et al. [LJCM13] presented an open-source mobile eye tracking device. The Gaze Interaction Association COGAIN provides a list of open-source systems, which contained — at the time of writing — nine entries and a list of low-cost eye tracking providers with six entries.[7]

[4] http://www.tobii.com/eye-experience, last visited December 27, 2013
[5] http://www.omron.com, last visited May 8, 2014
[6] http://www.gazegroup.org, last visited May 8, 2014
[7] http://wiki.cogain.org/index.php/Eye_Trackers last visited January 22, 2014

The cell phone Samsung Galaxy S4 provides, as the first mobile device, a simplified gaze control system. The idea is called "Smart Scroll" and is designed for changing the behavior of the mobile phone application whether the user is looking at the phone or not. When the user is facing the phone, browsing is performed when tilting the head or tilting the device. "Smart Pause" pauses the playing of a video on the smart phone when the user is not facing it any more.

A patent revealed by Apple on "Electronic Devices With Gaze Detection Capabilities" describes several actions depending on the viewing direction of users, for example, the idea to dim a display screen when the user is not looking at the it. Another patent about the usage of gaze information supports the assumption of wider-spread of eye tracking techniques in the future: Google's patent on the usage of eye tracking technology describes a system for using eye gestures for unlocking head-mounted displays. And the fast development of sensors in IT hardware in the last years is still continuing. For example, future mobile devices can be equipped with high-resolution cameras or even infrared sensor and light sources.

2.2.3 Acceptance and Privacy Concerns

Eye tracking devices based on video recordings are non intrusive and can be used without wearing contact lenses or fixating the participant's head as it was necessary in the beginning of eye tracking systems. Infrared light is invisible, thus, it does not irritate the user while his gaze is recorded. However, calibration is needed and it can be necessary to repeat the calibration during a longer period of use, although the movements of the participants in front of the eye tracker are limited.

In addition, it can only deliver a glimpse on how users would accept an eye tracking device in their daily lives, their impression of using such a device in the experiments conducted for this thesis was investigated in questionnaires. Eighty-seven times experiment participants gave statements on their feelings during the usage of an eye tracking device on a Likert scale from 1 ("I felt uncomfortable while my eye movements were recorded") to 5 ("I did not feel uncomfortable while my eye movements were recorded"). An average rating of 4.5 (SD: 1) showed that the majority of the participants did not feel uncomfortable while their gaze was recorded. Furthermore, eye tracking technology fascinates the users as a new input device; for some of them, the control by eye tracking felt like "magic" because they can control a device without using a computer mouse or keyboard.

Privacy concerns are of importance when dealing with personal data: this includes the analysis of user behavior. For that reason, all data gathered for this thesis was recorded anonymously under user IDs that were not linked to the participants. Because the data can deliver sensitive information, for example, on the relationships between users (e.g., who fixated which person longest or how long does a participant fixate his own photo), no analysis was performed in this direction. The experiment on photo selection by gaze was approved by the Institutional Review Board (IRB) of the University of California, Irvine.

When using gaze data in everyday life, privacy has to be protected in the same way as other data on personal behavior such as web logs. Although gaze data delivers additional information, often the more sensible information is which content was viewed, not how. For example, the more sensitive information is which photo was viewed, instead of which part of the photo was fixated. If a photo with compromising content was displayed, even without eye tracking information, nobody would assume that only neutral parts of photos were fixated.

2.2.4 Technical Equipment Used in this Work

As the underlying idea of this work is to benefit from gaze data in everyday situations, the usage of a non invasive and not head-mounted tracking device was appropriate.

The experiments that work is based on were performed with a Tobii X60 device, providing an accuracy of 0.5° and a precision of 0.19° under ideal conditions (see Table 2.1 and Appendix A.3). This device makes use of both the bright and the dark pupil approach. It has a data rate of 60 Hz, thus, a raw eye movement data point is recorded each 16.7 milliseconds. The fixation points are calculated for both eyes and the average is built for obtaining one POR. For the calibration performed in the experiments, data of 5 calibration points was recorded and analyzed. The device is an external, table-mounted eye tracker, and it was placed in front of a computer monitor. An overview of a typical experiment setup can be seen in Figure 2.8. The table in our experiment laboratory was height adjustable; thus, it can be adapted to the participant's body height for bringing the eye in the head box area of the eye tracking device.

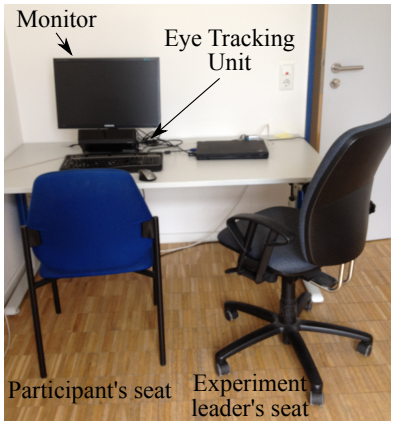


Figure 2.8: Typical setup for an eye tracking experiment as conducted for this thesis.

2.3 Saliency and Gaze Prediction

Because the visual attention is influenced by the low-level features of the stimulus, that is, the viewed photo, it can be predicted which parts of the photos are of high saliency. Privitera and Stark [PS00] developed an algorithm based on different image processing algorithms for identifying visual regions of interest. They compared these fully automatically calculated regions of interest with those generated from recoded gaze data. They showed that the loci of the human fixations can be predicted but the temporal order in which fixations take place cannot be predicted.

More research was done on investigating how to calculate so-called saliency maps, which predict areas on an image that attract human visual attention. Itti and Koch [IK00] presented the model of a saliency map based on orientation, intensity, and color information.Other approaches consider object-based visual attention, as, for example, presented by Sun and Fisher [SF03], or combine low-, middle- and high-level image features for creating a saliency map that does not concentrate solely on bottom-up computation, as introduced by Judd et al. [JEDT09]. Saliency maps calculate the regions in an image attracting the most visual attention. These methods have limitations on the image complexity, the placement of the objects, and the discrimination of different objects in one image.

2.4 Gaze Data

Eye tracking devices deliver POR on the computer screen at a point in time. Humans either gaze at a stationary point (fixations) or move their gaze quickly between the fixations (saccades). Humans perceive the maximum visual information during the fixations. First, the fixations have to be extracted from the raw eye tracking data and can be used as an indicator for human attention. Then, eye tracking measures are applied for analyzing the data. A eye tracking measure is a function on the users' fixations. The measure is calculated for predefined image regions, for example, the measure "fixation count" calculated for each region how many fixations were positioned on this region. The "fixation duration" indicated how long an image region was fixated by summing up the durations of all fixations on this region. Several measures can be used to analyze the gaze paths and will be presented in the following.

2.4.1 Preprocessing of Raw Eye Tracking Data

Raw eye tracking data consist simply of a list of look-at positions at a certain point in time. With a sampling rate of 60 Hz, each 16.7 millisecond a POR is recorded. In this stream of data, it has to be identified which raw data points belong together and build one fixation. The differentiation between two fixations is performed by applying a velocity threshold. When the eye movement velocity rises above this threshold, the movement of the eye to a new fixation is assumed. In addition, a duration or a distance threshold can be applied for an improved identification of single fixations. For the determination of the position

of a fixation, the median of all raw data points that belongs to one fixation is built.

In this work, an algorithm integrated in the software Tobii Studio is used for filtering the raw eye tracking data. The algorithm was presented by Olsson [Ols07] and it combines a velocity threshold as well as a distance threshold. The threshold can be freely chosen; in this work, the default threshold of 35 pixels for the velocity threshold (which corresponds to a 0.42 pixels/ms threshold) and 35 pixels for the distance threshold, which means that two fixations are merged into a single fixation if their Euclidean distance is below this threshold, were used.

2.4.2 Eye Tracking Measures

A eye tracking measure is a function on the users' gaze path. It is calculated for a specific region r on the computer screen. This region has to be defined, which can be manually by drawing forms or automatically based, for example, on segmented image. It can be calculated for one single gaze path and its fixations or it can be calculated by aggregating several gaze paths and using all fixations from the gaze paths of several users. State-of-the-art eye tracking measures are described below, and an overview is presented in Table 2.2, including their units of measurement. Most of the measures are a standard in eye tracking analysis, and here they were taken from the gaze analysis performed by Tobii Studio. Other measures come from related work.

The standard measure firstFixation (min count) computes the number of fixations on the image before fixating on a region r. The favorite is the region that was fixated first, that means the region with no previous fixations on the image. A modification of the firstFixation measure called lastFixation [Kla10] (min count) counts the fixations on the image after the last fixation on the examined region. The measure fixationDuration (max millisecond) describes the sum of the duration of all fixations on a region r. The measure firstFixationDuration (max millisecond) considers the order of the fixations and describes the duration of only the first fixation on a region r. The measure lastFixationDuration (max millisecond) provides the duration of the last fixation on the region. The standard measure fixationCount (max count) counts the fixations on a region r. The three measures maxVisitDuration (max millisecond), meanVisitDuration (max millisecond), and visitCount (max count) are based on visits. A visit describes the time between the first fixation on a region and the next fixation outside. The last measure saccLength (max centimeter) [KKK09] provided good results for the relevance feedback in image search. The assumption is that moving the gaze focus over a long distance (i.e., long saccade) to reach an image region r shows high interest in a region. In this work, results for single measures as well as the combination of eye tracking measures are investigated and described in the particular sections. Dome additional measures are introduced and evaluated.

Name	Description	Favorite	Origin
firstFixation	Number of times the participant fixates on the image before fixating on region r for the first time	min count	Tobii
lastFixation	Number of times the participant fixates on the image after last fixation on region r	min count	[Kla10]
fixationDuration	Sum of the duration of all fixations on r	max seconds	Tobii
firstFixationDuration	Duration of the first fixation on r	max seconds	Tobii
lastFixationDuration	Duration of the last fixation on r	max seconds	New
fixationCount	Number of times the participant fixates on r	max count	Tobii
maxVisitDuration	Maximum visit length on r	max seconds	Tobii
meanVisitDuration	Mean visit length on r	max seconds	Tobii
visitCount	Number of visits within r	max count	Tobii
saccLength	Length of saccade before fixation on r	max centimeter	[KKK09]

Table 2.2: Eye tracking measures applied to an image region r.

2.5 Conclusions from Background

The psychological background on the human visual perception process showed the complexity and the diversity of influencing factors on the visual perception, ranging from low-level pixel information to complex mental processes including knowledge and emotions. Although no apparent strategies for scene viewing are known, humans have an impressing ability to interpret complex scenes in real-time. The human visual perception is an interaction between bottom-up and top-down processing. Related work showed that the influence of bottom-up processing is strong enough to draw inferences from it, providing information on the perception process. However, other research showed the strong influence of interest (in free-viewing situations) and given tasks. Although it can be that an object is fixated but not perceived or the other way around, it is not very likely that this happens. When objects are in the focus of the visual attention, they are usually fixated in the center.

Eye tracking systems deliver fixation data with accuracy errors of a few pixels. The rapid development in eye tracking hardware will presumably allow the usage of gaze information in everyday tasks in the near future. Recent eye trackers allow the users to move, although limited to a specific area. Interacting in front of an eye tracking device was not perceived as uncomfortable.

Chapter 3

Related Work on Eye Tracking Applications, Region Labeling, and Photo Selection

Related work on the two core themes of this work, the image region labeling and the creation of photo selections, is discussed first in this section. Subsequently, the usage of eye tracking in existing applications and research is presented. Interactive and diagnostic are distinguished, and a new category of eye tracking applications — the exploitative applications — is introduced. Finally, the related work is summarized.

3.1 Creation of Image Region Annotation

The description of image contents is important for numerous applications such as search. The creation of image descriptions or labels at pixel level can be performed manually, semi automatically, or automatically. In the following, existing work on these approaches is presented.

3.1.1 Manual Image Region Labeling

The simplest approach for annotating image regions is manual labeling. The photo sharing platform Flickr[1] allows its users to manually mark image regions by drawing rectangle boxes on it and by assigning a text to it. Jeong [JHL11] found that region labeling is not very frequently used in Flickr; only 27% of the analyzed photos had region labels at all. Mostly, the notes were used for comments expressing feelings or emotions. Only 14% of the comments included information on concrete objects on the photos. Other web platforms such as

[1] http://www.flickr.com, last visited January 3, 2014

LabelMe [RTMF08] allow for a more precise creation of regions by drawing polygons on the images. These regions are annotated with a tag. The goal of LabelMe is to create ground truth data for object detection algorithms. The LabelMe community members have a strong interest in the data set as they need the data for their own research, for example, on object detection, and thus, they are willing to contribute to it. In general, the manual tagging of regions is tedious [Rod99], and the users rarely perform region labeling.

Games with a Purpose

A variation of the manual labeling are games with a purpose (GWAPs). GWAPs are computer games that have the goal to obtain information from humans in an entertaining way. The acquired information is usually easy to be created for humans but challenging or impossible to be created by fully automatic approaches. An example of a GWAP is the game *Peekaboom*, presented by von Ahn et al. [vALB06]. Two users play together; one of the users can see a given photo and a word related to the photo. He/she has to click on the photo for making a specific area visible for the other player, who has to guess what the given word is. The user can give extra hints to the guesser by additionally clicking on the photo and by giving hints on how the word is related to the photo. The players collect points for each correctly named word. In another game, named Squigl,[2] two randomly selected users team up to mark regions on an image without seeing the markings of the partner. The goal is to mark the same image regions for a given word. The highest score is obtained for congruent regions. Ni et al. [NDFY12] introduced a game for explicitly labeling image regions. The users look for specific objects in photos and mark them by drawing bounding boxes. Known objects are added to the photos for measuring the quality of the drawn bounding boxes.

Huang et al. [HCC09] presented a collaborative benchmark for region of interest (ROI) detection in images. They collected a large number of annotations by means of a game called Photoshoot. With the data gained from the game, they were able to evaluate different detection algorithms for ROI. In Photoshoot, two anonymous players are grouped together. One user has to draw rectangles over the image using drag-and-drop. The other player has to guess which region was highlighted by the first player by clicking on the image. For agreements, the players collect points, and it turned out the regions on which the two players agree are usually the salient regions of a photo. Salvador et al. [SCGiN$^+$13] presented the game Ask'nSeek, which serves the improvement of image segmentations and the identification of objects in images in an entertaining way. Two players play together; one marks a specific object on a photo while the other has to guess which object was marked by clicking on the image. The first player gives hints to the person who tries to find the object.

Although the idea of the GWAPs is to entertain the users while tagging image regions, the user still has to spend time in playing the games, which is contrary to the goal not to put a strain on the users at all.

[2] `http://www.gwap.com/squigl-a/`, last visited December 8, 2013

3.1.2 Automatic Region Labeling

Much work was done on the automatic assignment of *tags to images*. Li et al. [LSW09] made use of visual similarity for tag recommendation. They presented an approach that recommends tags for an unlabeled image by using low-level similarity with already tagged images and by obtaining relevant tags from these images. Tsai et al. [TJL+11] performed large-scale annotation of web images by considering images that are visually similar and correspond to the same semantic concept. They showed that their approach facilitates a better prediction performance, compared with competing methods. Tang et al. [TYH+09] extracted semantic concepts from community-contributed images and tags. They succeed in providing a more robust and discriminative approach compared with other semi-supervised learning approaches. In addition, they proposed a label refinement strategy that removed tag noise. However, these approaches did not address the problem of assigning the tags to image regions but to the image as a whole.

The automatic identification of concrete objects and their position in the images is still a challenging task. There are different approaches based on computer vision or saliency calculation. One approach is object detection with computer vision techniques. A large amount of training data — consisting of images and labeled image regions — is needed for such a purpose, e.g., [CF01, VJ01]. The identification of objects is limited to the learned concepts and to the visual similarity to the learned concepts, e.g.[SK00].

Different approaches were investigated to make use of *salient image regions* in region labeling. Rowe [Row02] presented an approach for finding the visual focus of an image by applying image processing in terms of segmentation and low-level features. The idea was to link the visual focus with the image caption. This approach was designed for images with a single object only [Row02]. In addition, it has limitations concerning the position and characteristics of the shown object. Duygulu et al. [DBDFF06] performed a mapping between region types and keywords supplied with the images by learning a fixed image vocabulary. Liu et al. [LCY+09] proposed a method to automatically assign labels at image level to image regions. The method was based on local image patches, gained from image over-segmentation, each of which may partially characterize one image label. They exploited the fact that two images with the same labels are likely to contain some similar patches. The images used in their experiment were simple, with an average of 2 to 3.5 labels per image.

Itti et al. [IKN98] presented a visual attention model based on multi scale image features (colors, intensity, and orientation), which delivered salient points in order of decreasing saliency. Their system is offered in a toolkit, which is used in this work as saliency-based baseline.[3] Navalpakkam and Itti [NI05] introduced a model that also took the influence of tasks into account. Besides the usage of low-level features, their system considered a manual initialization by the user by explicitly giving keywords and the relevance of these keywords. The prediction of visual saliency was then biased by visual information relevant to the given keywords. For their approach, a hand-coded ontology as well as manually created classifications of images showing the same objects were

[3] `http://ilab.usc.edu/toolkit/`, last visited May 8, 2014

needed. Privitera and Stark [PS00] compared the identification of ROIs by gaze information and by image processing algorithms. They showed that the algorithms cannot predict the sequential ordering of the loci of human fixations. Yuan et al. [YLZ07] made use of spatial context constraints for solving the region labeling task. Besides image features, four types of spatial relations were considered (i.e., left, top, right, and bottom) when assigning semantic labels to image regions. Liu et al. [LYL+10] performed label-to-image-region assignments by means of online image search. For each tag that was assigned to an image, they performed an online image search for obtaining visually similar images. These images were analyzed for obtaining salient and descriptive features for the tags. Finally, the assignment of the labels to regions was performed based on these features.

The first difficulty of automatic region annotation is the lack of training set with region-level ground truth as also indicated by Yuan et al. [YLZ07]. Automatic techniques are limited in their abilities to understand photo contents and they are based on visual similarity. These approaches are dependent from training data and a trained model, what makes them inflexible concerning the number of concepts and new concepts.

3.2 Automatic Creation of Photo Selections

Manually selecting subsets of photos from large collections in order to present them to friends or colleagues or to print them as photo books can be a tedious task. Research was performed for supporting users in these tasks. Content-based approaches make use of pixel information extracted from the images while context-based approaches analyze contextual information, such as capture time and focal aperture, or use both to determine a proper subset of photos.

3.2.1 Content-Based Approaches

The pixel information of photos are analyzed for these approaches with the goal to identify photos that are representative for a set of photos.

Chu and Lin [CL08] selected photos by identifying near-duplicate photos in given photo clusters. In the clusters, near-duplicate photo pairs were selected first, and the relationships between these photos were modeled by a graph, from which the most representative photos were selected by identifying the most important note in this graph. They concluded their evaluation results with human subjective judgment as satisfactory but also recognized high variances in the human judgments. A semiautomatic collage creating tool that created a selection of photos purely based on content-based information, such as color histograms from which a sharpness score is calculated, was presented by Xiao et al. [XZC+08]. They identified near-duplicate photos by means of binary classifiers working with similarity measures from literature and time stamp. Their system additionally offered an auto-crop algorithm and automatic lighting/color enhancement. Zhao et al. [ZTL+06] introduced an approach for the automatic labeling of persons in family photo albums. They made use of face and body information derived from image analysis of the photos. In addition, social con-

text information was used when analyzing which persons were depicted together more often. When persons were labeled in photos, the photos depicting, for example, the in total most often depicted persons could be selected. The most interesting photos in a sequence were identified by an algorithm proposed by Grabner et al. [GNDVG13], which made use of computer vision techniques. An evaluation showed that already four basic interestingness cues (emotion, complexity, novelty, and learned) and their combination delivered good results, even without considering semantic interpretations. However, the characteristics of their data set have to be considered — image sequences recorded by a static video camera — and make the application of the results to other, more diverse, data sets difficult.

There are very few approaches in related work using content information alone. Often, the context-based approaches that are described in the following section also rely on content analysis.

3.2.2 Context-based approaches

The *content-based approaches* were followed by *context-based approaches*, which exclusively or additionally analyze the contextual information of photos. Contextual information can be technical parameters of the digital camera, such as capture times or GPS coordinates, or information gained from social networks like blogs.

Platt [Pla00] clustered photos concerning capture time and/or photo content, based on a probabilistic model that identifies similar image characteristics. Li et al. [LLT03] created summaries of photo collections based on time stamps and facial features. Their photo summarization application had the aim to facilitate browsing and to offer summarizations in two steps: in the first step, the whole photo set was divided into partitions based on the capture time. In the second step, key photos were selected for the partitions by means of a "face criterion" and on the "temporal importance" of the photos. Following the face criterion, photos that depicted large and center-positioned frontal faces were preferentially selected. For the "temporal importance," photos were selected that were part of time periods with a high concentration of photos, based on the assumption that something interesting happened when many photos were taken.

A framework for generating representative subsets of photos from large personal photo collections by using multidimensional content and context measures was introduced by Sinha et al. [SJ11, SMJ11]. They solved the selection task by maximizing the interest in the selected photos, the distance between the photos (thus, the most diverse photos should be selected), and how strong the selected photos represent the whole set. They proposed a concept space that has five dimensions, which were visual (scene type such as outdoor day or sunset), temporal (time stamps), event type (predefined event types such as birthday or trip), location (city names), and people (unique faces from face recognition) for computing these properties. They were able to show in an experiment that the suggested summarization algorithm, based on this concept space, outperforms baseline algorithms.

Rabbath et al. [RSB11] used content and context information from blogs to automatically create photo books. In their work, photos were collected in a social network, and a subset of these photos was selected. The collection and the selection took place based on photo content information as well as people-based information (the persons depicted on the photos) and textual-based information (from textual media attached to the images). Beginning with a seed photo, which was a photo that could be easily found as it was well annotated, more relevant photos were found from the social network by the above described characteristics. From these photos, important photos had to be selected in order to create a photo book with a limited number of photos. The criteria for selecting important photos were that the photos had a high resolution; the photos achieved a specific ratio between those which show people and those which show landscapes (as background images in a photo book); the photos were subject to user interaction like tag or caption assignment; the photos contained important persons. The information on important persons was derived from a people rank, which considered how often the user was depicted with this person.

Mor et al. [NYGMP05] made use only of contextual information that is automatically available when a photo is taken, such as the capture time and the location a photo was taken at. From this data, they derived events and location clusters. From photos with labeled persons, labels for the photos with no annotations were generated. Their system was evaluated with four different personal photo collections and users who manually labeled the depicted persons. They can successfully annotate up to 90% of the photos, even when only 10% of the photos were previously labeled.

A framework for automatically selecting photos as a summary of a bigger set of photos was presented by Jaffe et al. [JNTD06]. The provided algorithm made use of location metadata (the location a photo was taken at), capture time, photographer, textual labels, the photo quality, and relevance (which is expressed by a relevance factor that measures an arbitrary bias concerning possible factors such as recency or specific user attributes) for creating a ranking of all photos in a set. An evaluation with 25 participants showed that the summarization based on these features outperformed baseline approaches, which were randomly created or based only on single features (capture time for recent photos and ratings coming from Flickr for interesting photos). Boll et al. [BSST07] presented a component-based framework for the automatic selection of a subset of photos from a large collection based on both content-based and context-based information.

3.3 Eye Tracking Applications

Eye tracking data is used in diverse applications. The applications are often classified into two main directions — *interactive* and *diagnostic* systems [Duc07]. In interactive applications, gaze data is used to alter the runtime behavior of software. One common field of application is to provide user interfaces for users with disabilities, who have problems to use devices such as computer mice or keyboards. But with the wider spread of eye tracking hardware, more and more applications appear for diverse user groups. In diagnostic applications, gaze

data is used as a measure for the human visual attention. The recently most common use case is in usability studies in observation laboratories. Diagnostic applications serve the goal either to understand and improve the visual saliency of the viewed stimulus, such as web pages or user interfaces for software applications, or to understand the viewing behavior itself, for example, in psychological research.

During the last years, a new direction appeared in eye tracking applications; in this work, they are called *exploitative applications*. These applications serve neither the goal of improving design nor the goal of understanding the viewer but of exploiting the users' viewing behavior for getting information about the viewed content. This information is in turn used to support the users. Because of these two components, *exploitative* applications are categorized to be in between the interactive and the diagnostic applications, as depicted in Figure 3.1.

In this section, state of the art work in interactive systems, diagnostic and exploitative systems are reviewed.

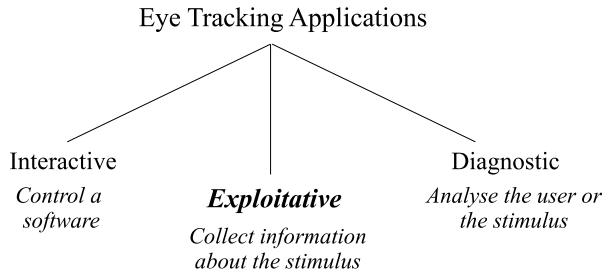


Figure 3.1: Classification of eye tracking applications with the newly introduced "Exploitative Systems."

3.3.1 Interactive Applications

In interactive eye tracking applications, the movement of the eyes is used to control a software. This can be typing, drawing, or browsing. The areas of application are as numerous as for other input devices. Usually, the applications offer gaze-sensitive areas, which have to be fixated for causing a reaction. But control can also be realized by gaze gestures, that is, predefined eye movements that trigger actions.

Gaze control is often used for supporting handicapped users but can also be of advantage for other users. For example, for hygienic reasons, as no device has to be touched in public area and there is no need for a mechanical input device, which can be damaged. In addition, gaze control can be very fast [SJ00], and users are often fascinated and entertained by the gaze interaction [QZ05, LBS07].

Gaze control can be challenging, as the eyes are usually an input device for perceiving information. When commands should be given by moving the eyes, the movements have to be strongly controlled. This kind of control is unusual to the users and can be tiresome. The control of the eye movements can be difficult

because they are influenced by many factors such as visual distraction and the eyes are constantly moving (see Section 2.1 for the diverse factors influencing the eye movements). Thus, errors and inaccuracies always have to be considered.

Jacob [Jac91] introduced the "Midas touch" problem, which deals with another specific problem in gaze control. It describes the problem that the two modes "viewing for perceiving information" and "fixating with the goal to control a software" cannot be easily distinguished. For example, a user has to fixate a button for understanding its function but the button should not be activated automatically from this fixation. Solutions for this problem are dwell-time approaches, where the user has to fixate, for example, a button for a specific period of time for activating it, or gaze gesture, where specific eye movements have to be performed for triggering a control command. Drewes and Schmidt [DS07] showed that users can perform complex gaze gestures intentionally and that even standard applications can be controlled by it. Heikkilä [Hei13] compared these two forms of interactions — dwell buttons and gaze gestures — for drawing applications, with the results that in her application both methods performed equally fast, while the gaze gestures were the preferred method of the study participants. Gaze can also be combined with other input devices such as mice or keyboards for avoiding the Midas touch problem. Zhai et al.[ZMI99] introduced the MAGIC pointer, which combined gaze information and manual control by moving the cursor toward the fixated area on the screen while selections take place by means of a pointing stick. Kumar et al. [KPW07] combined gaze and keyboard control. In their study, different tasks had to be performed by means of a computer mouse and different variations of *EyePoint*. The results indicated that control by *EyePoint* has a performance similar to the one of mouse control.

Typing is an important application in the area of gaze controlled applications, and several typing systems have been presented, for example, by Majaranta et al. [MAŠ09]. Räihä and Ovaska [RO12] introduced an extensive study on eye typing. They showed the importance of variable dwell-time threshold, allowing the users to adapt the software to their own experience or need. The development of drawing systems was also the focus of some research. For example, Hornof and Cavender [HC05] suggested a drawing system based on dwell time, which allows children with disabilities to draw digital images. Bartelma [Bar04] investigated the combination of gaze control and image segmentation. He implemented a system that was controlled by gaze to manually segment images. The gaze was exclusively used as a mouse replacement, and the subjects were instructed to outline a given object with their gaze.

The use of gaze control in gaming was investigated by Smith and Graham [SG06]. They showed that gaze control can increase the playing experience for the investigated games, as the users felt more immersed in the gaming world. However, problems also occurred from gaze control, such as that for some games the learning was easier with the conventional mouse control. Another use case is the active presentation of information depending on the look-at position of the user. For example, the eyeBook framework from Biedert et al. [BBS+10b, BBS+10a, BBD10] displays additional information such as sound or modifies the text when the reader's gaze reaches a specific trigger position in the text. The user can be supported when skimming a text, or translations are offered for foreign-language readers. Milekic [Mil03] introduced

a conceptual framework for a museum application. The interface is controlled by gaze gestures, for example, for zooming. GazeSpace, a browsing application controlled by eye movements, was introduced by Laqua et al. [LBS07]. In the presented study, users had to browse pictorial and textual information for answering some information tasks. The results showed that the users enjoyed using the gaze-based interaction systems. Mollenbach et al. [MSH08] conducted an experiment with users performing simulated search and browse as well as simulated target-selection tasks. The experiment showed that selection tasks can be performed faster with gaze control compared to mouse control. An interactive systems for city trip planning that offers visual and audio information about different locations on a map, depending on the user's behavior, was presented by Qvarfordt and Zhai [QZ05]. In *iTourist*, interest was derived from predefined gaze patterns. The evaluation of the prototype system showed that it can offer required information to the users and that it caused positive reaction.

Gaze information is also of use for adapting an application to the users' mental capacities. The goal for some applications is to improve the human learning process by taking into account the user attention and to adapt the content presentation to it. Among others, one framework with the name AdeLE was presented by Pivec et al. [PTP06]. The information about the position of the eye was used to adapt the content to the learner, for example, to provide an abstract of a text to a reader who is only skimming a text. In the e-learning environment of Porta [Por08], an "emotion recognizer" was implemented. It used mainly pupil size, blink rate, saccade length, and occurrence rate to identify the user's current capability to work with e-learning content. Another approach was to continue with the presentation of content only when the user is looking on the screen, invented as eyeLook by Dickie et al. [DVSC05]. The information if a user is attentive was also used in an application presented by Nakano and Ishii [NI10]. The disengagement of users in human-agent conversations was identified, and questions were asked for bringing the user's attention back to the conversation.

3.3.2 Diagnostic Applications

In diagnostic use cases, eye movements are recorded and analyzed concerning a given stimulus. This stimulus is not modified by the viewing behavior, thus the analysis can take place offline. The stimulus as well as the human observers can be subject of the studies.

In psychological research, human behavior is analyzed based on gaze data. In Section 2.1, the human visual perception process is described and many of the insights into the visual perception are based on eye tracking research. Eye movements also provide insight into the memorability of humans. Bulling and Roggen [BR11] showed that familiar and unfamiliar abstract pictures can be discriminated by means of eye movement analysis. In these use cases, the goal is to better understand humans and their visual perception.

The most common usage of gaze information in a diagnostic context is in design and usability studies. Here, the stimulus is in the focus of the investigations as graphical user interfaces are subjects to be improved and optimized. In these evaluations, the areas of highest user attention are identified and compared with

the intentions of the designers, to identify design or usability problems. For detailed analysis, ROI are marked on the investigated medium, for example, a web page or a commercial. Based on these ROIs, the users' attention was analyzed in order to optimize the object that is under examination [CJP10, GSL+02]. Work in this direction was presented, for example, by Bruenau, Sasse, and McCarthy [BSM02]. Often, the design of a web page or the placement of advertisements is analyzed.

Other studies address more general research questions on viewing behavior. Pan et al. [PHG+04] investigated patterns in web page viewing. They found that scan paths depend on both individual characteristics of the viewers and the stimuli. Influencing factors seemed to be, for example, the gender of the participants, the order of the viewed stimuli, and the complexity of the viewed web pages. Duggan and Payne [DP11] examined the reading behavior of users searching for information in text documents. As they were on time pressure, they had to skim parts of the texts. The aim of this work was to better understand how text is read for supporting producers of online content. Findings were, for example, that the texts at the start paragraphs are more likely to be read and that also skimming can be effective. Cutrell and Guan [CG07] analyzed the viewing behavior of users interacting with a web search service with the goal to better understand how much information should be displayed. They investigated how search results lists are used to find information, that is, how these lists are read. By analyzing the gaze behavior, they found, for example, that increasing the length of snippets in the search results list increases the performance of users searching for information. In a study investigating the viewing of web search engine result pages, Buscher et al. [BDC10] showed that the given task type, the quality of the displayed advertisements, as well as the sequence of advertisements of different quality influence the scan paths. Marcos et al. [MNST12] introduced a user navigation search model with five navigation patterns for identifying users with a need of support in their search. Martínez-Gómez and Aizawa [MGA14] derived knowledge on users' language skills by analyzing the reading behavior and Toker et al. [TSG+] identified the users' skills in understanding visualizations in the form of diagrams.

3.3.3 Exploitative Applications

In exploitative applications, eye monitoring is used to generate information about the viewed stimulus, not with the goal to modify it or to adapt it but for annotating it.

One research direction is the identification of performed activities in real-life situations. Kunze et al. [KUS+13] analyzed gaze data, recorded with a mobile head-mounted eye tracker, for the classification of documents the participants have read. The results of an evaluation showed that by means of the reading behavior, the six investigated text document types can be distinguished. According to their own statement, this work was a first step in the direction of creating read logs from a scene camera where texts cab be identified and annotated with the determined document type. Different activities performed by users in office work situations are identified from eye movement by Bulling et al. [BWGT09]. Six activities like reading, browsing, or writing were distin-

guished with an average precision of 76.1%. EyeContext, a system presented by Bulling et al. [BWG13], identifies human behavior from gaze analysis. This information is valuable, for example, in the creation of life logs.

Another approach is to obtain information about a text from reading behavior analysis with the goal to annotate text parts, for example, as important. These important parts can be used in the automatic generation of text abstracts, as suggested by Buscher et al. [BDEM08]. Xu et al. [XJL09] provided individual document summarizations from gaze data. The quality of the summaries was comparable with manually created summaries, as shown by the presented evaluation. Putze et al. [PHK+13] combined eye tracking information with electroencephalogram (EEG) data to identify events in video streams. The gaze data was used for finding the location of the perceived event (with an accuracy of 86.3 %), while EEG identified the temporal occurrence of an event. The study was performed in a controlled setting with simulated video sequences.

Implicit Relevance Feedback in Search

In search applications, eye tracking data can be used as implicit relevance feedback. The visual attention delivers information about the relevance of a document to the user.

Among others, the application of implicit relevance feedback to text search was presented by Salojärvi et al. [SKSK03, SPK05]. They showed that relevance can be inferred from human attention patterns. Buscher et al. [BDvE08] investigated eye movement measures for detecting reading behavior. Their preliminary results suggested that relevant and irrelevant text documents can be discriminated by means of gaze analysis.

In image search, several approaches made use of eye tracking data to identify images in a search results list as important and used this information as implicit user feedback to adapt the search in subsequent retrieval steps. For example, Klami et al. [KSDK08] performed implicit user feedback on image search results lists by means of gaze information. They showed that it is possible to use gaze information in the detection of image relevance in a controlled setup, with four images on each experiment page. For each set, the participants had to decide whether it contains a relevant photo for the search task "sport" by pressing a key on the keyboard. Solely from the eye movement, Klami et al. were able to identify relevant sets with an average area under the curve (AUC) score of 0.81 (random would be 0.5) and relevant photos in the sets with an AUC score of 67.7 (random: 0.25). Kozma et al. [KKK09] showed that a comparison of the implicit gaze feedback with explicit user feedback by clicking on relevant images and a random baseline are promising for quality of the gaze approach. They presented GaZIR, a gaze-based interface for browsing and searching for images. In the GaZIR system, images were presented in a circular order according to a search query. The gaze data of the user viewing these images were recorded and analyzed. Based on this implicit relevance feedback, the search was continued whereby the images that obtained the most attention were considered relevant. For six investigated categories, the gaze relevance feedback always performed better than a random baseline.

Pasupa et al. [PSS+09] applied a support vector machine (SVM) algorithm using eye tracking information together with content-based features to rank images. The participants had to choose and rank 5 of 10 photos concerning their relevance for the topic "transport." The results showed that the combination of simple image features and implicit gaze feedback improves the search for relevant images. Hardoon and Pasupa [HPS10] have extended this approach by using images with gaze data as training set for ranking images when no eye tracking data is available. The ranking is conducted using tensor kernels in an SVM. Essig [Ess08] also took user-relevance feedback, gained from gaze information, into account to improve the content-based image search. The feedback is calculated on the basis of image regions. He showed that the retrieval results of his approach received significantly higher similarity values than those of the standard approach, which is based only on automatically derived image features.

Gaze-based Image Labeling

The labeling of image regions is a difficult but important task in multimedia, as pointed out in the introduction. Gaze data was analyzed in related work with the goal to perform this task. Jaimes et al. [Jai01] carried out a preliminary analysis on identifying common gaze trajectories in order to classify images into five predefined semantic categories. These semantic categories were handshake, crowd, landscape, main object in uncluttered background, and miscellaneous. The general assumption was that similar viewing patterns occur when different subjects view different images in the same category. To this end, a generic object-definition model was provided that allowed the users to specify the relation of objects in the images, such as persons and hands, in an image showing a handshake situation. The results of this work were encouraging, and the researchers determined that it may be possible to construct an automatic image category classifier from the approach. However, the construction of the object-definition model was tedious, and an object classifier needed to be provided for each object category in the definition model in order to actually be able to classify new images.

Hajimirza and Izquierdo [HI10] used eye tracking information in a semi-automatic image annotation system to annotate a selection of images with tags based on gaze visit time and revisits. In their experiment, a concept was presented to the user. Then a list of images was presented and all images that attract the user's attention were annotated with the given concept. Preliminary results showed an average annotation precision of 80% and a recall of between 60% and 80%. Hajimirza et al. [HPI12] also introduced a real-time user-adaptive framework that offered a user interest score that can be used for image annotation.

A framework that assigned a *user interest level* between 0 and 1 to viewed photos was introduced by Haji et al. [HMPI11]. In their experiment, users viewed photos with the task to select a photo for the cover of a magazine with a predefined content, given by an example photo. Twenty-one gaze-based features were used for calculating the user interest level. Haji et al. declared that this information can be used as a source of information for user adapted

image annotation (when the user had a concept in mind, for example, during search) and retrieval. Soleymani et al. [SKP13] used EEG, facial expressions, and eye gaze for labeling images. In the experiment they conducted, images were shown to users with correct and incorrect tags. From the users' reaction, the correctness of the shown tags can be derived with an F-measure of 0.59. The results also showed that the gaze approach outperformed the two other modalities.

Gaze-Based Image Region Labeling

Gaze information is also used for obtaining information on viewed photos at region level. Papadopoulos et al. [PAD13] made use of implicit feedback from gaze data in region-based image retrieval. Besides presenting features for assigning relevance assessment to image regions, they also showed the application of this data to an image-retrieval application. For the investigation of gaze features, the users viewed photos in sets of tens and in a zoom-in-image mode in a first experiment part. The experiment application was controlled by eye movements, using the dwell-time approach. The task was to observe the images, taking a given semantic concept, such as building, street, or desert, into account. Subsequently, the viewed objects were rated concerning their relevance for a given concept. The recorded gaze data was analyzed to identify image regions that were of interest to the participants. By means of a SVM, the degree of relevance was calculated for all regions. Papadopoulos and colleagues could show that the proposed features significantly outperformed features from related work. In a second experiment part, the users performed image search sessions with five successive iterations. The iterations were performed by taking the relevant image regions into account, using the proposed features. The results showed that gaze-based relevance feedback at image region level can improve the image retrieval results compared with concepts from related work. However, the amount of improvement varied between the concepts. The researchers' interpretation of that result was that the variance in the low-level visual features for some concepts (such as car or desert) hinders further improvements.

Santella et al. [SAD+06] presented a method for semi-automatic image cropping using gaze information in combination with image segmentation. The goal of this work was to find the most important image regions for adapting the image cropping process. The users in their evaluation first viewed in total 50 photos with the task to "find the important subject matter in the photo." The participants knew that the gaze data was recorded and that it would be used in photo cropping. Afterward, sets of two crops (one based on saliency, one based on gaze data) were presented to the subjects who had to decide which crop was the better one. For the gaze-based crop, first, the image was segmented and then the most important image regions were identified from the gaze trajectories. The results of this evaluation showed that the image cropping approach based on gaze information was preferred by the users to fully automatic cropping in 58.4% of the cases. Sawada et al. [STM13] presented a system called iMap, which made use of the knowledge of salient image regions in the creation of film comics. The goal of their work was to find the best positions for speech bubbles (those regions with less important content that could be covered by the

bubbles) and to find good parameters for photo trimming. The gaze data was collected from users just viewing the movies. Doug and Santella [DS02, SD02] used eye tracking data to identify meaningful regions in photographies. The gaze data was collected from users viewing photos for five seconds. This information was used to perform a transformation of a photo into an artistic image in line-drawing style.

Ramanathan et al. [RKS+10] made use of gaze information to improve the segmentation of digital images. Their idea was to analyze the fixation data for identifying good seeds for the segmentation algorithm. The gaze information was collected from users free-viewing images, that is, without a concrete task or interest in a specific object. The images used in their analysis showed only one salient object against the image background. Their gaze-base approach performed the segmentation 10% better, compared with the segmentation without gaze information. Klami et al. [Kla10] presented an approach for identifying image regions relevant in a specific task using gaze information. In their work, relevance was calculated only from the gaze information represented in a Gaussian mixture model, which resembles heat maps. The model was built based on several gaze paths for identifying ROI. In an evaluation with 25 participants, two different tasks were given to the subjects. The two tasks were to inspect photos in the role of a burglar or in the role of a house buyer. This work revealed that the regions identified depend on the task given to the subject before viewing the image. They showed that their classifier clearly outperformed simple gaze measures such as the number of fixations on a specific image part. The work of Ramanathan et al. [RKH+09] aimed at localizing affective objects and actions in images by using gaze information. Image caption localization was performed based on the segmented image and the gaze fixations. An affect model for world concepts was derived from fixation patterns. Experimental results showed that an accuracy of 80% was achieved for the labeling of affective concepts with the image caption texts.

3.4 Summary of Related Work

Many publications from related work aim at assigning tags to images. The goal of this work is to assign tags to image regions at pixel level. Automatic assignments of tags to regions in related work are usually based on the visual similarity, a given training set, and a number of learned concepts. But as Grabner et al. [GGVG11] constituted, objects are often identified by human observers based on their function, not on their visual appearance. This shows the limitation of the visual-similarity approaches. Humans are able to identify objects based on but not limited to, their visual appearance.

Despite all research on the creation of photo selections, the task is still challenging. Work was done on defining criteria that should be fulfilled by photo selections. Sinha et al. [SMJ11] stated that an effective subset summary should satisfy the criteria quality, diversity, and coverage. Savakis et al. [SEL00] investigated how humans select photos from a collection with the result that the selections are subjective and differ between the users. They also determined that it is hard to identify the attributes on which the decisions are based,

as it is part of a high-level human cognitive process. The automatic photo selection approaches have to solve two problems: on the one hand, the diverse and individual criteria of individual users have to be reduced to a model or computational criteria. On the other hand, automatic techniques from, for example, computer vision have, to be used for identifying photos fulfilling the criteria. More complex criteria such as personal preferences or interest are insoluble problems for these approaches.

The diversity of gaze controlled applications points out the numerous possible fields of applications for eye tracking technology. Advantages are that the user does not have to physically touch a device. The gaze data can also be used additionally and in combination with standard input devices. The information on visual attention and the mental state of the users can be very valuable, and it can hardly be measured with other input devices. Last but not least, the control by eye movements is entertaining and can even be perceived as "magic" by the users.

The expected spread of eye tracking hardware and the corresponding increase of eye tracking applications support the approach presented in this thesis. However, the idea behind this thesis is very different from most approaches presented in the last section, as for these interactive applications, the goal is to control or adapt applications. Diagnostic applications analyze either the viewed stimulus or the persons viewing the stimulus. The goal is to adapt the stimulus for getting an intended viewing behavior (e.g., in usability studies) or to understand the human (e.g., in psychological research). Both aims are different from those of the applications called exploitative in this thesis. Exploitative applications have the goal to obtain information from analyzing the viewing behavior for annotating the stimulus. Existing work shows that gaze data can be meaningfully used in different application, for example, for labeling logs with specific activities or for identifying important parts of texts from reading behavior. Visual attention derived from gaze data is also a valuable source of information as implicit relevance feedback in search. The related work on this kind of feedback showed that for concrete search tasks, relevant photos can be identified. However, no work deals with the identification of photos that are of personal importance in a free-viewing task, as it is investigated in this work and described in Section 8.

Related work also has shown that the identification of important image regions, for example, for improving search, can be performed and the value of this data was proven in some studies. But the concrete labeling of image parts with a marked region at pixel and a tag explicitly assigned to this region was not investigated in the presented related work. This region labeling and its evaluation is part of this work and is described in Sections 4 to 7.

Chapter 4

Image Region Tagging with Given Tags and Given Object Regions

Manually providing image annotations in the form of labels is a tedious task for the users. This becomes even more cumbersome when objects shall be annotated in the images. In addition, the automatic labeling of image regions is far from fulfilling the human needs. Such region-based annotations are of value in various areas of application such as similarity search or as training set for object detection algorithms. The labeling of image regions by means of gaze analysis is the goal of this thesis. The gaze data of users is recorded with an eye tracking device and used as implicit source of information.

In this section, a first step in the direction of labeling image regions by means of gaze data is presented. By means of gaze data, one object region r is selected from a set of manually created high-quality object regions by means of gaze analysis. The gaze data was recorded in a controlled experiment conducted with 30 participants. In the experiments, subjects had to decide whether an object, described by a given tag, can be seen on an image or not. A sequence of 51 tag-image-pairs was viewed by each participant while their gaze information was recorded. The experiment consisted of three steps: first the tag was presented, then a red dot concentrates the attention on one starting point, and finally the image was shown and the decision was made. The images and labeled image regions were taken from the LabelMe data set [RTMF08]. About 50% of the given tags were correct, the others were incorrect.

In total 799 gaze paths were analyzed to calculate the tag-to-region assignments. These assignments allocated the given tag a favorite object region, if the users pressed the button for indicating that an object described by the tag is depicted on the photo. 13 eye tracking measures were considered in the analysis and further parameters regarding the extension of region boundaries and weighting of smaller regions were investigated. In addition, it was investigated if different object regions were selected as favorite region when gaze data from users with different primings (different tags) were analyzed. Furthermore, an

in-depth analysis of the obtained results is provided, in which the size and position of the correctly respectively incorrectly assigned regions analyzed with the goal to identify typical characteristics that can restrict the gaze-based approach. Gaze data of several users with the same primings are aggregated during the analysis. The influence of the number of subjects in an aggregated analysis on the precision of the tag-to-region assignments was investigated. The impact of a previous fixation on the first fixations on an image with respect to identifying the correct image region is explored. And finally, a closer look into differentiating two objects shown in the same image by analyzing gaze paths with different primings is taken.

The experiment setup in this part of the work simulates the situation of a user viewing images while being interested in a specific object. The users did not know for which reason their gaze paths were recorded and that the shown tag was assigned to an image region. Thus, the users were concentrated on solving the given tasks to decide on whether they can see the given object or not and scanned the photos with this goal, without trying to control the gaze in any way. This viewing behavior simulates situation that occurs in everyday live, when users are dealing with photo, for example, when searching for a photo with by giving search term in an image search scenario. This scenarios may include further challenges, such as possible distractions from the surrounding web search page or smaller image size in the search results lists. Because of such additional challenges, the approach to the overall research question of assigning tags to image regions based on eye tracking data is broken down into a series of distinct steps as presented in Figure 4.1. The first step, presented in this Section 4, is the analysis of gaze data gained in a controlled experiment, with given tags, and the usage of predefined high-quality segmentation (manually drawn in polygons in LabelMe). The manually created object segments were chosen in this first step, to excluded additional challenges and sources of error from image segmentation algorithms. The goal in this strongly controlled experiment is to investigate if the gaze indeed concentrates on given objects. The research questions are extended in the following sections.

In this section, three research questions are addressed:

RQ 1.1 *Is it possible to identify an object, from a given set of objects, by means of gaze data from users who had decided if they can see that specific object on a photo?*

RQ 1.2 *Can the identification be improved when considering inaccurate data?*

RQ 1.3 *Does the aggregation of gaze data gained from several users improve the region identification results?*

In Section 4.1, the conducted experiment is presented, including the experiment data set. The analysis of the recorded gaze data is subject of Section 4.2. Results on user feedback are presented in Section 4.3 while the presentation of the gaze analysis results takes place in Section 4.4. Detailed analysis on specific object region characteristics and the possibilities of distinguishing several objects in one photo are presented in Sections 4.5 and 4.6 before this part of the

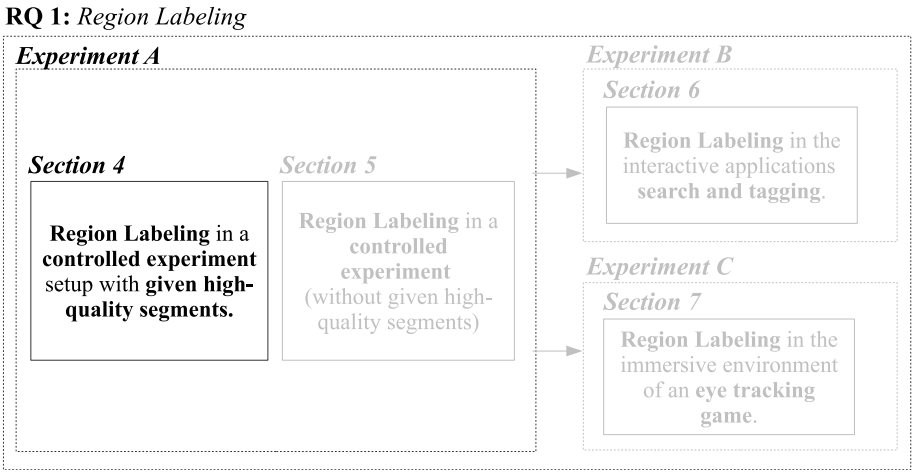


Figure 4.1: Embedding of this experiment (A) in the context of this thesis. The gaze analysis based on given high-quality object region is the first step in the analysis of gaze data with the goal to label image regions.

work is concluded in 4.7. Two papers [WSS13a] and [WSS12] were published about this part of the work. The experiment images, gaze data and results were published under `http://west.uni-koblenz.de/Research/DataSets/gaze`.

4.1 Experiment Setup

In the experiment application image-tag-pairs were presented to the subjects with the task to decide whether an object, described by the tag, is depicted on the image or not. The experiment application has been designed such that first a tag and subsequently an image was shown to the subjects.

4.1.1 Participants

30 participants (9 of them female) attended the experiment. The age of the subjects was between 22 and 45 years (average: 28.7, SD: 6.78). They were undergraduate students (10), PhD students (17), or worked in other professions (3). The subjects received a small present for participating.

4.1.2 Data Set

The LabelMe data set [RTMF08], with in total 182,657 user contributed images (download August 2010), is used as experiment data set. It provides images of complex indoor and outdoor scenes. The LabelMe community has manually created image regions by drawing polygons into the images and tagging them. The labels were used as tags and the regions as a manual, thus high-quality image segmentation. The annotated regions were used as ground truth in the analysis.

For the experiment, images from the LabelMe data set with a minimum resolution of 1000×700 pixels and at least two labeled regions were randomly selected. In average, every image in the selection is labeled with 18.4 tagged regions (SD: 22.4, min: 3, max: 152). 72% of the image areas are covered by the manually drawn polygons in average (SD: 32%, min: 1%, max: 100%). From these images, three sets of 51 images, each with an assigned tag, were created. For each image, a "true" or "false" tag was randomly selected. "True" means that an object described by the tag was labeled on the image. "False" means that no label with the tag was given for the image. These "false" tags had been randomly selected from other LabelMe images. The purpose of creating true and false image-tag-pairs was to keep the participants' attention during the experiment. Some images had to be removed manually from the selected ones when a) the randomly selected false tags by coincidence correlated to some actually visible parts of the image and thus were true tags. Also images where b) the tags where incomprehensible or expert knowledge was required had to be removed. In some cases there were c) not all instances of an object labeled on the image.

4.1.3 Experiment Setup

The experiment was performed on a screen with a resolution of 1680×1050 pixels. The experiment application was implemented as a simple web application running in Microsoft's Internet Explorer. The participants' gaze paths were recorded with a Tobii X60 eye tracker at a data rate of 60Hz and an accuracy of 0.5°. For each image-tag-pair, the following three steps were conducted:

1. First, the tag with the question "Can you see the following thing on the image?" was presented to the participants (see Figure 4.2, left). After pressing the "space" button, the application continued with the next screen.

2. In this screen, a small blinking dot in the upper middle was displayed for one second (see Figure 4.2, center). The participants were asked to look at that point. The red dot animated all participants to start viewing the image (which has been shown next) from the same gaze position. It was placed above the actual image that is shown in the third screen.

3. Finally, the image was shown to the participants (see Figure 4.2, right). While viewing the image, the participants had to judge whether the tag shown in the first screen would have an object counterpart in the image or not. The decision was made by pressing the "y" (yes) or "n" (no) key.

The first image-tag-pair was used to introduce the application to the participants. It has not been used in the analysis. Each participant evaluated one of the three sets consisting of 51 image-tag-pairs from the data set described above. The participants were told that the goal of the experiment was not to measure their efficiency in conducting the experiment task. No time constraints were given.

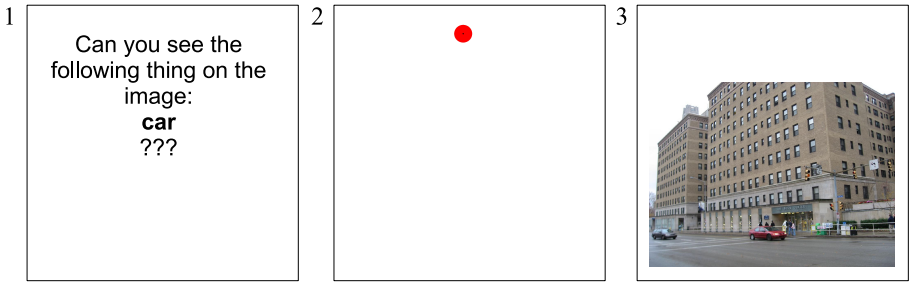


Figure 4.2: The three pages in the experiment setup. 1. Declaration of the object, 2. Fixation point, 3. Decision page.

4.2 Analysis

In this section, the analysis of the gaze data, the baseline approaches and the calculation of precision P for all approaches are described.

4.2.1 Gaze Analysis

In the analysis of the gaze data, only fixations on the images are considered. Fixations on the experiment screen but outside the evaluated image are ignored.

Eye tracking measures are applied to the experiment photo. They are calculated for each given image region over given fixations. The given tag is assigned to the region with the highest fixation measure value. This region is called the favorite region. In the analysis it is investigated which measure provides the highest number of correct aggregations between tag and image region. In total, 13 eye tracking measures are applied. 10 of them are standard measures, integrated in common gaze analysis software such as Tobii Studio, or they were introduced in related work. These measures are described in Section 2.4.2. Three additional measures are introduced. As a variation of (1) firstFixation, the measure (2) secondFixation (min count) ignores the very first fixation. This measure was introduced because the very first fixation can be strongly influenced by the position the user fixated before the image was displayed. Two other measures take the specific application characteristics into account and considers the moment of the decision making. (4) fixationsBeforeDecision (min count) considers the fixations before the moment the key was pressed by the user. Gaze paths can contain fixations after the making of the decision by pressing the button on the keyboard, due to the inherent reaction time of the experiment application. The measure (5) fixationsAfterDecision (min count) analysis these fixations. An overview of all measures investigated in this part of the thesis is given in Table 4.1.

Aggregation of Gaze Paths

The eye tracking measures can be calculated for the fixations of a single gaze path, recorded from one single user. Cumulative interest in a location is often a

No	Name	Description	Favor.	Origin
1	firstFixation	Number of times the participant fixates on the image before fixating on region r for the first time	min count	Tobii
2	secondFixation	Number of times the participant fixates on the image before fixating on region r for the first time without the first fixation on the image	min count	New
3	lastFixation	Number of times the participant fixates on the image after last fixation on region r	min count	[Kla10]
4	fixationsBeforeDecision	Number of times the participant fixates on the image after the last fixation on r and before the decision	min count	New
5	fixationsAfterDecision	Number of times the participant fixates on the image after the decision and before the fixation on region r	min count	New
6	fixationDuration	Sum of the duration of all fixations on r	max seconds	Tobii
7	firstFixationDuration	Duration of the first fixation on r	max seconds	Tobii
8	lastFixationDuration	Duration of the last fixation on r	max seconds	New
9	fixationCount	Number of times the participant fixates on r	max count	Tobii
10	maxVisitDuration	Maximum visit length on r	max seconds	Tobii
11	meanVisitDuration	Mean visit length on r	max seconds	Tobii
12	visitCount	Number of visits within r	max count	Tobii
13	saccLength	Length of saccade before fixation on r	max centim.	[KKK09]

Table 4.1: Applied eye tracking measures f_m including three new measures.

valuable measurement. Particularly when the problem of distinguishing between the scanning of a photo and the fixating of an important object on the photo has to be solved. In the experiment design, several users viewed the same image with the same tag given. When the region showing an object described by this tag should be identified in the photo, the gaze paths of all users can be aggregated and the eye tracking measures can be calculated based on all this data. Here, only gaze paths of users who correctly identified a tag as correct are considered. It can be that participants that gave an incorrect answer, did not see the object. In a real-world scenario, these incorrect answers either has to be identified or the gaze paths of all users have to be included in the analysis.

Extending Object Boundaries

Two additional parameters for identifying correlations between tags and image regions are investigated, dealing with the specific characteristics of eye tracking data.

The first parameter is an extension of the region boundaries to deal with the inaccuracy of eye tracking data. One obstacle in the identification of image regions from gaze information is the inaccuracy of the eye tracker (see Section 2.2.1). It is investigate if this measurement uncertainty can be diminished by extending the region boundaries. By this, fixations near to a region are also considered belonging to the region. Values for the region extension $d = 1 \dots 35$ pixels are analyzed.

Weighting Small Objects

The second parameter deals with the fact that larger image regions are likely to be fixated by coincidence than smaller regions while the participant is scanning the image on the search for an object. It is analyzed, if the tag-to-region assignment quality can be improved by adding a linear weighting function to support smaller regions. The weighting depends on the image region size in relation to the total image size. $f_m(r)$ with $m = 1 \dots 13$ is a measure functions applied on region r as described in Table 4.1.

In the following, the linear weighting function $weighted\text{-}f_m$ on an image region r is considered:

$$weighted\text{-}f_m(r) = \begin{cases} f_m(r) \cdot weight(s_r) & \text{if } s_r \leq T \\ f_m(r) & \text{else} \end{cases} \tag{4.1}$$

with

$$weight(s_r) = \frac{1 - M}{T} s_r + M$$

The relative region size s_r is calculated from the size of the region in pixels divided by the image size in pixels. The measure is weighted with a factor only when $s_r \leq T$, where T is a predefined threshold. Thus, only image regions up to a specific size gain from the weighting function. The weighting factor itself is calculated depending on the threshold T and the maximum weighting value M. In the analysis, the parameters M and T of the weighting function by

calculating the precision results of all images for $T = 0 \ldots 1$ and $M = 1 \ldots 50$ are investigated. An example of applying the *weighted-f_m* for $T = 0.05$ and $M = 4$ is shown in Figure 4.3. Here regions of size between 0% and 5% of the actual image size are weighted with a factor between 1 and 4.

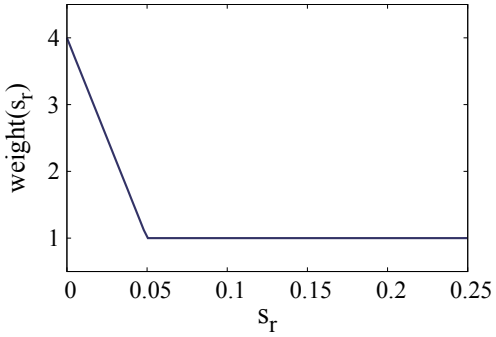


Figure 4.3: Example weighting function for $T = 0.05$ and $M = 4$.

4.2.2 Baselines

The baseline results and the gaze results are computed and compared for the same input images. The baseline approaches use exclusively the same data as the gaze approach — excluding the eye tracking data. This is the set of images, their tags, and the manually created image regions obtained from the LabelMe data. The use of methods based on a training set, methods requiring a training period, or methods that support a limited number of pre-defined concepts only (such as typical object detection algorithms) are hard to compare to the proposed gaze approach as they require additional input data or a bigger data set.

Three baselines are applied for comparing the gaze-based approach to other approaches that are not based on the usage of eye tracking information. These baselines are (a) a "random" baseline [KSDK08], (b) a baseline based on the calculation of the most salient points on the image [NI05] [Row02], and (c) a "naive" baseline [KKK09]. The random baseline (a) randomly selects one of the labeled regions of the image as favorite. The saliency baseline (b) assumes the depicted object at the most salient points on the images. The salient points were calculated by the toolbox offered by Itti et al. [IKN98]. The favorite region is selected by using the salient points and their ordering as computed by Itti et al. and interpreting them as simulated gaze paths for the gaze analysis method. The eye tracking measures introduced above and shown in Table 4.1 are used to compute the favorite region from the saliency map. The naive baseline (c) makes the assumption that the area in the center of an image should be the favorite one. It was chosen because it is common that photographers position important motives in the middle of the image.

4.2.3 Calculating the Precision of Tag-to-Region Assignments

The whole procedure for calculating the tag-to-region assignments is illustrated in Figure 4.4. The single steps for each fixation measure are:

1. For each region in an image b) a value for a fixation measure is calculated for each gaze path c).

2. For each region, the fixation measure results for each gaze path are summed up. From this, an ordered list of image regions for a fixation measure that determines the favorite region d) as described before is obtained.

3. The label of the favorite region is compared with the tag a) that was given to the participant in the experiment. If label and tag match, the assignment is true positive tp, otherwise it is a false positive fp. The total number of correct and incorrect assignments is summed up over all images and the precision P for the whole image set is calculated using the following formula:

$$P = \frac{tp}{tp + fp} \tag{4.2}$$

For the baseline approaches, the selected favorite region is also evaluated by calculating precision P. The results of the gaze-based and the baseline approaches are compared subsequently.

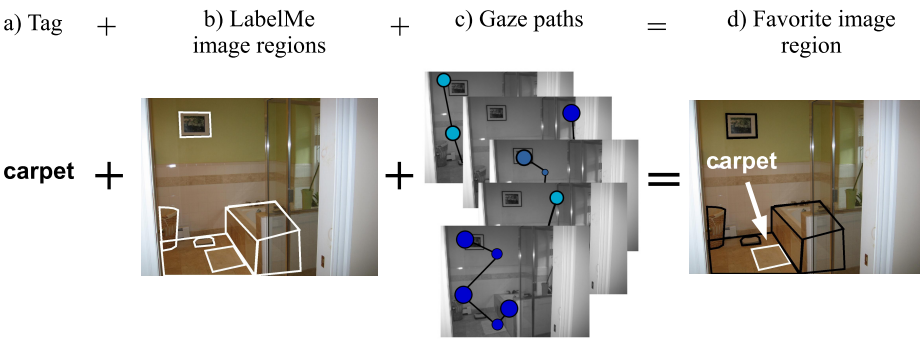


Figure 4.4: Overview of calculating the tag-to-region assignments.

4.3 Evaluation Results for Effectiveness, Efficiency, and Satisfaction

Besides recording the raw gaze data, also the time the participants took to make a decision per image and the correctness of the answers was taken. In addition, the participants were asked to express their emotions during the experiment on a 5-point-Likert scale where a value of 1 means strong disagreement and a value of 5 stands for strong agreement.

4.3.1 Effectiveness

It was measured how many image-tag-pairs have been correctly classified by the participants. Correctly classified means that a true tag is confirmed with "yes" and that a false tag is decided with "no" in the experiment application. In total, 1,500 answers, 10 answers per image-tag-pair, were given. 5.4% of the given answers of all participants were incorrect. The proportion of wrong answers for true (5.8%) tags is close to the value for false tags (4.8%). The highest number of wrong answers for one image-tag-pair is 8, that is, most of the users did not correctly identify whether the tag given was true or false. In this work, only the gaze paths of participants having successfully identified a tag as true or false. Only image-tag-pairs with a true tag and a given the correct answer were analyzed.

4.3.2 Efficiency

The average answer time over all images is 3.00 ms (shortest answer time is 204 ms and the longest is 25.16 ms). 50% of the answers are given in a time between 1.42 ms and 3.92 ms. For true tags, the average answer time over all participants and all images is 2.82 ms, for false tags it is almost twice as long with 3.85 ms. Also the number of fixations on the image is higher for false tags (13 fixations in average) than for true tags (9.6 fixations). In an independent-samples Mann-Whitney U Test the answer durations and number of fixations measured for true and false tags were compared. For both tests a significant difference with $p < .0001$ was obtained. This means that the participants look longer and more precisely on images when there is no object related to the provided tag.

4.3.3 Satisfaction

Concerning the statement "It was easy to decide on an answer.", the participants answered on average with a score of 3.85 (SD: 0.59). 15 participants agreed or strongly agreed with the statement. Most of the participants felt comfortable during the evaluation (average: 4.4, SD: 0.75). 11 strongly agreed and 6 agreed to the statement. Thus, it can be assumed that the results obtained from the experiment application are not influenced by side effects such as users feeling discomforted in front of the eye tracker.

4.4 Results of Finding Objects in Images

In total 1,500 gaze paths were recorded (30 users, each viewed 50 images) during the experiment. Each of them contains fixations on the presented image. An average number of fixations per image over all images and all users is 10.9 (SD: 9.2, min: 1, max: 112). The gaze information was also recorded during and after the decision making by pressing of the "y"- or "n"-button on the keyboard. 88% of the records contain fixations after the decision before the next page of the experiment application was shown.

Only the gaze paths from images with a true tag and a correct answer given by the user were used in the analysis (see Section 4.1.3). In cases where the participants gave incorrect answers, it cannot be known if a participant did not took enough time to examine the image, if he/she did not understand the given tag, or if other problems occurred. 799 gaze paths were collected during the experiment that fulfill the requirement. 656 (82%) of these gaze paths have at least one fixation inside or near (10 pixels) a correct region.

The preprocessing of the raw eye tracking data for identifying fixations is performed with the fixation filter offered by Tobii Studio with the default velocity threshold of 35 pixels and a distance threshold of 35 pixels (see Section 2.4.1 on preprocessing of the raw eye tracking data).

4.4.1 Best Eye Tracking Measures

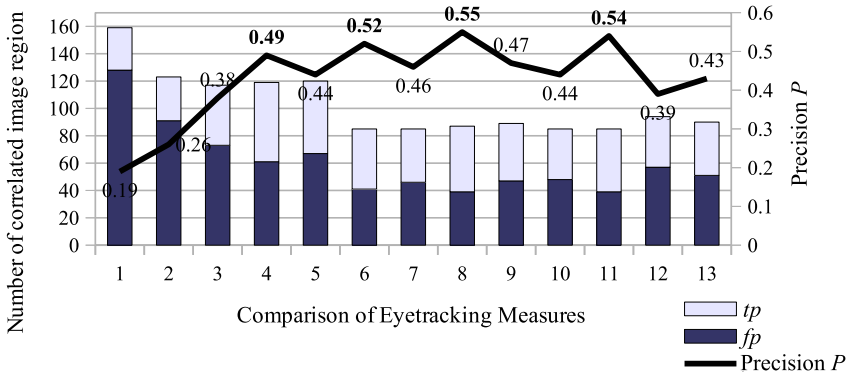


Figure 4.5: Precision for the eye tracking measures from Section 4.2.1 calculated from tp (true positive) and fp (false positive).

The results for all measures are presented in Figure 4.5. For each measure the tp and fp results and the precision P, calculated as described in Section 4.2.3, are depicted. The best result was obtained for the measure (8) lastFixationDuration with precision $P = 0.55$. That means, 55% of the image regions selected by the gaze analysis are described by the tag shown to the participants. The second best value with $P = 0.54$ is (11) meanVisitDuration, followed by (6) fixationDuration with precision $P = 0.52$. The fourth best result is $P = 0.52$ for (4) fixationsBeforeDecision. Among the top four measures, two measures take the moment of decision into account: (8) lastFixationDuration, (4) fixationsBeforeDecision. The lowest precision results were 0.19, and 0.26 for (1) firstFixation and (2) secondFixation. These measures are using the first fixations on an image and the fp values are very high. This problem is further examined in Section 4.5.4.

Figure 4.6 shows some examples of successfully identified tag-to-region assignments. A closer look at the image region characteristics and a qualitative description of the incorrect correlations can be found in the detailed analysis presented in Section 4.5.

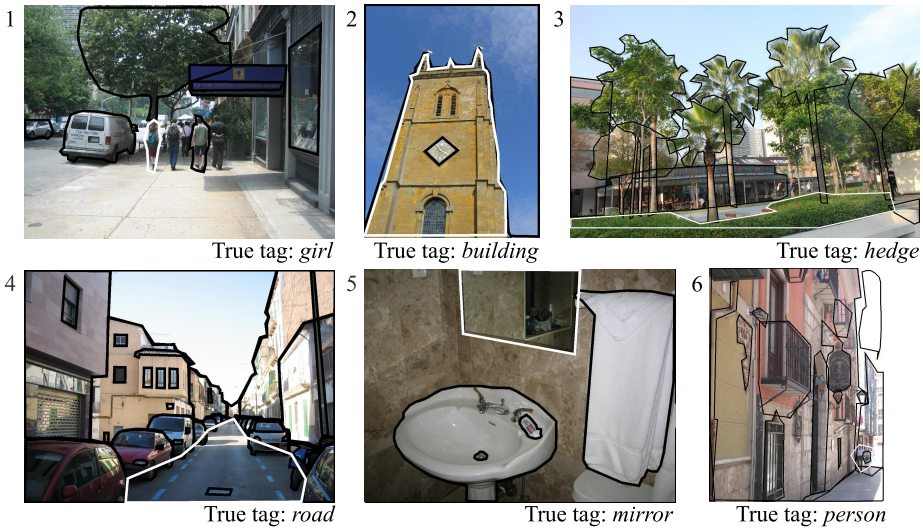


Figure 4.6: All labeled regions (black borders) and correctly identified favorite objects (white borders).

A possible influence of the complexity of a scene on fixations measures was analyzed. As measure for the complexity of a scene, the number of tagged regions per image was used. The number of tagged regions nt is clustered according to the three quartiles ($Q_1 = 6.25$, $Q_2 = 11.5$, $Q_3 = 21$). The maximum difference $diff$ between the precision results for different quartiles for one measure is also calculated. The results are depicted in Figure 4.7. For each measure from (1) firstFixation to (13) saccLength the precision P is calculated separately for images with a number of tagged regions between 0 and Q_1, Q_1 and Q_2, etc . In general, more correct assignments were performed for images with less tagged regions. This finding is not surprising as it is easier to perform a correct assignment by chance for less complex scenes with less regions. The influence of the scene complexity is varying between the measures. The three best performing measures have an average $diff$ value between 0.33 and 0.35. The measure (5) fixationsAfterDecision with the smallest result $diff = 0.21$ shows an average precision performance.

4.4.2 Extension of Region Boundaries

The influence of the extension on the precision of the three best performing measure (8) lastFixationDuration, (11) meanVisitDuration and (6) fixationDuration is described in this section. The precision increases when applying the extension parameter. The best result was precision $P = 0.6$ for (8) lastFixationDuration with $d = 18$ as shown in Figure 4.8. This corresponds to an

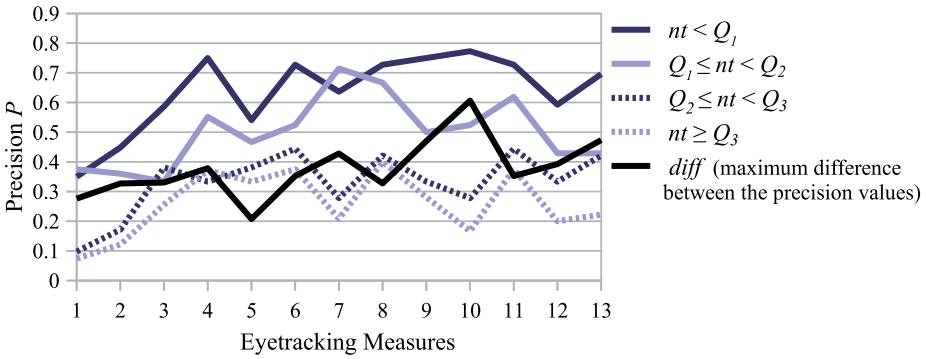


Figure 4.7: The precision P compared for different levels of scene complexity (measured by the number of tagged regions nt).

improvement of about 9%, compared with the result of $P = 0.55$ without extension. A baseline is added to each diagram, displaying the precision results without extension. The precision is even below ($> 1\%$) the threshold for $d < 6$ for (8) lastFixationDuration, $d > 32$ for (6) fixationDuration, and $d > 29$ for (11) meanVisitDuration.

The results suggest that it is reasonable to include the extension of region boundaries in the calculation of tag-to-region assignments. The precision is fluctuating depending on the chosen extension value d. In the investigations, the best results were obtained for $6 \leq d \leq 29$.

4.4.3 Weighting function

The best precision applying the weighting function on the fixation measure (8) lastFixationDuration is $P = 0.56$, the worst result is $P = 0.47$. (for (11) meanVisitDuration best $P = 0.6$ and worst $P = 0.53$, for (6) fixationDuration: best $P = 0.54$ and worst $P = 0.48$). These results were obtained from different combination of M and T. In Figure 4.9, the results for the two weighting parameters are displayed. As baseline precision, the precision results obtained without extension and weighting (see Section 4.2.1) are considered. Values equal to this baseline are marked in white. Values higher than and lower than the baseline precision are highlighted in the figure in red respectively blue.

From the results depicted in Figure 4.9, one can see that the influence of parameter T is higher than the influence of M. The precision is strongly varying. Every chart (a) to (c) shows an area of highest values for $0.04 < T < 0.1$. The precision decreases for every measure from $T > 0.13$ but also here good precision results can appear for higher T.

The usage of the weighting function can improve the results. However, the precision can also decrease. Further investigations are necessary to better explain the fluctuation of the graph.

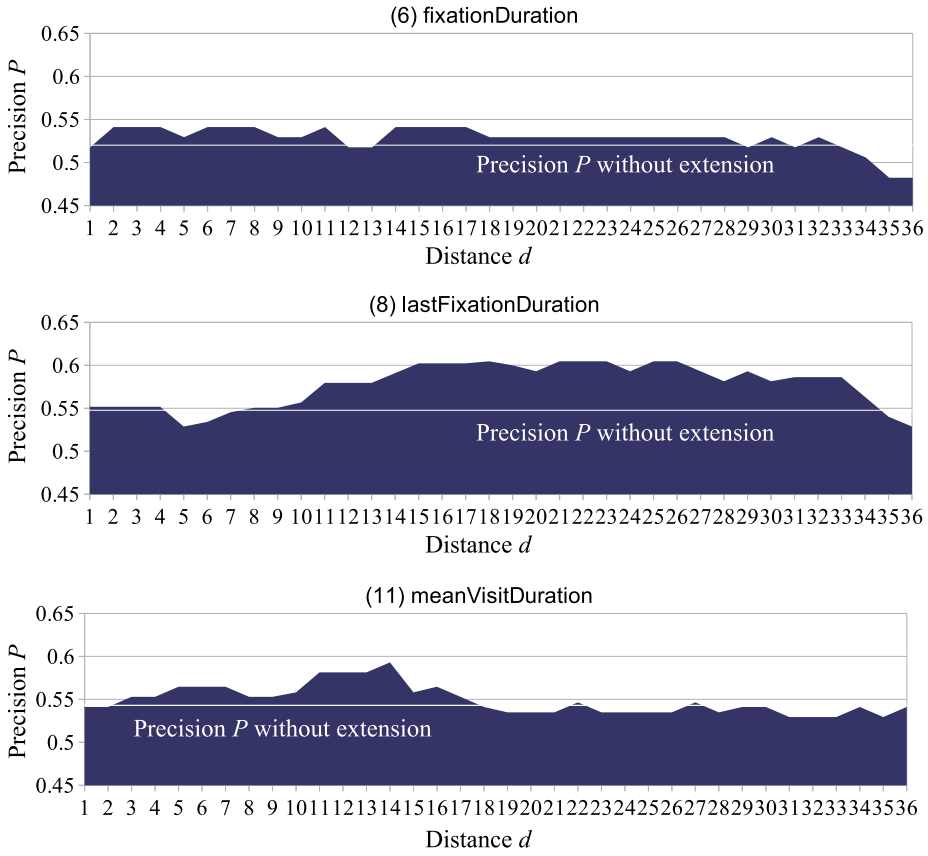
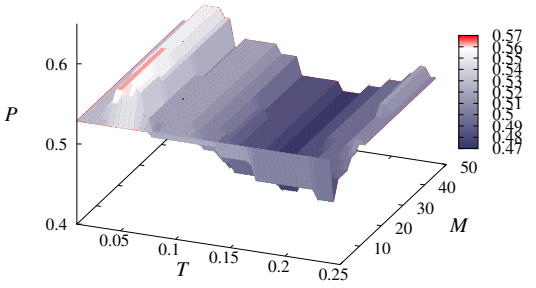
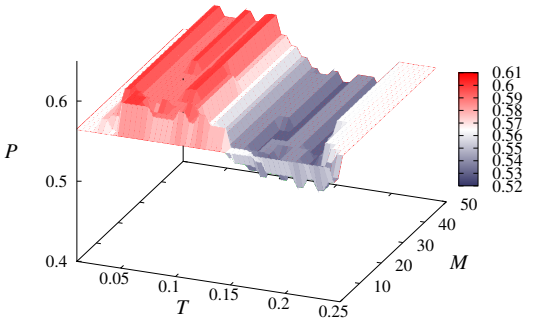


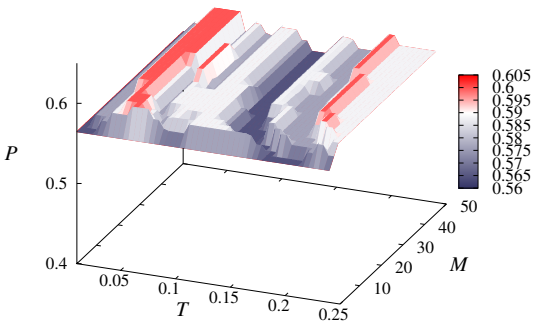
Figure 4.8: Influence of different extension parameters d on the precision results for three eye tracking measures.



(a) (8) lastFixationDuration



(b) (11) meanVisitDuration



(c) (6) fixationDuration

Figure 4.9: Influence of the weighting function on precision P for three different eye tracking measures (white: baseline without weighting).

4.4.4 Combination of Region Extension and Weighting Function

Finally, the three best performing eye tracking measures were combined with both parameters, the region extension and the weighting function. The best precision for fixation measure (8) lastFixationDuration was $P = 0.62$, the worst result is $P = 0.49$. The best result was delivered by (11) meanVisitDuration with $P = 0.63$, including extension $d = 10$ and weighting (e.g., $T = 0.05$, $M = 4$). For (6) fixationDuration the best result was $P = 0.61$, the worst $P = 0.51$.

4.4.5 Comparison of the Eye tracking Approach with three Baselines

The precision results obtained by the gaze approach are compared with the three baselines described in Section 4.2.2. The results in Figure 4.10 show that the random baseline has an average precision of 0.17 over 30 samples (SD: 0.04, min: 0.1, max: 0.26). The saliency approach has a best precision of 0.21 for the measure (11) meanVisitDuration, followed by a precision of 0.20 for (1) firstFixation. The worst result was obtained with a precision of 0.15 for the measure (2) secondFixation. The naive approach achieves a precision of also 0.21. These baseline results are compared with the gaze-based approach with precisions between 0.52 and 0.55 for the measures (6), (8), and (11), and between 0.61 and 0.63 for the measures with extension and weighting. The identification of assignments based on gaze or on gaze including extension and weighting performs better than the baseline approaches. 18 Chi-square tests were performed to investigate significant differences between the approaches. They all show a statistical significance at level $\alpha < 0.05$. The least significant result with $\chi^2(1, N = 124) = 10.723, p < 0.0015, \phi = 0.162$ was obtained for the naive baseline and measure (6) fixationDuration without extension and weighting.

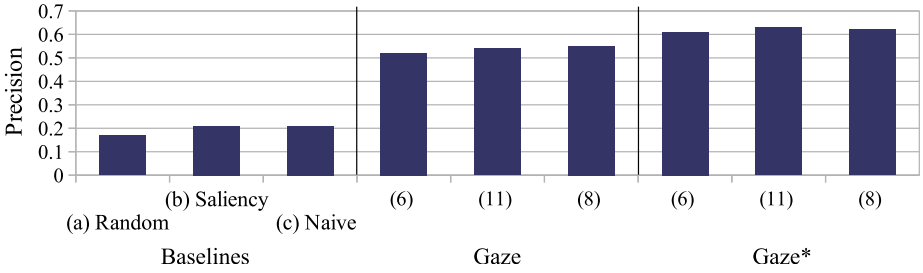


Figure 4.10: Precision for three baselines approaches and gaze based analysis.

4.5 Detailed Analysis of Image Region Characteristics and Gaze Paths Patterns

The best precision $P = 0.63$ for measure (11) meanVisitDuration (including extension and weighting) was obtained from 54 tp and 32 fp assignments. First

in this section, a qualitative analysis of the fp assignments is presented. Subsequently, it is investigated if there are typical characteristics concerning region sizes or positions of image regions for correct and incorrect tag-to-region assignments, followed by a look into typical patterns for the first fixations. Finally, the effect of aggregating gaze paths of several participants is investigated.

4.5.1 Qualitative Analysis of Incorrect Assignments

Some examples of incorrect assignments can be seen in Figure 4.11. The white boundaries show the objects that corresponds to the tag given to the participants. The black boundaries show the objects determined as favorite from the gaze information. The correlations are calculated with measure (11) meanVisitDuration including extension and weighting. From an qualitative analysis of the 32 wrongly assigned tags, the following characteristics were identified:

- Some images show scenes with a small correct object also had small wrongly selected favorite object which were located next to the correct object (cf. images 1 and 2). Six images belonged to this category. These wrong assignments can be caused by the inaccuracy of the eye tracker.

- In some images, the correct object is displayed within another object (cf. image 3, *lamp* inside *wall*). In these cases, the outer region is identified as favorite. That means the weighting function does not work for all occurrences of smaller regions. Eight images belonged to this category.

- Further images show scenes with an object that seems to be very easy to identify. For example, larger objects such as *road* (cf. image 4), *sky* or *tree* might be perceived even in the corner of the human eye or based on context knowledge (e.g., sky is above sea is above sand in a beach scene). Nine images belonged to this category. This is a basic limitation of the provided approach but it appears infrequently in comparison to the number of all shown images.

4.5.2 Comparing the Region Size for Correct vs. Incorrect Assignments

The average size of the LabelMe regions in the images used in the experiment is 66,3811 pixels. The average region size for correctly assigned regions tp is 123,609, for incorrectly assigned regions fp 214,704 pixels. The region size of the selected favorite regions (tp or fp) is clearly larger than the average region size. Thus, larger regions are selected with a higher probability for tag-to-region assignments by the gaze-based approach. It is also interesting to notice that the average region size of fp assignments is about 70% larger than the region size of tp assignments.

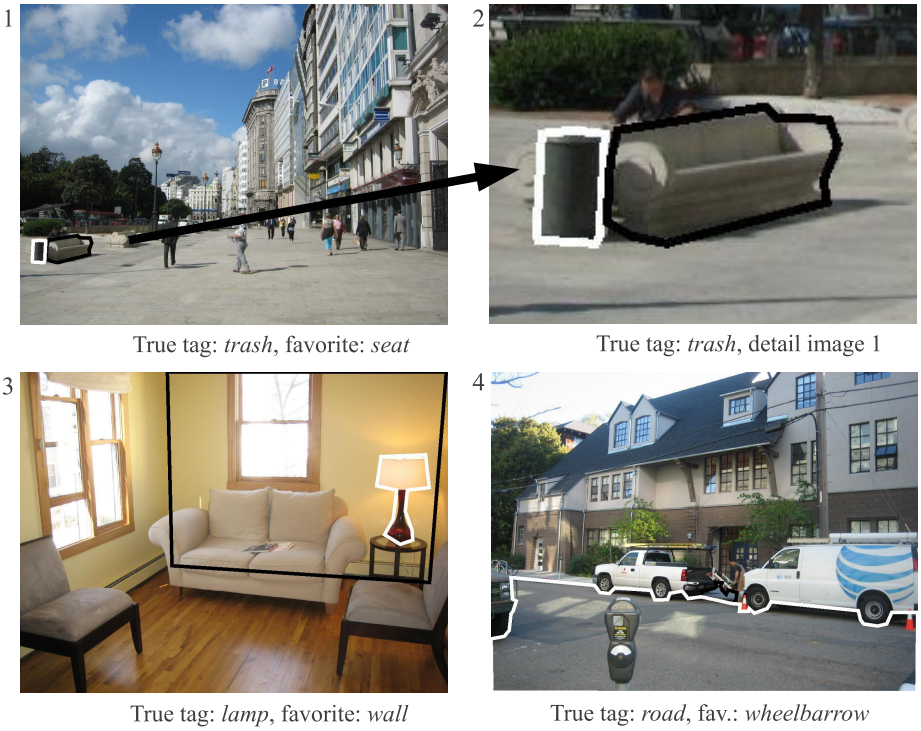


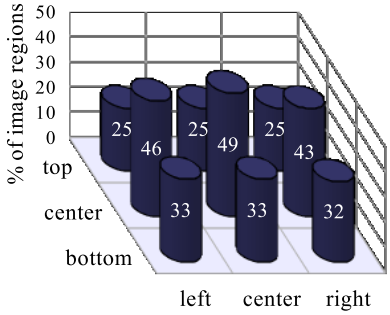
Figure 4.11: Examples of image-tag-pairs with given tags (white shape) and incorrectly identified favorites (black shape).

4.5.3 Comparing the Region Positions for Correct vs. Incorrect Assignments

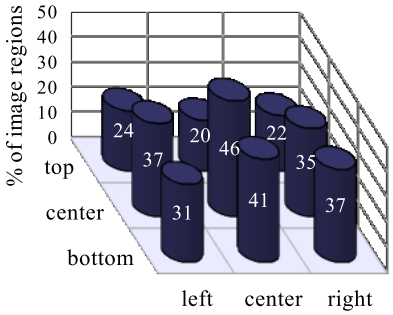
The images were divided into nine uniform areas. Based on these areas, the positions of the assigned regions were investigated. The percentage of image regions having an overlap with the particular area is calculated. In Figure 4.12(a), the positions of all regions in the data set corresponding to true tags are depicted. 49% of the regions overlap with the center field of the image. In the upper third of the images only one fourth of the regions is located. In the lower areas it is about one third. This can be explained by how people take images, for example, with the object in the center of the image and sky or the ceiling in the upper areas. The differences between the left and the right areas are very small. In Figure 4.12(b) and (c), the positions of correctly and incorrectly assigned regions are depicted. One can see in Figure 4.12(b) that the positions of the correctly assigned regions are distributed over the image areas in a similar way as the true-tag image regions (cf. Figure 4.12(a)). For the correct assignments it is not possible to identify a privileged area on the image.

For the incorrect assignments in Figure 4.12(c), one can notice that the positions of the regions concentrate in the center of the image. One can further observe that in the center top part the value is also increased compared with the true-tag image regions. The total number of touched areas is bigger for (c)

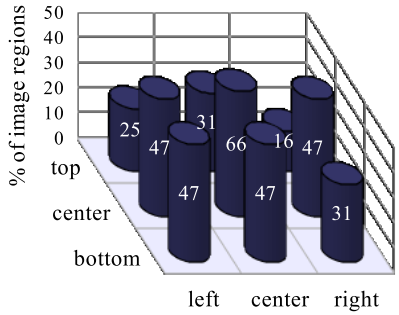
compared with (a) and (b). This finding is based on the bigger size of incorrect areas, as described in Section 4.5.2. This higher percentage of wrongly assigned regions might be caused by a concentration of fixations in the center of the images. This concentration has been observed during the first fixations on the images as shown in the next section.



(a) Positions of all tagged regions



(b) Positions of correct favorite regions



(c) Positions of incorrect favorite regions

Figure 4.12: Percentage of regions located in image areas for (a) all labeled image region in the experiment data set, (b) only correctly identified object regions, and (c) only incorrectly identified object regions.

4.5.4 Bias in the First Fixations

Figure 4.13 shows an illustration of the first five fixations over all participants and all images. One can see that the first fixations are concentrated in the center of the images. Later, the fixations are more distributed over the whole image. This effect is knows as center bias and was described in Section 2.1.3. In the related work, the eye tracking information showing the center bias is collected in free-viewing scenarios (i. e., no specific task was given to the users, they were asked to just view the images). The influence of this bias was not clear in task-driven viewing and a fixated starting point outside the image itself. As can be seen in Figure 4.13, the center bias is highly distinct only for the very first fixations. This is a valuable finding, as the fixation order is considered in the analysis. The weak results for the measure (1) firstFixation and (2) sec-

ondFixation show this problem. In the experiment setup, the participants were asked to look at a red dot — placed above the image position — before the image appeared on the screen (see Section 4.1.3). The influence of this point can be seen in the illustration of the first and second fixations because of the fixations in the upper center of the images. This also provides an explanation for the high value of incorrect aggregations in the center of the images in the previous Section 4.5.3.

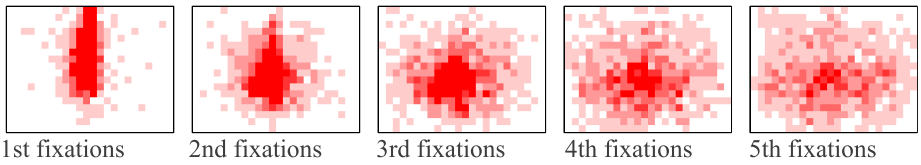


Figure 4.13: Positions of the first five fixations accumulated over all participants and all images.

4.5.5 Effect of Aggregation of Gaze Paths on Precision

Finally, it is interesting to know how many users are needed to accomplish a certain level of reliability in assigning a tag to the correct region. Thus, it was investigated which precision can be reached when aggregating an increasing number of users' gaze paths. Precision results for aggregations from 1 to 10 participants for the measure meanVisitDuration, including extension and weighting, are presented. Precision P was calculated for each possible subset of participants and averaged for all subgroups of the same size. As shown in Figure 4.14, the number of users has a high influence on the precision. With the gaze paths of only single users, an average precision (over all users and all images) of $P = 0.25$ (SD: 0.1, min: 0.16, max: 0.53) was obtained. For the aggregated data of all 10 users the precision increased to $P = 0.63$. This corresponds to an improvement of 152%. The biggest improvements took place between the first group sizes. For example, between one and two users per group an improvement of 46% in average was measured. Between nine users and ten users per group, only an improvement of 7% was observed.

In addition, the range between minimum and maximum precision is depicted in Figure 4.14. The range decreases from the single user results to the multiple user results. Even for single users and single images a good precision can be achieved for some images and regions, respectively. For 10 users there is only one set, therefore no range can be indicated. The big step for the minimum values between the subgroups of 8 and 9 users can also be caused by the small number of only 10 subsets for 9 users. The results based on multiple gaze paths are considerably better than the ones calculated from only a few gaze paths.

The results based on multiple gaze paths are considerably better than the ones calculated from only a few gaze paths. However, the improvement of the precision gets lower when aggregating more gaze paths. Compared with the two baselines from Section 4.4.5, the results for single users are still significantly

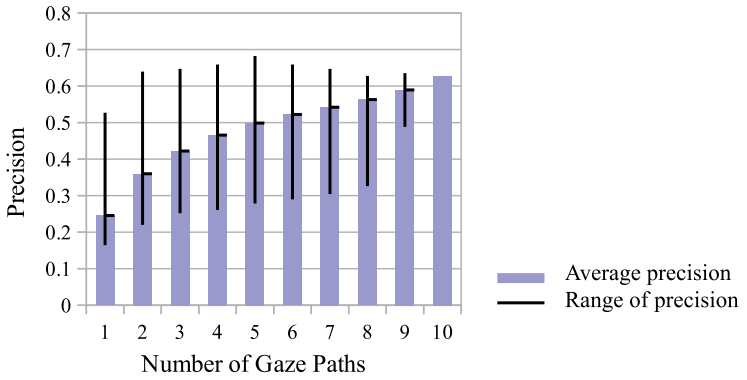


Figure 4.14: Influence of gaze paths aggregation on precision P for numbers of users between 1 (no aggregation) to 10.

better than the naive or random baseline. The Chi-square test provides for the naive approach $\alpha < 0.001$ and for the random approach $\alpha < 0.002$.

4.6 Discriminating Different Objects in One Image

It was investigated if it is possible to differentiate objects in one depicted scene by analyzing the users' gaze paths. Two of the three image data sets from Section 4.1.2 were composed from the same image subsets, which allows to perform this analysis. As a result, two sets of 51 image-tag-pairs each, sharing the same images but different tags, were obtained. All combinations of correct and incorrect tags appear: images with a correct tag for both sets, images with one correct, one incorrect tag and images with two incorrect tags. The data set included 16 true-true image-tag-pairs (tags for both groups were true), 24 true-false image-tag-pairs (one tag was true, one tag was false), and 10 false-false image-tag-pairs. In this section, the investigated measure is again (11) meanVisitDuration, including extension and weighting.

4.6.1 Proportion of Correctly Discriminating Two Objects

For the 16 images with two correct tags, the favorite image regions were calculated. In 6 images, two correct image regions were identified. This is a proportion of 38%. In Figure 4.15, some examples with two correctly identified regions are shown. As the figure shows, the two tags *sky* and *sea* can be distinguished in the upper image. Also the tags *water pot* and *teas* in the lower image can be identified using gaze information. The average probability to identify the correct region in one image is 63% (see Section 4.4). Therefore, the probability to obtain two correct tag-to-region assignments in two different images is 40%. With a value of 38% for two image regions in one image, the probability is close to the probability for two image regions in two different images. Thus, it is

possible to identify different image regions in one image with an accuracy very close to the accuracy of the single assignments. The 16 images with two correct tags provided to the participants has in average 15 tagged regions (SD: 16, min: 3, max: 62). The six images with two correctly identified favorite regions have an average of 17 tagged regions (SD: 22, min: 5, max: 62), whereas the 10 images with one or two incorrect favorite regions has in average 14 tagged regions (SD: 11, min: 3, max: 37). These results indicate that the rate of successfully assigned tags is not or only weakly influenced by the complexity of the depicted scene. An accumulation of the error for detecting multiple objects in one image can lead to an overall low precision of the tag-to-region assignments.

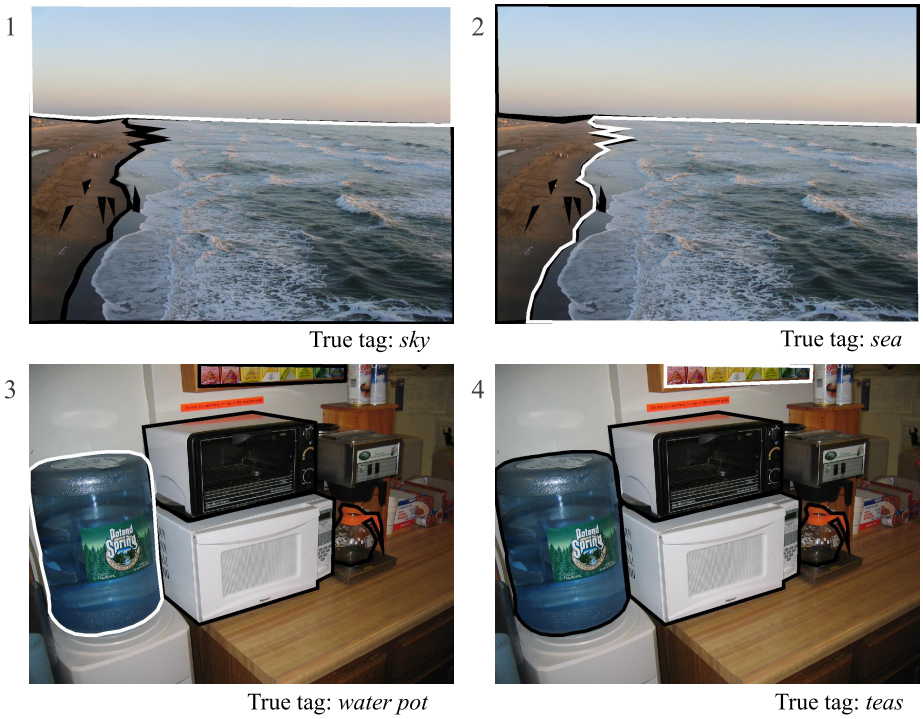


Figure 4.15: Example images with two correctly identified object regions (white borders). Black borders: all given object regions.

4.6.2 Influence of Different Tag Primings on Tag-to-region Assignments

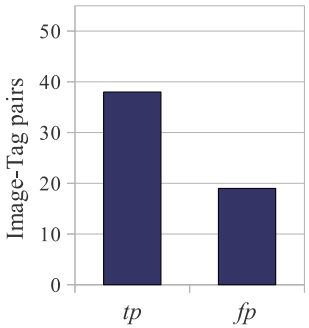
In this section, the influence of the priming by the given tags on the tag-to-region assignments is investigated. Every true tag t_r, assigned to an image, describes one or multiple image regions r. Here, the results for the gaze-approach for users with a provided true tag t_r are compared with results for users, viewing the same image but given a false tag or a tag describing another region on the image. It is measured how often region r is determined as favorite region,

although not tag t_r but another tag was presented in advance. The calculation was performed based on the 16 true-true and 24 true-false image-tag-pairs.

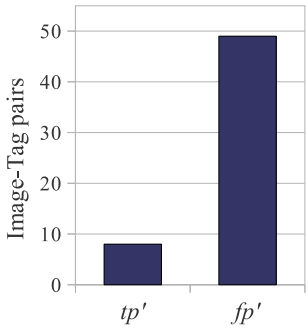
The tp and fp values in Figure 4.16(a) show the results for the assignment of tag t_r to region r from the analysis presented in Section 4.4. A tp assignment means that the favorite region was described by the tag presented to the user. The fp assignments describe the incorrect correlations, that is, when a favorite region was selected that was not r.

tp' in Figure 4.16(b) shows how often a region r was determined as favorite region from the gaze path analysis that did not belong to tag t_r, that is, where a tag is provided to the users that did not refer to region r. Rather, the tag given to the users could be incorrect (true-false image-tag-pairs) or correct for another region in the image (true-true image-tag-pairs). In case of fp', the region r was not identified as favorite. Thus the fp' assignments mean that the investigated region was not described by the tag presented to the user and the region r was not determined as favorite. A low precision of $P = 0.12$ was obtained in the calculations from tp' and fp'. That means that the region r referring Figure 4.16(a) was rarely selected when a tag was shown to the user that did not correlate to the image at all or correlated to a region different from r.

The results show that the assignments to region r by providing a tag referring to a different region or not referring to any region at all are significantly lower compared with the assignments based on true tags of region r. The result of a Chi-square test shows that the difference is significant with $\chi^2(1, N = 114) = 32.8005, p < 0.0001, \phi = 0.5364$. The correct assignments did not appear coincidentally but strongly depend on the gaze paths, guided by the given tag.



(a) Results for users given a tag t_r describing region r (see Section 4.4)



(b) Results for users NOT given tag t_r for the same images as in (a)

Figure 4.16: Comparing the identification of region r as favorite from gaze paths (a) corresponding and (b) not corresponding to tag t_r.

4.7 Conclusion

The assignment of tags to given object regions by means of gaze data alone was investigated in a first experiment. The gaze data was collected in a controlled experiment where the users had to decide whether they can see a given object on presented photo. It could be shown that 82% of these gaze paths had at least one fixation inside or near (10 pixels) an region region, showing the asked object. 13 eye tracking measures were investigated for analyzing the gaze data and for assigning the tags to an image region. The best result was obtained for the measure (8) lastFixationDuration with precision $P = 0.55$. Taking the extensions of region boundaries into account as well as weighting of smaller regions improves the results. The best performing fixation measure correctly assign tags to regions for 63% of the image-tag-pairs and significantly outperformed three baselines (random, saliency-based, and naive).

No limitation on typical visual objects' appearances (such as size) or positions were observed. Investigating the first fixations on the image explains the low precision results of measures such as firstFixation and shows the center bias. More incorrect tag-to-region assignments are made in the center than correct assignments, what can also be caused by the center bias. The result showed the potential of gaze path aggregation, which means that the gaze data of several users is aggregated in the gaze analysis. An improvement of 152% was measured when comparing the results for single gaze path analysis with the results for 10 aggregated gaze paths. The potential of discriminating different objects in the same image was studied. Here, it could be shown that two regions in the same image with different primings can be identified with an accuracy of 38%.

Summarizing the findings of this first experiment, it could be shown that the identified tag-to-region assignments were not a matter of chance but are the results of analyzing the users' gaze path, as shown by evaluating the effect of different primings such as providing different tags.

Chapter 5

Image Region Tagging with Given Tags

The understanding of photo content is still a challenge in automatic image processing. Often, tags are used to manually describe the content of images. Another approach is to analyze the text surrounding an image, for example, on web pages and to draw conclusions about the depicted scene. A better understanding of the objects depicted in an image can improve the handling of images in many ways, for example, by allowing similarity search based on regions [KY08] or by serving as ground truth for computer vision algorithms [RTMF08].

Eye tracking data can be used to assign given tags to given object regions in order to describe the depicted scene in detail. This is the result of the first step in the direction of region labeling that was described in the previous Section 4. The analysis took place based on given high-quality object segments, which are usually not available for photos. Thus, the next step is to avoid using these segments in the analysis. The manually labeled objects are still used as ground truth data for evaluating the labeling results. Two novel eye-tracking-based measures for conducting tag-to-region-assignments are introduced and compared in this section. The first measure is the *eye-tracking-based measure I Segmentation Gaze*. It is based on a standard image segmentation algorithm [AMFM11] and selects the image segment as most relevant for the given tag by means of fixation information. The second measure is the *eye-tracking-based measure II Heat Map Gaze*. It is based on a traditional eye tracking heat map. Both measures are applied on gaze data obtained from the experiment with 30 subjects described in the previous Section 4. The experiment details like the setup, the data set and the participants are described in the Section 4.1.

How this section is embedded in the whole context of this work, is depicted in Figure 5.1. As an extension of the work presented in Section 4 it does relies on manually created polygons describing depicted objects. However, the data was still collected in a strongly controlled experiment. This work pave the way for region labeling performed in real-world applications. In this section, the following research question is tackled:

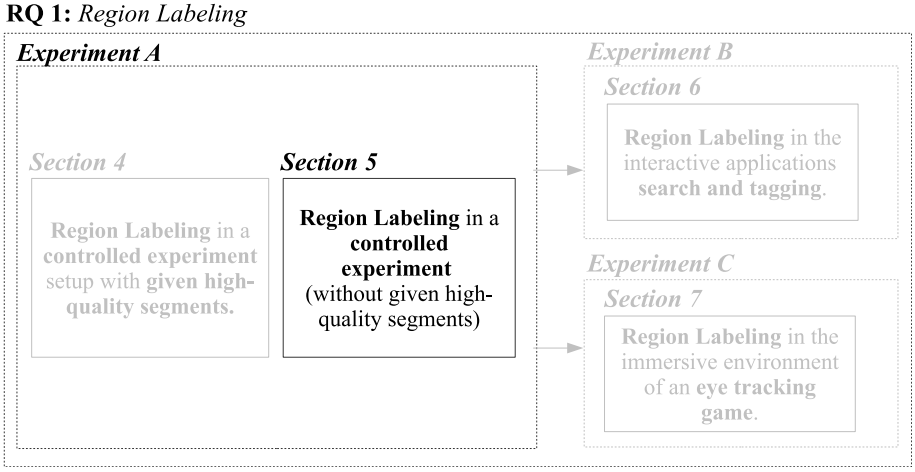


Figure 5.1: Embedding of the analysis presented in Section 5 in the context of this thesis. This second step in the region labeling approach does not rely on high-quality segments.

RQ 1.4 *Can objects on photos be identified from gaze analysis when no high-quality object regions are given?*

In detail, the question is answered by treating the following two questions facing different evaluation measures:

- To which extent may the two new eye-tracking-based measures identify the correct position of an object in the image for a given tag (maximum precision)?

- To which extent does the area determined by the two new measures cover the actual object depicted in the image (maximum F-measure)?

It can be shown that the *I Segmentation Gaze* measure performs better for both questions, although the difference to the *II Heat Map Gaze* measure is not significant. The I Segmentation Gaze measure delivers significantly better results for precision and F-measure than the baseline approaches.

It can be shown that the labeling is indeed possible without the usage of given object regions and still baseline approaches can be outperformed. In Section 5.1, the two novel eye-tracking-based measures and the baselines are introduced. The examination of the best parameters determined on a subset of the images is presented in Section 5.2 followed by the results obtained from the experiments in Section 5.3. The section is concluded in 5.4 The results of this research were published in [WSS13b].

5.1 Gaze Analysis and Baselines

Two methods for assigning tags to image regions, thus identifying objects that correspond to a predefined tag, are proposed in this work. Both methods proceed using the following input:

1. A photo o is a set of pixels $p(x, y)$, $0 \leq x < width$, $0 \leq y < height$
2. A tag t, describing an object depicted in o
3. A set of users U that have viewed the images during the experiment
4. Set of gaze paths provided by users $u \in U$, to which the tag t was shown and who had to decide whether an object described by t can be seen in the image or not

The baseline methods perform the assignments without the usage of gaze data. A Gaze path G consist of fixations and saccades. A fixations f is a short stop that constitute the phases of the highest visual perception, while saccades are quick movements between the fixations. Each gaze path G_t consists of a set of fixations F, provided by user $u \in U$. Every fixation $f = (x_f, y_f, d)$ is described by a fixated point in the image (x_f, y_f) and a duration d. To measure the human visual attention, the fixations are analyzed by so-called eye tracking measures. From these eye tracking measures, a measure $f_m(r)$ is calculated for given regions r of a photo o. Example eye tracking measures are the fixationCount, a standard measure which counts the number of fixations on a region and the lastFixationDuration, which sums up the duration of the last fixation on an image region. 13 eye tracking measures were compared in the previous section (see 4.4) with respect to their ability to identify a concrete image region for a tag t given to the users. Derived from the results of this work, the measure lastFixationDuration is used in the analysis of this section.

5.1.1 Eye-tracking-based Measure I Segmentation Gaze

The idea of the first approach is to calculate $f_m(r)$ for the eye tracking measure lastFixationDuration for all regions $r \in R$ gained from the automatically segmented image. $f_m(r, u)$ is calculated for every user $u \in U$ viewing the image. The values f_m are summed up for every region over all users and the favorite region r_{fav} is determined by the highest value:

$$r_{fav} = \arg\max_{r \in R} \sum_{u \in U} f_m(r, u) \tag{5.1}$$

5.1.2 Eye-tracking-based Measure II Heat Map Gaze

Heat maps are two-dimensional graphical representations of a number of gaze information. They visualize the frequency of fixations for every pixel $p = (x, y)$ in an image. Different colors symbolize how many times or how long a pixel was fixated. The advantage of heat maps is that they can summarize a large quantity of data and are easy to comprehend by humans. Thus, they are often

used in usability experiments to visualize users' attention. Different kinds of heat maps can be created based on different eye tracking measures, for example, a fixationCount or an absoluteDuration heat map [Boj09]. As the lastFixationDuration was the best measurement for the region identification in section 4, this measure is used as basis for this approach. A radius rd has to be defined for the creation of a heat map. A default value of 50 pixels is used, taken from Tobii Studio [tob10]. A maximum value of $h_{max} = 100$ is assigned to the pixel fixated by a fixation $f = (x_f, y_f, d)$. Starting from this point, values are added to the pixel in the surrounding of the fixation, based on a linear interpolation between h_{max} and 0. The result is multiplied by the fixation duration d. An example is visualized in Figure 5.2(a). For a single fixation, the heat map values h are calculated for all pixels $P = (x, y)$ in the surrounding of the fixation:

$$h(P,f) = \begin{cases} d * (h_{max} - \left(dist(P,f) * \frac{h_{max}}{rd}\right)) & \text{, if } dist(P,f) \leq s \\ 0 & \text{, otherwise} \end{cases} \tag{5.2}$$

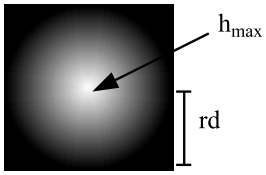
All last fixations f_{last} of all gaze paths provided by the users $u \in U$ are summed up in the final heat map H:

$$H(P) = \sum_{u \in U} h(P, f_{last}) \tag{5.3}$$

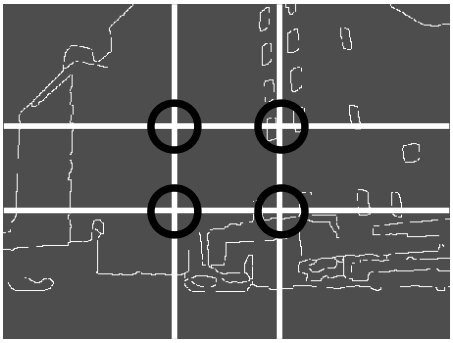
From all heat map values H, the highest value $max(H)$ is determined. To obtain the favorite region from the heat map, a threshold $0 < t \leq 100\%$ is set. For example, $t = 5\%$ means that only heat map values are considered that belong to the highest 5% of all values. This procedure can be described by an analogy of a flooded region with valleys and elevations. The threshold t symbolizes the water level. With a level of $t = 5\%$, only the highest 5% of the landscape are visible above the water level or here all pixels with $H(p) > 0.95 * max(H)$ are determined as possible favorite regions. The biggest area of connected pixels is selected as favorite region r_{fav}. An illustration of this thresholding is presented in Figure 5.6.

5.1.3 Baselines

Initially, a random baseline approach as used in the previous section 4.2.2, which randomly selects one segment of an automatically segmented image as favorite region. As the results of this baseline were very weak, the baseline approach was improved by taking into account the position of the segments in the image in two different ways. As the photos used in the experiment were taken by humans, an inherent photographic bias can be supposed. The golden ratio rule is a very basic rule in photography [Fre07]. Taking images based on this rule can improve the aesthetics of a photograph and it is often met instinctively to achieve aesthetically appealing photos. According to the golden ratio, width and height of an image are divided into two parts in the ratio 1 to 1.618. This results into four intersections, at which important objects in the photos are



(a) Heat map creation for one fixation.



(b) Identification of the four golden section points.

Figure 5.2: Visualization of heat map and golden ratio baseline calculation.

often placed. In Figure 5.2(b), the golden sections are highlighted by black circles. Another typical bias is to position the important object in the center of the image. For each photo, the golden ratio and the center baselines are calculated. The segment placed at the golden section respectively the center point is selected as favorite region r_{fav}.

5.1.4 Evaluation Measures

After obtaining favorite regions with one of the two new measures or the baseline measures, the results have to be evaluated by means of comparing them with ground truth object labels. In information retrieval, precision, recall, and F-measure are standard approaches to measure the relevance of search results.

$$\text{precision} = \frac{tp}{tp + fp} \tag{5.4}$$

$$\text{recall} = \frac{tp}{tp + fn} \tag{5.5}$$

$$\text{F-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{5.6}$$

These measures are used in evaluating the coverage of the ground truth object region r_{gt} by the favorite region r_{fav} at pixel level. The algorithm runs through the image and classifies every pixel as tp (true positive), fp (false positive), fn (false negative), and tn (true negative) as described in Table 5.1.

5.2 Determining Best Parameter Settings

The data set is split into two subsets: a training set for the parameter fitting (56 images-tag-pairs each viewed by 10 users) and a test set for the evaluation of the

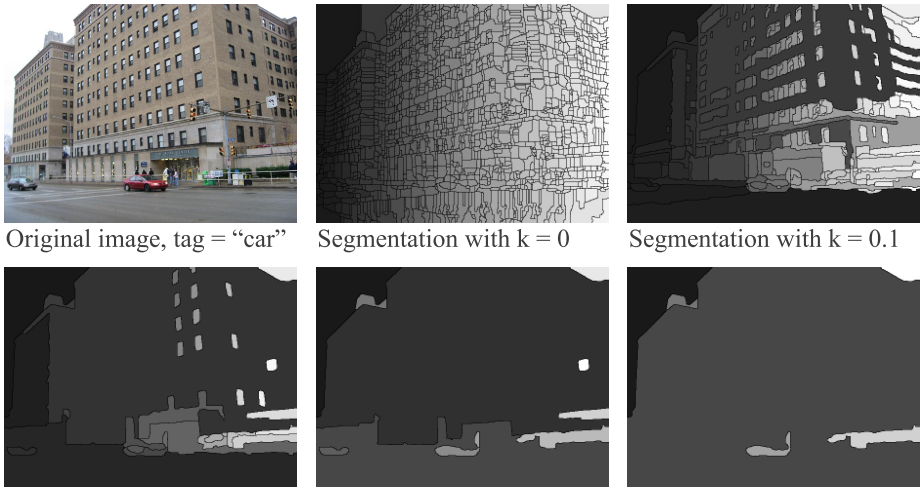


Figure 5.3: Image and its segmentations with different parameters k.

		r_{gt} from the ground truth image	
		Pixel belongs to r_{gt}	Pixel does not belong to r_{gt}
r_{fav} calculated from heat map, segmentation or baseline measure	Pixel belongs to r_{fav}	tp	fp
	Pixel does not belong to r_{fav}	fn	tn

Table 5.1: Calculation of tp, fp, fn, and tn.

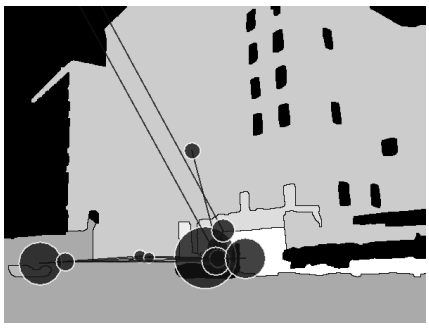
approaches (29 images-tag-pairs each viewed by 10 users). In this section, three different parameters are investigated for the gaze-based approaches and identify the parameters leading to the best results. The outcome is applied to the test data set and used for comparing the different measures from Section 5.1.

5.2.1 Eye-Tracking-Based Measure I Segmentation Gaze

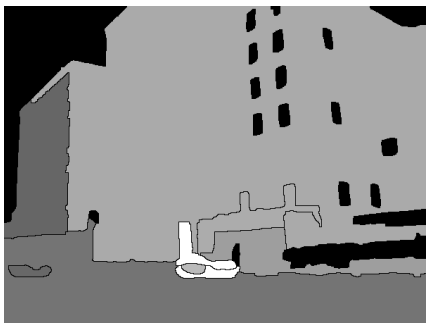
The segmentation is performed by using the gPb-owt-ucm algorithm [AMFM11]. Different hierarchy levels for $k = 0 \ldots 1$ are calculated, each representing a different level of detail. An example is presented in Figure 5.3, showing the

segmentation results for different k-values. The first segmentation level $k = 0$ delivers 1831 segments, the segmentation with $k = 0.4$ the least number of segments, namely six.

Applying eye-tracking-based measures I Segmentation Gaze to those segmentations provides the favorite region r_{fav} from all segments, as described in Section 5.1. In Figure 5.4(a), an example for a gaze path of a single user is shown. The fixations are displayed as circles, the fixation duration is presented by the diameter of the circles. The saccades are depicted as lines between the fixations. The brightness of the image segments encodes the eye tracking measure values f_m. The order of the viewed regions is encoded from the favorite region in white to the segments with few fixations in dark gray. The black segments have not been fixated at all. Figure 5.4(b) shows the results for one image aggregating the gaze paths of all users. To determine the best hierarchy level k, the results for different levels $k = 0 \dots 1$ are compared by calculating precision, recall, and F-measure. For $k > 0.4$, the number of segments is too low to obtain a reasonable favorite region r_{fav}. Basically the result is one very large segment, covering almost the entire image plus a few very small segments.



(a) Gaze path with intersected regions for $k = 0.2$. Fixations are depicted as circles.



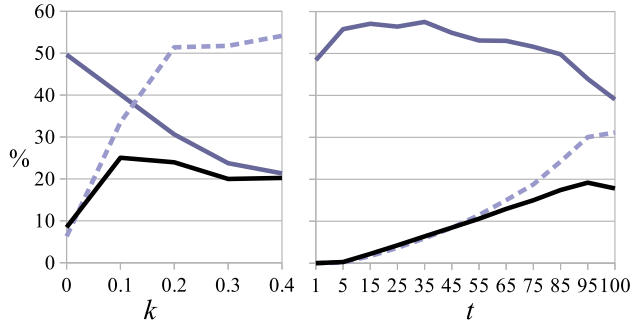
(b) Favorite regions over all users for $k = 0.2$.

Figure 5.4: Identification of r_{fav} for one user (a) and aggregated for 10 users (b) with measure I Segmentation Gaze.

The results for all investigated k values are depicted in Figure 5.5(a). The best precision with 50% was obtained for the smallest sizes of segments for $k = 0$ and the best recall with 54% for $k = 0.4$. The maximum F-measure of 25% was reached with $k = 0.1$. It was calculated from a precision of 4% and a recall of 34%. One can see that the F-measure was relatively stable between $k = 0.1$ and $k = 0.4$ because of the rising recall and the falling precision values.

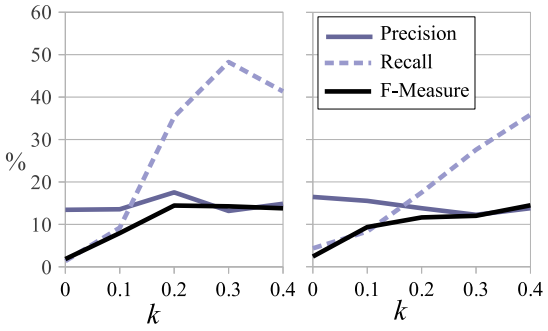
5.2.2 Eye-Tracking-Based Measure II Heat Map Gaze

For measure II Heat Map Gaze, described in Section 5.1, different thresholds $t = 1 \dots 100\%$ were investigated. Some examples are depicted in Figure 5.6. It shows the original image, next to a classical heat map visualization of gaze



(a) Measure I Segmentation Gaze with k.

(b) Measure II Heat Map Gaze with parameter t.



(c) BL Golden Section with k.

(d) BL Center with k.

Figure 5.5: Precision, recall, and F-measure for the two gaze-based and the two baseline measures (BL).

information from all 10 users. The next four images show different potential favorite areas after applying the threshold t to the heat map. If multiple areas appear, the biggest one (i.e., the one with the most pixels) is supposed to be the favorite region r_{fav}.

Precision and F-measure are calculated, comparing the computed favorite region r_{fav} with the ground truth object region r_{gt}. An overview of the results is presented in Figure 5.5(b). The highest precision value is obtained for $t = 35\%$ with 57%. Even with constantly high precision values of more than 44% the F-measure values cannot get very high because of the poor recall results (maximum: 31%). The best F-measure result is 19% with $t = 95\%$.

5.2.3 Baseline Measures

For the baseline measures, the segmentation using the gPb-owt-ucm algorithm is computed. [AMFM11] For both baselines, the best parameters $k = 0 \ldots 0.4$ were investigated by means of the training set. For the golden section baseline, the highest precision value over all images with 18% for $k = 0.2$ and the highest

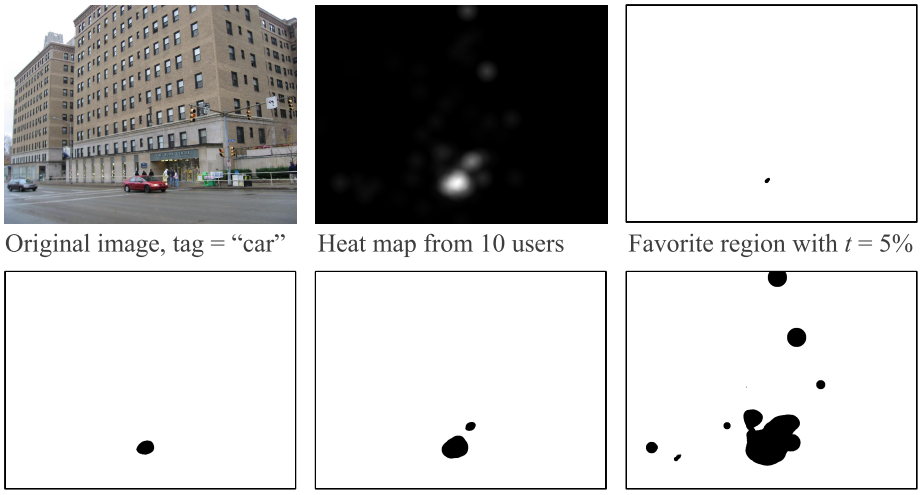


Figure 5.6: Visualization example for measure II Heat Map Gaze.

the F-Measure with 14% for $k = 0.2$ was obtained. The best results for the center baseline are a precision of 16% for $k = 0.1$ and a F-Measure of 13 % for $k = 0.4$.

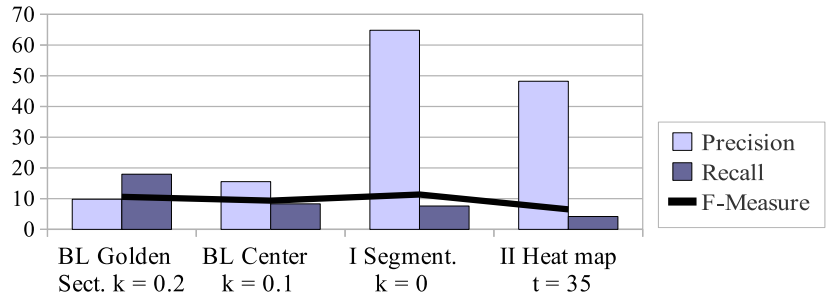


Figure 5.7: Comparison of the two gaze-based measures I Segmentation Gaze (I Segment.) and II Heat Map Gaze and the baseline measures BL Golden and BL Center — best precision results.

5.3 Results

The best performing parameters from the training data set for each of the measures are applied to the test data set. For each measure, values for precision and F-measure for each image were obtained. For comparing the different measures, a Kolmogorov-Smirnov test was conducted to determine if the precision values

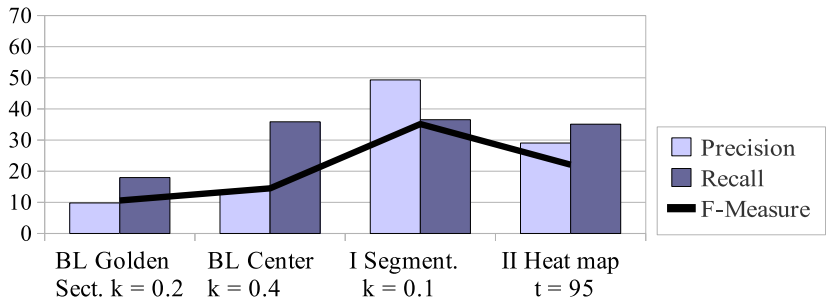


Figure 5.8: Comparison of the two gaze-based measures I Segmentation Gaze (I Segment.) and II Heat Map Gaze and the baseline measures BL Golden and BL Center — best F-measure results.

and F-measure values exhibit a normal distribution. As most of the computed values do not exhibit a normal distribution, a Friedman test was conducted to investigate for a statistical significance in the difference of the obtained precision values and F-measure values. The differences between the four assignment measures (I Segmentation Gaze, II Heat Map Gaze, and two baselines) were significant ($\alpha < .05$) for precision ($\chi^2(3) = 32.668$, $p = .000$) and F-measure ($\chi^2(3) = 15.891$, $p = .001$). Thus, post-hoc analyses with pairwise Wilcoxon Tests were conducted with a Bonferroni correction for the significance level (now: $\alpha < .017$). The values used in the pairwise Wilcoxon Tests are presented in Figure 5.7. The best precision with 65% was obtained for I Segmentation Gaze and the second best with 48% for the measure II Heat Map Gaze. These results significantly outperform the two baselines with $Z = -4,059, p = .000$ for the measure I Segmentation Gaze compared with the golden section baseline, respectively $Z = -4,090, p = .000$ for the center baseline. The results for II Heat Map Gaze are $Z = -3,438, p = .001$ and $Z = -3,286, p = .001$, respectively. There was a weakly significant difference between the two eye-tracking-based measures ($Z = -1.905, p = .057$). For 12 of 29 images, r_{fav} lies completely inside r_{gt}. For 20 images at least 1% of r_{fav} intersects the ground truth object region r_{gt}. The highest F-measure was obtained again by the measure I Segmentation Gaze with 35%. All results are depicted in Figure 5.8. The result for the heat map measure was 22% and for the baselines 11% (golden section) and 14% (center). A significant difference was recognized between the segmentation measure and the baselines with $Z = -2,943, p = .003$ for both baseline. The other results did not show significance (I Segmentation Gaze — II Heat Map Gaze: $Z = -.934, p = .350$, II Heat Map Gaze — golden section: $Z = -2,345, p = .019$, II Heat Map Gaze - center: $Z = -2,186, p = .029$).

5.4 Conclusion

In this section was shown that the labeling of image region is possible, even without the usage of high-quality segmentation. The assignment of tags to regions becomes much harder without the given, manually created regions as they were used before. The reason are additionally inaccuracies caused by the automatic image segmentation. Two measures were presented in this section for performing the labeling. For both measures, best parameters for obtaining a maximum precision and a maximum F-measure were determined on a training data set. The measures and the best performing parameters were applied to a test data set for evaluating the approach. A maximum average precision of 65% at pixel level was obtained by the measure I Segmentation Gaze, which is base on the segmented photo. The other proposed measure, the measure II Heat Map Gaze, can deliver a maximum precision of 48%. The second measure does not use any low-level image information at all but is exclusively based on the gaze data. The best 'coverage' of an given object is obtained by the I Segmentation Gaze measure with a F-measure of 35%. For measure II Heat Map Gaze, this results was 19%. Overall, both newly introduced gaze-based measures deliver better results than baseline measures which select a segment based on the golden ratio of photography or the center position of an object region in the image. The eye-tracking-based segmentation measure I Segmentation Gaze significantly outperforms the baselines for precision and F-measure. By means of the parameters for both measures, it can be controlled if a high precision result is intended (the position of an object can be determined but the selected region is small and does not cover the object) or if a high F-measure should be obtained (good coverage of the primed object). The decision can be made based on the application.

Chapter 6

Image Region Tagging during Search

The aim of Sections 4 and 5 was to create labeled image regions by assigning tags or object names to image region by means of gaze analysis with the aim to describe the content of the photos. In these first steps, the gaze data was collected in a controlled classification experiment. In this experiment, the users had to decide whether they can see a given object on a photo by pressing a button on the keyboard. It was shown that the gaze analysis delivers correct image regions, depicting the given object, with a precision of up to 65% at pixel level.

It is intuitive for humans to automatically identify objects depicted in an image and this identification is very fast. Humans can easily compensate perspective distortions, occlusions, and they can also identify objects with an unusual appearance. The work described in this section goes a step further in benefiting from these skills. Gaze data from users who viewed photos in the results list of an image search application were analyzed with the goal to automatically perform the labeling of images at region level. The gaze paths of users searching for images were recorded by an eye tracking device. Subsequently, the gaze paths were analyzed and regions of the photos in the search results that caught most attention were identified. The search terms entered by the user was assigned to the most viewed image regions for describing the photo content. The gaze paths of several users were aggregated when they viewed the same photos with the same search term. The labeled image regions were evaluated by comparing them to ground truth regions, which were part of the experiment data sets. The work presented in this section addresses the following research question:

RQ 1.5 *Can the region labeling approach be applied to daily routine tasks such as online image search, with users searching for photos in a simulated web search application?*

The context of this work can be seen in the overview, depicted in Figure 6.1. This work was published in [WNS14].

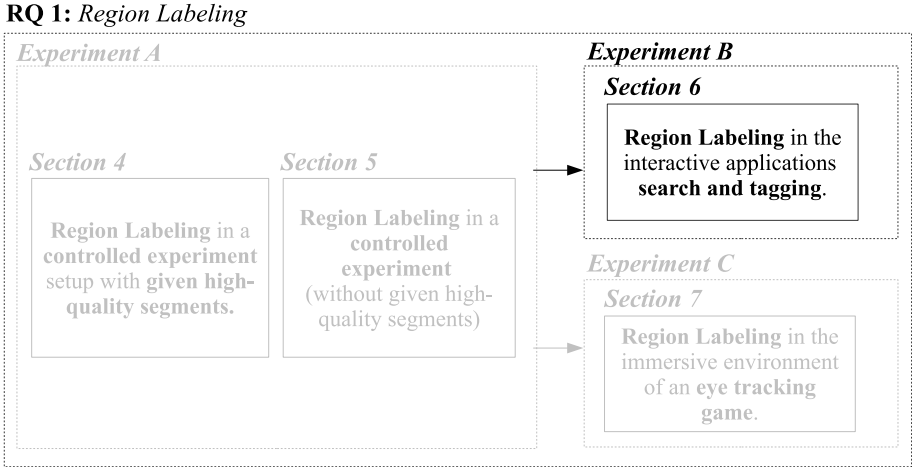


Figure 6.1: Embedding of this experiment in the context of this thesis. After a first proof of the feasibility of gaze-based region labeling in a first, controlled experiment, the application to a search scenario is shown.

First, the performed gaze analysis, the applied baseline measures and the evaluation measures are presented in Section 6.1. Subsequently, in Section 6.2, the experiment design and the experiment data sets are described. The results of the analysis are presented and discussed in Section 6.3, before this part of the thesis in concluded in 6.4.

6.1 Analysis

The work in this section is part of the overall goal to use human gaze information in the annotation of image region. The gaze analysis performed in this work was first presented in Section 5. An experiment was conducted for collection data in a less controlled application, which is more alike to a real-world scenario.

6.1.1 Gaze Analysis

Two gaze-based predictors are applied for labeling image regions. The two gaze-based predictors are the I Segmentation Gaze and the II Heat Map Gaze approach, both presented in the previous Section 5. By means of these approaches, a given search term is assigned to an image region for labeling it. The measures were modified for allowing the selection of several object regions. An overview of the calculation of both measures with one sample image from the experiment data set is depicted in Figure 6.3. For all photos belonging to a search set, the input for the gaze analysis was (i) the given search term and (ii) the gaze paths of all users who fixated the photo (Figure 6.2). The I Segmentation Gaze measure additionally took (iii) (hierarchical) photo segments as input data. The photo segments for measure I Segmentation Gaze were obtained from applying the gPb-owt-ucm algorithm [AMFM11]. The different hierarchy levels describe

different levels of detail and are controlled by the parameter $k = 0, 0.1 \ldots 0.7$, with $k = 0$ as highest level of detail. Please refer to the original publication by Arbeláez et al. [AMFM11] for details of the gPb-owt-ucm algorithm.

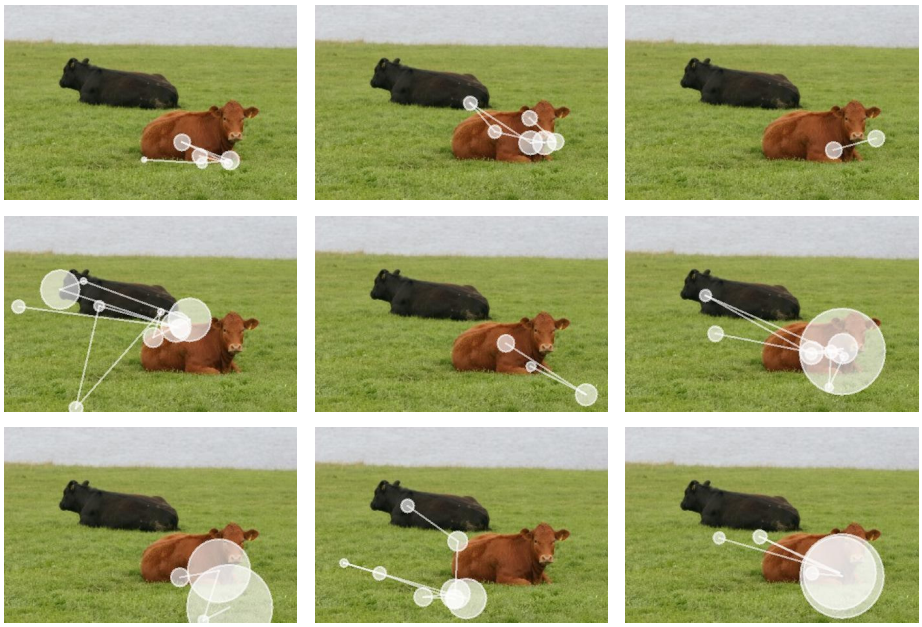


Figure 6.2: Example gaze paths (ii) of nine different users searching for a "brown cow" viewing one photo of the search results list.

The I Segmentation Gaze measure can be performed based on several eye tracking measures (cf. Section 5.1.1). A subset of the measures presented in Table 4.1 was selected for the analysis based on their capability in preliminary tests. The measure (1) fixationCount counted the number of fixations on a segment. (2) fixationDuration calculated the sum of the duration of all fixations on a segment. The measure (3) firstFixationDuration also considered the duration of a fixation but it only took the very first fixation on a segment into account. Accordingly, (4) lastFixationDuration measured the fixation duration of the very last fixation on a segment. A visit describes the time between the first fixation on a region and the next fixation outside. (5) visitCount counted the number of visits on a segment and (6) meanVisitDuration calculated the average duration of these visits. In the previous section, only the segment with the highest eye tracking measure results was selected. Here, the segments with the highest 10 % of the measure values were selected. They were assumed to show an object or several objects described by the search query. The search term was assigned to these regions. The measure results for all participants which viewed the same photos are summed up. In order to take the inaccuracies in the eye tracking data into account, the region extension introduced in Section 4 was also applied. The region extension considers fixations in the surrounding of up to 13 pixels of a segment as belonging to the segment.

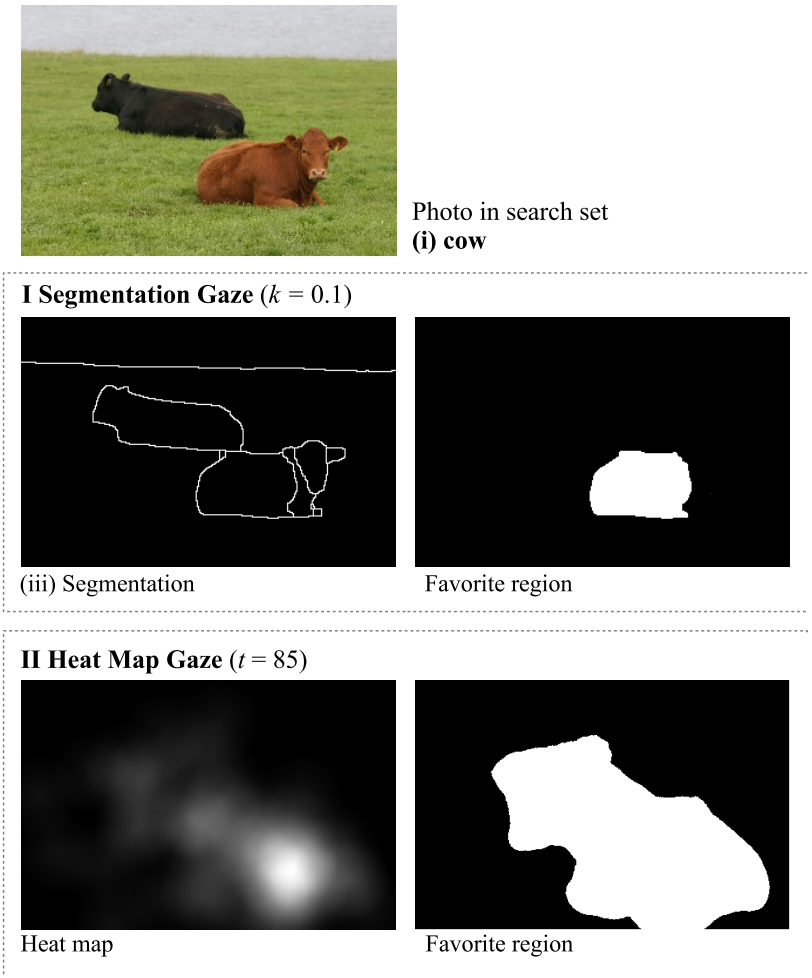


Figure 6.3: Gaze-based region labeling with predictors I Segmentation Gaze and II Heat Map Gaze. Input data is (i) the given search category, and (iii) the segmented image (only for I).

The II Heat Map Gaze approach identified intensively viewed photo regions by summing up the fixations of all gaze paths at pixel level (cf. Section 5.1.2). A value of 100 was applied to the center of each fixation. In a radius of 50 pixels, linear decreasing values were applied to the surrounding pixels. The value of all fixations were summed up for all pixels of the image for building the so-called heat map. From the created heat map, the assumed object region was calculated by applying a threshold to the data, identifying the mostly viewed pixels. The parameter t indicates the percentage of viewing intensity (e.g., $t = 10$ indicates the 10 % of all pixels with the highest values). The investigated parameters in this work were $t = 1$ and $t = 10 \ldots 100$ in steps of 10. In the previous section, an additional step was performed for selecting only one favorite region from all remaining regions after the application of the threshold. This step is skipped in the analysis here, thus more than one region can be the result.

6.1.2 Baselines

Two baseline approaches were compared with the gaze-based ones. The baseline approaches did not make use eye tracking data. Furthermore, the baselines did not need training data nor a training period, exactly like the gaze-based approaches. Both baseline were introduced in the previous sections.

The saliency baseline is based on the assumption that the important objects of a photo are the most salient points on an image. The saliency baseline was presented in the previous section and is described in Section 5.1.3. The saline points were calculated by the toolbox offered by Itti et al. [IKN98]. The favorite region was selected by using the salient points and their ordering as input data. This saliency paths were interpreted as simulated gaze paths. Subsequently, the same methods as for the gaze analysis approach, described in the previous section, were used to analyze them. Thus, the investigated baseline approaches are called the III Segmentation Saliency approach and the IV Heat Map Saliency approach

Finally, for the baseline V Random, the photo was first segmented by the algorithm published by Arbeláez et al. [AMFM11]. Subsequently, one of the segments was selected randomly and the search term was assigned to this segment. The random baseline was also used in Section 4.2.2. This very naive baseline serves as measure for how difficult the task of selecting one favorite region was.

6.1.3 Calculating Precision, Recall, and F-measure

By means of ground truth data for all images and assigned labels (cf. Section Photo Sets described above), the computed object regions can be evaluated. For each pixel, the ground truth was compared with the labels obtained from the gaze analysis by calculating precision, recall, and F-measure, with F-measure $= 2 \cdot \frac{precision \cdot recall}{precision + recall}$, as presented in Section 5.1.4. An example photo with two object regions and their evaluation can be found in Figure 6.4.

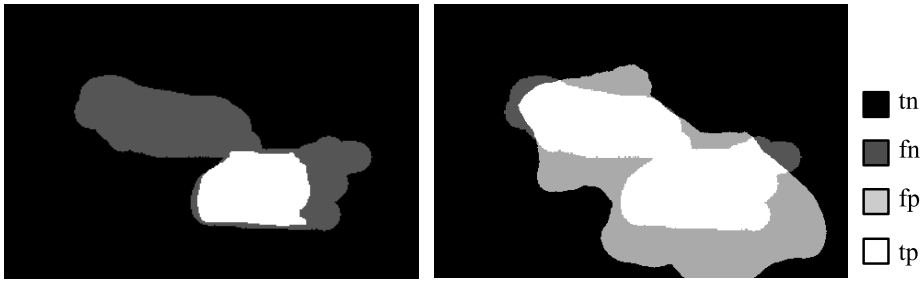


Figure 6.4: Comparing labeled image regions and ground truth regions at pixel level.

6.2 Experiment Setup

An experiment for investigating the potential of photo region labeling during image search was conducted. Therefore, participants used a simulated search page for performing different search tasks.

Figure 6.5: Sample search tasks and images not fulfilling and fulfilling the exact search task.

6.2.1 Participants

23 volunteers participated in the experiment, 11 of them were female. Their average age was 23.3 (SD: 2.09) with the youngest person being 20 and the oldest 29. Most of the participants were computer science students but there were also students of other subjects, such as mechanical engineering, biology, geology, and educational science.

Figure 6.6: Cropped and scaled screen shots of the three experiment steps: A Search task and start search, B Search results, C Photo selection. The arrows show interaction options.

6.2.2 Photo Sets

Photographs of natural scenes were presented to the users. These photos were taken from three data sets. All sets provided ground truth region labeling data. The VOC2012 data set [EVGW+] was made available for the Visual Object Classes Challenge. The segmentation set, which contains ground truth region labels at pixel level, contains 2913 photos and 20 classes of objects such as "airplane," "sofa," and "dog." The MSRC [WCM05] data set, published by Microsoft Research, consists of 592 photos and 23 labeled object classes. The objects belong to simple concepts like in the VOC2012 set, e. g., "bird," "sky," and "sheep." The LabelMe [RTMF08] data set with 182,657 user contributed images and 291,841 labels (download August 2010) provides images of complex indoor and outdoor scenes. The LabelMe community has manually created region labels by drawing polygons into the images and by tagging them.

The photos for the experiment data set were selected based on their labels. The labels were taken from the "All time most popular tags" of the online photo sharing page Flickr[1]. Among the most frequently used tags, 23 occur in at least two of the three data sets. These labels were selected for the use in the experiment application. For each label, a random number of photos between 9 and 24 was chosen from the two resp. three data sets. 10 labels occur in all three data sets, whereas 13 labels are present in only two sets. The label-sets were composed in equal parts of the data sets. In total, the experiment data set consists of 361 photos, with 103 photos taken from MSRC, 112 from VOC2012, and 146 from LabelMe.

6.2.3 Tasks

For each search set, consisting of a label and a set of photos, a search task was defined with the goal to simulate an online image search and to motivate the users to scan the image search results lists. The tasks request the participants to find an object with specific characteristics. For example, for the label "bus," the search task was *"Search for a green bus"*. The tasks were created in a way that at least one photo fulfills the task. Often, even more than one photo could

[1] http://www.flickr.com/photos/tags/ (last visited Sept. 29, 2013)

be selected. Also, fore some tasks the answer could depend on the subjective impression of the user. For example, a participant might chose an image showing a bird with an orange bill for the task *"Search for a bird with a red bill"*. Some more examples of search tasks can be found in Figure 6.5. This figure also shows examples of photos fulfilling and not fulfilling the given search task. 10 of the search tasks ask for a specific color as characteristic (e.g., *Search for a green bus*), 4 for animals with a specific coat color or pattern (e.g., *Search for a dog with black spots*), 5 tasks concentrate on other characteristics (e.g., *Search for a building with balcony*), and 4 ask for objects in specific situations (e.g., *Search for a horse with bridle*). In the analysis, the named object was assigned to an image region in all photos of the search results list that were fixated, ignoring the specific characteristics. Possible differences in region labeling results for photos fulfilling the search task (the photos with the *green* bus) and photos not fulfilling the task (photos depicting a bus but not a green one) were investigated.

6.2.4 Procedure and Experiment Application

Before starting the experiment application, the participants were introduced to the experiment tasks and the eye tracking device. A calibration of the eye tracker was performed by fixating five dots on the computer screen.

The experiment application was designed to resemble online image search pages. It consists of three pages. Screen shots of the application can be found in Figure 6.6. On the first page of the experiment application, page A in Figure 6.6, the search task was presented to the user. The user had to enter a search term as free text into the search input field. By pressing the OK button the simulated search was started. It was not allowed to start the search with an empty text field but no further checks with regard to its meaning were performed on the given search query. On the second page B, the photos of the experiment data set were displayed in rows of three photos each. The photos were scaled to a maximum width and height of 450 pixels. The page was scrollable as not all photos could be shown on a static page. The user could go back to the search page by pressing the "Back" button. By clicking on the photos, page C opened. On this page, the user could select a photo by pressing the "Select" button for completing the search task. It was possible to go back to the search result page by clicking on the "Back" button.

Eye tracking data was recorded while the user performed the tasks. No time limitations were given for the 23 search tasks. The order of the tasks was randomly alternated for each participant. Also the order of the photos on the search result pages was randomized. At the end of the experiment, each user filled out a questionnaire. It comprised questions about demographic information (age, profession) and some ratings about the experiment application and tasks.

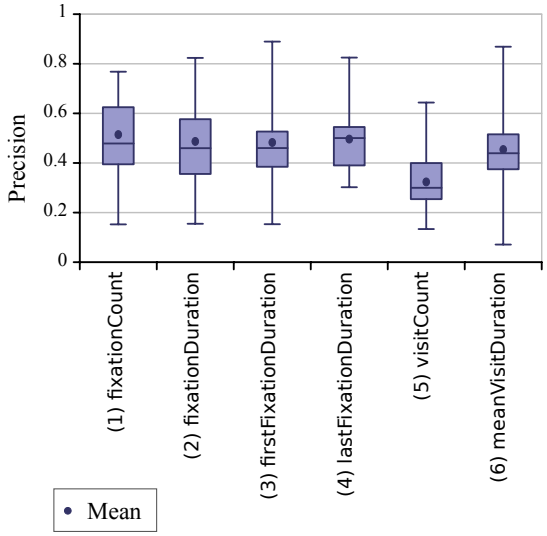


Figure 6.7: Precision for I Segmentation Gaze with $k = 0$ for six different eye tracking measures.

6.3 Results

In this section, the labeling results are presented and the gaze-based methods are compare to the baseline methods. Also the results for the three different data sets are compared. In addition, the differences for photos fulfilling or not fulfilling the search task were investigated.

6.3.1 User Feedback and Behavior

The participants did not feel uncomfortable while their eye movements were recorded by the eye tracking device. Most participants gave an answer of 5 (M: 4.92, SD: 0.28) on a Likert scale from 1 ("I felt uncomfortable while my eye movements were recorded") to 5 ("I did not feel uncomfortable while my eye movements were recorded"). The users' comfort was asked in the questionnaire to check if there was a strong influence of the eye tracker recording on the participants' well-being and thus their gaze. As the users did not feel uncomfortable such an influence is not very likely.

The users did not have problems controlling the application as shown by an average answer of 1.04 (SD: 0.2) on a scale from 1 ("The application was easy to control") to 5 ("It was hard to control the application"). Also the tasks were not to difficult to perform, as the level of difficulty was in average rated with 1.33 (SD: 0.62) on a scale between 1 ("The search for images was easy") to 5 ("The search for images was difficult").

The average time the users spent on a search task was 14.6 s. The longest average search time was obtained for the search task *"Search for a road with median strip."* with 23.3 s. The shortest average time was 8.8 s for task *"Search for*

a chair with a red seating surface." The searching behavior of the participants showed that in 99.98 % of all cases the photo selection page was opened only once, namely for the final selection. Nine times participants went from photo selection page C back to search page B before they chose an image according to the search query. With regard to the final selections, a percentage of 98.03 % correctly selected images reveals the high quality of the results.

On average, each user fixated 11.63 photos per search query. The average number of fixations over all users per photo is 2.88 (SD: 1.63). The average number of fixations on an image was highest for the search set "bottle" with 6.42 (SD: 1.91). In contrast, for the search set "car," the number of fixations on an image on average was the lowest with only 1.94 fixations (SD: 0.91).

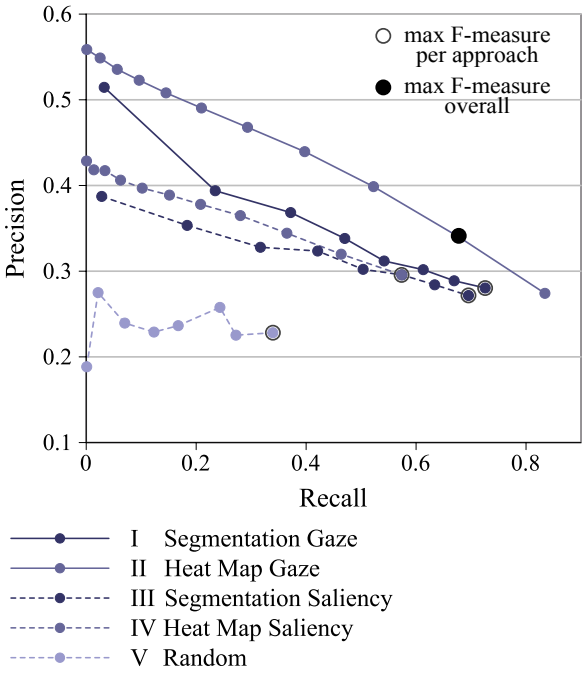


Figure 6.8: Precision and recall for the two gaze-based measures I and II, the two saliency-based measures III and IV, and the V Baseline measure.

6.3.2 Comparison of Eye Tracking Measures

First, the six eye tracking measures are compared for the I Segmentation Gaze predictor. As parameter for this approach, the smallest segmentation size $k = 0$ was chosen. Figure 6.7 depicts the detailed results. For each eye tracking measure the average precision results for each search term are depicted. The box plot diagram shows the first and third quartiles as boxes, the median is displayed inside the boxes as horizontal line, the mean as small circle, and the vertical lines

show the range of all values. The measure (5) visitCount clearly performed worse than the other measures. (6) meanVisitDuration and (3) firstFixationDuration had good mean results but a big spread in the results over the different search terms. The measures (1) fixationCount and (2) fixationDuration performed best. As the measure (1) fixationCount provided the best average result ($M = 0.48$, $SD = 0.13$) over all search terms compared with (2) fixationDuration ($M = 0.47$, $SD = 0.13$), (1) fixationCount is used in the following analysis.

6.3.3 Region Labeling Results

The results for the five region labeling approaches are compared in Figure 6.8. The precision and F-measure results are depicted for different parameters $k = 0 \ldots 0.7$ and $t = 1 \ldots 100$ (see Section Analysis above). Both gaze-based approaches I Segmentation Gaze and II Heat Map Gaze performed better than the baseline approaches. The saliency approach already showed better results than the random baseline. The II Heat Map Gaze approach clearly delivered the best precision and recall results over all parameters. The best F-measure was obtained for II Heat Map Gaze with 0.38 (marked as black circle in Figure 6.8) with $t = 90$. The overall best precision was obtained for the same measure and parameter with 0.56. The best performing baseline approach with a F-measure result of 0.33 is IV Heat Map Saliency with $t = 100$.

A Wilcoxon signed-rank test showed a statistically significant difference with $\alpha < 0.05$ when comparing the average precision and F-measure results per search category for the best performing predictor II Heat Map Gaze with $t = 90$ and the best performing baseline predictor IV Heat Map Saliency with $t = 100$ (precision: $N = 23, Z = -3.194, p = .001$, F-measure: $N = 23, Z = -3.346, p = .001$).

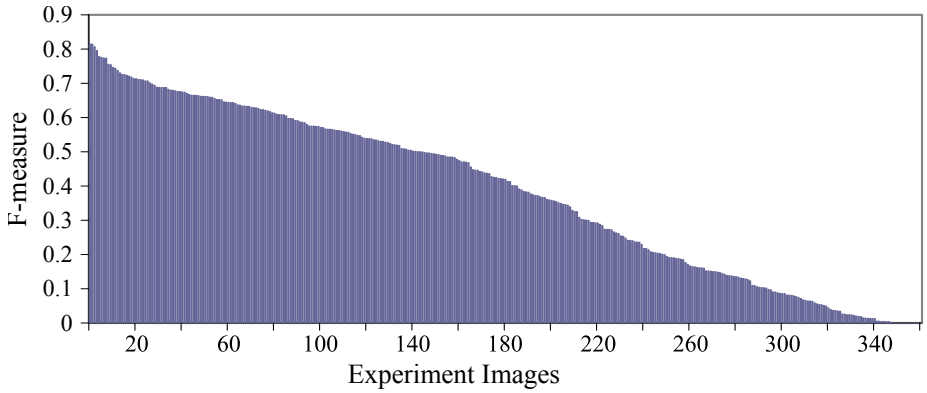


Figure 6.9: F-measure results for all images of the experiment data set calculated with II Heat Map Gaze with $t = 90$. The images were sorted according to their F-measure value in descending order.

Figure 6.10: Example image with results for II Heat Map Gaze with $t = 90$ with evaluation of the labeled image regions.

6.3.4 Example Photos

The F-measure results for all photos are depicted in Figure 6.9, sorted by the F-measure values. No correlations between the number of fixations on a photo and the precision nor F-measure results were found. Only 9 of the 361 photos had a precision result of 0, that is, not a single pixel of the labeled area covered a correct object.

The three photos with the best F-measure results are depicted in Figure 6.10. Some negative examples with low F-measure results are shown in Figure 6.11. Besides the original photo, also the region the search tag was assigned to, as well as the ground truth regions for the given object, are depicted. Regarding the average number of fixations for the best labeling predictions one can observe that 1.47 fixations on that image were obtained by 15 participants (the other ones did not fixate the image). In contrast, the second ranked image was fixated on average 9.90 times by 20 participants. The image placed on rank three was fixated 2.15 times by 13 participants.

6.3.5 Example Sets

In Figure 6.12, the precision and F-measure results for approach II Heat Map Gaze with $t = 90$ were split up for the different search tasks. In the diagrams, the results for all photos in each task are displayed (boxes show the area between the first and the third quartile, median as horizontal line, and the range of all photo results as vertical line). One can see that the range in the results is high. This means that the labeling results strongly depend on the given photos. The highest average precision value over all photos of one search task was obtained

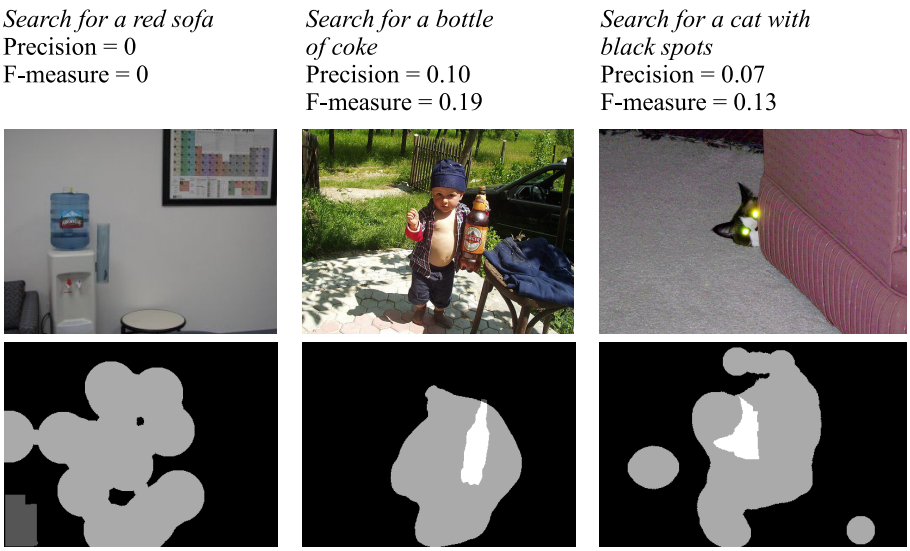


Figure 6.11: Negative example image with results for II Heat Map Gaze with $t = 90$ with evaluation of the labeled image regions.

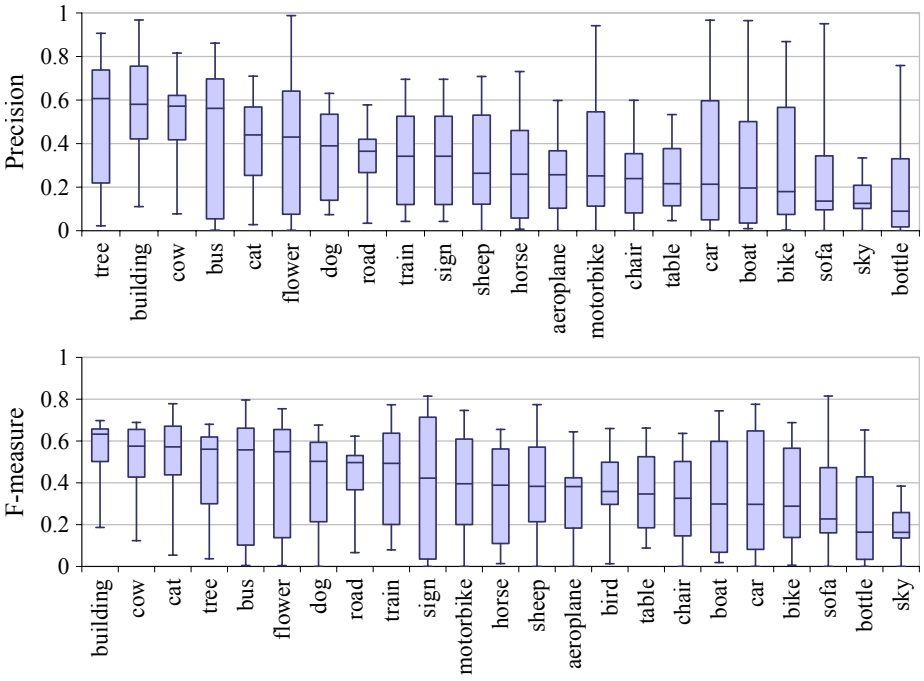


Figure 6.12: Detailed precision and F-measure region labeling results for each search task for approach II Heat Map Gaze with $t = 90$. The terms are sorted in descending order by their median precision value (above) and F-measure value (below), respectively.

for "tree" with $P = 0.61$, the worst for "bottle" with $P = 0.09$. The best average F-measure value was obtained for "building" with $P = 0.63$, the worst for "sky" with $P = 0.16$.

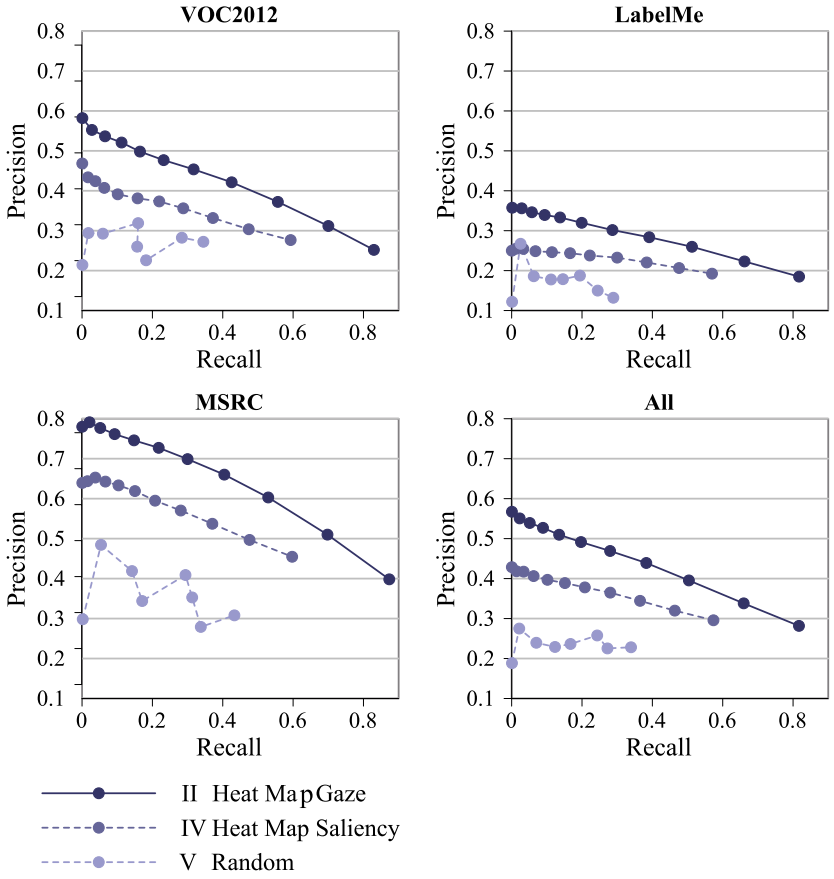


Figure 6.13: Compare results for the different data sets.

6.3.6 Comparison of the Data Sets

The experiment data was composed of photos from three different data sets, as described in Section Photo Sets (6.2.2). For the best performing approach II Heat Map Gaze, the best performing baseline approach IV Heat Map Saliency, and the V Random Baseline, the precision and recall results were split up for the three data sets VOC2012, MSRC, and LabelMe in Figure 6.13. Already the random baseline shows differences in the level of difficulty for the segmentation approach. In total, the results are much better for the MSRC data set containing scenes of low complexity, compared with the most challenging data set LabelMe which includes images showing scenes of high complexity (i. e., many different objects). However, it can be observed that the gaze-based approach improves

the results for all data sets over the saliency baseline. The results of II Heat Map Gaze always lie above IV Heat Map Saliency.

6.3.7 Comparison of True and False Images

In the experiment application, a search task was given to the participants asking for an object with specific characteristics, e. g., *"Search for a green bus"*. In the search results list, photos showing an object which was asked for (e. g., "bus") were displayed. But only a few photos showed the object with the specific characteristics (e. g., "green bus"). In total, 97 of the 361 photos fulfilled the search task, 264 did not. For the approaches II Heat Map Gaze and IV Heat Map Saliency, labeling results for photos fulfilling the search task and not fulfilling the task are compared. Precision and recall results are depicted in Figure 6.14. As can be seen in the figure, the curves lie close to each other. The results for the photos fulfilling the tasks were slightly better. A Wilcoxon signed-rank test was applied to the data. The results were computed using the values obtained from the approach II Heat Map Gaze with $t = 90$. The differences in the results are not significant with $\alpha < 0.05$ for precision ($N = 23, Z = -.487, p = .626$) and F-measure ($N = 23, Z = -3.346, p = .001$). This suggests that the approach also works for objects that do not exactly fulfill the task, that is, where the photos show the object asked for but the object does not match the additional characteristics such as the color. With other words, the results imply that the labeling of objects is agnostic to characteristics of the objects the user is looking for.

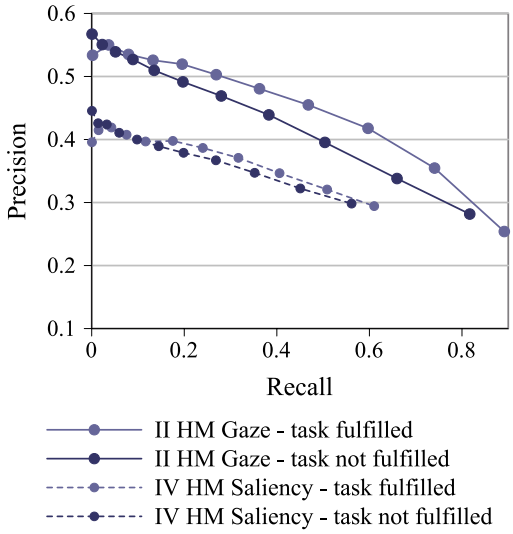


Figure 6.14: Precision and F-measure results for II Heat Map Gaze with $t = 90$ and IV Heat Map Saliency with $t = 100$ for photos fulfilling the search task versus not fulfilling the search task.

6.4 Conclusion

The experiment results presented in this section suggest that the labeling of image regions by means of gaze data is possible in a scenario such as the search for images. Comparing the best precision and F-measure results ($P = 0.56$, F-measure = 0.33) of this work shows slightly lower results compared with the ones obtained in previous work, presented in Section 5 ($P = 0.65$, F-measure = 0.35). The results strongly vary for different search terms and photos. There are usually two reasons for difficulties in identifying objects in photos: One reason is caused by the characteristics of human visual perception. Big objects and objects that can easily be identified in the corner of ones eyes. Here, the user does not have to fixate it directly. One of the weak categories, "sky," is very likely to belong to this group of things. Another challenge are very small objects, due to inaccuracy of the eye tracking data and the segmentations of the photos. This problem also occurred in the previous experiment (cf. Section 4.5.1).

A detailed analysis of the factors influencing the results (such as how many details are depicted on a photo) can be subject of a future study. More data and different photos might be needed for such a study. Only "correct" photos were selected for the search sets. Correct means that on each photo at least one correct object is depicted, even though the object did not have the specific characteristics. In a real-world application, search engines reach a very high quality for simple search queries. Thus, it can be assumed that the results may be transferred to a real search engine. However, when applying the gaze-based method to real image search, this question has to be handled and wrong photos in the result set have to be considered. From the two approaches for the gaze-based (I, II) and the saliency-based (III, IV) methods, the heat map approach performs better. An additional advantage of this approach is – compared with the segmentation-based approach – that no segmentations have to be calculated. The computation of high-quality segmentations can be time-consuming. By varying the parameters of the II Heat Map Measure Gaze approach, the focus can be moved from good F-measures results (a higher parameter t which leads to bigger selected areas) to good precision values (small t values).

The results presented in this section shows that it is possible to assign search terms to image regions by means of gaze paths recorded while users are searching for images. The usage of gaze data significantly improves the labeling results over a baseline approach using only saliency information. The method works even for photos depicting an object that was asked for but did not fulfill the specific characteristic mentioned in the search task. With a performance time of 14.6 s per search query, including the scanning of numerous photos, the labeling of image regions is very fast compared with the manual drawing of polygons. Also, no more effort is needed by the users than viewing search engine results.

Chapter 7

EyeGrab — A game with a purpose

Metadata, describing the content of photos at pixel level are of high importance for applications such as image search or as part of training sets for object detection algorithm. Up to this point, the previous sections showed the potential of labeling image region by means of gaze analysis in two experiments. First, the gaze data was collected in a strongly controlled classification experiment (Sections 4 and 5). A second experiment, presented in Section 6, showed that even in an experiment setup which resembles a real-world application, the search for images, the labeling of image region is feasible and outperforms baseline approaches.

In this section, tags are again applied to image regions for a more detailed description of the photo semantics. The region labeling is again performed without any additional effort for the user, just from analyzing eye tracking data. Here, the data is recorded while users are playing a gaze controlled game. In the game *EyeGrab*, users classify and rate photos falling down the screen. The photos are classified according to a given category under time pressure. The game has been evaluated in a study with 54 subjects. 91% of the users enjoyed playing the game, thus the requirement to attract the users attention in the game is fulfilled. Only 7% of the shown images passed without classification and 90% of the classifications were correct with respect to the given category. The region labeling results, based on the analysis of the fixations on the images, show that it is possible to assign the given categories to image regions with a precision of up to 61% at pixel level. Thus, an almost equally good region labeling using gaze information can be performed even in an immersive environment like in *EyeGrab* compared with a previous classification experiment that was much more controlled. The contribution of this section to the overall goal of this thesis is shown in Figure 7.1. The specific research question, tackled in this section is:

RQ 1.6 *Does the approach perform well in a distracting situation?*

For answering this question, gaze-based region labeling is compared with baseline approaches and to previous labeling results from other experiments. The work presented in this section was published in [WNS12] and [WSS14a]. The recorded data, leading to the results of this section, can be downloaded under `http://west.uni-koblenz.de/Research/DataSets/gaze`.

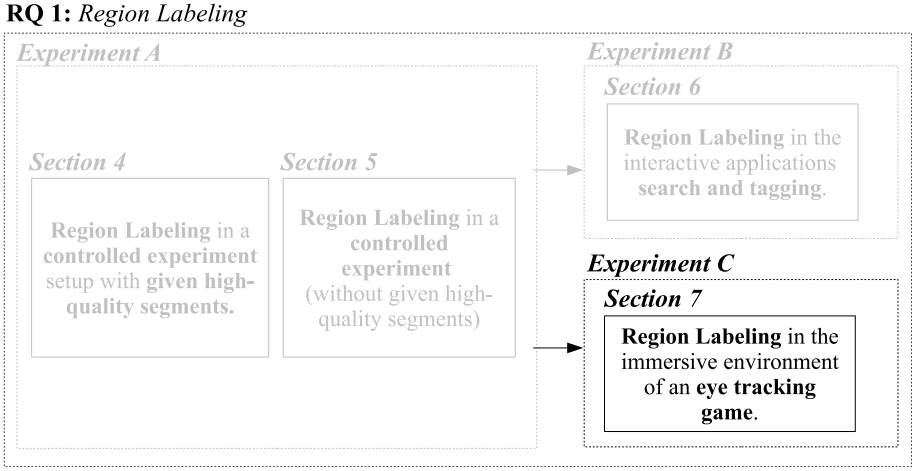


Figure 7.1: Embedding of this experiment in the context of this thesis. The labeling of image region by means of gaze data was shown before in a controlled experiment and in an image search scenario. In this Section 7, the labeling in the strongly distracting scenario of a gaze controlled computer game is investigated.

Related work is shortly discussed in the subsequent section before the approach for region labeling based on gaze data in is presented in Section 7.1. The game *EyeGrab* is introduced in Section 7.2, and the experiment setup is described in Section 7.3. The results concerning the image classification are presented and discussed in Section 7.4. In Section 7.5, the photo region labeling results are presented before the section is concluded.

7.1 Approach

Games with a purpose (GWAPs) are computer games that have the goal to obtain information from humans in an entertaining way. The information is usually easy to create for humans but challenging or impossible to be created by fully automatic approaches. In Section 3.1.1, some games are presented which have the aim to deliver image labels.

Smith and Graham [SG06] described the advantages of gaze control in video games. They stated that the use of gaze control can improve the game play experience. An example is *EyeAsteroids*[1], an eye-controlled arcade game pre-

[1] http://www.tobii.com/en/gaze-interaction/global/demo-room/tobii-eyeasteroids/, last visited May 15, 2012

sented by Tobii. The game is entertaining but does not have the goal to exploit the users' activities while playing.

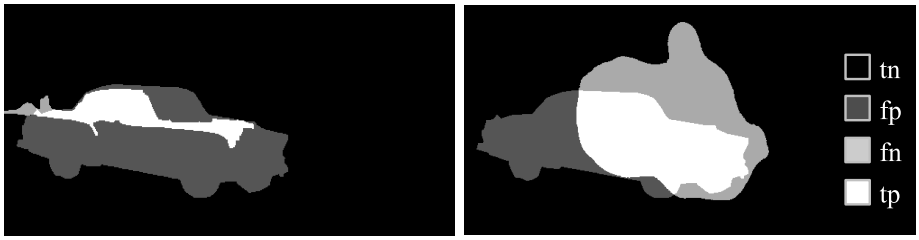
In *EyeGrab*, users classify and rate photos falling down the screen. Photos are selected by fixating them. Subsequently, the classification is performed by fixating specific objects on the screen, which represents different classes. In the classification is considered, if a photo belongs to the given category and if a player likes a photo or not. By analyzing the recorded gaze paths, given categories, which describes a specific object such as "car" or "tree," are automatically assign to image regions. All photos used in the evaluation had ground truth information concerning the depicted objects.

In order to assign a given category to an image region, the two gaze measures (presented in Section 5.1) and a baseline are applied to the data. The two gaze-based measures are the segmentation measure (I) and the heat map measure (II). All fixations on the classified photos are analyzed for performing the region labeling. The measures predict which region of the photo is assumed to show an object, belonging to the given category.

In the segmentation approach, the fixations on every region of the segmented photo are counted, which corresponds to the fixation measure *fixationCount*. The segment with the highest outcome is assumed to show the object for the given category. For the I Segmentation Gaze measure the results for different parameter $k = 0 \ldots 0.5$ are investigated. The heat map approach (II Heat Map Gaze) identifies intensively viewed photo regions by summing up the fixations of all gaze paths at pixel level. A value of 100 is applied to the center of each fixation. In a radius of 50 pixels, linear decreasing values are applied to the surrounding pixels. From the created heat map, the object region is calculated by applying a threshold to the data, identifying the mostly viewed pixels. The parameter t indicates the percentage of viewing intensity (e.g., $t = 5$ indicates the 5% of all pixels with the highest values). After the thresholding, the biggest area of connected pixels is assumed to depict the object. The concrete parameter values for both approaches are determined based on the findings in previous work presented in Section 5. Also the center baseline approach from this work is also applied to the data.

Only fixations on correctly classified images are part of the analysis. The gaze data of all users on the same image with the same category and a correct classification are aggregated, as the results in Section 4 showed the potential of gaze aggregation. In order to take potential inaccuracies in the eye tracking data into account, region extension and weighting, also introduced in Section 4, were applied. The region extension considers fixations in the surrounding of up to 13 pixels of an segment as being on the segment. Due to the weighting results for segments that are smaller than 5% of the photo are multiplied by a factor up to 4. Different segmentation levels $k = 0 \ldots 0.5$ are considered in the analysis.

By means of ground truth data for the image regions and labels (cf. Section 7.3), the computed object regions can be evaluated. For each pixel, the ground truth was compared with the label obtained from the measures by calculating precision, recall, and F-measure. An example photo with two object regions and their evaluation can be found in Figure 7.2. Details about the calculation can be found in Section 5.1.4.

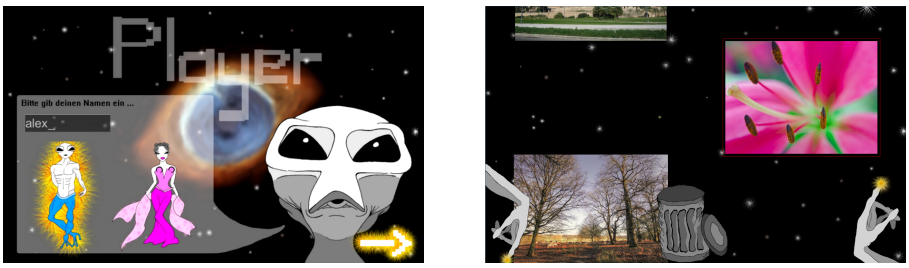


Segmentation Approach ($k = 0.2$) Heat Map Approach ($t = 85$)

Figure 7.2: Comparing labeled image regions and ground truth regions at pixel level.

7.2 The *EyeGrab* Game

The task in *EyeGrab* to "clean up an aliens' universe" by categorizing and rating photos. Before starting the game, the user has to calibrate the eye tracking device by fixating several points on the screen. Subsequently, a small introduction to the game's rules is given to the gamer. In addition, he/she has to choose a user name and to indicate his/her gender. Besides entering the gamer's nickname, the game is solely controlled by eye movements. Gaze-based interactions are triggered after a dwell time of 450 m. With a normal dwell time for fixations of between 200 and 400 ms (see Section 2.1), the selection dwell time lies above this value to avoid random selections. For example, the selection of the gender is done by focusing on a male or female character as shown in Figure 7.3(a).



(a) Entering of personal information (b) Playing screen

Figure 7.3: Screen shots of *EyeGrab*. (a) Starting page with player's name and gender input, (b) Playing screen with three photos.

A game consists of several rounds. In each round, a set of photos has to be classified concerning a given category such as "car," "person," and "sky." First, the category is presented to the user for 6 s. Subsequently, the photos fall down the screen as depicted in Figure 7.3(b) and are classified by the gamers. Each round has a different speed level at which the photos move. Several photos can appear on the screen at the same time. The player selects an image by fixating

it for longer than the dwell time of 450 ms. As soon as a photo is selected, it is highlighted by a thin frame, and the user can classify it into one of three categories. The classification takes place by fixating symbols on the screen as shown in Figure 7.4. The categories are "not relevant" (symbolized by a trash can), "relevant & like" (symbolized by a hand pointing upward), and "relevant & dislike" (symbolized by a hand pointing downward). Playing *EyeGrab*, the gamer scores for each correctly categorized image, receives negative points for each wrong one, and no points for images that fell off the screen without classification. No scores are obtained for the ratings of "like" and "dislike." An acoustic feedback is given for each classification. An applause is played for correct classifications, while a booing sound signals incorrect classifications and missed photos. A high score list is presented to the user at the end of the game.

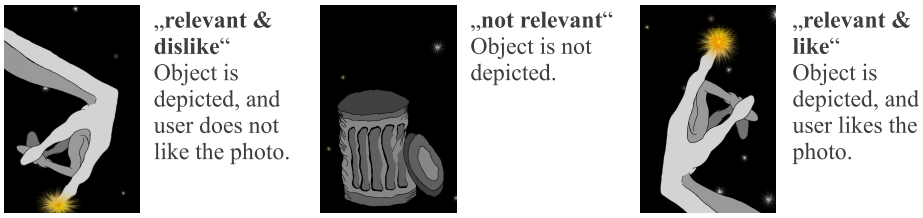


Figure 7.4: Symbols representing the classification options.

7.3 Experiment Description

EyeGrab has been evaluated with 54 subjects (with 19 female). The subjects' ages were between 17 and 56 years (avg = 30 years, SD = 7.7). The majority of the participants were students or research fellows in computer science (70%) but students from other fields of study or members of other professional groups such as restorers or psychotherapists participated in the experiment as well.

As pointed out by von Ahn [VAD04], games with a purpose have to fulfill two requirements. On the one hand, they have to offer solutions for the addressed problem and on the other hand the playing has to be fun. Most subjects enjoyed playing the game *EyeGrab*. In a questionnaire, 49 of the 54 subjects rated the statement "The game is fun." with a 4 or a 5 on a standard 5-point Likert scale (avg = 4.22, SD = 0.72). The level of difficulty playing *EyeGrab* seems to have been adequate, as most of the participants did not agree with the statement "The game overexerts me." (M = 2.54, SD = 1). Most of the participants did not feel uncomfortable using the eye tracking device as shown by the low average agreement of 2.24 (SD = 1.15) to the statement "The eye tracker has a negative impact on my well-being."

7.3.1 Procedure

Every participant played four rounds of *EyeGrab*. The first round was a short test round consisting of only 12 photos. This test round with the category

"tree" served as an introduction to the game. The data collected during this round was not used in the later analysis. The other three rounds with the categories "car," "person," and "sky" consisted of 24 photos each. The photos of each round were displayed in a randomized order. Different falling speeds were applied to each round. In the slowest pace (speed 1) the photos were falling with 3.6 pixels/ms, and they were visible on the screen for 5,200 ms. In the medium pace (speed 2), the photos were visible for 4,500 ms (pace = 4.3 pixels/ms). In the most challenging speed (speed 3) the photos were falling down within only 3,800 ms (5 pixels/ms). A complete round took between 64.4 s (speed 1) and 50 s (speed 3). A Latin Square design was applied in order to randomize the order of the three categories with the three speed levels. The participants were asked to express their agreement to several statements on a 5-point Likert scale between 1 (strongly disagree) and 5 (strongly agree) in a questionnaire at the end of the experiment.

7.3.2 Data Set: Categories and Photos

The categories used in *EyeGrab* were taken from the top six of the list with the mostly used tags in LabelMe [RTMF08]. The LabelMe data set consists of photos, uploaded by the community, and has manually drawn region labels. The first two categories of this list ("window" and "building") are not taken into account because often not all instances of these objects are labeled on the photos. This can cause problems during the evaluation of the approach, as Ground truth data with a complete labeling of all occurring objects belonging to the given category is needed. Thus, the next top categories were taken, which are the above-mentioned categories of "car," "person," and "sky."

In total, 84 photos (24 for each round and 12 for the test round) were selected from the image hosting page Flickr[2] and from LabelMe [RTMF08]. To create a challenge for the gamers, only 50% of the selected photos actually belonged to the given category. Thus, half of the photos were randomly chosen from the photos tagged with the given category, the other half from all other photos. An additional criterion for the selected photos was a minimum size of 450 pixels for one of the photo dimensions. All photos were scaled such that the longer edge has a length of 450 pixels. The 46 photos from Flickr belonged to the ones labeled as the most "interesting." For all photos in the experiment, ground truth information regarding the region labels was required. For the LabelMe images, manually drawn polygons describing the shapes of the depicted objects are part of the data set. Some photos had to be replaced after a manual check because not all occurrences of an object were labeled or an object described by the given category was depicted, although the photo was not labeled with it. For the Flickr images, the ground truth region labels were manually created by a volunteer not involved in the research.

[2] http://www.flickr.com/

7.4 Photo Classification Results

Excluding the test round, 72 photos in the three rounds were viewed by each subject. This makes a total of 3,888 photo views. In 260 cases (7%), a photo passed without classification, resulting in a total of 3,628 classified photos. 3,279 images (90%) were correctly classified. Overall, 1,624 correct classifications for photos belonging to the given category (true-positive), 1,655 correct classifications for photos not belonging to the given category (true-negative) were obtained. Meanwhile, 241 classifications were false-negative (photo belonged to the category but was classified as not), and 108 classifications were false-positive, which leads to a precision of 94% and a recall of 87% over all users. The number of incorrect assignments per image lies between 2 and 40 with an average of 4. The three photos with the lowest error rate and the three photos with the highest error rate are depicted in Figure 7.5.

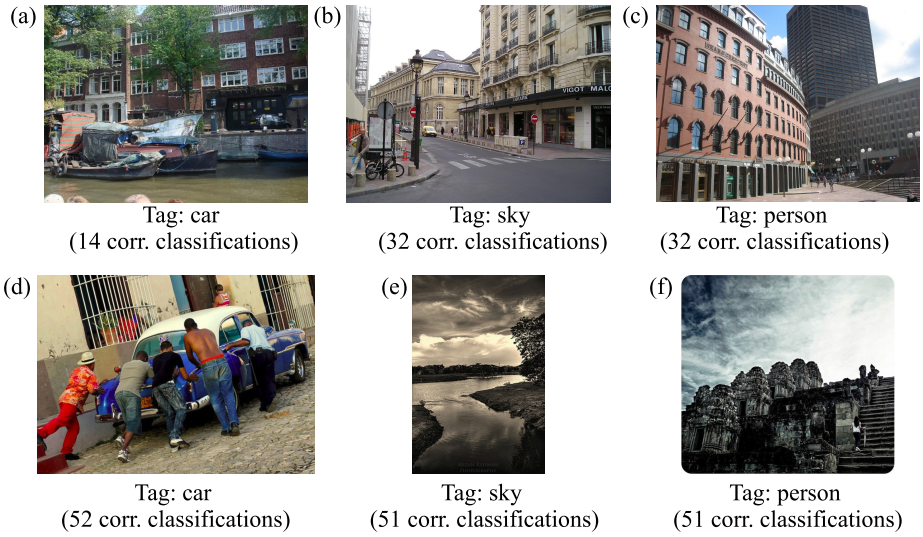


Figure 7.5: Upper row: the three photos with the lowest number of correct classifications. Lower row: the photos with the highest number of correct classifications. All photos show an object described by the given category.

When comparing the error rates for different speed levels, one can see that the number of unassigned or incorrectly assigned photos is increasing with the falling speed of the photos. See Figure 7.6 for the results. The number of not-assigned photos is increasing from 7% to 12%. The number of incorrectly assigned photos is increasing from 4% to 11%. The number of unassigned photos is increasing stronger than the incorrectly assigned photos. Thus, the subjects were still capable of deciding if an image belongs into a category or not, even with a higher speed level. However, they run out of time to focus each image for classification.

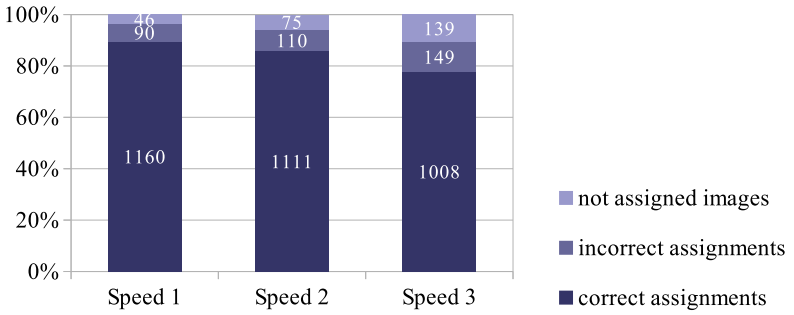


Figure 7.6: Distribution of correctly assigned, incorrectly assigned, and unassigned images for different speed levels. The total numbers of assignments are given inside the bars.

Also the improvement of the gamers while playing the game was investigated. The classification error rate per round was analyzed, for all participants, categories, and speed levels. An overview of the classifications per round over all user can be found in Figure 7.7. Regarding the first and the last round played by each subject, it is to see that there is a small error rate decrease of 2.78%. Of this rate 2.39% are due to an decrease of a reduction of not assigned images, whereas only 0.39% reduction are reflected by the incorrect assignments. However, the second round does not show an improvement as its values worsen the total error rate for 1.31% to a total of 17.82%. The obtained results show that a small learning effect regarding unassigned images takes place after the second round is played. For the classification, however, the values vary only between 8.33% and 9.72%.

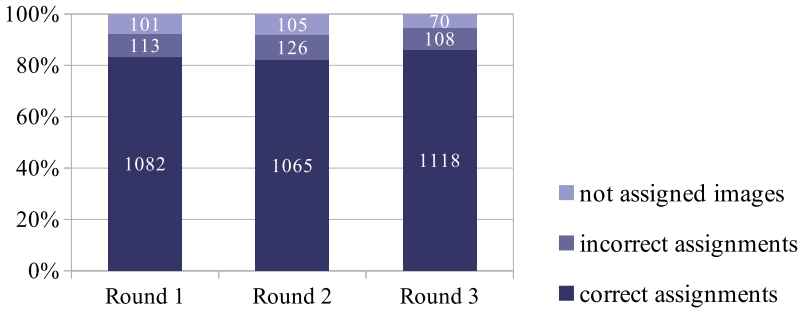


Figure 7.7: Distribution of correctly assigned, incorrectly assigned, and unassigned images separated for each of the three main rounds. The total numbers of assignments are given inside the bars.

The classification results of *EyeGrab* were compared with results from the photo classifications in the strongly controlled experiment which was conducted before and presented in Section 4.1. In the previous experiment, a specific tag was first presented to the subjects. Subsequently, a photo was presented to the user who had to decide whether an object described by the given tag is depicted. The decision was made by pressing a key on the keyboard. Of all classification, 5.4% were incorrect. In this work, 10% of all classifications are incorrect over all speeds. The slowest speed level with an error rate of 7% is close to the results observed in previous work.

In the questionnaire, the subjects were asked how much effort they put into the subjective classification of the photos into "like," and "dislike" (0 = no effort, 5 = much effort). They answered this question with a mean value of 3.43 (SD = 1.35), which points out that their effort was not very high. Of all classified photos, 62% were rated as "like," the rest as "dislike." Of the Flickr images, 70%, were liked in comparison with 56% of the LabelMe images. As the Flickr photos were selected from the most interesting, it can be assumed that they are more attractive to most viewers than the LabelMe photos. This assumption is only reflected slightly in the rating results. In summary, the user gave a rating but it does not seem to be of high quality. Thus, the rating information is not further considered in the remainder of the work.

7.5 Photo Labeling Results

The region labeling results for all photos using the aggregated data of all users who correctly classified a photo were analyzed. In Figure 7.8, the results for the region labeling using the different measures are depicted by comparing precision and recall, as well as precision and F-measure. The best precision with 61% was obtained for the segmentation measure with parameter $k = 0$, which corresponds to very small segments. The highest precision for the heat map measure was obtained for $t = 1$ with 59%. For the baseline approach the best precision was only 19% ($k = 0$). The best recall results were 96% for the heat map measure with $t = 100$, 70% for segmentation measure with $k = 0.5$, and 53% for the baseline with also $k = 0.5$. The F-measure was also calculated, considering both, precision and recall. The overall best F-measure was obtained by the segmentation approach with 32% ($k = 85$), followed by the heat map approach with 31% ($k = 0.5$). The baseline approach clearly performed weaker, with a maximum result of 21% ($k = 0.5$). A Friedman test was applied to compare the results for the best performing parameters. The test showed that the differences are significant ($\alpha < .05$) for precision ($\chi^2(2) = 15.436$, $p = .000$) and F-measure ($\chi^2(2) = 18.048$, $p = .000$). A post-hoc analysis with pairwise Wilcoxon tests with a Bonferroni correction ($\alpha < .017$) showed two significant results for precision between heat map and baseline ($Z = -3.527, p = .000$) and segmentation and baseline ($Z = -3.704, p = .000$). No significance was measured in the post-hoc test for F-measures.

The results vary for the three categories "car," "person," and "sky." For example, the precision values for $k = 0$ are $p_{car} = 0.79$, $p_{person} = 0.28$, and $p_{sky} = 0.76$. This range of results seems to be caused by the sizes of the

objects. The average size of the ground truth objects of the different categories are (compared with the whole image size) as follows: $size_{car} = 11.5\%$ (SD = 8.3%) , $size_{person} = 11.7\%$ (SD = 19.9%), and $size_{sky} = 42.8\%$ (SD = 23.1%). Although the $size_{car}$ and $size_{person}$ are similar, the high standard derivation for "person" points out that the object sizes vary strongly. Very small objects are known to be difficult in the region labeling (cf. Section 4.5.1).

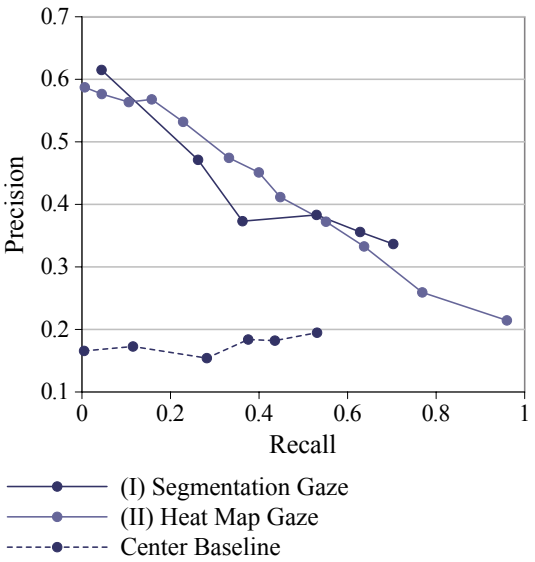


Figure 7.8: Precision and recall results for the three labeling approaches. The curves are limited by the investigated parameters (e.g., the Center Baseline by the number of segmentation levels).

In addition, the region labeling results for the different falling speeds were analyzed. A faster falling speed increases the pressure on the user to perform the classification. An overview of the results for the different speed levels can be found in Figure 7.9. It shows that the falling speed does not have a high impact on precision and F-measure. For both eye tracking measures, the medium speed level delivers the best results. However, only minor differences can be noticed. Please note that the results for all speeds are not the average of all speed levels as the region labeling for the different speed levels is done with only one-third of the data. This is caused by the fact that every user played the game in three different speed levels (cf. Section 7.3). Thus, the influence of the speed on the region labeling results is, at the least, not strong.

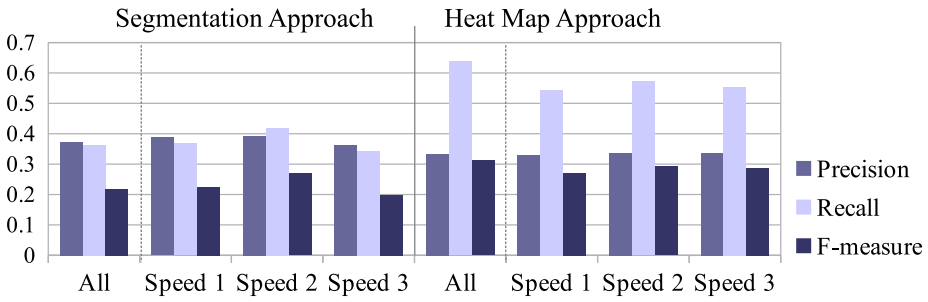


Figure 7.9: Region labeling results for different falling speeds.

7.6 Comparison of *EyeGrab* Results with Those from the More Controlled Experiment

The results in terms of precision and F-measure from the *EyeGrab* experiment with the results obtained from previous work, presented in Section 5 were compared. The best performing parameters were determined in the previous work by means of a training set and applied to the test set of the previous work and to the *EyeGrab* data. The parameters are $k = 0.1$ for the segmentation measure, $t = 95$ for the heat map measures, and $k = 0.4$ for the baseline. The results are depicted in Figure 7.10.

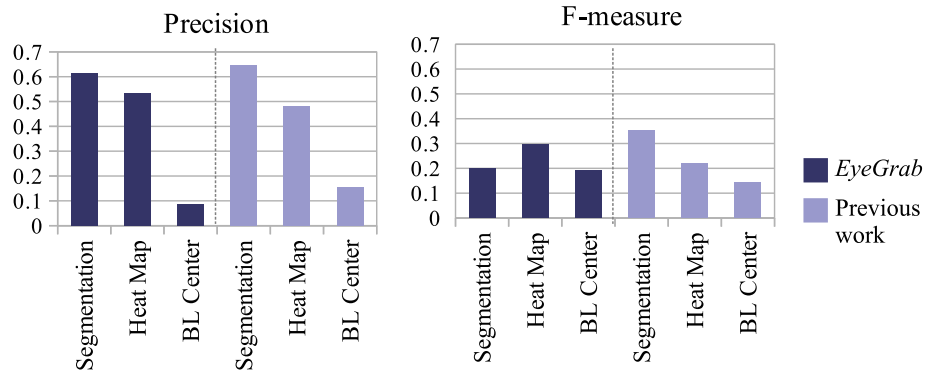


Figure 7.10: Region labeling results for *EyeGrab* and previous work (Section 5).

The segmentation measure performs best, while the baseline approach delivers clearly weaker results than both eye tracking methods. The F-measure results are more diverse. The differences between the two gaze-based measures and the baseline are less distinct for the *EyeGrab* data than for the data from the previous experiment presented in Section 5 (i. e., the results between the measures and baseline in the earlier experiment differ more). Using the parameters from earlier work for the *EyeGrab* analysis delivers only slightly better results for

the segmentation approach than the baseline, whereas the heat map approach performs clearly better. The center baseline results for photos of the first experiment and *EyeGrab* data were compared in a Mann-Whitney U test and do not obtain a significant difference, neither for precision ($U = 467, p = .291$) nor for F-measure ($U = 446, p = .302$). Thus, it can be concluded that the photo sets are comparable concerning the center baseline results and infer that the region labeling results can be compared. No statistically significant differences can be found comparing the results from *EyeGrab* and the previous work with regard to the segmentation measure and the heat map measure, neither for precision (segmentation: $U = 528, p = .909$; heat map: $U = 480, p = .467$), nor for F-measure (segmentation: $U = 436, p = .19$; heat map: $U = 468, p = .376$). Thus, similar results in region labeling in *EyeGrab* and the previous, simplified experiment were obtained.

7.7 Conclusion

The work in this section showed that the labeling of image regions is possible by means of data collected from subjects playing the immersive game-with-a-purpose *EyeGrab*. A precision of 61% of correctly labeled image region pixels was obtained. For one of two gaze-based measures, the results were comparable with those from the previous, much less immersive experiment, described in Section 5. This is quite interesting as the conditions for obtaining the gaze data are more difficult due to factors such as time pressure and distraction caused by the gaming environment in *EyeGrab*. The region labeling results are only slightly influenced by different speed levels, which are forcing the subjects to make decisions on the photo classifications faster. The possibility to offer *EyeGrab* as a game for a wide public is appealing, as a big number of region labels could be achieved from crowd sourcing.

Chapter 8

Photo Selection by Gaze Analysis

Users easily take hundreds of photos during vacation or personal events such as weddings or birthday parties. The amount of digital images makes the creation of selections an essential task. Only few are worth to be kept in a photo book or to show them to friends and family. Often, selections of "good" photos are created to reduce the amount of photos stored or shared with others [FKP+02, KSRW06, NF09, RW03]. The manual selection of interesting photos is possible but a very labor-intensive approach. While users enjoy certain photo activities like the creation of collages for special occasions such as anniversaries or weddings, these tasks are seen as "complex and time consuming" for normal collections [FKP+02].

In the section on Related Work 3.2, different content- and context-based approaches for the creation of selections were presented. While acknowledging the achievements made by content- and context-based approaches, they miss an important factor in the photo selection process: the user's interests. In the first sections of this thesis, the gaze data was analyzed with the goal to label image regions. Here, the information gained from the eye tracking data is interest. Capturing the user's interests is important as the human photo selection process is assumed to be guided by very individual factors and is highly subjective [SEL00].

First in this thesis, the capability of gaze analysis in the labeling of image regions was investigated in Sections 4 to 7. The approach of exploitative eye tracking analysis is applied to the photo selection problem for supporting the users. It is assumed that interesting photos catch the human attention already at first glimpse and the catch it longer than less interesting photos. The visual attention is measured through gaze data recorded with eye tracking devices during the photo viewing process. As such devices become cheaper and more ubiquitous and it is expected that they become part of future standard computer hardware (cf. Section 2.2.2), the opportunity to record gaze data during everyday tasks like photo viewing can be assumed.

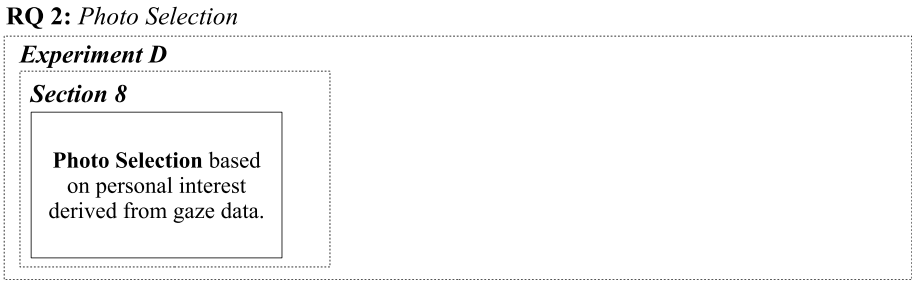


Figure 8.1: Embedding of this experiment in the context of this thesis. After the first part of this theses dealt with the labeling of image regions, the work presented in this Section 8 investigated the exploitation of visual attention in the creation of photo selections.

The main research question addressed in this part of the thesis is:

RQ 2 *Can important photos in a collection be identified from gaze analysis, and is this information worthwhile in the creation of individual photo selections?*

Human gaze paths are influenced by different factors from low-level information such as contrasts and colors to high-level factors like a given task, see Section 2.1.3. In this section it is investigated, if the influence of "interest" on the viewing behavior is high enough to derive valuable information from the gaze paths. The aim is to identify the users' preferences when viewing photos *without* a specific task, besides the instruction to get an overview of a photo collection. The visual attention while viewing the photos is analyzed for creating personalized photo selections. Photos with the highest attention, that is, those that are fixated longest, are assumed being most interesting to the user and should be part of a selection. In the evaluation, gaze-based selections are compared with selections based on related work. The analysis aims to answer the question:

RQ 2.1 *Does a gaze-based selection outperform objective selections based on content and context analysis when comparing the selections with those created manually by the users?*

Eye movements are strongly influenced by interest (see Section 2.1 for the background on human visual perception). Here, it is also investigated how the personal relevance of viewed photo sets influences the quality of the gaze-based photo selection results. To this end, photos of an event the user took part in or in which the user knew the participants ("home collection") were shown as well as photos of an event the user was not personally involved in ("foreign collection"). Thus, selection results for both kinds of photos can be compared for answering the research question:

RQ 2.2 *Does the personal interest in a viewed photo set have an impact on the obtained selection results?*

First in this section, the experiment design including the data set and the experiment application are presented in Section 8.1. Subsequently the applied methods for creating photo selections are introduced in 8.2. In Section 8.3, the results of analyzing the participants' behavior when viewing and selecting photos as well as the distribution of the selected photos are shown. Finally, the gaze selection results are presented and discussed in 8.4 before the section is concluded. The work presented in this section was published in [WNS+13] and [WSS14b].

8.1 Experiment

An experiment application was developed that allowed the participants to view and select photos from a collection $C = \{o_1, o_2, \dots o_n\}$. In the first part of the experiment, eye tracking data was collected from the participants while viewing photos. Subsequently, ground truth data was collected by asking the participants to manually create three personal selections of these photos.

8.1.1 Participants

A total of 33 participants (12 of them female) completed the first part of the experiment. Twelve were associated with a research lab A in North America and 21 with institute B in Europe. Members of institute A and institute B did not know one another. Their age ranged between 25 and 62 years (M: 33.5, SD: 9.57). Twenty of them were graduate students and 4 post-docs. The remaining 9 participants worked in other professions, such as secretaries or veterinary assistants. Eighteen of the 33 participants (7 of them female) completed the second part of experiment. Six of them were associated with institute A and 12 of them with institute B. Their average age was 31.7 (SD: 8.74).

8.1.2 Materials

The experiment photo collection C consisted of two collections of photos taken during two social events, one organized by each of the two research institutes the participants were associated with. The activities during the events included teamwork situations, group meals, as well as leisure activities such as bowling and hiking. Event A lasted half a day and event B three days. The photos were taken by different people: three people for collection C_A and two for collection C_B. The photos were not preselected but taken directly from the camera. Only two extremely blurry photos were removed. The photo collection of the participants' own institute is called "home collection" and the other one "foreign collection" (cf. Figure 8.2). Collection C_A (photos were taken during the event of institute A) consisted of 162 photos and C_B 126 photos. The photo collection $C = C_A \cup C_B$ was split chronologically into sets of nine photos $c_i = \{p_{i \cdot 9+1}, \dots, p_{i \cdot 9+9}\}$. Each set c_i contained only photos of one of the

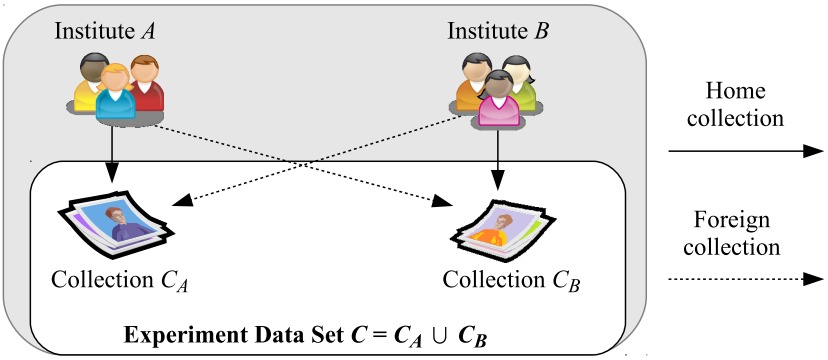


Figure 8.2: Composition of the experiment data set.

collections. The complete collection of 288 photos was thus split into 32 sets (18 sets of C_A and 14 sets of C_B).

It is assumed that the photos of the home collection are of higher personal interest for the participants than the photos of the foreign collection. This assumption is supported by results from the questionnaire. The participants were asked to indicate how interesting the two photo collections were using a Likert scale from 1 ("Not interesting") to 5 ("Very interesting"). For the home collections, the question was answered with an average of 4.36 (SD: 0.6) and for the foreign collection with an average of 2.72 (SD: 1.14). A chi-square test was applied for testing the significance of the differences as the data was not normally distributed (shown by a Shapiro-Wilk test of normality with $p < .001$ for the home set ratings and $p < .018$ for the foreign set ratings). The chi-square test showed a statistically significant difference between the answers, $\chi^2(5, N = 66) = 34.594, p < .001$.

8.1.3 Apparatus

The experiment was performed either on a 22-inch or a 24-inch monitor for the two research groups (cf. section Participants). The participants' gazes were recorded with a Tobii X60 eye tracker at a data rate of 60 Hz and an accuracy of 0.5°. The distance between the participants and the computer screen was about 60 cm. The setup (including a laptop, the eye tracking device, and a standard computer mouse) was the same for both groups.

8.1.4 Procedure

The experiment consisted of four steps, one viewing and three selection steps. In the first step ("Photo Viewing" in Figure 2.8), the participants were asked to view all photos of collection C with the goal "to get an overview." Eye tracking data was recorded only during this photo viewing step. This was crucial to avoid an impact of the selection process on the viewing behavior. The order in which the two collections C_A and C_B were presented to the participants in the experiment was alternated. No time limit was given for viewing the photos.

The participants were told that they would afterward, in the second step, create selections of the photos. No details about the selection process were given at this stage of the experiment.

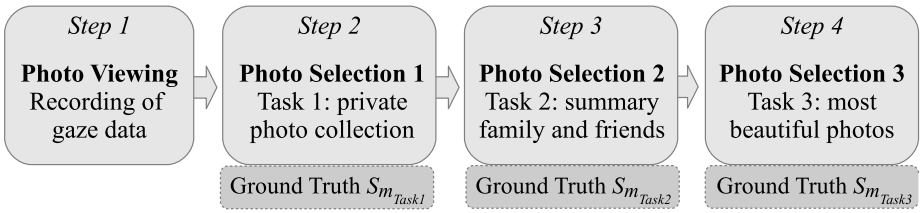


Figure 8.3: Experiment setup with the photo viewing step and the three selection steps.

Each photo set c_i was presented on a distinct page; the photos were arranged in 3×3 grids in the center of the screen. The photos' maximum height and width were set to 330 pixels, corresponding to about 9° at the visual angle. The minimum distance between the photo was 22 pixels (0.6°). By clicking on a button, the next set of nine photos was presented. The photos in each set c_i were arranged in random order, whereas the sets themselves and the order of the sets remained the same to preserve the chronological order of the events.

After having viewed all photos, the participants were asked to select exactly three photos of each set c_i in the second step ("Photo Selection 1" in Figure 8.3). The photos were selected by means of a drag-and-drop interface as depicted in Figure 8.4. The same sets as in the viewing step were again presented to the participants but the photos were rearranged in a new random order. The participants were asked in this second step to select the photos as they would do for their private photo collection. No specific instructions were given regarding the selection criteria for choosing the photos. Thus, the participants could apply their own (perhaps even unconscious) criteria. Also, in the third and fourth steps ("Photo Selection 2" and "Photo Selection 3" in Figure 8.3), the participants performed manual selections. In the third step (Task 2), the participants were asked to "select photos for their friends or family that provide a detailed summary of the event." The fourth step (Task 3) was to "select the most beautiful photos for presenting them on the web, for example, on Flickr."

In the experiment steps 3 and 4, the users performed the manual selections only for the photo sets belonging to their home collections, not the complete collection C. Eighteen of the participants completed these tasks. The manual selections served as ground truth in the later analysis. Finally, the participants filled in a questionnaire. It comprised questions about demographical user data (age, profession), the experiment data set, and the experiment task as well as a rating on different selection criteria.

Figure 8.4: Photo selection interface with one selected photo.

8.2 Methods for Creating Photo Selections

The aim of the photo selection methods is to create a subset $S \subset C$ that best suits the user's preferences. The capabilities of each method are evaluated by comparing the calculated selection with the manual selection for each set c_i created during the experiment. A "perfect" selection would be a selection identical to the manual selection. The photos C were displayed in sets of nine photos c_i. Selections of $j = 3$ photos are created for each set. An overview of the different photo selection approaches is shown in Figure 8.5. They are presented in detail in the following sections. First, the content-based and context-based measures for photo analysis used in the baseline system are described. Subsequently, eye tracking based measures and then the combination of different measures by means of logistic regression are presented. Finally, the calculation of precision P for comparing the selections with the ground truth selections $S_{m_{Task1}}$, $S_{m_{Task2}}$, and $S_{m_{Task3}}$ is described.

8.2.1 Content and Context Analysis Baselines

Six measures that analyze the context or the content of photos are used as baselines. An overview is shown in Table 8.1. The measures are motivated from related work, and details on their implementations can be found in the cited papers.

The first measure, (1) concentrationTime, relies on the assumption that many photos are taken within a short period of time when something interesting happens during an event [LLT03]. This measure is context-based as the information when a photo was taken is obtained from the photos' meta-information. Li et al. [LLT03] created a function f_k indicating the number of photos taken for a point in time. By means of the first derivation of this function, a temporal representative value for each photo is calculated. The next four measures are content-based as they analyze the photos' content at pixel level. The photos'

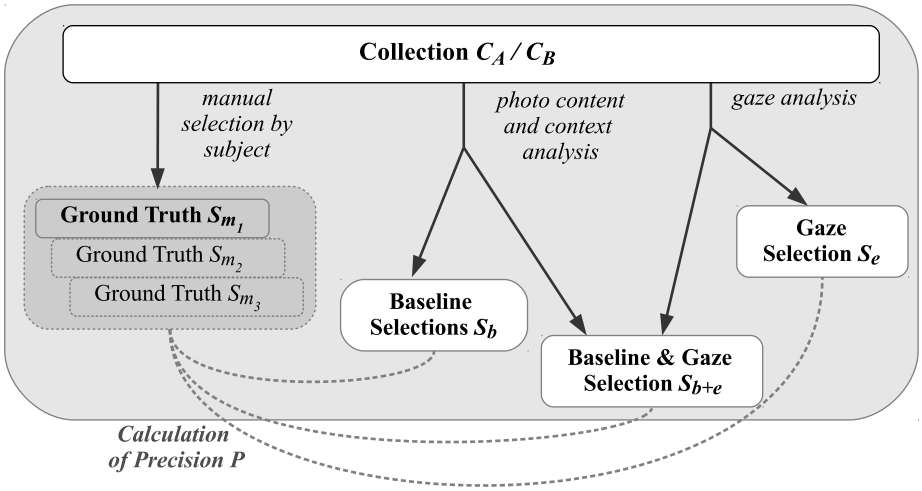


Figure 8.5: Overview of the investigated photo selection approaches and calculation of precision P.

No	Measure	Description
1	concentrationTime	o was taken with other photos in a short period of time [LLT03]
2	sharpness	Sharpness score [XZC+08]
3	numberOfFaces	Number of faces
4	faceGaussian	Size and position of faces [LLT03]
5	personsPopularity	Popularity of the depicted persons [ZTL+06]
6	faceArea	Areas in pixels covered by faces

Table 8.1: Baseline measures based on content and context analysis for photo o.

quality is considered in measure (2) sharpness by calculating a sharpness score as presented by Xiao et al. [XZC+08]. The score is calculated as $Q = \frac{strength(e)}{entropy(h)}$ with $strength(e)$ as the average gradient edge strength of the top 10% strongest edges and $entropy(h)$ as the entropy of the normalized gradient edge strength histogram. The edge strength is calculated by the well-known Sobel operator from computer vision.[1]

Related work, presented by Boll et al. [SB11], showed that depicted persons play an important role in the selection of photos. The four measures (3) to (6) are based on the analysis of depicted persons. Measure (3) numberOfFaces simply counts the number of faces on a photo. The detection of faces is done using OpenCV's Haar Cascades[1]. Also, measure (6) faceArea is based on this calculation. It considers the size in pixels of the photo areas covered by human faces. A Gaussian distribution of the face areas as proposed by Li et al. [LLT03]

[1] http://opencv.org/, last visited September 17, 2013

is considered by measure (4) faceGaussian, identifying photos with large depicted faces in the photo's center. Measure (5) personsPopularity considers a persons' popularity in the data set as presented by Zhao et al. [ZTL+06]. It assumes that faces appearing frequently are more important than the ones appearing less often. The calculation is performed by the OpenCV's face recognition algorithm and considers persons appearing in each set c_i of nine photos. This measure is context-based as well as content-based.

8.2.2 Gaze Analysis

A visualization of a sample gaze path, recorded in the experiment, can be found in Figure 8.6. Fixations are visualized as circles, and the diameter indicates the duration of a fixation. The gaze paths has to be filtered for extracting fixations and the fixations are analyzed by means of eye tracking measures, as described in Section 2.4. An overview of all measures used in this section can be found in Table 2.2. To compensate the inaccuracy of the eye tracking data, fixations in the surrounding of 11 pixels (0.3° at the visual angle) of a photo are also considered as being on a photo (the smallest distance between two photos is 22 pixels, or 0.6°). This extension as introduced in Section 4.2.1. In previous sections, a weighting factor was applied to small regions. In this analysis, it was not apply as all photos were of about the same size.

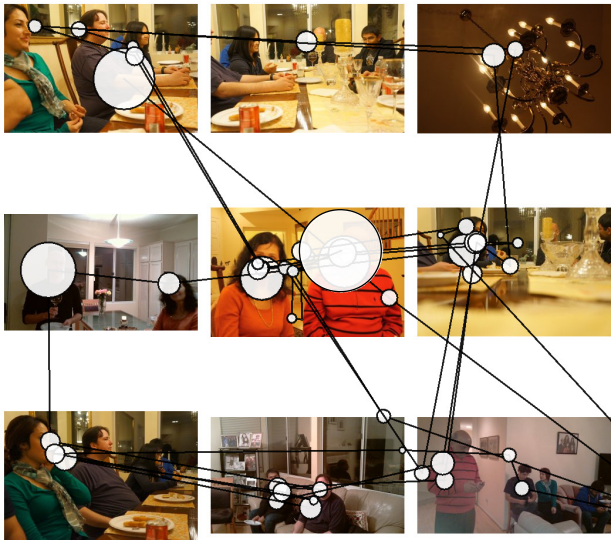


Figure 8.6: Visualization of a gaze path on a photo set.

8.2.3 Combining Measures Using Logistic Regression

Different combinations of the content-based and context-based measures and eye tracking measures are investigated. To this end, all measure values are normalized per set c_i by subtracting the mean of the nine values per set and

No	Measure	Description
7	fixated	Indicates if photo o was fixated or not
8	fixationCount	Counts the fixations on o
9	fixationDuration	Sum of the duration of all fixations on o
10	firstFixationDuration	Duration of the first fixation on o
11	lastFixationDuration	Duration of the last fixation on o
12	avgFixationDuration	Average of the durations of all fixations on o
13	maxVisitDuration	Maximum visit length on o
14	meanVisitDuration	Mean visit length on o
15	visitCount	Number of visits within o
16	saccLength	Mean length of the saccades before fixating on o
17	pupilMax	Maximum pupil diameter while fixating on o
18	pupilMaxChange	Maximum pupil diameter change while fixating on o
19	pupilAvg	Average pupil diameter while fixating on o

Table 8.2: Eye tracking measures for photo o.

dividing it by the standard derivation σ. The measures are combined by means of a model learned from logistic regression as presented by Fan et al. [FCH+08]. The data of all users is split into a training set and a test set. About 15% of the data are selected as test data, which correspond to five sets of nine photos for every user as test data and 27 sets of nine photos as training data. The test sets are randomly chosen. Only complete sets c_i are selected for training and testing, respectively. When analyzing subsets of the data (e.g., when analyzing only the photos that are part of the home collection for each user) less data is available. The test data size is reduced to three sets of nine photos. The model is trained with the training data of all 33 users. That corresponds to $33 * 27 * 9 = 8,019$ training samples, when using the whole data set C. This number reduces to 3,699 samples when training the model only with those photos of the home sets. 1,998 samples were used when performing the training for the data from the experiment steps 3 and 4, which were completed by less participants. The default parameter settings of the LIBLINEAR library [FCH+08] are used for training. For every analysis, 30 iterations with different random splits are performed and the average results of all iterations are presented in this section.

Three different measure combinations are investigated. Selection S_b takes only the baseline measures (1) to (6) into account. For the selection S_{b+e}, all 19 measures are considered in the logistic regression. For S_e exclusively the gaze measures (7) to (19) are used in the learning algorithm. The logistic regression predicts a probability of being selected for each photo in set c_i of nine photos.

The three photos with the highest probability are chosen for the selection and compared with the ground truth selections $S_{m_{Task1}}$ to $S_{m_{Task3}}$.

8.2.4 Computing Precision P

For comparing a computed selection to the ground truth, the percentage of correctly selected photos of all selected photos is calculated (precision P). This calculation is conducted for each set c_i. Precision P for a selection approach is the average precision over all sets c_i. As three of nine photos are selected, a random baseline selecting three photos by chance would have an average precision of $P_{rand} = 0.\overline{3}$. Figure 8.7 shows an example page with two selections and corresponding precision results. For both selections S_e and S_b the same precision with $P = 0.667$ is obtained, as for both selection two of three photos are part of the baseline selection $S_{m_{Task1}}$. Recall is always the same as precision because of the fixed number of selected photos.

Figure 8.7: Examples of different selections and evaluation results.

8.3 Users' Photo Viewing and Selection Behavior

In this section, the users' photo viewing time and photo selection time in the experiment are investigated. Subsequently, the distribution of the manual photo selections of the participants is presented. Finally, the users' rating regarding the importance of different photo selection criteria are given.

8.3.1 Viewing and Selection Durations

The sets c_i of nine photos were viewed on average for 12.6 s (SD: 11.9 s). The shortest viewing time was below a second and the longest 121.1 s. The viewing duration were on average higher for the sets belonging to the home collection with 13.3 s (SD: 12.2 s) compared with 11.8 s (SD: 11.5 s) for the foreign collection. These values are calculated from the time the participants looked at the photo viewing pages in the experiment application. The distribution of the viewing durations significantly deviated from a normal distribution (shown by a Shapiro–Wilk test of normality with $p < .001$ for the home set and foreign set, respectively). Thus, a Mann–Whitney U test is applied in comparing the viewing durations for the sets belonging to the home collection and the foreign collection. The result is that the viewing durations are significantly longer for the home sets compared with the foreign sets ($U = 138462, Z = -3.194, p = .001$).

The average selection time per set was 20.9 s (SD: 11.6 s) for Task 1. The selection durations were slightly shorter for the foreign sets with an average of 20.1 s (SD: 10.5 s) compared with those of the home collection with an average of 21.7 s (SD: 12.6 s). Like the viewing durations, the distribution of the selection durations also significantly deviated from a normal distribution (shown by a Shapiro–Wilk test with $p < .001$ for the home set and foreign set, respectively). Applying a Mann–Whitney U test on the selection durations showed that the differences are not statistically significant ($U = 125877, Z = -1.013, p = .311$). The selection process clearly took longer than the viewing step (+66%). Although the selection process was different from selections usually performed in daily life, it shows that the selection of photos is more time-consuming than the viewing.

The participants rated how difficult the creation of the selection was on a Likert scale from 1 ("It was hard to select the photos") to 5 ("It was easy to select the photos"). The ratings were performed separately for the home collection and the foreign collection. The results show that the ratings were on average higher for the home set with 3.85 (SD: 0.94) versus 3.06 (SD: 0.94). Shapiro–Wilk tests revealed that the data was not normally distributed ($p < .001$ for the home set ratings and $p < .015$ for the foreign set ratings). A chi-square test was applied, which showed that the difference is significant ($\alpha < 0.05$) with $\chi^2(4, N = 66) = 9.714, p < .046$.

8.3.2 Distribution of the Users' Manual Photo Selections

In Figure 8.8, the numbers of selections for all photos are displayed. On average, every photo was selected 3.7 times. The highest number of selections was 24. Approximately 75% of the photos were selected five times or less. Thus, most of the photos were selected only by a minority of the participants. There are very few overall favorites, those that were selected by most of the participants. The two photos with the highest number of selections are shown in Figure 8.9. The left photo was selected by 24 users and the right one by 21. Thus, the photo selections were very individual in the experiment. This finding confirms results from previous work [SEL00].

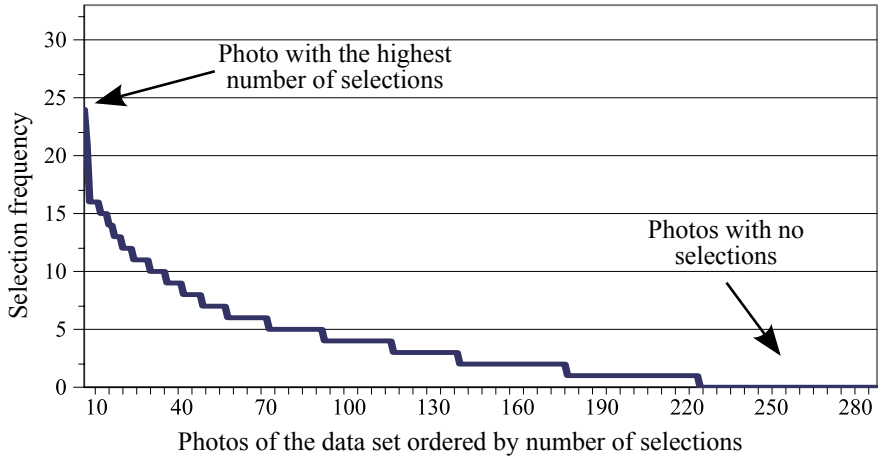


Figure 8.8: The number of selections for all photos in data set C, ordered by the number of selection.

Figure 8.9: The two most frequently selected photos.

Cohen's kappa k was calculated for all possible user pairs with $k = \frac{q_x - q_r}{1 - q_r}$. In this formula, q_x is the observed agreement between two users. This corresponds to the percentage of photos that were selected by both users. The value $q_r = 0.556$ is the probability of a by-chance agreement of two users on their photo selections. As the number of selected photos is high compared with the total number of photos per page (three out of nine), the value for q_r is already quite high. The obtained results for Cohen's kappa comparing all user selections have a minimum of $k = 0.5$ and a maximum of $k = 0.757$. The average Cohen's kappa over all users is $k = 0.625$. The average result lies only about 12% above the by-chance probability of $q_r = 0.556$. This further confirms that the photo selections are very diverse.

8.3.3 Ratings of Photo Selection Criteria

In the second experiment step, where a manual selection was created for Task 1, no specific criteria regarding the selection of photos were given to the participants. They were just asked to create selections for their private photo collection and could apply their own criteria. In the questionnaire, the participants were asked to indicate how important different criteria were for their selections. Nine criteria were rated on a five-point Likert scale. In addition, the users were given the option to add criteria as free text. The selection criteria were taken from related work [SMJ11, SB11, XZC+08, KSRW06, RW03]. An overview of the criteria rated by the participants can be found in the following list:

1. Attractiveness — the photo is appealing
2. Quality — the photo is of high quality (e.g., it is clear, not blurry, good exposure)
3. Interestingness — the photo is interesting to you
4. Diversity — there are no redundant photos
5. Coverage — all locations/activities of the event are represented
6. Depiction of the persons most important to me
7. Depiction of all participants of the event
8. Refreshment of the memory of the event
9. Representation of the atmosphere of the event

Figure 8.10 shows the results of the ratings on a Likert scale between 1 ("Not important") and 5 ("Very important"). The criteria are ordered by their mean results. One can see that some of the criteria have a wide range of ratings, from 1 to 5. Every criterion has at least one rating with five points.

The criteria were classified as "rather objective" (striped bars) and "rather subjective" (solid bars), expressing if a criterion is an objective measure and can (theoretically) be calculated by computer algorithms. Although this classification can be a subject of discussion, it serves the goal to better understand the nature of selection criteria. In Figure 8.10, it can be seen that three of the five criteria with the largest range in the answers (8, 4, 6, 7, 5) belong to the objective criteria. Also, the two criteria with the lowest mean results are rather objective criteria. It is remarkable that the two criteria with the highest average rating and the smallest deviation, *3. Interestingness* and *1. Attractiveness*, are rather subjective criteria. Also, four of the five highest-rated criteria are subjective. Eight participants provided additional criteria as free comments such as "the photo makes me laugh" or "the photo is telling a story." All criteria added by the participants were very subjective.

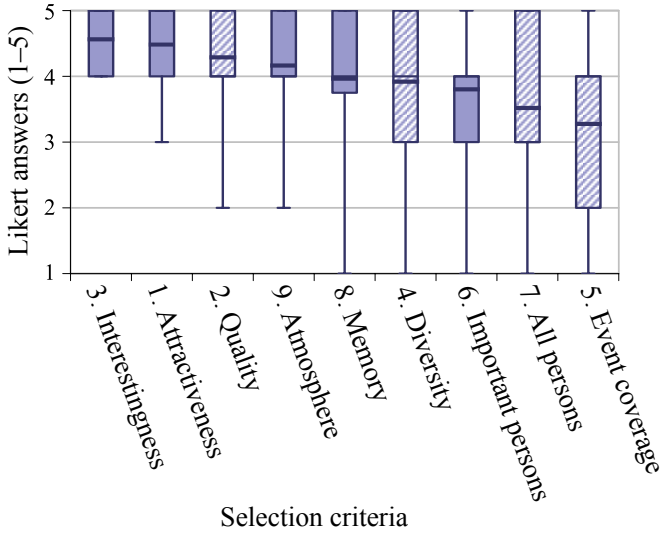


Figure 8.10: Selection criteria sorted by mean value.

8.4 Gaze Selection Results

The results for selections based on single measures are presented, followed by the results from combining the measures with machine learning. Subsequently, the influence of personal interest in the photos on the selection results is shown. Finally, the weak influence of different selection tasks is revealed.

8.4.1 Correlation between Measures and Manual Selections

In order to investigate how well the measures work in terms of predicting which photos are selected by participants, the correlations between the measure results and the number of correctly selected photos by these measures are analyzed. To this end, the measure results (i. e., number of photos selected) were binned in the range of $\pm 1\sigma$ in equidistant steps of 0.2σ. This results 11 bins, for which the photos selected are calculated by the measures as shown in Figure 8.12. The baseline measures are depicted as dashed lines and the eye tracking measures as solid lines. For example, for the bin $\sigma = 0.2$ and the measure (1) concentrationTime the percentage of correctly selected photos is 0.312. A correlation between an increasing measure value and an increasing percentage of selected photos can be seen for the eye tracking measures. For the baseline measures, a high fluctuation can be observed. A Pearson's correlation analysis shows a very strong, positive correlation between the percentage of selected photos and the eye tracking measures, which were statistically significant: (9) fixationDuration ($r(9) = .983, p = .000$), (13) maxVisitDuration ($r(9) = .928, p = .000$), and (14) meanVisitDuration ($r(9) = .937, p = .000$). The correlations for the baseline measures were much weaker and not significant, (1) concentrationTime

Figure 8.11: Sample photos with the highest and lowest results for three of the baseline measures.

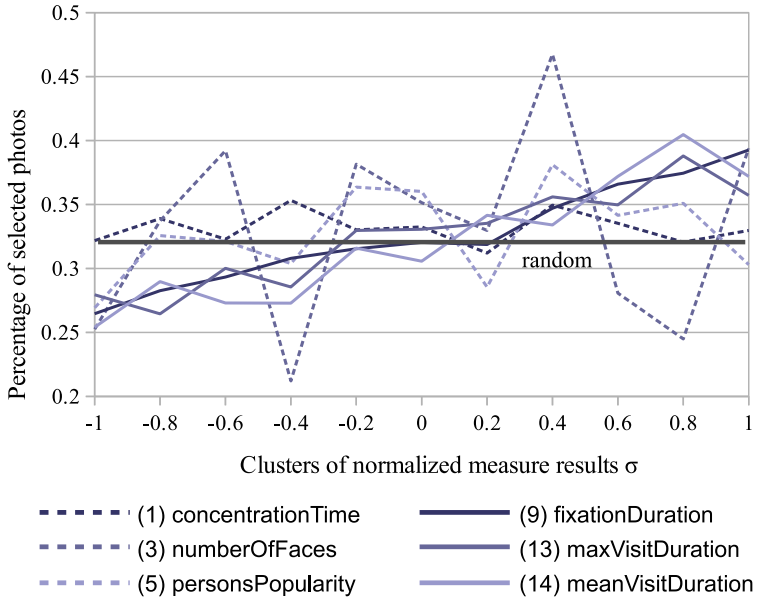


Figure 8.12: Correlation between percentage of correctly selected photos and measure values for gaze (solid lines) and baseline measures (dashed lines). The results show that the eye tracking measures (9), (13), and (14) are a good linear estimator for the selected photos. The baseline measures show a high variance and thus can hardly predict the selected photos.

($r(9) = -.054, p = .874$), (3) numberOfFaces ($r(9) = .181, p = .595$), and (5) personsPopularity ($r(9) = .349, p = .292$). Overall, this shows that the eye tracking measures are a good approach to predict which photos are picked out by the participants for their photo selections.

8.4.2 Selection Results for Single Measures

Figure 8.11 shows some sample photos with the highest and lowest measure results for three baseline measures. The samples show that the measures basically succeeded in analyzing the content of the photos. For example, the first row shows the most blurred photo (left) and the photo with the highest sharpness (right). But it also shows the limitations of today's computer vision approaches as, e. g., the photo with the highest number of faces is determined with 7 faces, although almost 20 people are captured in this shot. Please note that for measure (4) faceGaussian the examples with the lowest result of 0 (no faces) are not considered in this overview.

As described in the previous section, the data set is randomly split into a training set and a test set in 30 iterations. For the analysis of the performance of single measures in this section, no training was needed. Thus, the training data set was not considered but for ensuring compatibility to the following sections, the measures are applied only on the test data sets. Figure 8.13 shows the

average results for each user over the 30 iterations. Precision P was calculated by using only a single measure for creating the selections of the test data sets. The photos in the selections were the three photos with the highest measure values. The results strongly vary between $P = 0.202$ and $P = 0.56$ for different users and measures. Of all baseline measures, (6) faceArea performed best with a mean precision of $P = 0.365$. One can see that the face-based baseline measures (3) to (6) show high variances in precision P. (2) sharpness delivered a mean result that lies with $P = 0.344$, which is close to random selection with $P_{rand} = 0.\overline{3}$. It is interesting that photo quality as a selection criterion was ranked very high by the users (third important measure, see previous section) but the sharpness score, considering the photo quality, did not deliver good results. On average, 29.3 fixations were recorded per set (SD: 19.97). The average fixation number per photo is 3.25 (SD: 3.15). The highest median precision results are obtained by the three eye tracking measures (9) fixationDuration ($P = 0.419$), (13) maxVisitDuration ($P = 0.42$), and (14) meanVisitDuration ($P = 0.421$). The pupil-based eye tracking measures (17) to (19) did not deliver good results. They are close to the precision results for a random selection $P_{rand} = 0.\overline{3}$ or even slightly below for (19) pupilAvg with $P = 0.32$.

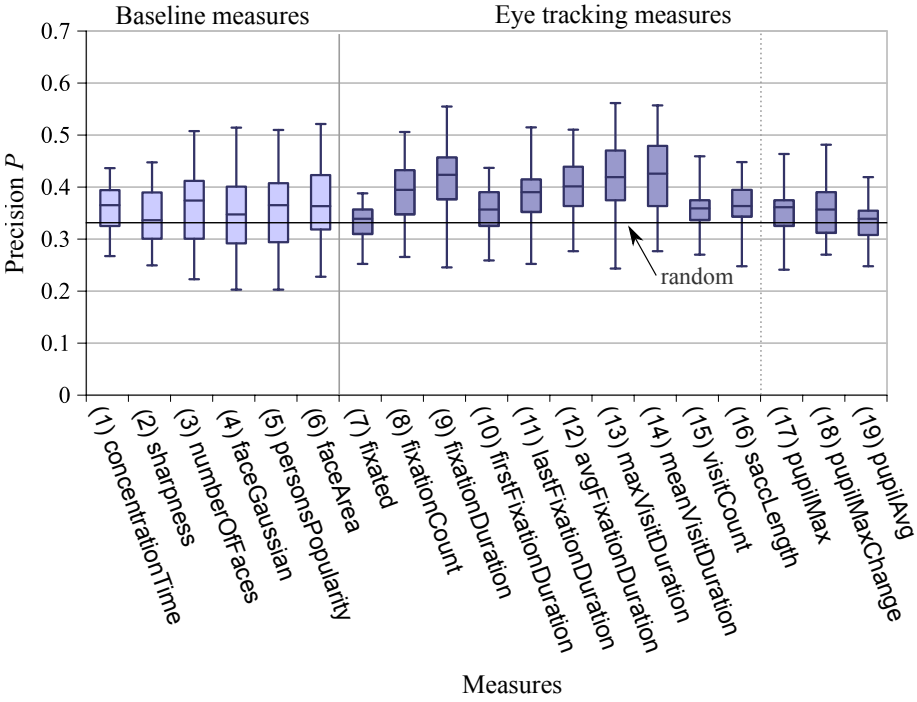


Figure 8.13: Precision results for all users averaged over 30 random test sets when selecting the photos based on single measures.

8.4.3 Selection Results for Combined Measures

The measures are combined by means of logistic regression. Pairwise Pearson correlation tests showed that all correlation coefficients were below 0.8. Thus, the correlations between the single measures were not too high, and we, therefore, decided not to exclude measures from the logistic regression. The best average precision result of $P = 0.428$ is obtained for S_{b+e}, the selections created based on baseline measures and eye tracking measures. The result for S_e (only eye tracking measures) is $P = 0.426$ and $P = 0.365$ for S_b (only baseline measures). Using gaze information improves the baseline selection by 17%. The results of all users averaged over 30 iterations are shown in Figure 8.14. Statistical tests were applied on the average precision values obtained from the 30 random splits for each user for investigating the significance of the results. A Mauchly's test showed that sphericity had been violated ($\chi^2(2) = 27.141, p = .001$). Consequently, the nonparametric Friedman was used for the analysis. The differences between the three selections are significant ($\alpha < 0.05$) for P with $\chi^2(2) = 49.939, p = .001$, $n = 33$. For post hoc analysis, pairwise Wilcoxon tests were conducted, with a Bonferroni correction for the significance level ($\alpha < 0.017$). The tests showed that baseline selection S_b was significantly outperformed by the gaze including selections S_{b+e}, $Z = -4.297, p = .001$, and S_e, $Z = -3.600, p = .001$. No significant difference was detected between S_{b+e} and S_e, $Z = -0.019, p = .496$.

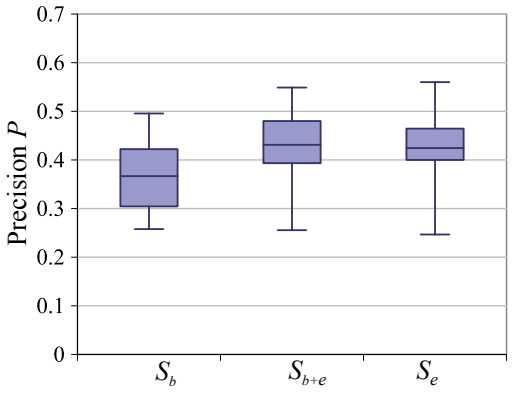


Figure 8.14: Precision results for all users averaged over 30 random splits obtained from combining measures by logistic regression. The results are based on baseline measures S_b, eye tracking measures S_e, and all measures S_{b+e}.

Figure 8.15 shows the results for the 30 random splits for one single user. Precision results are between $P = 0.267$ and $P = 0.6$ and point out the strong influence of the training data and test data splits. The user selected for this example is the one with the precision result closest to the average precision over all users.

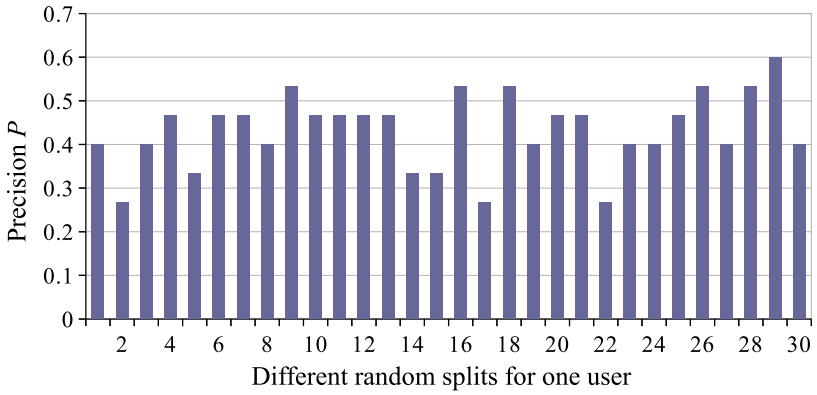


Figure 8.15: Precision results for S_{b+e} over 30 different random splits for one user.

8.4.4 Influence of Personal Involvement

For each user, it is distinguished between photo sets c_i that were part of the home collection and those that were part of the foreign collection as described in the section Experiment. Precision of selection S_{b+e} was calculated separately for both collections. The results can be found in Figure 8.16. They show that P results for the foreign photo set have a larger range, and the average precision is lower with $P = 0.404$ compared with $P = 0.446$ for the home set. Comparing the precision result for the home sets with the results for S_b leads to an improvement of 22%. A Wilcoxon test showed a significant difference between the precision values of all users for the home and foreign photo sets, $Z = -2.842, p < .004$.

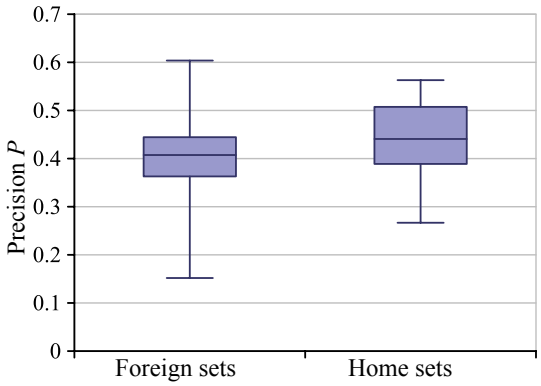


Figure 8.16: Results for S_{b+e} for foreign and home sets.

8.4.5 Influence of the Selection Task

In the experiment, the participants were first asked to create a "selection for their private photo collection" (Task 1). Subsequently, they were asked to perform further selections for the task: "Select photos for giving your friends or family a detailed summary of the event" (Task 2) as well as "Select the most beautiful photos for presenting them on the web, for example, on Flickr" (Task 3). The participants created the selections in Task 2 and Task 3 only for the photo sets of personal interest (the "home sets"), which were taken during the event they participated in.

Precision results of the selections under each task (Tasks 1 to 3) for investigating the performance of the gaze selections in the context of different applications. The results are shown in Figure 8.17. The average precision results for the 18 participants that took part in this part of the experiment are $P = 0.456$ for Task 1, $P = 0.432$ for Task 2, and $P = 0.415$ for Task 3. A Friedman test revealed no statistical significance between the three tasks with $\alpha < 0.05$ for P, $\chi^2(2) = 0.778, p = .678$, $n = 18$.

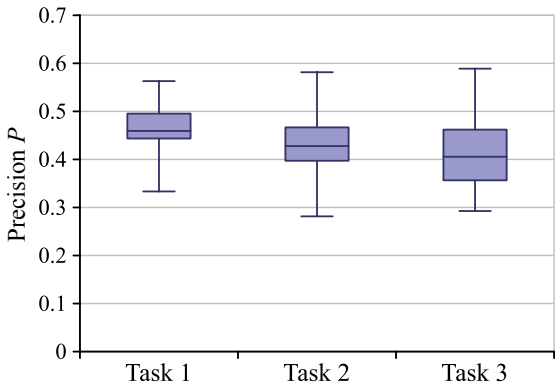


Figure 8.17: Results for S_{b+e} for different selection tasks.

8.5 Conclusion

In this section individual photo selections were created by means of gaze analysis. It was shown that users created highly individual photo selections based on very individual selection criteria in the experiment. From the analysis of the selection criteria, it can be concluded that the criteria judged as most important by the users are rather subjective. At the same time, the more objective criteria which can at least theoretically be calculated by algorithms, such as the number of faces depicted or the sharpness of a photo, are less important to most users. In addition, the manually created selections are very diverse; only few photos were selected by most of the users. Thus, there is no "universal" selection that fits the preferences of all users.

Previous attempts to automatically select photos solely based on content information and context information are not sufficient. Rather, a system supporting users in automatic photo selections by applying eye tracking data significantly outperformed these approach by 17%. Considering only photo sets that were of personal interest, the improvement increased to 22% over the baseline approach. Thus, the approach performed better for photos that are personally related to the user viewing them. The overall best selection result with a mean precision of 0.428 were obtained when combining all measures (content, context, and gaze) by machine learning. It is noteworthy that a single eye tracking measure already delivered competitive results with a mean precision of 0.421 without any machine learning.

In the experiment application, users viewed sets of nine photos and navigated through the sets by clicking on a "Next" button to avoid scrolling. This viewing behavior is different from real life photo viewing, where it is more likely that photos are viewed in a file viewer environment or in full screen mode. It could be that the analysis of viewing behavior in these settings has to be adapted. Bias effects such as the concentration on the first photo of a page would be necessary to be considered.

The results strongly vary between users and between different partitions of the data into training set and test set for the machine learning. It is possible that this effect depends on the users and their individual viewing behavior or on the characteristics of the viewed photo sets. For example, for sets including many interesting and good photos the viewing behavior is less obvious because it is likely that several photo are intensively fixated, and it is more difficult to create a selection.

Automatic approaches, even when including gaze data, may probably be not sufficient for a "perfect photo selection," because of the complexity of human decision processes. The decision on how much support a gaze-based system should offer has to be made by the user. Assistance in the creation of selections by suggesting photos is an option as well as applications that fully automatically create photo selections for the user without additional interaction. One participant in the study concluded: "Dealing with only half of the photos of a collection would already be an improvement."

The viewing durations and the selection durations were longer for photo sets of personal interest. At the same time, the ratings from the questionnaire showed that the selection was rated as being less difficult for the photos of personal interest. This indicates that on the one hand, users like viewing photos of personal interest but on the other hand, the selection process seems to be even more time-consuming for these sets. The suggested approach delivers significantly better results for photo collections of personal interest than for photo sets of less personal interest. With other words, the prediction of the photo selections performs better when the photos' content is personally related to the users. This suggests that the proposed approach works even better in real life with users viewing photos of strong personal interest, e. g., one's wedding, summer vacation, or a family gathering, compared with the data set in this experiment, which is taken from a working group situation. Finally, the results for different manual selections created under different selection tasks were compared. The obtained results are about the same. This result indicates that

the information gained from eye movements can be useful in diverse scenarios where photo selections are needed.

Based on the results, others features such as photo cropping based on gaze data [SAD+06] may be integrated into future research. The findings may be implemented in authoring tools such as miCollage [XZC+08] to enhance an automatic photo selection for creating multimedia collections.

Chapter 9

Conclusions

Eye tracking data delivers information on where humans fixate their gaze during the visual perception process. This information is very unique and allows an insight into the human attention. In this thesis, eye tracking data was used for deriving information on the viewed stimulus, here digital images. The gained information can support users in the management of photo collections. The new terminology of "exploitative eye tracking applications" was introduced for describing the approach of exploiting the visual attention in the annotation of photos while the user is performing other tasks.

One research direction was the labeling of image region, described in Sections 4 to 7. The aim of this part of the research was to assign object names to image regions at pixel level for describing the depicted scene. In three experiments with a total of 100 unique participants and 108 test executions, it was demonstrated that in diverse contexts, the gaze-based labeling outperformed baseline approaches. The labeling was performed by assigning names to objects in a controlled classification experiment, by assigning search terms to image regions in an image search scenario and by assigning object categories to object regions in a classification game. The most important contributions of this part of the work are as follows:

- Gaze data can significantly improve the assignments of tags to regions compared with baseline approaches. This applies to different scenarios such as image search or game playing.

- The analysis of gaze data aggregated from several users delivers better results compared with analysis based on the data of single users.

- The possible inaccuracy of gaze data can be alleviated by considering the surrounding of a fixation as possible fixation target (region extension).

The second line of research investigates the usage of visual interest as a selection criterion in the automatic creation of photo selections, described in Section 8. An experiment with 33 participants showed that the gaze data made a valuable contribution to an automatic selection process. The most important contributions in this part of the work are as follows:

- Gaze-based photo selection performs better than baseline approaches.

- Gaze-based selections perform even better for photo collections of personal interest.

The presented approach of using gaze data in the generation of information results has some considerable advantages. First of all, no additional effort is required for the users. During their usual work with images in everyday life — like the viewing of photos or the search for photos — they are fixating photos. This information is usually unused but can be recorded and exploited by the presented eye tracking approach. The application of machine learning techniques delivers only small improvements compared with simple gaze-analysis approaches. Thus, even without training data and training period, gaze data can be used and information can directly be gained from it. Usually information on photos are derived from the analysis of the pixel information (content) or the surrounding information (context). The presented approach provides an additional source of information, which adds a new dimension of information to a photo. For example, in the labeling process, regions can be labeled as depicting the same objects, even if the visual features of these regions are very different. In the selection process, photos with visual features that usually indicate that a photo is of low quality (e.g., a blurry photo or bad lighting condition) can be selected because a user is interested in it.

9.1 Lessons Learned

The main challenge in the analysis of gaze data is that on the one hand, the data can be inaccurate because of technical limitations. On the other hand, the gaze paths are not strictly focused. They can be rather spread over a stimulus because of anatomical impossibility to fixate the eyes at one static point for a longer time. In addition, the gaze paths also included the scanning process. The information on which areas are interesting in a photo when looking for a specific object gets much more stable when the data of several user is aggregated and eye tracking measures are applied to the aggregated data. In Section 4, it has been shown that the aggregation led to an improvement of 152% compared with non aggregated data. The Internet and the big number of users who share and view photos support the aggregation of gaze paths. The inaccuracy of gaze data that usually occurs in the dimension of a few pixels can be handled by assigning a fixation to all objects in a certain radius around the fixation itself. This region extension was used in the gaze analysis performed in the context of this work.

The experiments have shown that the success of the gaze-based labeling and selection approach strongly depends on the viewed stimuli. The random measures and baseline measures in this work showed the level of difficulty of the labeling and selection tasks. The comparison of the results for these measures and the gaze-based measures was analyzed and interpreted but the comparison of results for different data sets from different experiments was difficult because of the different characteristics of the data, that is, photos and object names.

The variances between the users were also strong in some cases. This is caused by different personalities (e.g., is a decision made fast or does a user scan a photo several times before deciding?) and diverse other factors (e.g., is a participant looking for an object that is randomly fixated very fast?). A detailed analysis of these factors was not possible within the scope because of the big diversity and the high number of these possible factors. The goal was not to consider these details but to evaluate the potential of gaze data over all users and randomly selected stimuli.

During the conduction of the experiments, it became apparent that many participants were fascinated by eye tracking technology as an unusual input device. The technical limitations such as a limited freedom of movements or potential problems caused by the lighting situation did not have a strong influence on the wellbeing of the participants as questionnaires and observations during the experiments showed. The fascination of eye tracking is not only based on the unknown technology but also on the fact that the user does not have the feeling to explicitly control a software (because the hands as most important body part when usually controlling a computer does not have to be moved).

9.2 Outlook

Some potential future research directions arise from the work done in this thesis.

The steps forward in the development of eye tracking hardware are big, and state-of-the-art open-source solutions are developing rapidly (see 2.2.2). One interesting next step in the research on exploitative usage of gaze data is to use devices based on low-cost hardware or even hardware already integrated in devices such as webcams, for the recording of the eye movements. Another step further is the stronger use of image content information. This combination can be promising, as the information gained from gaze and from low-level features are very different and can potentially augment the information gain.

As mentioned before, one advantage of the introduced method of gaze-based region labeling is that the visual appearance of an object is not of importance and even unusual objects can be labeled as long as they can be identified by the users. An extension of this advantage is the assignment of more abstract concepts such as “speed” and “love,” which can be represented by various objects and depicted scenes. This kind of labels are extremely challenging for automatic labeling algorithms as “emotional semantics of an image lies on the highest level of the abstraction”[JHL11].

The possibility to identify personal preferences from gaze information can also be adapted to other domains. One application could be the recommendation of products based on previous fixations on photos or objects in photos. An example would be the suggestion of products with specific characteristics to user who fixated on these characteristics before (e.g., specific carrying straps of hand bags). In social media content, it could be possible to identify persons that are important to the user and provide specific information about these users. Frohlich et al. [FWK13] showed that the forgetting of photos is becoming a serious problem because of the size of digital photo collections. Gaze-based

selections could support users in refinding photos they viewed in the past. An application could be a viewing history, which highlights the mostly viewed photos or photo parts. A browsing history as a browser extension could also support users in organizing photos they have seen before. The eye tracking data delivers more information than which photos are interesting. For example, the information on which parts are interesting. This information could be used in the creation of slide shows and photo books, where photos not only have to be selected but also combined, cropped, or scaled down. The interest in photo book creations is still increasing.

When including gaze support to applications, it could be important to allow the users to decide how much support he/she wishes. For the photo selection, an application could provide a ranked list, facilitating the selection but avoiding a dictation by the software on which photos are important to the user.

I hope that the results of my research can support users in their photo tasks and management and that it allows them to find more meaningful photos during the image search and to spend more time on the pleasurable aspects of viewing photos and creating photo products such as slide shows or collages.

Bibliography

[ABD06] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM, 2006.

[AMFM11] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, May 2011.

[Bar04] Jeffrey M Bartelma. *Flycatcher: Fusion of gaze with hierarchical image segmentation for robust object detection.* PhD thesis, Massachusetts Institute of Technology, 2004.

[BBD10] R. Biedert, G. Buscher, and A. Dengel. The eyebook - using eye tracking to enhance the reading experience. *Informatik-Spektrum*, 33(3):272–281, June 2010.

[BBS+10a] R. Biedert, G. Buscher, S. Schwarz, J. Hees, and A. Dengel. Text 2.0. In *CHI '10 extended abstracts on human factors in computing systems*, pages 4003–4008, New York, NY, USA, 2010. ACM Press.

[BBS+10b] R. Biedert, G. Buscher, S. Schwarz, M. Möller, A. Dengel, and T. Lottermann. The text 2.0 framework – writing web-based gaze-controlled realtime applications quickly and easily. In *Proceedings of the International Workshop on Eye Gaze in Intelligent Human Machine Interaction (EGIHMI) held in conjunction with IUI 2010*, 2010.

[BDC10] Georg Buscher, Susan T Dumais, and Edward Cutrell. The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 2010.

[BDEM08] G. Buscher, A. Dengel, L. Elst, and F. Mittag. Generating and using gaze-based document annotations. In *CHI'08 extended abstracts on Human factors in computing systems*, pages 3045–3050, 2008.

[BDvE08] Georg Buscher, Andreas Dengel, and Ludger van Elst. Eye movements as implicit relevance feedback. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, pages 2991–2996. ACM, 2008.

[BMEL08] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, 2008.

[Boj09] Agnieszka Aga Bojko. Informative or misleading? heatmaps deconstructed. In *Human-Computer Interaction. New Trends*, pages 30–39. Springer, 2009.

[BR11] Andreas Bulling and Daniel Roggen. Recognition of visual memory recall processes using eye movement analysis. In *UbiComp*, pages 455–464, 2011.

[BSM02] D. Bruneau, M.A. Sasse, and J.D. McCarthy. The eyes never lie: The use of eye tracking data in HCI research. In *Proceedings of the CHI*, volume 2, 2002.

[BSST07] Susanne Boll, Philipp Sandhaus, Ansgar Scherp, and Sabine Thieme. Metaxa - context- and content-driven metadata enhancement for personal photo books. In *Advances in Multimedia Modeling, 13th International Multimedia Modeling Conference, MMM 2007, Singapore, January 9-12, 2007. Proceedings, Part I*, pages 332–343. Springer, 2007.

[BWG13] Andreas Bulling, Christian Weichel, and Hans Gellersen. Eyecontext: recognition of high-level contextual cues from human visual behaviour. In *CHI*, pages 305–308, 2013.

[BWGT09] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Tröster. Eye movement analysis for activity recognition. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 41–50. ACM, 2009.

[CF01] Richard J. Campbell and Patrick J. Flynn. A survey of free-form object representation and recognition techniques. *CVIU*, 81(2):166–210, 2001.

[CG07] Edward Cutrell and Zhiwei Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 407–416. ACM, 2007.

[CJP10] S. Castagnos, N. Jones, and P. Pu. Eye-tracking product recommenders' usage. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 29–36. ACM, 2010.

[CL04] MG Calvo and PJ Lang. Gaze Patterns When Looking at Emotional Pictures : Motivationally Biased Attention. *Motivation and Emotion*, 28(3), 2004.

[CL08] Wei-Ta Chu and Chia-Hung Lin. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection. *ACM Multimedia*, page 829, 2008.

[CLWB01] Mark Claypool, Phong Le, Makoto Wased, and David Brown. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pages 33–40. ACM, 2001.

[CZ07] Ondrej Chum and Andrew Zisserman. An exemplar model for learning object classes. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[DBDFF06] P. Duygulu, K. Barnard, J. De Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Computer Vision–ECCV 2002*, pages 349–354, 2006.

[DP11] Geoffrey B Duggan and Stephen J Payne. Skim reading by satisficing: evidence from eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1141–1150. ACM, 2011.

[DS02] Doug DeCarlo and Anthony Santella. Stylization and abstraction of photographs. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 769–776. ACM, 2002.

[DS07] Heiko Drewes and Albrecht Schmidt. Interacting with the computer using gaze gestures. In *Human-Computer Interaction–INTERACT 2007*, pages 475–488. Springer, 2007.

[Duc07] A.T. Duchowski. *Eye tracking methodology: Theory and practice*. Springer-Verlag New York Inc, 2007.

[DVSC05] C. Dickie, R. Vertegaal, C. Sohn, and D. Cheng. eyeLook: using attention to facilitate mobile media consumption. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*, page 106. ACM, 2005.

[ERK08] W. Einhäuser, U. Rutishauser, and C. Koch. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 8(2), 2008.

[Ess08] Kai Essig. Vision-based image retrieval (vbir)-a new approach for natural and intuitive image retrieval. 2008.

[EVGW+] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[FCH+08] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[FK13] Tom Foulsham and Alan Kingstone. Optimal and preferred eye landing positions in objects and scenes. *The Quarterly Journal of Experimental Psychology*, pages 1–22, 2013.

[FKP+02] David Frohlich, Allan Kuchinsky, Celine Pering, Abbe Don, and Steven Ariss. Requirements for photoware. In *Computer supported cooperative work*, pages 166–175. ACM, 2002.

[Fre07] M. Freeman. *The Photographer's Eye: Composition and Design for Better Digital Photos*. Focal Press, 2007.

[FWK13] David M. Frohlich, Steven Wall, and Graham Kiddle. Rediscovery of forgotten images in domestic photo collections. *Personal and Ubiquitous Computing*, 17(4):729–740, 2013.

[GGVG11] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1529–1536. IEEE, 2011.

[GNDVG13] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool. Visual interestingness in image sequences. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 1017–1026. ACM, 2013.

[Gol13] E Bruce Goldstein. *Sensation and perception*. Cengage Learning, 2013.

[GSL+02] Joseph H Goldberg, Mark J Stimson, Marion Lewenstein, Neil Scott, and Anna M Wichansky. Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 51–58. ACM, 2002.

[HBCM07] J.M. Henderson, J.R. Brockmole, M.S. Castelhano, and M. Mack. Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye movements: A window on mind and brain*, pages 537–562, 2007.

[HC05] Anthony J Hornof and Anna Cavender. Eyedraw: enabling children with severe motor impairments to draw with their eyes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 161–170. ACM, 2005.

[HCC09] T.H. Huang, K.Y. Cheng, and Y.Y. Chuang. A collaborative benchmark for region of interest detection algorithms. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 296–303. IEEE, 2009.

[Hei13] Henna Heikkilä. Tools for a gaze-controlled drawing application–comparing gaze gestures against dwell buttons. In *Human-Computer Interaction–INTERACT 2013*, pages 187–201. Springer, 2013.

[HI10] S.N. Hajimirza and E. Izquierdo. Gaze movement inference for implicit image annotation. In *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2010.

[HMPI11] S Navid H Haji Mirza, Michael Proulx, and Ebroul Izquierdo. Gaze movement inference for user adapted image annotation and retrieval. In *Proceedings of the 2011 ACM workshop on Social and behavioural networked media access*, pages 27–32. ACM, 2011.

[HNA+11] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[HPI12] S Navid Hajimirza, Michael J Proulx, and Ebroul Izquierdo. Reading users' minds from their eyes: A method for implicit image annotation. *Multimedia, IEEE Transactions on*, 14(3):805–815, 2012.

[HPS10] D. Hardoon, K. Pasupa, and S. Szedmak. Image ranking with implicit feedback from eye movements. In *In Proceedings of the 6th Biennial Symposium on Eye Tracking Research & Applications (ETRA'2010)*, 2010.

[IK00] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

[IK01] L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[Itt03] L. Itti. Modeling primate visual attention. In J. Feng, editor, *Computational Neuroscience: A Comprehensive Approach*, pages 635–655. CRC Press, Boca Raton, 2003.

[Jac91] Robert JK Jacob. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems (TOIS)*, 9(2):152–169, 1991.

[Jai01] Alejandro Jaimes. Using human observer eye movements in automatic image classifiers. *SPIE*, 2001.

[JEDT09] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*. Citeseer, 2009.

[JGP+05] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. ACM, 2005.

[JHL11] Jin-Woo Jeong, Hyun-Ki Hong, and Dong-Ho Lee. Exploiting of flickr note and its applications for social image sharing and search. In *Multimedia (ISM), 2011 IEEE International Symposium on*, pages 165–170. IEEE, 2011.

[JHW07] Seikyung Jung, Jonathan L Herlocker, and Janet Webster. Click data as implicit relevance feedback in web search. *Information Processing & Management*, 43(3):791–807, 2007.

[JNTD06] Alexandar Jaffe, Mor Naaman, Tamir Tassa, and Marc Davis. Generating summaries and visualization for large collections of geo-referenced photographs. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval - MIR '06*, page 89, 2006.

[KKK09] L. Kozma, A. Klami, and S. Kaski. GaZIR: gaze-based zooming interface for image retrieval. In *Multimodal interfaces*. ACM, 2009.

[Kla10] A. Klami. Inferring task-relevant image regions from gaze data. In *Workshop on Machine Learning for Signal Processing*. IEEE, 2010.

[KPW07] Manu Kumar, Andreas Paepcke, and Terry Winograd. Eyepoint: practical pointing and selection using gaze and keyboard. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430. ACM, 2007.

[KSDK08] A. Klami, C. Saunders, T.E. De Campos, and S. Kaski. Can relevance of images be inferred from eye movements? In *Multimedia information retrieval*. ACM, 2008.

[KSRW06] David Kirk, Abigail Sellen, Carsten Rother, and Ken Wood. Understanding photowork. In *CHI*, pages 761–770. ACM, 2006.

[KTS01] Ioannis Kompatsiaris, Evangelia Triantafyllou, and Michael G Strintzis. A world wide web region-based image search engine. In *Image Analysis and Processing, 2001. Proceedings. 11th International Conference on*, pages 392–397. IEEE, 2001.

[KUS+13] Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. I know what you are reading: recognition of document types using mobile eye tracking. In *Proceedings of the 17th annual*

international symposium on International symposium on wearable computers, pages 113–116. ACM, 2013.

[KY08] D.H. Kim and S.H. Yu. A new region filtering and region weighting approach to relevance feedback in content-based image retrieval. *Journal of Systems and Software*, 81(9):1525–1538, 2008.

[LAGA14] Dmitry Lagun, Mikhail Ageev, Qi Guo, and Eugene Agichtein. Discovering common motifs in cursor movement data for improving web search. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 183–192. ACM, 2014.

[LBS07] S. Laqua, S.U. Bandara, and M.A. Sasse. GazeSpace: eye gaze controlled content spaces. In *Proceedings of the 21st British HCI Group Annual Conference on HCI 2008: People and Computers XXI: HCI... but not as we know it-Volume 2*, pages 55–58. British Computer Society, 2007.

[LCY+09] Xiaobai Liu, Bin Cheng, Shuicheng Yan, Jinhui Tang, Tat Seng Chua, and Hai Jin. Label to region by bi-layer sparsity priors. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 115–124. ACM, 2009.

[LJCM13] Kristian Lukander, Sharman Jagadeesan, Huageng Chi, and Kiti Müller. Omg!: A new robust, wearable and affordable open source mobile gaze tracker. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, pages 408–411. ACM, 2013.

[LLLL12] Yu-Tzu Lin, Ruei-Yan Lin, Yu-Chih Lin, and Greg C. Lee. Real-time eye-gaze estimation using a low-resolution webcam. *Multimedia Tools and Applications*, 65(3):543–568, August 2012.

[LLT03] Jun Li, Joo Hwee Lim, and Qi Tian. Automatic summarization for personal digital photos. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 3, pages 1536–1540. IEEE, 2003.

[LSW09] Xirong Li, Cees G. M. Snoek, and Marcel Worring. Annotating images by harnessing worldwide user-tagged photos. In *Acoustics, Speech, and Signal Processing*, pages 3717–3720. IEEE, 2009.

[LYL+10] Xiaobai Liu, Shuicheng Yan, Jiebo Luo, Jinhui Tang, Zhongyang Huango, and Hai Jin. Nonparametric label-to-region by search. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3320–3327. IEEE, 2010.

[MAŠ09] P. Majaranta, U.K. Ahola, and O. Špakov. Fast gaze typing with an adjustable dwell time. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 357–360. ACM, 2009.

[MGA14] Pascual Martínez-Gómez and Akiko Aizawa. Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 95–104. ACM, 2014.

[Mil03] Slavko Milekic. The more you look the more you get: Intention-based interface using gaze-tracking. In *Museums and the Web*, pages 57–72, 2003.

[MNST12] Mari-Carmen Marcos, David F Nettleton, and Diego Sáez-Trumper. A user study of web search session behaviour using eye tracking data. In *Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers*, pages 262–267. British Computer Society, 2012.

[MR98] Arien Mack and Irvin Rock. *Inattentional blindness*. The MIT Press, 1998.

[MSH08] E. Mollenbach, T. Stefansson, and J.P. Hansen. All eyes on the monitor: gaze based interaction in zoomable, multi-scaled information-spaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 373–376, Gran Canaria, Spain, 2008. ACM.

[NDFY12] Yuzhao Ni, Jian Dong, Jiashi Feng, and Shuicheng Yan. Purposive Hidden-Object-Game: Embedding Human Computation in Popular Game. *IEEE Transactions on Multimedia*, 14(5):1496–1507, October 2012.

[NF09] Carman Neustaedter and Elena Fedorovskaya. Understanding and improving flow in digital photo ecosystems. In *Proceedings of Graphics Interface 2009*, pages 191–198. Canadian Information Processing Society, 2009.

[NH10] Antje Nuthmann and John M Henderson. Object-based attentional selection in scene viewing. *Journal of vision*, 10(8), 2010.

[NI05] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005.

[NI10] Y.I. Nakano and R. Ishii. men conversations. *Proceeding of the 14th international conference on Intelligent user interfaces*, pages 139–148, 2010.

[NYGMP05] Mor Naaman, Ron B. Yeh, Hector Garcia-Molina, and Andreas Paepcke. Leveraging context to resolve identity in photo albums. In Mary Marlino, Tamara Sumner, and Frank M. Shipman III, editors, *JCDL*, pages 178–187. ACM, 2005.

[Ols07] Pontus Olsson. Real-time and Offline Filters for Eye Tracking. *KTH Royal Institute of Technology*, (Msc thesis), 2007.

[PAD13] G. Papadopoulos, K. Apostolakis, and P. Daras. Gaze-based relevance feedback for realizing region-based image retrieval. *Multimedia, IEEE Transactions on*, PP(99):1–1, 2013.

[PHG+04] Bing Pan, Helene A Hembrooke, Geri K Gay, Laura A Granka, Matthew K Feusner, and Jill K Newman. The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, pages 147–154. ACM, 2004.

[PHK+13] Felix Putze, Jutta Hild, Rainer Kärgel, Christian Herff, Alexander Redmann, Jürgen Beyerer, and Tanja Schultz. Locating user attention using eye tracking and eeg for spatio-temporal event selection. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 129–136. ACM, 2013.

[PIKI05] Robert J Peters, Asha Iyer, Christof Koch, and Laurent Itti. Components of bottom-up gaze allocation in natural scenes. *Journal of Vision*, 5(8):692–692, 2005.

[Pla00] John C Platt. Autoalbum: Clustering digital photographs using probabilistic model merging. In *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on*, pages 96–100. IEEE, 2000.

[Por08] M. Porta. Implementing eye-based user-aware e-learning. *CHI'08 extended abstracts on Human factors in computing systems*, pages 3087–3092, 2008.

[PS00] C. M Privitera and L. W Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on pattern analysis and machine intelligence*, 22(9):970–982, 2000.

[PS03] T. Partala and V. Surakka. Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, 59(1-2):185–198, 2003.

[PSS+09] K. Pasupa, C. Saunders, S. Szedmak, A. Klami, S. Kaski, and S. Gunn. Learning to rank images from eye movements. In *Workshops on Human-Computer Interaction*, 2009.

[PTP06] M. Pivec, C. Trummer, and J. Pripfl. Eye-Tracking Adaptable e-Learning and Content Authoring Support. *Special Issue: Hot Topics in European Agent*, 3:83–86, 2006.

[QT09] A Quattoni and A Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009.

[QZ05] P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–230. ACM, 2005.

[RKH+09] Subramanian Ramanathan, Harish Katti, Raymond Huang, Tat-Seng Chua, and Mohan Kankanhalli. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In *Multimedia*, New York, New York, USA, 2009. ACM.

[RKS+10] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.S. Chua. An eye fixation database for saliency detection in images. *Computer Vision–ECCV 2010*, pages 30–43, 2010.

[RO12] Kari-Jouko Räihä and Saila Ovaska. An exploratory study of eye typing fundamentals: dwell time, text entry rate, errors, and workload. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 3001–3010. ACM, 2012.

[Rod99] Kerry Rodden. How do people organise their photographs?. In *BCS-IRSG Annual Colloquium on IR Research*. Citeseer, 1999.

[Row02] N.C. Rowe. Finding and labeling the subject of a captioned depictive natural photograph. *IEEE Transactions on Knowledge and Data Engineering*, pages 202–207, 2002.

[RSB11] Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. Automatic creation of photo books from stories in social media. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 7(1):27, 2011.

[RTMF08] B. C Russell, A. Torralba, K. P Murphy, and W. T Freeman. LabelMe: a database and web-based tool for image annotation. *J. of Comp. Vision*, 77(1):157–173, 2008.

[RW03] K. Rodden and K.R. Wood. How do people manage their digital photographs? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 409–416. ACM, 2003.

[SAD+06] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *ACM Human Factors in Computing Systems (CHI)*, pages 771–780. ACM, 2006.

[SASHH09] J. San Agustin, H. Skovsgaard, J.P. Hansen, and D.W. Hansen. Low-cost gaze interaction: ready to deliver the promises. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 4453–4458. ACM, 2009.

[SB11] Philipp Sandhaus and Susanne Boll. Social aspects of photobooks: Improving photobook authoring from large-scale multimedia analysis. In *Social Media Modeling and Computing*, pages 257–277. Springer, 2011.

[SBL+09] Line Sæther, Werner Van Belle, Bruno Laeng, Tim Brennen, and Morten Øvervoll. Anchoring gaze when categorizing faces' sex: evidence from eye-tracking data. *Vision research*, 49(23):2870–2880, 2009.

[SC99] Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception-London*, 28(9):1059–1074, 1999.

[SCGiN+13] Amaia Salvador, Axel Carlier, Xavier Giro-i Nieto, Oge Marques, and Vincent Charvillat. Crowdsourced object segmentation with a game. In *Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, pages 15–20, New York, NY, USA, 2013. ACM.

[SD02] Anthony Santella and Doug DeCarlo. Abstracted painterly renderings using eye-tracking data. In *Proceedings of the 2nd international symposium on Non-photorealistic animation and rendering*, pages 75–ff. ACM, 2002.

[SEL00] Andreas E Savakis, Stephen P Etz, and Alexander CP Loui. Evaluation of image appeal in consumer photography. In *Electronic Imaging*, pages 111–120. SPIE, 2000.

[SF03] Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1):77–123, 2003.

[SG06] J David Smith and TC Graham. Use of eye movements for video game control. In *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*, page 20. ACM, 2006.

[SJ00] Linda E Sibert and Robert JK Jacob. Evaluation of eye gaze interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–288. ACM, 2000.

[SJ11] P Sinha and R Jain. Extractive summarization of personal photos from life events. *Multimedia and Expo (ICME), 2011 IEEE*, 2011.

[SK00] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 746–751. IEEE, 2000.

[SK10] W. Sewell and O. Komogortsev. Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In *Proceedings of the 28th of the international conference extended*

abstracts on Human factors in computing systems, pages 3739–3744. ACM, 2010.

[SKP13] Mohammad Soleymani, Sebastian Kaltwang, and Maja Pantic. Human behavior sensing for tag relevance assessment. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 657–660. ACM, 2013.

[SKSK03] Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel Kaski. Can relevance be inferred from eye movements in information retrieval. In *Proceedings of WSOM*, volume 3, pages 261–266, 2003.

[SMJ11] Pinaki Sinha, Sharad Mehrotra, and Ramesh Jain. Summarization of personal photologs using multidimensional content and context. *Multimedia Retrieval*, pages 1–8, 2011.

[SPK05] J. Salojärvi, K. Puolamäki, and S. Kaski. Implicit relevance feedback from eye movements. *Artificial Neural Networks: Biological Inspirations–ICANN 2005*, pages 513–518, 2005.

[STM13] Tomoya Sawada, Masahiro Toyoura, and Xiaoyang Mao. Film comic generation with eye tracking. In *Advances in Multimedia Modeling*, pages 467–478. Springer, 2013.

[SWS+00] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[TBHS03] J. Triesch, D.H. Ballard, M.M. Hayhoe, and B.T. Sullivan. What you see is what you need. *Journal of Vision*, 3(1), 2003.

[TG80] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[TJL+11] D. Tsai, Y. Jing, Y. Liu, H.A. Rowley, S. Ioffe, and J.M. Rehg. Large-scale image annotation using visual synset. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 611–618. IEEE, 2011.

[tob10] Tobii studio 2.x - user manual, 2010. http://www.tobii.com.

[TOCH06] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review*, 113(4):766–786, 2006.

[TS08] Dian Tjondronegoro and Amanda Spink. Web search engine multimedia functionality. *Inf. Process. Manage.*, 44(1):340–357, 2008.

[TSG+] Dereck Toker, Ben Steichen, Matthew Gingerich, Cristina Conati, and Giuseppe Carenini. Towards facilitating user skill acquisition-identifying untrained visualization users through eye tracking. *System*, 16:19.

[TYH+09] J. Tang, S. Yan, R. Hong, G.J. Qi, and T.S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 223–232. ACM, 2009.

[VAD04] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.

[vALB06] Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: a game for locating objects in images. In *CHI*. ACM, 2006.

[Ver02] Roel Vertegaal. Designing attentive interfaces. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 23–30. ACM, 2002.

[VJ01] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[WCM05] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1800–1807. IEEE, 2005.

[WNS12] Tina Walber, Chantal Neuhaus, and Ansgar Scherp. Eyegrab: A gaze-based game with a purpose to enrich image context information [poster]. In *EuroHCIR - Workshop on Human-Computer Interaction and Information Retrieval*, 2012.

[WNS+13] Tina Walber, C. Neuhaus, Steffen Staab, Ansgar Scherp, and Ramesh Jain. Creation of individual photo selections: read preferences from the users' eyes. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 629–632. ACM, 2013.

[WNS14] Tina Walber, Chantal Neuhaus, and Ansgar Scherp. Tagging-by-search: automatic image region labeling using gaze information obtained from image search. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 257–266. ACM, 2014.

[WSS12] Tina Walber, Ansgar Scherp, and Steffen Staab. Identifying objects in images from analyzing the users' gaze movements for provided tags. In *Advances in Multimedia Modeling*, pages 138–148. Springer, 2012.

[WSS13a] Tina Walber, Ansgar Scherp, and Steffen Staab. Benefiting from users' gaze: selection of image regions from eye tracking information for provided tags. *Multimedia Tools and Applications*, pages 1–28, 2013.

[WSS13b] Tina Walber, Ansgar Scherp, and Steffen Staab. Can you see it? two novel eye-tracking-based measures for assigning tags to image regions. In *Advances in Multimedia Modeling*, pages 36–46. Springer, 2013.

[WSS14a] Tina Walber, Ansgar Scherp, and Steffen Staab. Exploitation of gaze data for photo region labeling in an immersive environment. In *MultiMedia Modeling*, pages 424–435. Springer, 2014.

[WSS14b] Tina Walber, Ansgar Scherp, and Steffen Staab. Smart photo selection: Interpret gaze as personal interest. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2065–2074, New York, NY, USA, 2014. ACM.

[XJL09] Songhua Xu, Hao Jiang, and Francis Lau. User-oriented document summarization through vision-based eye-tracking. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 7–16. ACM, 2009.

[XZC+08] Jun Xiao, Xuemei Zhang, Phil Cheatle, Yuli Gao, and C Brian Atkins. Mixed-initiative photo collage authoring. In *ACM Multimedia*, pages 509–518. ACM, 2008.

[Yar67] A.L. Yarbus. *Eye movements and vision*. Plenum press, 1967.

[YLZ07] Jinhui Yuan, Jianmin Li, and Bo Zhang. Exploiting spatial context constraints for automatic image region annotation. In *Proceedings of the 15th international conference on Multimedia*, pages 595–604. ACM, 2007.

[YMNL13] Ting Yao, Tao Mei, Chong-Wah Ngo, and Shipeng Li. Annotation for free: Video tagging by mining user search behavior. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 977–986. ACM, 2013.

[ZGS+10] Bangzuo Zhang, Yu Guan, Haichao Sun, Qingchao Liu, and Jun Kong. Survey of user behaviors as implicit feedback. In *Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference on*, volume 6, pages 345–348. IEEE, 2010.

[ZK11] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3), 2011.

[ZMI99] Shumin Zhai, Carlos Morimoto, and Steven Ihde. Manual and gaze input cascaded (magic) pointing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 246–253. ACM, 1999.

[ZTL+06] Ming Zhao, Yong Wei Teo, Siliang Liu, Tat-Seng Chua, and Ramesh Jain. Automatic person annotation of family photo album. In *Image and Video Retrieval*, pages 163–172. Springer, 2006.

Appendices

A.1 Nomenclatures

Eye tracker	A device for recording the human eye movements and for calculating fixated points on a computer screen
POR	Point of regard
Fixation	Part of a gaze path, when humans gaze at a stationary point
Saccade	Fast eye movements between the fixations
Heat map	Graphical representation of gaze data in a matrix with highlighting of intensively fixated areas
Gaze path	Graphical representation of fixations (as circles) and saccades (as lines connecting the fixations)
ROI	Region of interest

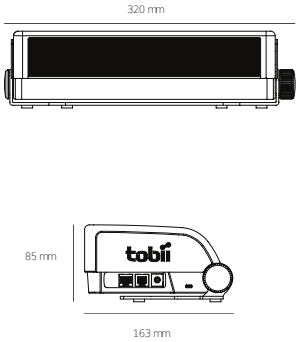
A.2 Glossary of Variables

Variable name	Description
o	Photo
r	Photo region, consists of connected pixels
C	Collection of photos
S	Selection of photos, subset $S \subset C$
$S_{[index]}$	$[index]$ shows how the selection was created (m = manual, b = baseline approach, e = eye tracking data)
P	Precision
G	Gaze path, consists of fixations and saccades
F	Set of fixations
f	Fixation, $f \in G$, $f \in F$
$f_m(r)$	Eye tracking measure on region r
t	Tag, keyword describing the content of a photo
t_r	Tag, describing a region r
s_r	Size of a region r
u	Participant in an experiment
U	All participants in an experiment, $u \in U$

A.3 Product Specification Tobii X Series

Tobii X Series Eye Trackers

Overall dimensions



Models	X60/X120*	X120*
Data rate	60 Hz	120Hz
Accuracy	typical 0.5 degrees	typical 0.5 degrees
Drift	typical 0.1 degrees	typical 0.1 degrees
Spatial resolution	typical 0.2 degrees	typical 0.3 degrees
Head movement error	typical 0.2 degrees	typical 0.2 degrees
Head movement box (width x hight)	44 x 22 cm at 70 cm	30 x 22 cm at 70 cm
Tracking distance	50-80 cm	50-80 cm
Max gaze angles	35 degrees	35 degrees
Top head-motion speed	25 cm/second	25 cm/second
Latency	maximum 33 ms	maximum 33 ms
Blink tracking recovery	maximum 17 ms	maximum 8 ms
Time to tracking recovery	typical 300 ms	typical 300 ms
Weight (excluding case)	~ 3 kg / 7 lbs	
Eye tracking technique	both bright and dark pupil tracking	
Eye tracking server	Embedded	
Screen size	-	
Screen resolution (Max)	-	
Display colors	-	
Vertical sync frequency	-	
Horizontal sync frequency	-	
TFT response time	-	
User camera	-	
Speakers	-	
Connectors	LAN, Power	

Average values over the screen measured at a distance of 63 cm in a controlled off ce environment.

* The Tobii X120 Eye Tracker can be run in 60 or 120 Hz mode.

The Tobii X60 & X120 Eye Trackers allows you to experience how people look at physical objects or scenes.

The Tobii X60 and X120 Eye Trackers are stand-alone eye tracking units designed for eye tracking studies relative to any surface. They enable a variety of stimuli setups such as a TV or other displays, a projection screen or a physical object or scene. They are our most f exible eye trackers, recommended for studies that require particular setups.

Curriculum Vitae

Tina Walber

Education

12/2012 – 03/2013	University of California Irvine, USA	Research stay
1/2006 – 09/2006	Fraunhofer Institute IAIS, Sankt Augustin, Germany	Diploma Thesis *Entwurf und Programmierung eines Gesichts-3D-Scanners*
9/2002 – 4/2003	Ecole des Mines de Nantes, France	Erasmus Scholarship
10/2000 – 09/2006	University of Koblenz-Landau, Germany	Studies of *Computational Visualistics*

Professional Experience

02/2010 – 05/2014	Institute WeST, University of Koblenz-Landau	Scholarship holder and research assistant
06/2008 – 9/2009	Vi-tu AG, Rotkreuz, Switzerland	Software Development, User Interface Design and Usability Analysis
06/2007 – 06/2008	Cromwell Business Resultancy, Rümlang, Switzerland	Consulting, Microsoft Sharepoint Server, Web Development

Teaching

Term	Course
WS 2013/14	Tutorial Algorithms and Data Structures
SS 2013	Practicum Eye Tracking Game “Schau Genau!”
SS 2011	Seminar Eye Tracking
SS 2011	Practicum Eye Tracking Game “EyeGrab”
WS 2010/11	Tutorial Interaktive Multimediasysteme
SS 2010	Tutorial Multimedia Databases

Supervised Theses

Thesis	Student Name and Title	Status
Master Thesis	Matthias Kuich: *Einfluss eines Ausrichtungswerkzeugs und der Kalibrierung auf die Qualität von Eyetrackingdaten in interaktiven Anwendungen*	Ongoing
Bachelor Thesis	Patrick Nitschke: *Verbesserung von Bildsegmentierung anhand von Eyetrackinginformationen*	In review
Bachelor Thesis	Tobias Schmidt: *Analyse von Eyetrackingdaten mit Support Vector Machines, Voting und Logistischer Regression*	In review
Master Thesis	Leon Kastler: *EyeVisionSearch Nutzung von Blickerfassungsgeräten zur Verbesserung der Bedienung von Bildersuchmaschinen*	Grade: 1.1
Master Thesis	Chantal Neuhaus: *EyeSelect — An Approach for Gaze-Based Image Selection from Large Photo Collections*	Grade: 1.0

Publications

Journal

- Tina Walber, Ansgar Scherp, and Steffen Staab. Benefiting from users' gaze: selection of image regions from eye tracking information for provided tags. *Multimedia Tools and Applications*, pages 1–28, 2013.

Conference Publications

- Tina Caroline Walber, Ansgar Scherp, and Steffen Staab. Smart photo selection: Interpret gaze as personal interest. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems,* CHI'14, pages 2065–2074, New York, NY, USA, 2014. ACM.
- Tina Walber, Chantal Neuhaus, and Ansgar Scherp. Tagging-by-search: automatic image region labeling using gaze information obtained from image search. In *Proceedings of the 19th international conference on Intelligent User Interfaces,* pages 257–266. ACM, 2014.
- Tina Walber, Ansgar Scherp, and Steffen Staab. Exploitation of gaze data for photo region labeling in an immersive environment. In *MultiMedia Modeling*, pages 424–435. Springer, 2014.
- Tina Walber, C. Neuhaus, Steffen Staab, Ansgar Scherp, and Ramesh Jain. Creation of individual photo selections: read preferences from the users' eyes. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 629–632. ACM, 2013.
- Tina Walber, Ansgar Scherp, and Steffen Staab. Can you see it? two novel eye-tracking-based measures for assigning tags to image regions. In *Advances in Multimedia Modeling*, pages 36–46. Springer, 2013.
- Tina Walber, Ansgar Scherp, and Steffen Staab. Identifying objects in images from analyzing the users' gaze movements for provided tags. In *Advances in Multimedia Modeling*, pages 138–148. Springer, 2012.

Other Publications

- Tina Walber, Chantal Neuhaus, and Ansgar Scherp. Eyegrab: A gaze-based game with a purpose to enrich image context information [poster]. In *EuroHCIR — Workshop on Human-Computer Interaction and Information Retrieval*, 2012.
- Tina Walber, Annika Wießgügel, and Ansgar Scherp. Image Region Labeling by Gaze Information during Image Search and Image Tagging. [Poster] In *17th European Conference on Eye Movements ECEM*, 2013.
- Tina Walber. Making use of eye tracking information in image collection creation and region annotation. In *Proceedings of the 20th ACM international conference on Multimedia*, Doctoral Symposium, pp. 1405-1408. ACM, 2012.