Multimodal Analysis of Image Search Intent

Intent Recognition in Image Search from User Behavior and Visual Content

Mohammad Soleymani
Swiss Center for Affective Sciences
University of Geneva
Geneva, Switzerland
mohammad.soleymani@unige.ch

Michael Riegler
Simula Research Laboratory and
University of Oslo
Oslo, Norway
michael@simula.no

Pål Halvorsen
Simula Research Laboratory and
University of Oslo
Oslo, Norway
paalh@simula.no

ABSTRACT

Users search for multimedia content with different underlying motivations or intentions. Study of user search intentions is an emerging topic in information retrieval since understanding *why* a user is searching for a content is crucial for satisfying the user's need. In this paper, we aimed at automatically recognizing a user's intent for image search in the early stage of a search session. We designed seven different search scenarios under the intent conditions of *finding* items, *re-finding* items and *entertainment*. We collected facial expressions, physiological responses, eye gaze and implicit user interactions from 51 participants who performed seven different search tasks on a custom-built image retrieval platform. We analyzed the users' spontaneous and explicit reactions under different intent conditions. Finally, we trained machine learning models to predict users' search intentions from the visual content of the visited images, the user interactions and the spontaneous responses. After fusing the visual and user interaction features, our system achieved the F-1 score of 0.722 for classifying three classes in a user-independent cross-validation. We found that eye gaze and implicit user interactions, including mouse movements and keystrokes are the most informative features. Given that the most promising results are obtained by modalities that can be captured unobtrusively and online, the results demonstrate the feasibility of deploying such methods for improving multimedia retrieval platforms.

CCS CONCEPTS

•Information systems → Image search; *Search interfaces;* **•Human-centered computing →** *Laboratory experiments;*

KEYWORDS

Search, multimedia, user interaction, intent, emotion, experiment, eye gaze, facial expression, computer vision

ACM Reference format:
Mohammad Soleymani, Michael Riegler, and Pål Halvorsen. 2017. Multimodal Analysis of Image Search Intent. In *Proceedings of ICMR '17, Bucharest, Romania, June 6–9, 2017,* 9 pages.
DOI: http://dx.doi.org/10.1145/3078971.3078995

ICMR '17, Bucharest, Romania

978-1-4503-4701-3/17/06…$15.00.
DOI: http://dx.doi.org/10.1145/3078971.3078995

1 INTRODUCTION

A multimedia retrieval system that takes users' intent into account can optimize its ranking and visualization methods to better satisfy its users. In this context, users' search intention has a profound effect on the way they interact with an information retrieval system. It has been shown that search tasks are not solely performed with the goal of information search, but also to re-find specific content [36] and for entertainment [13]. The success of multimedia retrieval does not only rely on the relevance of the content to the query and the way the content is framed and depicted is equally important to users, given their search intention. This is more evident for the case of multimedia retrieval in which entertainment is a major motivator behind multimedia search [22]. Understanding *why* a user is embarking on a multimedia search process is not possible with information retrieval systems simply accepting query terms. Therefore, search motivation shall be determined from the implicit indicators and the context.

Due to its nature, intent in multimedia search is different from the classic web search intentions, i.e., *informational*, *navigational* and *transactional* intentions [7, 22]. Moshfeghi *et al.* [30] used the categories of *seeking information*, *re-finding an item* and two *entertainment* categories to adjust arousal and mood. They further attempted to automatically identify video search intentions from users implicit feedback, e.g., mouse movements. They could demonstrate the feasibility of intent recognition from users' implicit feedback. Users implicit feedback and behavioral responses have been also used to automatically determine topical relevance in information retrieval [1].

In this work, we aimed at automatically recognizing search intent early in an image search session. We targeted three scenarios, i.e., *finding* images, *re-finding* images and *entertainment*. Our *finding* category includes *seeking information* and *transactions* since we asked the participants to identify images to be downloaded and used for a certain purpose. The re-find category is similar to the category of *mental image* proposed in [26] in which a user has a specific image in mind that they want to find. Unlike the work in [30], we let the participants decide how they want to entertain themselves through image search, thus our mood and arousal adjustments are mixed. It is also worth noting that it is not as easy to adjust for arousal through images in contrast to videos that are used in [30].

We recruited 51 healthy participants to participate in an experiment performing image search tasks under different intent conditions. We built a custom image search interface using the Flickr[1] API. Given the role of emotions in the search intent process [30],

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

we opted for recording and analyzing facial expressions and Galvanic Skin Response (GSR), in addition to the interactions analyzed in [30]. We also recorded eye gaze and pupil dilation which can capture interaction, attention and arousal [5, 20]. Additionally, we recorded implicit user interactions with the search interface, including key strokes, search logs and mouse movements. We analyzed these modalities and how they vary under different intent conditions. Moreover, we trained models to recognize intent early in the search session (first 30s) from the spontaneous reactions and implicit interactions. We also used the visual content from the visited images to recognize search intent sessions. In summary, our major contributions are as follows.

- We analyzed how facial expressions, eye gaze, queries and implicit user interaction vary by image search intent.
- We analyzed visual content features of the visited images and demonstrate how they can be utilized for search intent recognition.
- We built and evaluated a search intent recognition system through users spontaneous responses and interaction.

Thus, in contrast to the work presented in [30], which reported a user-dependent intent recognition in video search, we performed a user-independent intent recognition for image search. We also analyzed facial expressions, physiological responses, eye gaze, implicit user interactions and visual content of images and evaluated their effectiveness for intent recognition.

Using machine learning models to predict users' intentions from the collected information, our system achieved the average F-1 score of 0.722 for three classes in a user-independent cross-validation. Furthermore, experiments revealed that eye gaze and implicit user interactions are the most informative features. This is very promising because these modalities can be captured online, which demonstrates the potential of using such methods for improving multimedia retrieval platforms.

The remainder of this paper is organized as follows. The previous work is presented in Section 2. The experimental methodology and apparatus are explained in Section 3. Section 4 provides details about the image retrieval system in the experiments. Extracted features and statistical analysis on how they vary by intent conditions are reported in Section 5. Intent recognition and its results are given in Section 6, and the limitations of this work are discussed in Section 7. The work is finally concluded in Section 8.

2 BACKGROUND

There is a general belief in the multimedia community that intent in text search is different from the intent in multimedia search [22]. In addition to relevance, in multimedia search, the content and the presentation of multimedia items are also important to the user, i.e., *why* and for which purpose images and videos are taken, searched for and viewed. User intent in the area of multimedia can in general be divided in the domains of image and video search [22]. In each of the two domains, several models and categories of intent have been defined, which are conceptually similar, but named differently in different publications [14, 17, 24, 26].

In the image retrieval domain, the most important intent categories are *navigation* (find a specific image without knowing the content), *transaction* (find a specific image for further use), *knowledge orientation* (learn something by looking at an image) and *mental image* (know the content of the image beforehand) [26]. The following intent categories are identified in video search: *information* (obtain knowledge and gather new information), *experience learning* (acquire new skills or learn something), *experience exposure* (undergo a specific experience), *affect* (change mood or affective state) and *object* (the video itself as an object) [17]. Based on these intent categories, several researchers tried to perform intent classification from multimedia content. Hanjalic *et al.* [17] classified videos based on shot patterns, speech and metadata into three different categories (*information*, *experience* and *affect*) and reported a weighted F-1 score of 0.833 for the performance of the classification.

Lux et. al. [27] released and discussed an image dataset containing 1,309 images for five intent classes. The dataset and intent annotations for each image were collected in a user study, and for the intent classes, the categories presented in [26] were used. The intent dataset was used in [32] to explore if content-based visual features can be helpful to detect the photographers' intent (why the picture has been taken). This was done by performing unsupervised clustering of the images into five intent classes using global visual features. To evaluate the performance, the correlation between the human agreement with the clusters was compared, which revealed that content-based features can indeed be used to classify images based on the photographers' intent.

Moshfeghi *et al.* [30] studied the emotion and interaction in different video search intent scenarios. They recorded users' implicit feedback for four different intention classes, namely, *seeking information*, *re-finding* a particular information object, *entertainment* by adjusting arousal level, and *entertainment* by adjusting mood. They found significant differences between the emotional experiences of tasks in different intentions. They also found that the task difficulty and certainty perception varied by search intention. They could finally train a model to predict search intent, in a user-dependent cross-validation, with the accuracy of up to 57.29%.

Overall, the reviewed literature provides evidence for the usefulness of multimedia content and users' spontaneous reactions for recognizing intent in multimedia search. Nevertheless, predicting and using a user's intent during a multimedia search is still a rather unexplored field. Even though both content-based methods and user interactions features have been studied in such a context, to the best of our knowledge, this work is the first to combine both in a multimodal fusion and to explore the usefulness of additional modalities, i.e., eye gaze, physiological response and facial expression.

3 DATA COLLECTION

The experiment has received ethical approval from the ethical review board of the faculty of psychology and educational sciences, University of Geneva. 51 healthy participants with normal or corrected to normal vision were recruited through campus wide posters and Facebook. From these 51 participants, 18 were male and 33 were female, and the average age was 25.7 years (σ = 5.3). The participants were informed about their rights and the nature of

Figure 1: Our experimental setup including an eye gaze tracker, front-facing camera recording face videos and galvanic skin response.

the experiment. They then signed an informed consent form before the recordings. They received a monetary gratitude for their participation.

Our experiments were conducted in an acoustically isolated experimental booth with controlled lighting (shown in Figure 1). A video was recorded using an Allied Vision[2] Stingray camera at 60.03 frames/second with a 780 × 580 resolution. Stimuli were presented on a 23 inches screen (1920×1080), and the participants were seated approximately 60cm from the screen. Two Litepanels[3] daylight spot LED projectors were used to light up the participants' faces to reduce possible shadows. An infrared block filter was mounted on the lens to remove the reflection of the infrared light from the eye gaze tracker. Video was recorded using the Norpix Streampix software[4]. Eye gaze, pupil diameter and head distance were recorded using a Tobii[5] TX300 eye gaze tracker at 300Hz. The GSR was recorded using a Biopac[6] MP-36 at 125Hz through electrodes attached on distal phalanges of index and middle fingers. An experimental protocol was run by Tobii Studio, and the recordings were synchronized by a sound trigger that marked the frames before each stimulus for the camera. The same trigger was converted to a transistorfi?!transistor logic (TTL) trigger using a Brain Products StimTrak[7] and recorded alongside the GSR signals.

Our participants were first familiarized with the protocol and ratings in a test run, and then they performed search tasks under three different intent conditions. In this study, we were interested in assessing knowledge emotions in image search. Knowledge emotions are the emotions that arise as a result of the evaluation of one's knowledge, e.g., interest, confusion and surprise. Therefore, we asked the participants to self-report these emotions on a seven-point scale at the end of each session. In addition, we also asked the participants to report the level of control and boredom they felt using a similar seven-point scale rating.

The recorded database, except face videos are available for academic research[8]. For the face videos, we provide the landmarks, features and action units for the benefit of the community.

[2]https://www.alliedvision.com/
[3]http://www.litepanels.com
[4]https://www.norpix.com
[5]http://www.tobii.com/
[6]https://www.biopac.com/
[7]http://www.brainproducts.com/
[8]http://cvml.unige.ch/resources

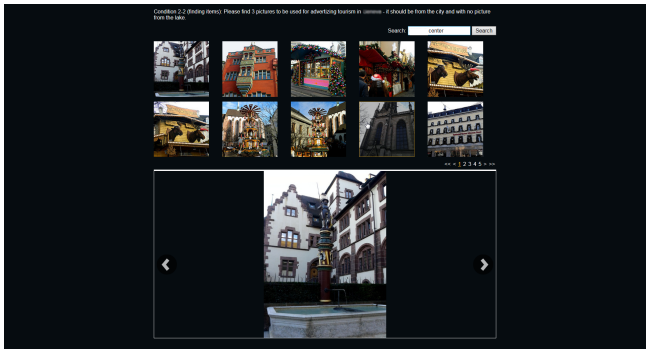

Figure 2: An example of a search performed with our image retrieval system. With a double click on the enlarged image, the user chose the goal image.

4 IMAGE RETRIEVAL SYSTEM

To conduct the search experiments, we created our own custom image search tool. A snapshot of our image retrieval system is shown in Figure 2. The design of the tool was made in a way that it could easily be customized to the different search intents. The images are retrieved using the Flickr API. When the user performs a search by submitting a query, the tool returns a ranked list of images. The images are presented to the user in an image gallery depicting the first ten images as thumbnails with the first image enlarged at the bottom.

The experiments were conducted in seven sessions, under three intent conditions, namely, *entertainment*, *finding* and *re-finding* items. Each session lasted three minutes and sessions were presented in random order. We had three sessions for finding images for a specific purpose; three sessions for re-finding images that were displayed for 15 seconds before the search session and one free viewing *entertainment* session. For two search sessions, one in *finding* and one in *re-finding*, we slightly modified the search terms on the fly to observe the effect of goal obstruction on users behavior. The modification was performed by adding and removing random query terms before they were sent to Flickr. The search tasks were mostly about Geneva where the experiments were conducted. The image search task instructions are given in Table 1.

We discarded the data from the participants who misunderstood the instructions, and the sessions in which the majority of eye gaze samples were lost (due to extreme head pose and, in one case, device failure). Therefore, out of 357 possible search sessions, we could analyze 299 session that were consistently and correctly performed and recorded. In the analysis of behavioral responses, the data from the first seven participants were also discarded since we made a small change in the layout of the interface, after recording them, which made the eye gaze responses inconsistent.

5 EXTRACTED FEATURES AND MULTIMODAL ANALYSIS

5.1 User interactions and responses

The following modalities were recorded from the participants interacting with the search interface: visual, physiological, eye gaze and implicit user interactions. In this section, we describe the processing and feature extraction performed on these modalities. For

Table 1: The list of image search task instructions given to the users during the experiment.

Task	Instruction
Condition 1 (*entertainment*):	You can search/browse freely to look for your content of interest.
Condition 2-1 (*finding* items):	Please find 3 pictures to be used for advertising tourism in Geneva; they should all show jet d'eau.
Condition 2-2 (*finding* items):	Please find 3 pictures to be used for advertising tourism in Geneva; it should be from the city and with no picture from the lake.
Condition 2-3 (*finding* items):	Please find 3 pictures to be used for advertising tourism in Geneva; it should show Nations square.
Condition 3-1 (*re-finding* items):	Please try to re-find the example image shown below. The example image is a boy on a beach playing with a bucket.
Condition 3-2 (*re-finding* items):	Please try to re-find the example image shown below. The example image is a Swiss flag swung on the edge of a lake.
Condition 3-3 (*re-finding* items):	Please try to re-find the example image shown below. The image shows passengers boarding a Vietnam airlines plane on a cloudy day.

Table 2: The list of 17 features extracted from the logged implicit user interaction with the search tool. Each feature is extracted per user and per session.

Feature type	Description	#
Mouse clicks	Clicking rate per second (mean and standard deviation), type of clicked object (search button, next image, etc.)	8
Mouse movement	Distance moved in pixels per second, speed (pixels per second) of movement for each second (mean and standard deviation)	4
Key strokes	Keystroke rate per second (mean and standard deviation)	2
Search terms	Number of different query terms, complexity of query terms derived from the synsets tree length in the WordNet [2, 29]	2
Images	Number of displayed images	1

the analysis in this section, we analyzed user interactions features extracted from the full sessions. The visual content features and the user interactions features from the first 30 seconds were used for the intent recognition in Section 6.

5.1.1 Implicit user interaction. All implicit user interactions performed with the search tool were logged (see Table 2). This includes the clicks, mouse movements, displayed images and thumbnails, selected images, search queries and keystrokes. Additionally, we implemented the possibility to manipulate search queries in the background. All the information was collected using Java scripts and stored in a MySQL server using PHP for later analysis. From the collected data, we extracted different types of features for further analysis, including mouse movements, mouse clicks and keystrokes.

5.1.2 Eye gaze. Optical eye gaze trackers track the direction of gaze and provide the projected gaze. The eye gaze pattern and pupil diameter have been used for detecting boredom, mind-wandering and interest [3, 11, 20, 25]. Eye gaze trackers often record head distance and pupil diameter which are both shown to be useful for recognizing affective and cognitive states [20]. Head distance is a measure of body posture. Pupillary reflex is modulated by emotional arousal through increase in sympathetic activities [5]. Eye gaze features such as fixations and saccades were extracted by the eye gaze analysis software (Tobii Studio). Fixations are the points where eye gaze is maintained in the same location for a minimum amount of time (around 100ms). The number of eye fixations, the presence of eye gaze in the area of interest (AOI) and the saccadic movements vary in different cognitive and emotional states. Saccades are the eye movements between fixations. The absolute direction of saccades (measured by their absolute angle) and the relative direction with regard to the last saccade were calculated by Tobii studio. With a simplifying assumption of straight saccadic movements, we defined the scan path as the direct path between the consecutive fixations. In eye gaze analysis, an area of interest is often defined to study the gaze pattern locally. We defined the area of the enlarged image as the AOI. Inspired by the relevant literature [3, 11], 60 features were extracted from eye gaze, pupil diameter and head distance (see Table 3).

Table 3: The list of 60 eye gaze features. The functionals or statistical descriptives are mean, standard deviation, first and third quartiles, median, maximum and minimum.

Feature type	Description	#
AOI	Number of fixations in the AOI, the proportion of gaze duration in AOI	2
Fixation	Number of fixations, statistical descriptives on fixation duration	8
Saccade	Number of saccedes, statistical descriptives of saccade duration, absolute and relative saccadic directions	22
Scan path	Statistical descriptives of scan path distances and their speed	14

5.1.3 Facial expressions. Facial expressions were analyzed using the Affdex SDK [28]. The Facial Action Coding System (FACS) [12] is a taxonomy of facial movements that can describe facial expressions, e.g., lip puller, brow raise and dimpler. The intensity of 19 facial action units were detected at frame level by Affdex. Additionally, head pose direction was also extracted. The following seven functionals were applied to the features in each session for pooling: mean, standard deviation, median, maximum, minimum, first and third quartiles. This resulted in a feature vector with 154 elements for each session.

5.1.4 Galvanic skin response (GSR). GSR is a measurement of electrical conductance on skin through a pair of electrodes. The skin's electrical resistance measured by GSR fluctuates with the activity of sweat glands which are driven by the sympathetic nervous system. GSR responses consists of tonic (slow) and phasic (fast and often event-related) responses. GSR varies by emotional arousal and has been extensively used in emotion sensing [8, 23]. GSR provides a measure for detecting the presence and intensity of emotions. We used the open source TEAP toolbox[9] [34] to extract nine features from the GSR signals. In order to capture the phasic

[9]https://github.com/Gijom/TEAP

Table 4: The list of nine GSR features.

Feature type	Description
Number of peaks	Number of peaks in resistance exceeding 100Ω
Amplitude of peaks	GSR peak amplitude from the saddle point preceding the peak
Rise time	The time it takes GSR to reach its peak from the saddle point in seconds
Statistical moments	Mean, first and third quartile & standard deviation (electrical resistance in Ω)
Trend	Intercept and slope for the linear trend

responses, we extracted the peaks that appears in GSR signals and calculated their frequency of occurrence, amplitude and rise time. Statistical descriptives were also extracted that captures both tonic and phasic characteristics of electrodermal responses. The list of features are given in Table 4.

5.2 Visual content features

To explore how the content of the visited images are related to the intent categories, we extracted three sets of visual content features from the images. We decided to use one feature-set that extracts the overall visual similarities based on colors and textures in images visited for the same intent and another set that is able to emulate the visual perception of a human to explore if the visited images in intent classes are of differentiable nature according to their visual perception. Additionally, sentiment expressed by images, i.e., visual sentiment [4], may carry information about the intent categories that can be useful. For each of these three feature-sets, we created one feature vector containing the visual features of all images visited in the first 10, 20 and 30 seconds of the search session for each search task. The visual content features were only used for the intent recognition experiments in Section 6. We decided not to directly use features extracted from a deep convolutional neural network trained on Imagenet [10] since such features are related to the concepts present in images that can be associated with the specific tasks rather than intentions.

5.2.1 Joint composite descriptor (JCD). JCD is a set of global visual features that represent the texture and color of an image. The JDC features are joint descriptors, which combine two compact composite descriptors (fuzzy color and texture histogram and the color and edge directivity) in one. The combination of the two descriptors is possible because their color information originates from the same fuzzy color system. The result of the combination is a descriptor, which contains fuzzy color information, texture information and edge information [9]. By extracting this feature, we wanted to study the visual similarities in images visited in the same intent condition.

5.2.2 Tamura. The Tamura features are defined based on the assumption that textural features correspond to the perception of the human eyes [35]. Tamura compared coarseness, contrast, directionality, line-likeness, regularity and roughness, which are six different texture features, with psychological measures taken from experiment subjects. The three features that achieved the best results in his evaluation are coarseness, contrast and orientation. Coarseness measures the size texture primitives (also called texture elements or texels) [18]. Larger textures have larger primitives and fine textures have smaller ones. The contrast measures how distinctive the differences between the textures in the images are. The contrast can be considered as clear to identify if all areas can easily be distinguished from each other. The orientation describes the dominant orientation of the textures in the image. A single image can have only one dominant orientation or several of them. Moreover, an image can also have no orientation at all, which then is called isotropic. For the Tamura global image feature, coarseness, contrast and orientation are extracted from an image and stored in a histogram representation [19]. Using this feature, we want to explore if the intent classes are related to the visual perception and if this can help to classify the intent class.

5.2.3 Visual sentiment descriptors. A visual sentiment concept detector [21] was applied to the images that were visited during the search process. This sentiment detector was trained to detect a refined set of a large-scale visual sentiment concept ontology (VSO) that detects the presence of adjective-noun pairs in images. These adjective-noun pairs (ANP) were selected from social media data based on their relevance in expressing sentiment. The model that we used is an improved version of the original work [4] that uses a residual deep convolutional neural network (CNN) that is trained to detect adjectives, nouns and ANPs simultaneously [21]. This CNN generates probability estimates (softmax output) for 553 adjective-noun pairs that are then combined for the visited images to form feature vectors.

5.3 Statistical analysis

In this Section, we analyze how the features extracted from behavior, reported emotions and interactions differ under different intent scenarios. Participants self-reported a set of emotions, e.g., interest, boredom and confusion, at the end of each short search session. We calculated the Spearman rank correlation between all the self-reported emotions (see Table 5). As expected, interest and boredom are inversely correlated. The significant correlation between surprise and confusion ratings demonstrate that these emotions were co-occurring in search scenarios. We further looked into whether

Table 5: Spearman rank correlation coefficient between the ratings. * implies significance ($p < 0.01$).

Rating	Interest	Boredom	Confusion	Control	Surprise
Interest	-	-0.49*	-0.03	0.31*	0.22*
Boredom	-	-	-0.05	-0.29*	-0.19*
Confusion	-	-	-	-0.30*	0.46*
Control	-	-	-	-	-0.06

the reported emotions varied in different intent conditions. We performed a one-way ANOVA test on the ratings in three conditions and found significant differences for interest ($p = 0.00$, $F = 12.10$), boredom ($p = 0.01$, $F = 5.25$) and control ($p = 0.00$, $F = 16.93$). The distribution of the reported emotions in three different intent conditions are given in Figure 3. We performed a two-tailed t-test to check whether the perception of control was changed with manipulating the query terms, i.e., adding and removing random terms before sending the query to Flickr in two out of seven sessions. We did not find any significant difference between manipulated and

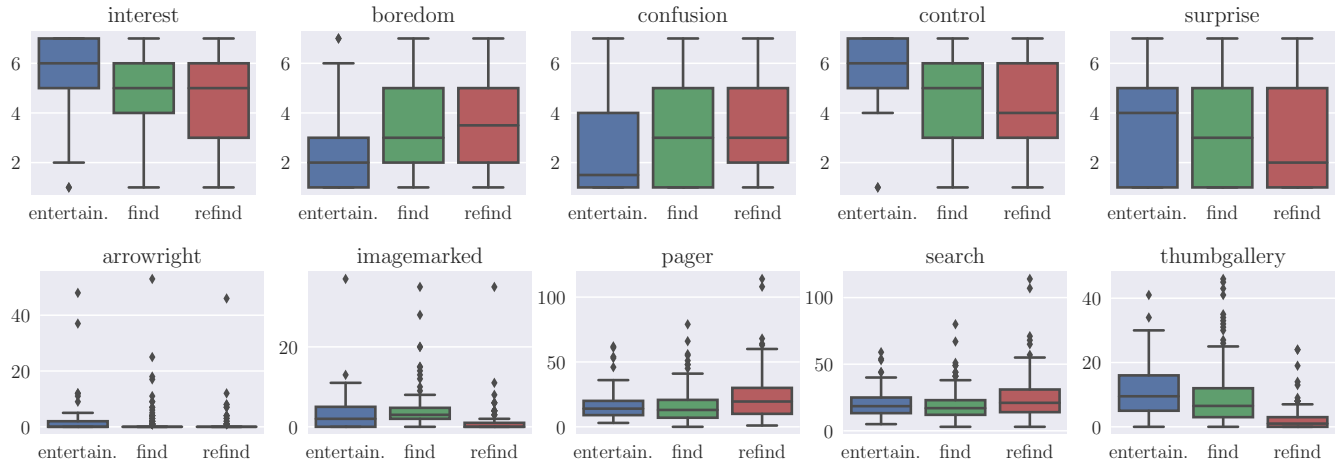

Figure 3: Self-reported emotions (top) and click frequency on different items (bottom) under different intent conditions.

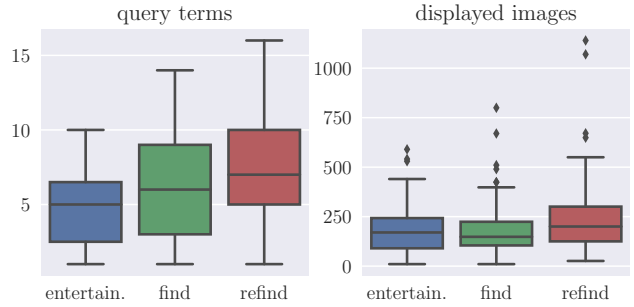

Figure 4: Box-plots showing the distributions of the number of images and the number query terms in each condition.

non-manipulated situations which demonstrated that our query manipulation was too weak and not effective in inducing the sense of goal obstruction.

Semantic complexity of a term is associated with its ambiguity, i.e., the number of different meanings that can be interpreted from the word. We calculated the query complexity using a synset tree; the deeper the vertex is in the tree the less ambiguous the term is [29]. The total complexity is calculated as follows. First, we get the synset tree for a given search term. In the tree, we iterate over all branches leading to the term. For all branches, we calculate the depth and sum it. The number and depth of branches is representative of the number of the meaning that can be derived and is a measure of complexity. For example, "cage" has three branches in the sysnet tree and each of them has a depth of one which makes its complexity score three. A one-way ANOVA test on the query term complexity showed a significant difference between different intent conditions. The *entertainment* condition queries had a much lower complexity compared to the *finding* and *re-finding* conditions. The *re-finding* condition had the highest complexity where the participants were trying to use generic terms describing the content they had in mind. Through fitting a linear model, we checked whether the complexity of the query terms varied over time. We found that, only in the *re-finding* condition, there was a significant increase in the complexity of the query terms ($slope = 0.05$, $p = 0.00$).

We also looked at the number of search terms used in each condition. A one-way ANOVA test showed significant differences between the number of query terms used in average ($p = 0.00$, $F = 17.75$). Similarly, the *entertainment* condition had the lowest average number of query terms followed by *finding* and *re-finding* conditions (see Figure 4).

To study the browsing pattern, we counted the number of browsed images per session. A one-way ANOVA test found a significant difference between the number of images displayed under three different conditions ($p = 0.00$, $F = 5.68$). Given the smaller number of query terms used in the *entertainment* condition, this demonstrates that the participants spent more time browsing rather than searching (see Figure 4).

We looked into two features from mouse movements under different intent conditions. The speed and the distance of mouse movement trajectories. A one-way ANOVA test showed significant differences for speech ($p = 0.00$, $F = 6.77$) and distance ($p = 0.01$, $F = 5.12$) under different intent conditions. Mouse movement was the fastest in the *entertainment* condition. However, the distance of the mouse trajectories was much shorter in average compared the *finding* and *re-finding* conditions. Therefore, participants spent less time exploring the interface with their mouse in the *entertainment* condition.

Mouse click features describe where on the interface the clicks occurred and at what time. We collected clicks based on the locations of the clicks, which are *arrowright* and *arrowleft*, *imagemarked* (enlarged image), *pager* (to change pages), *search* and *thumbgallery* (thumbnails). The distribution of the click frequencies is shown in Figure 3. A one-way ANOVA test showed significant differences between click frequencies for *imagemarked* ($p = 0.00$, $F = 22.66$), *pager* ($p = 0.00$, $F = 6.63$), *search* button ($p = 0.00$, $F = 7.00$) and *thumbgallery* ($p = 0.00$, $F = 43.78$). We also calculated the key stroke rates and found that the rate of key strokes were not different under the three intent conditions.

The aggregated heat maps of eye gaze on three different conditions are given in Figure 5. We can observe that users spent more time looking at the enlarged image in the *entertainment* conditions whereas they spent much less time in the *re-finding* condition. In

Figure 5: Eye gaze heat maps in different search intent conditions. From left to right: entertainment, finding and re-finding.

the re-find condition, the goal was to spot the image so they spent more time on queries and exploring the thumbnails.

Using the Affdex SDK [28], we also calculated the intensity of the following expressed emotions: sadness, joy, disgust, anger, contempt , surprise and engagement. The statistical analysis did not show any difference between the frequency of the expressed emotions under different conditions.

6 INTENT RECOGNITION

Predicting the search intent early on in a search session can help a retrieval system to adapt its results according to the user's need. Therefore, we attempted recognizing search intent from four different modalities, namely, facial expressions, physiological responses, eye gaze and implicit interactions, e.g., mouse movement. Inspired by the previous work that related visual content to the possible search intent [17, 26], we also extracted visual content features from the visited images and used them for intent recognition. We only extracted features from the first 10, 20 and 30 seconds of a search session to imitate a situation in which a retrieval system is predicting the search intent.

Using the extracted features from the first 10, 20 and 30 seconds, we trained Random Forest classifiers to tackle the classification. We opted for using an ensemble method due to its ability to perform well with lower number of samples and its robustness against over-fitting. The utilized features are described in Section 5. The performance of the classification was evaluated through a user-independent cross-validation, in which the samples from the same participants were never present in both train and test sets.

6.1 Experimental results and multimodal fusion

For all our classification experiments, we used the same Random Forest classifier with the respective features as input [6] with the following parameters: 300 trees, 100 iterations, unlimited maximal depth and a batch size of 100.

For the evaluation, we used user-independent cross-validation. Table 6 gives an overview of the classification results using the user interaction features, visual content features and fusions of them. We fused the modalities in two ways.

The first fusion method is *early fusion* which consists in concatenating different feature vectors into a single vector before being fed to a classifier. A problem with early fusion is that the combination of very diverse features, like user interaction features with visual features, can pose problems. For example, combining interaction features with a small number of dimensions with a visual feature-set with many thousand dimensions in an early fusion does not result to improved performance [16, 33].

The second utilized fusion method is *late fusion*. In late fusion or decision-level fusion, each modality has its own classifier. After these first classification steps, the output of all classifiers are combined to obtain a final result. This combination can be performed for example by simple majority vote, using another classifier or weighted sum of the scores. Because each feature is processed in a separate classifier, *late fusion* is more costly in terms of computational complexity. The choice of *late fusion* method depends on the dataset, the features and the metrics that are used to calculate the distances between the different features [15].

As shown in Table 6, the visual content features alone outperform the baseline, which we calculated using the ZeroR classifier. ZeroR assigns the label from the majority class to all the instances and provides a baseline. This is an indicator that the intent categories are somehow preserved in the visual features but single features are not very efficient to detect this. The same is true for all user interactions features. The performance of Tamura is particularly interesting since it should be able to give a similar perception to humans and indicates that different intent classes are perceived differently. The VSO features performed the best among the visual features. VSO features also performed better with more images, i.e., the longer the interaction window we analyzed. However, VSO features might have captured the concepts that are related to the tasks rather than intentions. Comparing visual content features and user interactions features, the user spontaneous and implicit interaction features perform better. Implicit interaction outperforms the visual content features. All sets of visual content features perform better than the face and GSR features.

Looking at the window length, the shorter segments yield better results for behavioral and interaction features whereas for content features longer segments mostly increased the performance. Overall, 30 seconds segments gives the best performance with late fusion reaching average F-1 score of 0.722 for three classes.

Early fusion does not lead to a large increase of the performance and for the user interactions features it even leads to a small decrease. Feature vectors from different modalities are of different size and simply concatenating them does not yield superior results (for example face compared to GSR). For the visual content features, the early fusion lead to a small increase, which was expected based on previous findings [16, 33]. Late fusion for all different combinations

Table 6: Classification performance based on the user interactions and visual content features in terms of weighted average of precision, recall and F-1 score. The results are reported for the first 10, 20 and 30 seconds. The best results are in boldface.

Features	Precision			Recall			F-1 score		
User interactions features	**10sec**	**20sec**	**30sec**	**10sec**	**20sec**	**30sec**	**10sec**	**20sec**	**30sec**
Face	**0.449**	0.444	0.392	**0.524**	0.520	0.458	**0.483**	0.479	0.421
Eye Gaze	**0.576**	0.542	0.562	**0.626**	0.587	0.572	**0.585**	0.548	0.536
GSR	**0.430**	0.363	0.421	**0.488**	0.413	0.438	**0.455**	0.386	0.422
Implicit user interaction	**0.693**	0.650	0.619	**0.724**	0.681	0.682	**0.698**	0.651	0.637
Visual content features	**10sec**	**20sec**	**30sec**	**10sec**	**20sec**	**30sec**	**10sec**	**20sec**	**30sec**
Tamura	0.392	0.396	**0.403**	0.478	0.482	**0.495**	0.43	0.433	**0.444**
JCD	**0.561**	0.532	0.46	0.512	**0.538**	0.468	0.475	**0.493**	0.463
VSO	0.502	0.617	**0.623**	0.583	0.627	**0.639**	0.539	0.601	**0.612**
Multimodal fusion	**10sec**	**20sec**	**30sec**	**10sec**	**20sec**	**30sec**	**10sec**	**20sec**	**30sec**
Early fusion user interactions and visual content	**0.581**	0.54	0.521	**0.677**	0.630	0.610	**0.625**	0.581	0.562
Late fusion user interactions and visual content	0.683	0.67	**0.743**	0.748	0.736	**0.748**	0.703	0.692	**0.722**
Baseline (ZeroR)	0.353	0.353	0.189	0.421	0.421	0.435	0.346	0.346	0.264

reaches the best results. In our experiments, we used the weighted sum of scores for late fusion. The scores were derived from the classifier by averaging the predicted class probability of the trees in a Random Forest (confidence score for each class). The overall best result is achieved fusing the user interactions features with the visual content features. We found that multimodal fusion achieves the best results for intent classification.

7 DISCUSSIONS

The intention categories in this work do not cover all possible search intent scenarios and are only three typical cases. To fully cover the whole spectrum of search intent, we need a finer granularity and additional intent categories specific to the image content. We did not witness significant emotional expressions, and this could be a result of the mental fatigue caused by the experiment or the artificial nature of the search tasks with less sense of ownership than genuine ones [31].

Another limitation of this work is the potential bias of the tasks cognitive load and behavioral responses elicited by instructions rather than intention. This might have resulted in different interaction and behavioral responses that are not necessarily related to intent. For example, the head pose or eye gaze position are more likely to be at the top to recall the task description in more difficult tasks compared to the easy one, i.e., entertainment.

The visual features used in this work show promising results. Nevertheless, further analysis is needed on content-based features. We hypothesize that emotional content of the query and retrieved results are related with search intent, e.g., positive sentiment might be associated with *entertainment* intent. However, what has been detected by visual sentiment features might not necessarily be related the intention, and the difference in the visual sentiment might be the result of the task at hand.

The intent recognition performance is similar to the one reported in [30], i.e., 51.4% in the early stage for four classes. However, our results are obtained using a user-independent cross-validation which is more generalizable and shows robustness to between-user variations. Nevertheless the performance of such a system shall be further evaluated in an operationalized system with a diverse set of scenarios in which intention can even change in the middle of a search session.

To the best of our knowledge, the relevance of the content to search intent has not been analyzed in text-content. Utilizing the language processing models to understand queries and retrieved content can be also useful for determining users search intent.

8 CONCLUSIONS

In this paper, we analyzed the spontaneous responses of users under different search intent conditions. We found that implicit user interaction including mouse movement to be the most informative channel of information for predicting the intent. Eye gaze behavior also showed promising power in differentiating between different intent conditions. However, despite the importance of emotions in the search process [30], facial expression and physiological responses underperformed in recognizing search intent compared to implicit user interactions. As we showed, the visual content of the browsed images can also provide hints on the underlying intention for image search. To summarize, we identified the best modalities for recognizing search intent and showed the feasibility of automatically identifying search intent early in a search session. If such automatic intent recognition method is deployed, the resulting intent-aware multimedia retrieval system can optimize its results by switching its ranking methods according to the user's underlying motivation.

9 ACKNOWLEDGMENTS

The work of Soleymani was supported by his Swiss National Science Foundation Ambizione grant. The work of Riegler and Halvorsen was supported by the Research Council of Norway as a part of the projects INTROMAT under grant agreement 259293 and EONS under grant agreement 231687. We thank David Sander for his kind assistance for the ethical review of the experiment. We also thank "Fondation Campus Biotech Genève" for providing access and support at their experimental facilities.

REFERENCES

[1] I. Arapakis, I. Konstas, and J. M. Jose. 2009. Using Facial Expressions and Peripheral Physiological Signals As Implicit Indicators of Topical Relevance. In *Proceedings of the 17th ACM International Conference on Multimedia (MM '09)*. ACM, New York, NY, USA, 461–470.

[2] S. Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*. 69–72.

[3] N. Blanchard, R. Bixler, T. Joyce, and S. D'Mello. 2014. Automated Physiological-Based Detection of Mind Wandering during Learning. In *Proceedings of ITS*. Springer International Publishing, 55–60.

[4] D. Borth, T. Chen, R. Ji, and S.-F. Chang. 2013. SentiBank: Large-scale Ontology and Classifiers for Detecting Sentiment and Emotions in Visual Content. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 459–460.

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

[6] L. Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[7] A. Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10.

[8] R. A. Calvo and S. D'Mello. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing* 1, 1 (jan 2010), 18–37.

[9] S. Chatzichristofis, Y. Boutalis, and M. Lux. 2009. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Procideeding of IASTED Int'l Conference on Signal Processing, Pattern Recognition and Applications*, Vol. 134643. 64.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[11] S. D'Mello, A. Olney, C. Williams, and P. Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies* 70, 5 (2012), 377–398.

[12] P. Ekman and W. Friesen. 1978. *The Facial Action Coding System (FACS)*. Consulting Psychologists Press, Stanford University, Palo Alto.

[13] D. Elsweiler, S. Mandl, and B. Kirkegaard Lunn. 2010. Understanding Casual-leisure Information Needs: A Diary Study in the Context of Television Viewing. In *Symposium on Information Interaction in Context (IIiX '10)*. ACM, New York, NY, USA, 25–34.

[14] R. Fidel. 1997. The image retrieval task: implications for the design and evaluation of image databases. *New Review of Hypermedia and Multimedia* 3, 1 (1997), 181–199.

[15] I. Gialampoukidis, A. Moumtzidou, D. Liparas, S. Vrochidis, and I. Kompatsiaris. 2016. A hybrid graph-based and non-linear late fusion approach for multimedia retrieval. In *Proceedings of the 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. 1–6.

[16] A. H. Gunatilaka and B. A. Baertlein. 2001. Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 6 (2001), 577–589.

[17] A. Hanjalic, C. Kofler, and M. Larson. 2012. Intent and Its Discontents: The User at the Wheel of the Online Video Search Engine. In *Proceedings of the 20th ACM International Conference on Multimedia (MM '12)*. ACM, New York, NY, USA, 1239–1248.

[18] R. M. Haralick. 1979. Statistical and structural approaches to texture. *Proc. IEEE* 67, 5 (May 1979), 786–804.

[19] P. Howarth and S. Rüger. 2004. Evaluation of texture features for content-based image retrieval. In *Image and Video Retrieval*. Springer, 326–334.

[20] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo. 2014. Predicting affect from gaze data during interaction with an intelligent tutoring system. In *Proc. of ITS*, Vol. 8474 LNCS. 29–38.

[21] B. Jou and S.-F. Chang. 2016. Deep Cross Residual Learning for Multitask Visual Recognition. In *Proceedings of the 2016 ACM on Multimedia Conference (MM '16)*. ACM, New York, NY, USA, 998–1007.

[22] C. Kofler, M. Larson, and A. Hanjalic. 2016. User Intent in Multimedia Search: A Survey of the State of the Art and Future Challenges. *Comput. Surveys* 49, 2, Article 36 (2016), 37 pages.

[23] S. D. Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological Psychology* 84, 3 (jul 2010), 394–421.

[24] C. Lagger, M. Lux, and O. Marques. 2012. What makes people watch online videos: An exploratory study. *Computer Entertainment* (2012).

[25] S. Lallé, C. Conati, and G. Carenini. 2016. Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 2529–2535.

[26] M. Lux, C. Kofler, and O. Marques. 2010. A classification scheme for user intentions in image search. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. 3913–3918.

[27] M. Lux, M. Taschwer, and O. Marques. 2012. A Closer Look at Photographers' Intentions: A Test Dataset. In *Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia (CrowdMM '12)*. ACM, New York, NY, USA, 17–18.

[28] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby. 2016. AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit. In *CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 3723–3726.

[29] G. A. Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[30] Y. Moshfeghi and J. M. Jose. 2013. On cognition, emotion, and interaction aspects of search tasks with different search intentions. In *International conference on World Wide Web - WWW '13*. ACM Press, New York, New York, USA, 931–942.

[31] A. Poddar and I. Ruthven. 2010. The Emotional Impact of Search Tasks. In *Proc. of the 3rd Symposium on Information Interaction in Context (IIiX '10)*. 35–44.

[32] M. Riegler, M. Larson, M. Lux, and C. Kofler. 2014. How 'How' Reflects What's What: Content-based Exploitation of How Users Frame Social Images. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 397–406.

[33] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. 2005. Early Versus Late Fusion in Semantic Video Analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)*. ACM, New York, NY, USA, 399–402.

[34] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel. 2017. Toolbox for Emotional fEAture extraction from Physiological signals (TEAP). *Frontiers in ICT* 4 (2017), 1.

[35] H. Tamura, S. Mori, and T. Yamawaki. 1978. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics* 8, 6 (1978), 460–473.

[36] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. 2007. Information Re-retrieval: Repeat Queries in Yahoo's Logs. In *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. 151–158.