Künstliche Intelligenz manuscript No.
(will be inserted by the editor)

Automatic Detection of Visual Search for the Elderly using Eye and Head Tracking Data

Michael Dietz · Daniel Schork · Ionut Damian · Anika Steinert · Marten Haesner · Elisabeth André

Received: date / Accepted: date

Abstract With increasing age we often find ourselves in situations where we search for certain items, such as keys or wallets, but cannot remember where we left them before. Since finding these objects usually results in a lengthy and frustrating process, we propose an approach for the automatic detection of visual search for older adults to identify the point in time when the users need assistance. In order to collect the necessary sensor data for the recognition of visual search, we develop a completely mobile eye and head tracking device specifically tailored to the requirements of older adults. Using this device, we conduct a user study with 30 participants aged between 65 and 80 years ($avg = 71.7$, 50% female) to collect training and test data. During the study, each participant is asked to perform several activities including the visual search for objects in a real-world setting. We use the recorded data to train a support vector machine (SVM) classifier and achieve a recognition rate of 97.55% with the leave-one-user-out evaluation method. The results indicate the feasibility of an approach towards the automatic detection of visual search in the wild.

Keywords visual search · activity recognition · machine learning · eye tracking · head tracking

M. Dietz, D. Schork, I. Damian, E. André
Human Centered Multimedia
Augsburg University
Universitätsstraße 6a
86159 Augsburg, Germany
E-mail: *lastname*@hcm-lab.de

A. Steinert, M. Haesner
Geriatrics Research Group
Charité - Universitätsmedizin Berlin
Reinickendorfer Straße 61
13347 Berlin, Germany
E-mail: *firstname.lastname*@charite.de

1 Introduction

With advancing age it is common to experience a decline of working memory [35]. This can lead to forgetfulness and is generally associated with an overall decrease in quality of life. Additionally, over 15% of elderly people can develop even more serious memory-related issues, including memory loss, confusion and other cognitive impairments [42]. As a consequence, forgetting a name, an object or an appointment can lead to very unpleasant circumstances. Within the Glassistant project, we therefore attempt to create an autonomous assistant using smart glasses and wearable sensors. The general aim of the project is to provide aid to the elderly by recognizing critical situations and offering appropriate support. One major use case for that is the automatic detection of instances during which the user searches for misplaced objects like keys or wallets. Such episodes are mostly experienced by elderly people due to the cognitive decline of their working memory, which can be very frustrating and time consuming. With the automatic detection of visual search we aim to recognize when the user needs assistance, in order to provide support for remembering the location of misplaced objects or even automatically directing the user to the location of the desired item. We see the main contributions of our current work as follows: (1) We introduce a novel concept for the automatic detection of visual search episodes in real-world scenarios based on the combined analysis of eye and head movement data. (2) We develop a completely mobile eye and head tracking device in order to capture the necessary sensor data, which is specifically designed to meet the requirements of elderly users. (3) We present a study setup which creates a natural environment for visual search activities and enables the collection of realistic data. (4) We propose a classification approach which is able to recognize the visual search behavior of elderly users in a realistic environment and distinguishes between certain search phases.

The remaining article is structured in the following way: First, we take a look at related work and discuss their findings and limitations. Afterwards we introduce our custom eye and head tracking device and describe the details of the user study which was conducted to collect training and test data to build our classification model. We then evaluate our approach and show that visual search can be detected in a real-world scenario, using the data from our mobile eye and head tracking device. Finally, we discuss the results and give an overview of possible future applications for our visual search detection approach.

2 Related Work

Visual search is commonly defined as the act of looking for a target object among several distractors [39]. During this process, attention is focused sequentially on each element of the visual scene, resulting in specific eye movement patterns [14]. The first one to analyze these patterns was Buswell [6] in 1935. He showed that eye movements differ distinctively during a visual search task on an image compared to a free viewing task with no instructions. Several years later, Yarbus [43] confirmed in 1967 that the visual task indeed plays an important role for the observed scan paths and patterns. Since then, a lot of research has been done regarding the analysis of eye movement patterns in visual search tasks. For example, Castelhano et al. [7] compared various eye movement measures, such as the fixation duration, saccade amplitude or percentage of fixated area, between a visual search and a memorization task. Thereby 35 photographs of real-world scenes were shown to the participants who were asked to either search for a certain target or to memorize the objects in the corresponding image. As the results show, most of the examined features yielded distinctive values for each of the tasks, enabling the usage of a binary classifier for their detection. Similarly, Mills et al. [30] examined the influence of a visual search, a memorization, a scene rating and a free viewing task on spatial and temporal characteristics of eye movements. For that, they conducted a user study with 53 participants and asked them to perform the four tasks on 67 images of computer-generated natural scenes. In compliance with [6,7,37,43] they identified several eye movement characteristics which can be used to distinguish between the tasks and are therefore considered in our work as well. Based on these findings, Henderson et al. [19] tried to infer the viewing task from eye movement measures with a naïve Bayes classifier. In their study they recorded the eye movements of 12 participants while performing a scene memorization and a visual search task on scene photographs presented on a monitor. As the results show, they were able to identify the viewing task with an accuracy of up to 83%. Likewise, Coco et al. [9] used eye movements to classify three visual activities. These consisted of a visual search, a scene description and an object naming task, which were performed on 24 photographs of indoor scenarios. Using a support vector machine (SVM), they achieved a maximum accuracy of 88% for the visual search task. Although these are promising results for the detection of visual search, most of the previous research has been conducted using static images on displays. Due to the restriction of the target area to a certain screen space compared to the wider view of a room or a building, these results might differ in a real-world scenario. Besides, head movements could also be a valuable indicator to identify the visual search process in such a setting, but were previously not considered because of the restricted target area. For these reasons we investigate the visual search task in a completely mobile and real-world scenario.

3 Tracking Device

Generally, there are two types of visual search: preattentive (parallel, efficient, uncontrolled) and attentive (serial, inefficient, controlled). In the first type, basic features like the color, shape and orientation of an object are perceived unconsciously while the second type requires the sequential allocation of attention from the observer for each element in the visual scene [38]. Since the first type occurs subliminally and can be hardly recognized, we focus on the detection of the attentive visual search. For that, we propose the usage of wearable sensors, which can capture specific behavioral patterns of the visual search activity. Combined with common machine learning techniques, the sensor data can be employed to train a binary classifier which is then able to detect the visual search process in real time. For the selection of suitable wearable sensors we make use of the findings from our previous work, in which we analyzed different modalities regarding their applicability for such a challenge [13]. As it turned out, a combination of eye and head movement data showed the most promising results for the detection of visual search, which is the reason why those modalities are used in our current work as well. In order to record the data, several commercially available devices, such as the Tobii Pro Glasses (50-100 Hz), the SMI Eye Tracking Glasses (120 Hz) or the Pupil Labs Headset (30-

120 Hz) could be used. However, these devices are not capable of providing feedback to the users and would require an additional output component to support them, which could be too intrusive for older adults. Since no commercially available device fulfilled this requirement, we decided to build our own prototype. Through that, we were able to consider the special conditions and requirements of our elderly user group. For example, the majority of older adults relies on prescription lenses. Therefore, it must be possible to wear the device in addition to glasses without disturbing the user. This also implies that the device should be as small and lightweight as possible. Furthermore, the prototype should not impact the mobility of the users and must work in a completely mobile setting to increase the acceptance of this technology.

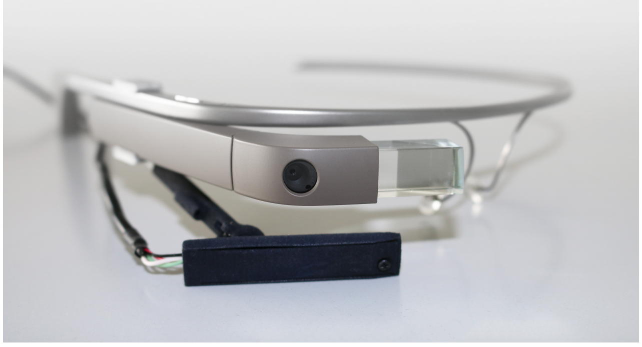

Fig. 1 Google Glass-based eye and head tracking device

Considering these requirements, we decided to use the Google Glass as basis for our prototypical device, since it is currently one of the lightest head mounted displays and can be worn on top of prescription lenses. Besides that, it already has a built-in accelerometer and gyroscope sensor which can be used to track the head movements of the users. In order to record the eye movements as well, we created a custom mount with a 3D printer and attached a small infrared camera (30 Hz, 640×480 resolution) taken from a Pupil Labs eye tracker to the frame of the smart glass as shown in Figure 1. The camera is connected to a Raspberry Pi 2, which can either record the eye video or stream the data to another processing unit. Afterwards, an algorithm based on the open-source Haytham Gaze Tracker[1] is applied to the video stream of the eye camera to determine the pupil position. Combined with the video from the scene camera of the Google Glass, we receive the same data as with a regular head mounted eye tracker, but with the added benefit of being able to support the user through instructions on the head mounted display.

4 User Study

Since the main goal of the Glassistant project is to support elderly users in critical situations, we conducted a large-scale study to collect test and training data for the automatic recognition of those situations. In order to achieve a rich dataset for user-independent machine learning models, we recruited 30 participants aged between 65 and 80 years ($avg = 71,7$) with a female ratio of 50%. During the study each of the subjects performed several activities, including the visual search for objects, while being equipped with our eye and head tracking device. Even though the study was not exclusively designed for the sole detection of visual search, the recorded data can be used for it because all other tasks were similar to day-to-day activities and thus can serve as a comprehensive baseline.

4.1 Tasks

Overall, the study involved five tasks, but since the aim of this work is the detection of visual search, we mainly focus on the search scenario and only give a brief overview of the other tasks. Before each task the participants received a detailed instruction and afterwards had to fill out a questionnaire regarding their experiences during it. In the first task, each participant was instructed to enter general demographic information into a smartphone app. Thereby the system vocally asked the subjects basic questions which they could answer using natural language. Due to the auditory nature of this interaction, the users were able to look around freely during this task. In the second one, the participants were asked to read and write texts on sheets of paper. After a fixed amount of time, an experimenter called them on a telephone and told them four terms which they should memorize and recall at the end of the session. For the following two tasks, the participants were then instructed to work with a computer. In task three, each user was asked to observe the screen for a certain visual condition and had to press a key every time it occurred. Similarly, in task four, an object was shown in the center of the screen for a few seconds while the users had to click on the corresponding button matching its condition. In between those tests, two videos were shown to the users.

Finally, the last task involved the visual search activity, which was investigated in the following two scenarios: the search for keys and the search for rooms. The reason why we chose these scenarios is that we wanted to capture the characteristics of visual search in a wide spectrum of occurrences, ranging from the search of a small item in a limited area to the search of a location

[1] http://eyeinfo.itu.dk

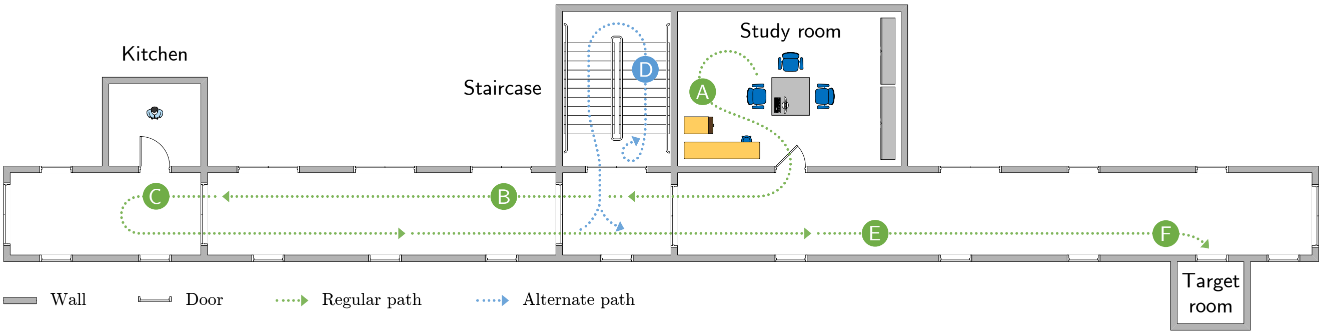

Fig. 2 Overview of the study location

in an open space. In order to create a realistic setting for both parts, we told the participants shortly before the end of the previous task, that we had to leave them to prepare the study for the next participant and that they should meet us in a certain room. Additionally, they were also asked to lock the door with a key located in one of the closets shown in Figure 2 Ⓐ, once they were finished. However, the hidden key did not match the lock on the door. This caused some subjects to continue the search even after finding the key. Eventually, after a certain amount of time, every participant gave up and started to search for the room in which they were supposed to meet the experimenter. Based on the room number we gave them, they assumed that it was located at the end of the hallway (Figure 2 Ⓒ), but upon arrival they realized that there was no room with that number. Instead, they only found a person standing in the kitchen nearby, who they asked for the right way. The person was instructed to tell the participants the number of the correct room (Figure 2 Ⓕ) and with that information most of them were quickly able to locate it. In spite of knowing the room number though, a few users got completely lost and used the staircase to search for the room on different floors (Figure 2 Ⓓ), which resulted in even more realistic search recordings. Nevertheless, all participants eventually found the target room which also marked the end of each session.

4.2 Sensor Setup

For the user study we employed a completely mobile and wearable sensor setup. At the core of the setup was the social signal processing framework SSJ [10], which is a mobile reimagination of the Social Signal Interpretation (SSI) framework [40]. It enabled us to interface with and extract data from multiple sensing devices in a synchronized fashion. Moreover, since SSJ has been designed and built specifically for mobile devices, the participants were able to freely move around the room and the building, increasing the authenticity of the search task. While our custom eye and head tracking device would have been sufficient to record the necessary data for the detection of visual search, additional sensors were used for the recognition of the other situations from our study. As a result, the complete setup consisted of two smartphones (Samsung Galaxy S4), our Google Glass-based eye tracking system, a Raspberry Pi 2 and an Empatica E3 sensor armband. All devices were synchronized to each other and communicated via WiFi. In order to not impact the mobility of the system, the WiFi hotspot was created using one of the two smartphones. The other one was handled by a researcher to control the entire sensor setup, i.e. synchronously starting and stopping the recording on all devices, triggering the calibration phase of the eye tracker or completely shutting down the devices. Moreover, the researcher also used this smartphone to label the start and end of the individual study tasks. The second smartphone was running an SSJ application (called pipeline) which extracted data from the device internal inertial measurement unit (IMU) and microphone as well as the Bluetooth-connected Empatica E3 armband, and stored it to the local SD card. Similarly, a second SSJ pipeline was running on the Google Glass

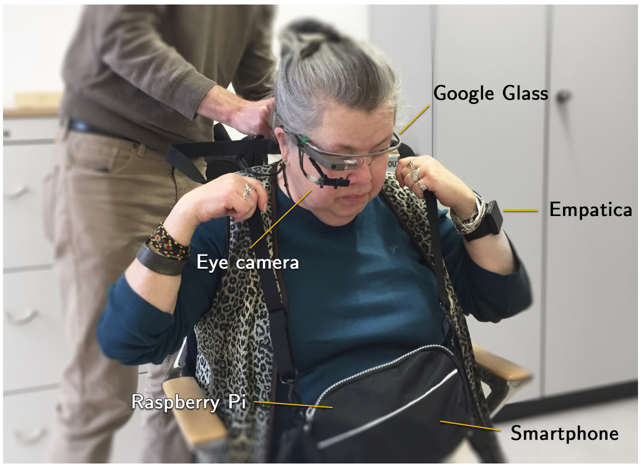

Fig. 3 Participant wearing the sensor setup

which was tasked with recording IMU, audio and video data. The eye tracking camera data was captured using a custom program running on a Raspberry Pi 2.

5 Data Analysis

Following the user study we analyzed the recorded data in order to prepare it for our classification approach. This was necessary to ensure that the sensors worked correctly in all sessions and provided reliable data sets in each case. Otherwise, false or missing data streams could have had a negative impact on the classification performance. Therefore, incomplete and corrupt session recordings had to be identified initially and removed before the data sets could be used in our classification approach for the automatic detection of visual search.

5.1 Signal Quality

Since the device used to record the gaze data consisted of a camera pointed at the participant's eye and another camera capturing the field of view, a calibration had to be conducted in order to map the pupil position on the eye camera to a gaze point in the field of view. After the calibration, it was important that the device stayed in the same position relative to the head. However, some participants treated the device like a pair of glasses and readjusted its position multiple times after calibration. In most cases this led to a shifted gaze point, which left most feature calculations unaffected. In some extreme cases there was so much readjustment that the eye was no longer visible in the field of view of the camera. This led to unusable data in the later part of the recordings. Another problem occurred because some users assumed that the study was concluded after filling out the last questionnaire following the fourth task. In these cases, they took the eye tracking device off before beginning the search part, so that no data could be recorded. One participant even required too much time to complete the tasks which led to the depletion of the Google Glass battery after one hour and forty minutes, resulting in an incomplete data set of the session. For these reasons, eight recordings had to be discarded, leaving 22 usable data sets (*avg. age* = 71.2, 50% female).

5.2 Task Annotation

Based on the recorded audio and video streams, we refined the task annotations for every remaining data set. During this process the first four tasks were labeled as *Baseline*, while both key and room search were annotated as *Search* to create a binary classification problem. Thereby, the annotation for the key search began once the participants approached the closets and ended as soon as they left the room and closed the door. This also marked the start of the room search which continued until the users arrived at the target location. The reason why we did not exclude certain phases, such as the short conversations when asking for the right way, is that even during these periods the participants were still looking around and tried to find the room. The resulting completion times for both search tasks are summarized in Figure 4.

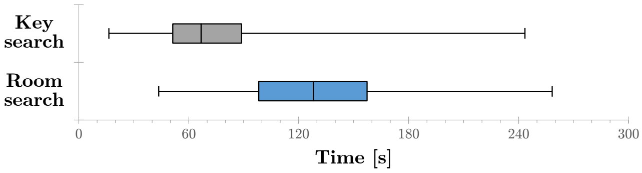

Fig. 4 Visual search task durations

In order to extend the baseline even further, we labeled one instruction phase where an experimenter explained an upcoming task to the participant with *Baseline* as well, since it resembled a regular conversation. Besides that, we also included one questionnaire phase which was similar to a common reading and writing task. As a result, the baseline consisted of the following day-to-day activities: reading, writing, speaking out loud, talking on a telephone, memorizing terms, holding a conversation, working on a computer and watching videos. We used this annotation set in our classification approach to accomplish the automatic detection of visual search with machine learning techniques. Additionally, we created a second annotation set with the same baseline but with individual labels for the key search and the room search. This enabled us to examine if there are any differences between these two scenarios.

6 Classification Approach

For the automatic detection of visual search we selected a support vector machine (SVM) as classifier (*linear kernel*, $C = 1$, $\epsilon = 0.1$, $\nu = 0.5$, $\gamma = 0.01$) since it is still one of the most popular algorithms in the field of machine learning [18] and also works efficiently on the current generation of mobile devices [11]. This is important because we aim to use the resulting classification model with our mobile eye and head tracking device in an online scenario. In order to also achieve a subject-independent classification model, all evaluations were conducted using the leave-one-user-out (LOUO) method. Thereby the classifiers are trained with the data from all users except one and are then tested on the remaining user. This process is repeated for every participant and afterwards the average values across all iterations are taken as result. A key benefit of this method is that it simulates a real time analysis based on the recorded data since the trained classifiers are always tested with the signals from an unknown user, which is also the case in an online classification. For the implementation of features, model training and evaluation we used the Social Signal Interpretation (SSI) framework [40]. It provides a variety of tools to support all phases of machine learning and enabled us to utilize the computational resources of our workstations and servers to accelerate this process. As shown in [12] its flexible architecture also allowed us to create a custom component for the calculation of features, which could be combined seamlessly with the provided modules for data processing and classification.

6.1 Feature Extraction

All gaze features are based on the raw sensor data from our mobile eye and head tracking device. For a given window length we process the data and calculate the *fixation duration*, *saccade duration* and *saccade length*. In our case, these metrics are defined as follows: *Fixation duration* is the time in seconds of a single fixation, *saccade duration* is the time in seconds between two subsequent fixations and *saccade length* is the Euclidean distance in pixels between two subsequent fixation points. For each of these three metrics we then compute the *mean*, *min*, *max*, *median*, *sum*, *standard deviation*, *skew*, *kurtosis* and *range* values, which were commonly used for visual search detection on displays and activity recognition in previous works [4,7,9,15,19,30,37]. In addition to that, we apply a wordbook analysis as proposed by Bulling et al. [5] to identify repetitive eye movement patterns. Furthermore, we analyze the spatial distribution of fixations by computing the *fixation dispersion*, *fixation coverage* and *number of fixation groups*. The *fixation dispersion* is calculated using the root mean square of the Euclidean distances between each fixation and the average position of all fixations within the current window [4]. For the *fixation coverage* we draw a circle with radius r based on the fixation duration around each fixation point and compute the ratio between covered area and total field of view [7]. Based on the fixation map from the previous feature we identify the connected areas which represent *fixation groups* and count their occurrences [34]. Besides that, we calculate the *number of saccades*, *fixations* and *blinks* as well as the *ratio between fixation and saccade duration* [4]. Combined with six movement independent features such as *sum*, *mean* and *variance* of the *blink duration* and *pupil size change* [5], this results in a total of 60 gaze features.

For the extraction of head movement features we directly use the raw accelerometer and gyroscope data from the Google Glass. Since both sensors share the same sample rate and provide the data for each axis (x, y, z) we apply the same features for both of them as suggested in [32]. While most features are computed for each individual axis, some are based on pairs of axes or even factor in all three of them. The features calculated for each axis include the *mean*, *variance*, *standard deviation*, *skew*, *kurtosis*, *interquartile range*, *mean absolute deviation*, *root mean square*, *energy* and *frequency domain entropy* values, which were previously used for activity recognition [1,2,8,20,26,33]. Additionally, we apply a *1D Haar-like filter* similar to [17]. Due to the variable filter parameters this feature has shown some promising results for various classification problems [17] and is therefore adopted in our work as well. Furthermore, we calculate the *crest-factor*, *spectral flux*, *spectral centroid* and *spectral roll-off* features, which are mainly used for the classification of audio signals [29,44]. However, as demonstrated by Rahman et al. [32], those features are also suitable to differentiate between activities based on acceleration and orientation data. For each pair of axes $\{(x, y), (y, z), (z, x)\}$ we then apply a *biaxial 1D Haar-like filter* [17] and calculate the *correlation* between the corresponding axes. The *correlation* can be computed by dividing the covariance through the product of the standard deviations and is especially helpful to detect activities that involve movements in a single direction [33]. Finally, we compute the *signal magnitude area*, which is defined as the sum of the absolute acceleration values from each of the three axes [23]. It is used since it has proven to be a suitable indicator to distinguish between stationary and movement-related activities [22]. Overall, 52 features are calculated for each of the two sensors, thus resulting in a total of 104 head movement features.

6.2 Feature Window Analysis

In order to explore the impact of window lengths on classification performance, we generated all features for different window sizes (1-10 seconds) and measured the accuracy of each feature set. For every window length we thereby also varied the overlap between each of the windows from 0 to 90%. As it turns out, our results did not reveal an overlap ratio with significantly better performances compared to the others. However, since previous works have shown the most success with a 50% overlap between each window [2,8,18,33], we selected it in our approach as well.

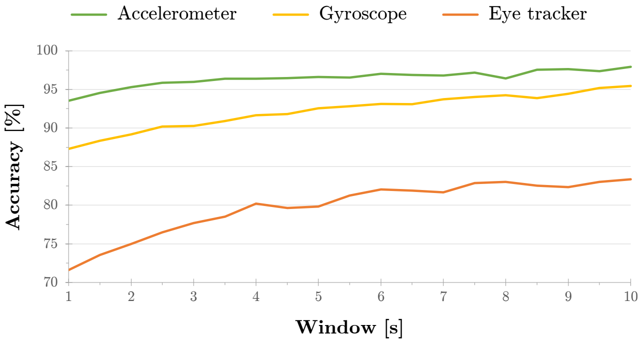

Fig. 5 Relation between window length and accuracy (50% window overlap)

Another interesting finding from our results is that the classification accuracy increases almost linearly with growing window sizes as shown in Figure 5. In order to achieve the highest possible detection rates, it therefore would make sense to use a longer window size as well. However, since our goal is to recognize the visual search behavior in real time, we can not use a very large window as it slows down the reaction time of our approach. Instead, we need to make a compromise between window size and detection rate, which is the reason why we chose a window length of four seconds.

6.3 Fusion and Feature Selection

After selecting a fixed window size, we applied various fusion techniques to combine the feature sets from the accelerometer, gyroscope and eye tracking sensors. During *early fusion* (feature level) the features from each modality are concatenated into a single feature vector before the classifier is trained [36]. As opposed to that, during *late fusion* (decision level) the individual classifiers for every modality are trained first and afterwards their predicted scores are combined [21]. For that, several methods can be applied including *AdaBoost*, *Borda count*, *Cascading Specialists*, *Dempster Shafer*, *Stacked Generalization*, *Weighted Majority Voting* or even simple rules such as the *Sum*, *Min*, *Max*, *Median* and *Product Rule* [24,25,27,28]. In our case the *Stacked Generalization* approach yielded the highest accuracy of those methods and is therefore used to achieve all further late fusion results. While both early and late fusion usually result in higher detection rates compared to the classification based on individual modalities, they also increase the required dimensionality of the input data. As a consequence all 164 features would need to be computed at the same time, which could cause performance bottlenecks in an online scenario. However, since not all features are equally useful in detecting the visual search activity, we employed the *sequential forward selection* (SFS) method to reduce the number of features and hence the required computational cost associated with it [41]. The feature selection was applied to the concatenated vector of all features for the early fusion as well as to each individual feature set for the late fusion. Using this technique we were able to reduce the number of required features by more than 50%. As shown in Figure 6, the feature distribution across all sensors stayed nearly the same after applying the feature selection, which indicates the importance of using a multimodal approach.

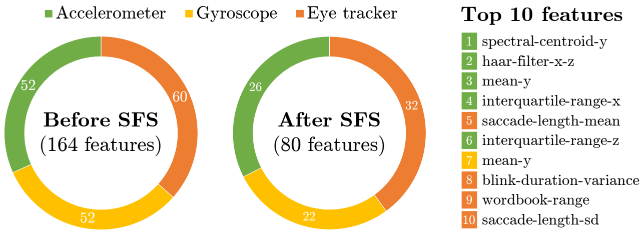

Fig. 6 Feature composition for early fusion

6.4 Classification Results

The final evaluation of our visual search detection approach is based on the reduced feature sets after applying the SFS feature selection. In compliance with all previous evaluations, we used the leave-one-user-out method to train and test our SVM models several times. Table 1 shows the average accuracy, precision and recall values for every modality as well as the results after early and late fusion. Overall, early fusion yielded the highest accuracy with 97.55%, closely followed by late fusion with a value of 97.39%. From the individual modalities the acceleration showed the highest accuracy, which is on par with the late fusion and only slightly lower than the early fusion results. Although

Source	**Accuracy**	**Precision**	**Recall**
Accelerometer	97.39%	97.65%	97.11%
Gyroscope	92.18%	94.14%	89.97%
Eye tracker	81.59%	82.67%	79.93%
Early Fusion	97.55%	98.11%	96.97%
Late Fusion	97.39%	97.47%	97.29%

Table 1 Classification results after feature selection for Baseline vs. Search

this might lead to the assumption that the accelerometer alone can be sufficient for the visual search detection, we still recommend using a combination of multiple modalities since it is more robust against signal fluctuations of individual sensors and therefore more reliable in real-world applications. Surprisingly, the gyroscope model yielded a five percent lower accuracy compared to the accelerometer even though both are based on the same initial feature set. The eye tracking model resulted in the lowest accuracy of 81.59%, which can be mostly attributed to the signal quality as described in section 5.1. Generally, all modalities and fusion methods showed high precision and recall values. This means that if visual search was detected, then it was usually correct (precision) and that almost all instances of visual search were recognized as such (recall).

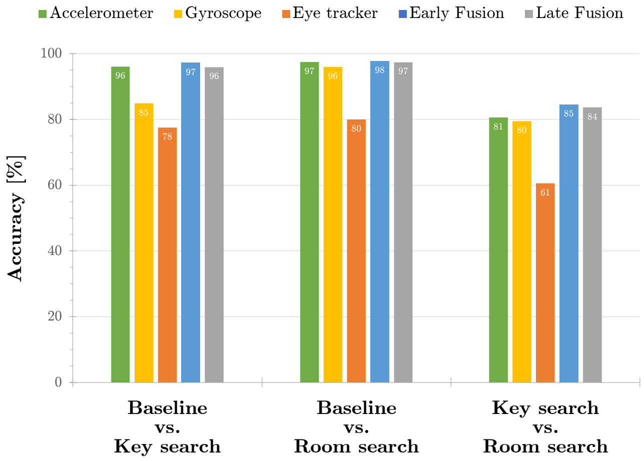

Fig. 7 Additional results for different search scenarios

In addition to the general detection of visual search we also investigated whether there are any differences when recognizing either of the two search scenarios from our user study and whether it is possible to distinguish between them. The results of this analysis are summarized in Figure 7. Interestingly, when trying to detect key search or room search individually, we achieve similar accuracies compared to the general detection of visual search. The only notable difference occurs in the accuracy of the gyroscope model which is seven percent lower for the key search and four percent higher for the room search. This could indicate that the head orientation is more distinctive when searching for large objects which might not fit into the field of view and require more head rotations, than when looking for smaller items such as keys. Using the same features as before we then tried to distinguish both scenarios from each other. As expected, the results were lower compared to the previous evaluations. However, we still achieved a reasonably high accuracy of 84.53% using the early fusion method, which could indicate that the target object type might have an influence on the search behavior. Information on the target object type would enable a system to provide more specific assistance to users after detecting visual search.

7 Discussion

Overall, our multimodal approach for the automatic detection of visual search proved to be successful and achieved recognition rates of up to 97.55%. Even though recognition rates for visual search depend a lot on the experimental setting and the nature of tasks, a look at recognition rates achieved for non-mobile and thus less challenging settings might be of interest. Henderson et al. [19] achieved accuracy rates of 83% for two tasks (search and memorization) involving scene stimuli presented on a monitor. As ours, their accuracy rates were clearly above chance level. When only considering the results based on eye movements, our approach yielded a two-percent lower accuracy rate than their approach. Due to the different nature of the search tasks (mobile vs. stationary setting), it is hard to compare results. Nevertheless, the use of a stationary eye tracker with a sample rate of 1000 Hz might have contributed to the better performance of their approach since it can record the eye movements a lot more precisely than our custom eye and head tracking device at 30 Hz.

Using a similar eye tracker as Henderson et al. [19] with a sample rate of 500 Hz, Coco et al. [9] achieved a recognition rate of 88% for the visual search task among three classes (*visual search*, *scene description* and *object naming*). They used the ten-fold cross-validation method to evaluate the classification performance which does not produce subject-independent results as opposed to the leave-one-user-out method applied in our work. When testing their classification model on new users, their approach is likely to suffer from a decrease in recognition rates. When considering the differentiation between both search phases, our approach yielded an accuracy of more than 84%. Similar classification results were reported by Haji-Abolhassani and Clark [16] for search tasks that involved distinguishing objects from distractors by a single feature (easy setting) or a combination of features (difficult setting).

Due to the very different experimental settings and nature of tasks, it is hard to conduct a comparison with other work on visual search. Summing up, it may be said, however, that the results we obtained "in the wild" are competitive with results achieved in stationary settings even when limiting ourselves to eye tracking features and excluding head movement features.

In order to examine how the key search and the room search could be distinguished effectively and why head movements resulted in higher accuracies (81%) than eye movements (61%), we now discuss the results of an exploratory analysis of both search phases using the recordings of participant #16. Generally, pupil dilation is influenced by a number of factors, such as the brightness and color of the surroundings. However, it also occurs during mental activity [3] and when a person is in an emotional state [31]. Since visual search is a confusing task that requires a lot of cognitive processing, we expected to see an increase in pupil size during the search phases. A higher amount of eye movement was also expected to be an indicator of search as well as the type of search, so we calculated the duration of saccades in a window of 4s. We also expected the participant to have different types of head movements while searching, so we used the y-axis values of the accelerometer data from our head mounted tracking device and calculated the standard deviation. A plot of the calculated data can be seen on Figure 8. When looking at the phases where no search (questionnaire segment at the beginning and Ⓕ) or a limited amount of search (Ⓒ) occurs, a very small pupil size, low amount of saccades and very little head movement can be observed. This is in contrast to the search periods (Ⓐ, Ⓑ and Ⓔ), where we can see a much larger pupil size, a higher amount of saccades and lots of head movement. The search types themselves can partially be distinguished by whether a search for a key (Ⓐ) or a search for a room in a hallway (Ⓑ and Ⓔ) was conducted. While pupil size during Ⓐ is very similar to Ⓑ and Ⓔ, overall saccade duration is about twice as high on average when searching for a room as opposed to a key. As expected, head movement during room search is much more prevalent than during key search, since the participants had to move their head to read the signs on the doors whereas the key was in a small area. This might give an explanation as to why the two types of search are easier to differentiate by head movement than by the given eye tracking data.

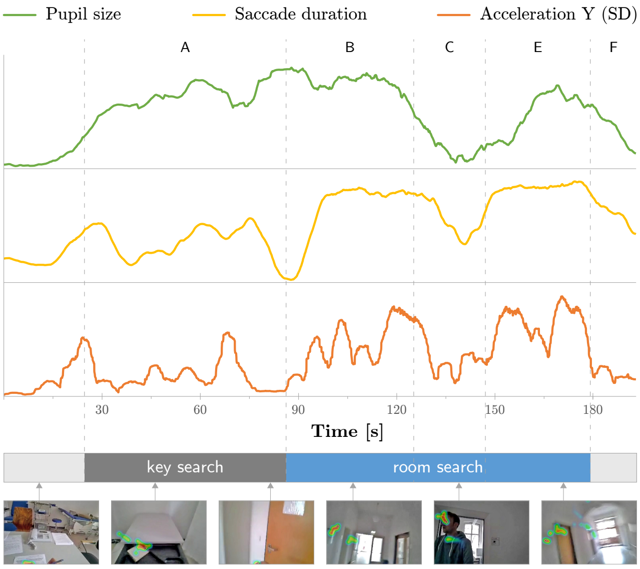

Fig. 8 Participant's pupil size (mvavg over 8s, shifted left by 4s), saccade duration in a 4s window (mvavg over 4s) and SD of the accelerometer's y-axis in a 4s window. Time Periods A-F as described in section 4.1

8 Conclusion

The aim of this work was the automatic detection of visual search for older adults in a completely mobile and real-world scenario. Since capturing sensor data in such a setting involves great challenges, we developed a custom eye and head tracking device to overcome these requirements. A study we conducted indicates the feasibility of tracking search-relevant behavioral data in the wild. Considering the usage in an online scenario, the obtained classification results are promising. One might argue that the classifier basically learns the characteristics of body movements, which are only indirectly related to the search. However, our experiment has shown that the eye reveals distinctive information on the different search tasks as well. Nevertheless, further studies are necessary to analyze specific characteristics of search tasks. In our future work, we will therefore focus on a broader range of visual search activities (e.g. searching for a room with a red door versus searching for a room with a specific number). Finally, we will also investigate different strategies to support the visual search process of older adults, after it has been detected. Possible options could be guiding the users with instructions or showing an image with the last position of the object on the head mounted display.

Acknowledgements This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under project Glassistant (FKZ 16SV7270).

References

1. Altun, K., Barshan, B.: Human activity recognition using inertial/magnetic sensor units. In: Human-Behavior Understanding, pp. 38–51. Springer (2010)
2. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Pervasive Computing, vol. 3001, pp. 1–17. Springer (2004)
3. Beatty, J.: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin **91**(2), 276 (1982)

4. Bixler, R., D'Mello, S.: Toward fully automated person-independent detection of mind wandering. In: User Modeling, Adaptation, and Personalization, vol. 8538, pp. 37–48. Springer (2014)
5. Bulling, A., Ward, J.A., Gellersen, H., Troster, G.: Eye movement analysis for activity recognition using electrooculography. IEEE transactions on pattern analysis and machine intelligence **33**(4), 741–753 (2011)
6. Buswell, G.T.: How people look at pictures: a study of the psychology and perception in art. University of Chicago Press Chicago (1935)
7. Castelhano, M.S., Mack, M.L., Henderson, J.M.: Viewing task influences eye movement control during active scene perception. Journal of vision **9**(3), 6.1–15 (2009)
8. Chen, Y.P., Yang, J.Y., Liou, S.N., Lee, G.Y., Wang, J.S.: Online classifier construction algorithm for human activity detection using a tri-axial accelerometer. Applied Mathematics and Computation **205**(2), 849–860 (2008)
9. Coco, M.I., Keller, F.: Classification of visual and linguistic tasks using eye-movement features. Journal of vision **14**(3), 11 (2014)
10. Damian, I., Baur, T., André, E.: Measuring the impact of behavioural feedback loops on social interactions. In: Proc. ICMI, pp. 201–208. ACM (2016)
11. Damian, I., Dietz, M., Gaibler, F., André, E.: Social signal processing for dummies. In: Proc. ICMI, pp. 394–395. ACM (2016)
12. Dietz, M., Garf, M.E., Damian, I., André, E.: Exploring eye-tracking-driven sonification for the visually impaired. In: Proc. AH, pp. 5:1–5:8. ACM (2016)
13. Dietz, M., Schork, D., André, E.: Exploring eye-tracking-based detection of visual search for elderly people. In: Proc. IE, pp. 151–154. IEEE (2016)
14. Findlay, J.M., Gilchrist, I.D.: Eye guidance and visual search. Eye Guidance in Reading and Scene Perception pp. 295–312 (1998)
15. Greene, M.R., Liu, T., Wolfe, J.M.: Reconsidering yarbus: a failure to predict observers' task from eye movement patterns. Vision research **62**, 1–8 (2012)
16. Haji-Abolhassani, A., Clark, J.J.: A computational model for task inference in visual search. Journal of Vision **13**(3), 29 (2013)
17. Hanai, Y., Nishimura, J., Kuroda, T.: Haar-like filtering for human activity recognition using 3d accelerometer. In: 13th Digital Signal Processing Workshop, pp. 675–678. IEEE (2009)
18. He, Z., Jin, L.: Activity recognition from acceleration data based on discrete consine transform and svm. In: Proc. SMC, pp. 5041–5044. IEEE (2009)
19. Henderson, J.M., Shinkareva, S.V., Wang, J., Luke, S.G., Olejarczyk, J.: Predicting cognitive state from eye movements. PloS one **8**(5), e64,937 (2013)
20. Huynh, T., Schiele, B.: Analyzing features for activity recognition. In: G. Bailly, J.L. Crowley (eds.) sOc-EUSAI, p. 159 (2005)
21. Kächele, M., Werner, P., Al-Hamadi, A., Palm, G., Walter, S., Schwenker, F.: Bio-visual fusion for person-independent recognition of pain intensity. In: F. Schwenker, F. Roli, J. Kittler (eds.) Proc. Multiple Classifier Systems, pp. 220–230. Springer (2015)
22. Khan, A.M., Lee, Y.K., Lee, S.Y.: Accelerometer's position free human activity recognition using a hierarchical recognition model. In: Proc. Healthcom, pp. 296–301. IEEE (2010)
23. Khan, A.M., Lee, Y.K., Lee, S.Y., Kim, T.S.: A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. IEEE T-ITB **14**(5), 1166–1172 (2010)
24. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
25. Knauer, U., Seiffert, U.: A comparison of late fusion methods for object detection. In: Proc. ICIP, pp. 3297–3301 (2013)
26. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE Communications Surveys & Tutorials **15**(3), 1192–1209 (2013)
27. Lingenfelser, F., Wagner, J., André, E.: A systematic discussion of fusion techniques for multi-modal affect recognition tasks. In: Proc. ICMI, pp. 19–26. ACM (2011)
28. Lingenfelser, F., Wagner, J., Vogt, T., Kim, J., André, E.: Age and gender classification from speech using decision level fusion and ensemble based techniques. In: Proc. INTERSPEECH, pp. 2798–2801 (2010)
29. Lu, H., Pan, W., Lane, N.D., Choudhury, T., Campbell, A.T.: Soundsense: Scalable sound sensing for people-centric applications on mobile phones. In: Proc. MobiSys, pp. 165–178. ACM (2009)
30. Mills, M., Hollingworth, A., van der Stigchel, S., Hoffman, L., Dodd, M.D.: Examining the influence of task set on eye movements and fixations. Journal of vision **11**(8), 17 (2011)
31. Partala, T., Surakka, V.: Pupil size variation as an indication of affective processing. International Journal of Human-Computer Studies **59**(1), 185–198 (2003)
32. Rahman, S.A., Merck, C., Huang, Y., Kleinberg, S.: Unintrusive eating recognition using google glass. In: Proc. PervasiveHealth, pp. 108–111. ICST (2015)
33. Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: Proc. IAAI, pp. 1541–1546. AAAI Press (2005)
34. Sadasivan, S., Greenstein, J.S., Gramopadhye, A.K., Duchowski, A.T.: Use of eye movements as feedforward training for a synthetic aircraft inspection task. In: Proc. CHI, pp. 141–149. ACM (2005)
35. Salthouse, T.A., Babcock, R.L.: Decomposing adult age differences in working memory. Developmental Psychology **27**(5), 763 (1991)
36. Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: Proc. MULTIMEDIA, pp. 399–402. ACM (2005)
37. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: 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)
38. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology **12**(1), 97 – 136 (1980)
39. Verghese, P.: Visual search and attention: A signal detection theory approach. Neuron **31**(4), 523–535 (2001)
40. Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., André, E.: The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time. In: Proc. MM, pp. 831–834. ACM (2013)
41. Webb, A.R.: Statistical pattern recognition. John Wiley & Sons (2002)
42. Weyerer, S., Bickel, H.: Epidemiologie psychischer Erkrankungen im höheren Lebensalter. Kohlhammer (2007)
43. Yarbus, A.L.: Eye movements during perception of complex objects. Springer (1967)
44. Yatani, K., Truong, K.N.: Bodyscope: A wearable acoustic sensor for activity recognition. In: Proc. UbiComp, pp. 341–350. ACM (2012)