Activity recognition from egocentric view

Emma Janssen
STUDENT NUMBER: 2029453

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Dr. Andrew Hendrickson
Dr. Pieter Spronck

Tilburg University
School of Humanities and Digital Sciences Department of Cognitive Science & Artificial Intelligence Tilburg, The Netherlands
December 2019

Preface

In September 2018 I started the master Data Science for business and governance. With a background in Communication- and Information sciences I was up for a challenge. My goal was to broaden my skills and knowledge in the beta field. Besides, gaining more experience in data science seemed very useful for my future career. The last 1.5 year has been hard work and it has tested my perseverance and patience. However, my interest in the field has definitely not decreased. I am looking forward to learning even more in practice.
I would like to thank my thesis supervisors Dr. Martijn van Otterlo and Dr. Andrew Hendrickson for their great support and feedback during the thesis process. In addition to the hard work, there was also a lot of fun during the group meetings. Besides, I am thankful for the help of Dr. Frouke Hermens and providing me with an interesting dataset. Last, I would like to thank my family and friends for the boost I needed during stressful times this year.

Emma Janssen

Tilburg, December 2019

Activity recognition from egocentric view

Emma Janssen

There has been considerable interest in the recognition of activities in day-to-day tasks. The coordination of movements and gaze play an important role in this process. With the development of wearable cameras, eye movements can be analysed from egocentric view when performing activities of daily life (ADL). More research on this subject could result in more knowledge on activity recognition which could contribute to research in the practical field (healthcare). This study aims to predict to what extent an activity is performed in ADLs, based on eye movements from egocentric view. This is done by the annotation and analysis of wearable videos from six participants while performing the activity of tea-making, which consisted of several smaller actions (e.g. pouring water, finding cup etc.). The Random Forest technique was used in order to find an answer to our research question. Analysis showed that there were no major differences in performance between models. However, the models performed better than the majority baseline score. It could be concluded that the models used in this study add value in predicting activities. In addition, analysis of the performance of actions separately was conducted. Analysis showed a difference between in the predictability of actions.

Keywords: activity recognition, egocentric view, classification, ADL, prediction

1. Introduction

Vision plays an important role in our daily lives. We need it to perform actions, such as preparing food, doing the dishes or reading a book. While performing these actions we locate objects and manipulate them. For example, when preparing food, we look for a plate and a knife. There has been considerable theoretical interest in understanding how movements and gaze are coordinated in day-to-day tasks, such as making tea, and classifying the types of monitoring action that the eyes perform. Land, Mennie and Rusted (1999) question whether these eye movements are random or related to the requirements of the motor task and if fixations are directed specifically to the places from which information is needed. This study is the first where movements of the eyes are studied in a natural setting instead of in a laboratory, due to new developments in recording devices. Land & Hayhoe (2001) studied the relation between on-going motor actions and the eye movements that accompany them as well. They conclude that the eye movements are not just passive responses to circumstances, but shift in advance of each action seeking the object and information. The researchers have suggested that individual eye-fixations fall into categories, called locating, directing, guiding and checking.

The identification of human activities is made possible by the improvement of wearable video devices. Examples of these devices are a GoProR or Tobii as shown in Figure 1. These devices track human behaviour from a self-centered view, which is called an egocentric view. This is different from the third person view; because of the different perspective of the camera and the change in camera motion profile. Eye-tracking research

can be conducted using this device, because eye movements and fixations can be measured.

Figure 1
Example of a wearable camera device, Tobii Pro glasses.

Several studies on activities of daily life (ADLs) identify all objects in egocentric videos that characterize the wearer's behaviour. According to Singh et al. (2017) there is a difference between activities and actions. Activities refer to a higher level of what the person does at a particular moment in time, while actions are usually smaller parts of an activity. One way to define behaviour in a video frame is by distinguishing certain activities. For example, the actions 'Finding a cup', 'putting a kettle on' and 'finding tea' can define the activity of 'Tea making'.

Fathi et al. (2011) describe three reasons why videos from an egocentric view could be advantageous for research that concerns object manipulation. First, the obstructed views of objects will be minimized, because the workspace containing these objects is visible. Next, viewing directions of objects are presented consistently. And last, objects tend to appear in the centre of the image, which results in high quality measurements. However, there is a difference in passive versus active eye tracking. In passive eye tracking, participants are shown stimuli without interacting in these actions, e.g. pictures or videos. Though in active eye tracking, eye movements of participants are tracked when performing an action themselves and interfere with objects (Blascheck, Kurzhals, Raschke, Burch, Weiskopf & Ertl, 2017). Intille, Larson, Tapia, Beaudin, Kaushik, Nawyn, and Rockinson (2006), state that there is a need for research on automatic recognition of activities of daily life (ADL) in home-settings. More specifically, the need for comprehensive testing, fully annotated datasets and complex, naturalistic environments.

There are also practical implications. People suffering from functional deficits from serious injury or neurodegenerative disorders often experience difficulties in ADL. Patients have a lack of autonomy and a high need for care. While much information on the cognitive status of the patients is identified in most cases, the consequences for ADLs are often uncertain for doctors and caregivers. Reduced functioning of learning, attention, concentration, information processing, orientation and memory affect the ability to perform ADLs. However, there is a need for more ways to evaluate patients' performance that approximate the real-world. In addition, Pirsiavash & Ramanan (2012) describe two reasons why research on ADLs can contribute to the medical field. Currently, evaluations of ADLs are done in hospitals. Developing computer-vision systems that can analyse these activities would make long-term monitoring at home possible (tele-rehabilitation). In this way patients can live in their residual areas for a longer time. Another possibility

would be the tracking of personal visual activities (life-logging), which could improve quality of living for patients with memory-loss.

It is clear that there is a need for more research done on activity recognition on ADLs. Using a new dataset could be relevant for the purpose of action recognition to improve understanding and representativeness of the targeted tasks (Gonzalez et al. 2015), the need for research in complex, naturalistic environments, comprehensive testing and annotated datasets (Intille et al. 2006). Besides, activity recognition in ADLs could contribute to research for people with disabilities, dementia, motor disabilities or other fields in healthcare that obstruct individuals from performing tasks in ADLs (Ramzaoui, Faure, Spotorno, 2018; Gerber, Müri, Mosimann, Nef and Urwyler, 2018; Gulde et al., 2014; Megret et al. 2010 and Gonzalez et al. 2015). Activity recognition can contribute to detecting patients' habits and understanding what activities lead to another. Research on healthy individuals can contribute to research on diseases, because understanding habits of healthy individuals is informative for research on patients with certain diseases (Ramzaoui, Faure, Spotorno, 2018).

Therefore it would be interesting to investigate activity recognition in ADLs in egocentric view on a new dataset. This could contribute to the reasons as described above. This leads to the following research question and sub-questions:

To what extent can actions performed in ADLs be predicted, based on eye movements from egocentric view?

- *What features are good predictors for actions?*
- *To what extent do actions differ in predictability?*

2. Related Work

There is a strong need for the automatic detection of activities in home-settings (Intille, Larson, Tapia, Beaudin, Kaushik, Nawyn, and Rockinson. 2006). However, research must overcome three challenges. The first challenge is the need for comprehensive testing. Testing in real-life settings is often costly and logistically difficult, making researchers choose for simulations of real-life settings in a laboratory and choosing small sample size. Besides, the need for annotated training datasets exists. Not many datasets that include participants performing activities of daily life are available. Moreover, the annotation of these datasets relies on participant recall or diary recordings, which are sensitive to errors. Last, complex, naturalistic environments are needed to design and develop applications that are aware of the home-context. When performing tasks, people naturally experience interaction with other people, dealing with other objects than those necessary for the task, multi-task and get interrupted. A laboratory environment does not make it possible to include these factors. The solution to these challenges is 'living laboratories'. These laboratories are natural environments that contain instruments such as sensors, which are useful for activity recognition. However, these laboratories need to compromise between two factors: realism of the environment and the quality of the sensors that are used. Yet, it improves experimental quality compared to traditional experiments and small studies.

Activity recognition of activities of daily life can be carried out with the use of sensors. These sensors collect sensor data streams which can provide information for activity recognition. This method is becoming more popular because it ensures privacy, low cost, fast deployment and flexibility (Jafari et al., 2005). According to Sarkar, Lee and Lee (2010) there are three ways in which activities can be recognized by using sensors. The first type of sensors is the universal and simple sensors which are deployed in the environment. Another type of sensors are video cameras that are built into the roof

or walls of the environment and last, wearable sensors that are worn by or attached to an individual (augmented reality devices (such as Google Glass). wearable health tracker devices, smartphones, smartwatches).

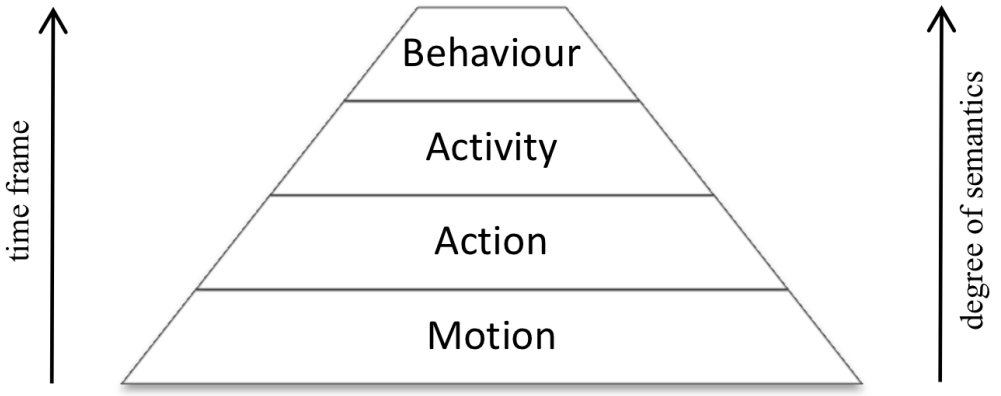
Intille et al. (2006) were the first to study activity recognition using universal and simple sensors. They implemented these sensors into a home environment called 'The PlaceLab' and used a Naïve Bayes classifier for activity recognition. However, this method showed low accuracy on recognition. The use of inbuilt cameras in the environment was studied by Wilson and Atkeson (2005); using binary sensors (contact switches, pressure mats, motion detectors and break-beam sensors). This study introduced the simultaneous tracking and activity recognition (STAR) method. Wilson and Atkeson studied whether participants were active or not and whether a room was occupied. Additionally, they counted participants in the room, tracked their movements and identified the participant. Zajdel et al. (2007) indicate how wearable sensors can be implemented in activity recognition. They used this method to recognize tasks performed during a meeting. Some of the challenges in this study were: the difficulties of signal analysis, participants were not always comfortable with wearing these sensors and the costs of the sensors.

Activity recognition is the understanding of how movements and gaze are coordinated in day-to-day tasks. Studies include many different approaches in activity recognition. Sensors such as GPS have been used to track people's driving behaviour (Zhou & Curry, 2015), find movement paths in buildings (Kang & Han, 2015), and recognizing tasks in an office environment (Oliver, Horvitz & Garg, 2002), such as interactions between individuals (Clarkson, Sawhney & Pentland, 1998). Wearable cameras are used to recognize behaviour of sitting, standing and walking (Ryoo & Matthies, 2013; Kitani et al., 2011), short-term and long-term actions (Poleg et al., 2015) and sign language as well as recognizing sports such as American football or basketball (Intille & Bobick 1999; Jug, Perš, Dežman & Kovačič, , 2003) .

2.1 Activity recognition

According to Padilla-López, Chaaraou and Flórez-Revuelta (2015) the analysis of tasks of human behavior can be classified into four dimensions, called motion, action, activity and behavior as shown in Figure 2. Events that consist of only a small amount of frames or seconds belong to the motion level, which includes for example saccades or head motion (further explained in section eye movements). In addition, object recognition, gaze estimation, foreground segmentation and hand detection belong to the motion level. One dimension higher is the action level, which consists of simple events of a longer timeframe. For example, "put kettle on", "finding cup", and "finding sugar". Sequences of actions that take up minutes or hours can be assigned to the activity level. This level also differs from the action level in the complexity of interactions between people and objects. Examples of these sequences are watching television or preparing food.

Figure 2
The visualization of the four dimensions of human behaviour according to Padilla-López et al. (2015).



Betancourt, Morerio, Regazzoni, & Rauterberg (2015) define two approaches for activity recognition: object-based and motion-based recognition. They state that information about activities is given by objects and their relation to hands. Objects and hands can be detected by analysing sequences of video frames that are acquired by an egocentric camera. Another way of recognizing objects is by allocating areas of interest (AOI). The location, frequency and poses of objects provide information for the action level. One way to detect the AOI is by tracking the focus of attention, based on the theory that people direct their attention in the area that they are manipulating. This can be done by using an eye tracker (Padilla-López, Chaaraou & Flórez-Revuelta, 2015).

2.2 Visual attention

The human eyes contain an element called the fovea; this part is sensitive to details in vision. By moving the eyes in the direction of the AOI the information of the AOI can be processed highly detailed. According to Bundesen et al. (2005) attention is viewed as 'selectivity in perception'. Through attention, it is chosen what information from the visual input is necessary for further processing. Due to limits of our brain capacity, attention is necessary to save energy available in the brain (Lennie, 2003). Attention enables a person to distinguish relevant information, objects and locations from less relevant information, objects and locations (Carrasco, 2011).

Several studies show that it is possible to link gaze of the eyes to attention (Deubel and Schneider, 1996; Corbetta, 1998; Henderson, 2003). This means that when an eye movement is directed to a different place, there is a shift in attention in that direction as well. Nevertheless, gaze is not always an indicator of attention. The attention can be on another location than the gaze (Vickers, 2009).

2.3 Eye movements

Land et al. (1999) define four ways in which eye movements provide useful information when performing tasks. The rapid movements of the eyes in between periods of fewer

movements of the eyes (fixation) are called ‘Saccades’. Both the number of saccades and the distribution of the intervals can be measured to acquire information of eye behaviour during certain actions (Land, Mennie & Rusted, 1999). Land et al. (1999) also specified different roles for fixations of the eyes. These are called: locating, directing, guiding and checking. Moreover, relocating objects is also a function of eye movements: how does the eye find the next object that is needed in the sequence? For example: If the object is located before and/or in the very recent past, gaze could be redirected faster (in one saccade) because of the place memory of the object. For other objects, it could take more saccades to find the object. Last, objects that are not fixated on offer information as well. Land et al. (1999) state that the eyes cannot fixate on everything relevant to the task. There are three frequent rules: Hands are barely fixated on. Second, objects that the hands already contacted are barely fixated on again. For example, when filling the kettle with tap water, the fixation is mostly on the stream of water instead of the tap. Certain familiar objects can be manipulated without visual involvement.

2.4 Different gaze behaviour

Three different types of gaze behaviors can be determined by cognitive psychologists: fixations, pursuit tracking and saccades. First, a fixation can be explained when the gaze is on a location or object for 100ms or longer (Optican, 1985; Carl and Gellman, 1987; Carpenter, 1988). 100ms is the threshold to recognize an object and when performing a simple movement, 180ms is needed. Second, pursuit tracking is the following of a moving object, where information can only be processed if the gaze is stabilized on the object. Last, when the eyes move from one location or object to another quickly, saccades occur. These rapid eye movements happen on average with three saccades per second with each lasting from 60ms to 100ms (Vickers, 2009).

2.5 Top down and bottom up attention

Attention can be divided into two types, *top-down control of attention* and *bottom-up control of attention*. Top-down control is described as goal driven attention. Yarbus (1967) stated that eye movements were different depending on the goals and were driven to parts of the visual scene that were related to the tasks being performed. Visual saliency is the aspect of bottom up control of attention. It can be defined as a result of visual contrast, due to parts that stick out from other parts. Several studies include visual saliency in their computational models on attention and eye movements (Bruce & Tsotsos, 2009; Rothenstein & Tsotsos, 2008).Top down and bottom up control can also interact. For example, when looking for an object with a certain colour, attention goes to all objects in the room with that specific colour. Folk et al. (1992) state that interaction of the two types can improve short term memory and attention capture.

2.6 Task relevance

Triesch, Ballard, Hayhoe and Sullivan (2003) present the importance of task requirements in the determination whether information is selected and stored in the memory or not. Moreover, tasks that require active participation of the person demand different information than tasks that require analysis of certain images or videos. Besides, the type of stimulus can create differences in which information is stored, such as viewing two-dimensional or three-dimensional displays. Depth information plays an important role in spatial complexity of the vision and results in greater demands for the visuomotor system. As stated in Xu and Nakayama (2003), when performing multiple actions, the eyes, head

and hands are all involved in the coordination and control of movements. Visual information from fixations is necessary to intend movement of the eyes and hands.

Furthermore, representation of spatial structure is necessary in an activity that consists of several fixations. Hayhoe, Shrivastavah, Mruczek, and Pelz (2003) showed this by analyzing eye- and hand coordination of subjects when they were instructed to make sandwiches. This spatial structure could contribute to the coordination of eye- and hand movements in sequences of actions. Aivar, Hayhoe, Chizk & Mruczek (2005) studied transsaccadic memory to understand what information is retained when guiding eye movements. They concluded that when an object is moved, the number of fixations that is required will increase in order to relocate the object after the change. This is in line with previous research (Hayhoe, Shrivastavah, Mruczek & Pelz 2003) which states that the spatial structure of an environment is retained across fixations and used for the guiding of eye movements.

2.7 Eye tracking research

Much research on activities of daily life (ADL) has been conducted using eye tracking devices. Research has focused on different locations and scenarios, such as tasks in an office, kitchen or other at home activities. Ogaki, Kitani, Sugano & Sato (2012) considered activity recognition in an office, using features extracted from an outside looking camera and features from an inside looking camera. By combining eye-motion and ego-motion classification for five different office tasks an average precision of 57 per cent was achieved. Taralova, De la Torre and Hebert (2011) studied recognizing egocentric activities in a kitchen, where participants had to prepare different types of food. Moreover, Poleg, Arora and Peleg (2014) also used egocentric camera video to analyse food preparation tasks. The researchers presented the learning of a hierarchical model of an activity that shows the hierarchical relationship between hands, objects and actions.

Fathi et al. (2011) studied the roles of objects and hands while performing daily tasks. Their research presents learning a hierarchical model of activities by using the appearance of objects, hands and actions from egocentric view. They conclude that combined modelling of objects, hands and actions results in a higher performance compared to the situation where they are modelled separately. While performing a task, objects naturally change states. In further research, Fathi et al. proved that changes in the state of the objects support action recognition as well. Their new model outperformed their previous model (32.4%) by 39.7%.

On the other hand, Singh et al. (2016 b) propose an uncombined model of hands and objects. Compared to publicly available datasets, they raised the performance of their classifier with over 11 percent. Their model can also be applied to videos where objects or hands are not visible. Ryoo and Matthies (2013) take a different approach on this subject. Their objective is to make sure an observer (wearing a wearable camera) understands what activities others are doing. They distinguish two types of interactions: friendly interactions (“a person hugging the observer”) and hostile interactions (“a person hitting the observer”). Eye movements were also used for activity recognition by Shiga et al. (2014). They studied six tasks in an office environment and received an accuracy of 90 %.

Gonzalez et al. (2015) have shown that activity recognition in egocentric view can be conducted, by classifying the combination of two sources of information. First,

active objects that are manipulated or observed by the user. Second, the context of where the actions take place matters. However, they state that activity recognition in unconstrained scenarios is still challenging. The fusion of additional sources of information is necessary, as well as the collection of more wearable video data. More wearable video data will provide a better understanding of the target tasks and will prevent high weighting factors. In this way the representativeness of the training sets increases.

In addition, several different techniques are compared in activity recognition research on ADLs. Spriggs et al. (2009), studied the classification of cooking and food preparation activities that were performed in a natural setting. They classified data from several video frames into actions and classified the overall task that was performed. Spriggs et al. found that their K nearest neighbour model (KNN) outperforms the Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) with 57.8 % when classifying frames, due to high dimensionality of the data. Pirsiavash and Ramanan (2012) described object use over time by proposing an activity representation based on temporal pyramids, which show activities as sequences of objects and locations. A Support Vector Machine (SVM) was used to train spatiotemporal interest points (STIPS). However, the model did not perform well with 16.5 % on pre-segmented video clips. The performance increases with 22.8 % when temporal pyramids are added to the model. The same applies to the bag- of- objects model that is trained. Bags-of-object models are trained by counting the occurrences of objects. This model increases performance to 32.7 %. According to Nguyen, Nebel and Florez-Revuelta (2016) SVMs have shown to be the most frequently used tool for training and recognizing objects. Kumar and Bhavani (2017) recognized human activity from egocentric view with comparing several approaches: probabilistic neural network (PNN), SVM, KNN and combined SVM + KNN classifiers. For feature extraction several methods are used: Gray-level co-occurrence matrix (GLCM), the local binary pattern (LBP) and Speeded up robust features (SURF). GLCM uses the spatial relationship of the pixels of the video. The LBP method defines the relationship between the pixel and its neighbours. Last, the SURF method detects local features used in computer vision tasks. Results showed the SVM + KNN classifier performed better for all methods (GLCM, LBP and SURF) than the other classifiers in this research.

2.8 Healthcare

There is a need for more research on ADLs in healthcare. Patients experience difficulties in performing ADLs due to functional deficits or neurodegenerative disorders. Reduced performance of ADLs is caused by the decrease in learning, attention, concentration, information processing, orientation and memory of the brain. Tele-rehabilitation would make it possible to monitor patients at home for a longer term. Nowadays ADLs are evaluated in hospitals, developing systems that analyze ADLs at home would give patients the freedom to live in their residual areas for a longer time. Besides, life-logging could improve performance of ADLs for patients suffering from memory-loss as described in Pirsiavash and Ramanan (2012).

Recently the use of virtual reality (VR) in neurorehabilitative therapies became one of the most promising new technologies. This technology shows a virtually simulated scenario that is a representation of a real world situation in a controlled and safe environment. This technology easily adjusts to the specific needs of patients, which is an advantage relative to real world care. Gerber, Müri, Mosimann, Nef and Urwyler (2018)

have shown that simulated ADL implemented in VR technology could contribute to diagnostic and rehabilitative purposes for patients with functional disabilities.

Gulde et al. (2014) studied eye movements, hand kinematics and the dynamics of manipulated objects during the task of tea-making for Cerebrovascular accident (CVA) patients, which suffer from compromised access to motor concepts relevant for ADLs. They state that information from gaze fixation could be an early indicator of unfit errors. This could help with preventing errors in the future. Mégret et al. (2010) first used egocentric video for recording ADLs on patients with dementia. In the early stage of dementia, patients' performance of their everyday activities decreases. Another challenge of dementia is that caregivers are continuously charged with the care for the patient. Moreover, researching eye behaviour in real-life settings could help patients suffering from Alzheimers' disease maintain their residual areas of functioning by adapting objects and living area to their specific needs. Ramzaoui, Faure & Spotorno (2018) studied the limitations of Alzheimer's disease (AD) on visual search, which means the scanning of the environment when looking for certain objects or locations. This deficit causes patients to have difficulties with finding objects efficiently and on time. This deficit is caused by attention and memory mechanisms. There has been some research on visual search in AD. However, more research in real-world scenes and settings is needed to find in what way these deficits affect the functional autonomy of the patient. Moreover, these researchers highlight the value of studying healthy individuals in the investigation on visual search in Alzheimers' disease.

2.9 Research gap

Previous research on the automatic detection of ADLs states that there is still a strong need for more research on this subject. Especially, conducting studies on ADLs in a real-life setting instead of laboratory environments is necessary, since there aren't many datasets available in this particular area. Besides, more annotated training datasets could contribute to the automatic detection of ADLs. Moreover, more research on the performance of specific tasks could improve the understanding of these tasks.

Furthermore, conducting more research on activity recognition of ADLs could benefit healthcare. Eye movements are an indicator for errors in the motor system. Better understanding of eye movements when performing ADLs could provide information of patients with functional disabilities or memory loss. This could serve diagnostic and rehabilitative purposes and keep functional autonomy of patients with dementia for a longer time. As stated in Ramzaoui, Faure & Spotorno (2018), studying activity recognition of ADLs on healthy individuals is important.

The annotation and analysis of a new dataset could contribute to this. By using the videos of Ioannidou, Hermens and Hodgson (2016), questions on eye-motion level and action level could be answered by analysing the movements and fixations of the eyes and analysing sequences of video frames. Moreover, research on activity recognition could be performed on this data set by the classification of the different tasks in the videos.

3. Experimental Setup

To recognize actions from eye movements a classification problem was to be solved. The goal is to accurately predict to which class new input belongs to. In this case those classes are the different actions performed while making tea.

3.1 Participants

Data from Ioannidou et al. (2016) was employed. They used forty-eight participants in their study. These participants were all students from the University of Lincoln. Data from forty-two participants was analyzed, of which 14 males and 28 females. Ages come from the interval [18, 46] (mean 21.38, SD 5.18). All of these participants had normal vision or corrected vision with contact lenses.

3.2 Design & Procedure

Apparatus

This study used a wearable video camera, called 'Tobii Pro 2 ultralight head mounted eye tracker'. This camera is a pair of glasses with a small recording component. This component analyses and stores eye movement data. The camera consists of a video resolution of 1920 to 1080 pixels. Scene views are sampled at 25 Hz and eye gaze data at 50 Hz. To calibrate the camera, the device needed to be held 1.5 meters in front of the participant. In this way, the system could calculate gaze position at different viewing distances.

Design

Ioannidou et al. (2016) instructed participants to perform three tasks, a navigation task, a tea making task and a card sorting task. For the present study, only video data from the tea making task was used. The study of Land et al. (1999) inspired Ioannidou et al. (2016) to use the tea-making task in their study. Participants were instructed to make a cup of tea in a kitchen. To perform this task, specific items were needed which one could find in the cupboards of the kitchen. These items included a mug showing coloured butterflies, a green jar with the word 'tea' written on it, a red jar with the word 'sugar' written on it, a small bottle of milk, which one could find in the fridge and a tea spoon. These items were to find amongst other kitchen items in a typical kitchen environment. Participants further received the instructions to act naturally and they could take as much time as needed for the task.

In order to predict what activity is performed during ADLs from egocentric view, a classification method will be used. A dataset that contains all activities (classes) and their attributes (features) needs to be constructed from the video-data. In order to do so, the video- data needs to be annotated to determine what activity was performed and what object was in sight during time frames of the video. By performing descriptive analysis on this data, features will be selected that could contribute to the prediction task. These features will be combined into a new dataset that represents each action (class) and its features. In this way, the classification algorithm can determine what attributes characterize certain classes in order to be able to predict to what class new input data belongs to. This study will take a multi-class classification approach, because there will

be more than two distinct classes (multiple smaller actions during the tea-making activity). Several models will be compared to find out what features are good predictors for actions. In addition, performance per action will be analysed in order to determine if there is a difference in predictability between actions.

3.3 Data

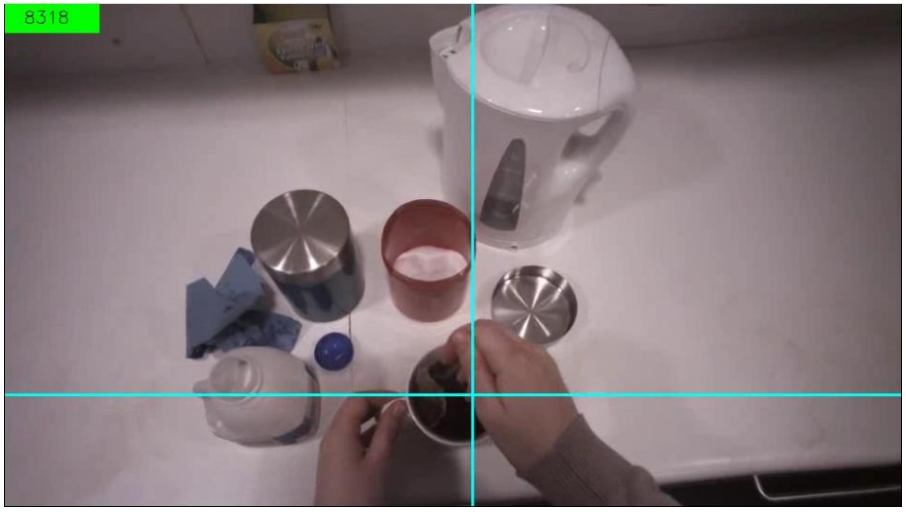
3.3.1 Dataset

To solve the classification problem in this study, several steps needed to be taken. First, the annotation of the raw data, which lead to the first matrix consisting of coded data. Second, the data needed to be pre-processed and engineered into useful features. Last, these features were combined into a new dataset that will be used for the classification task.

Annotation of the raw data

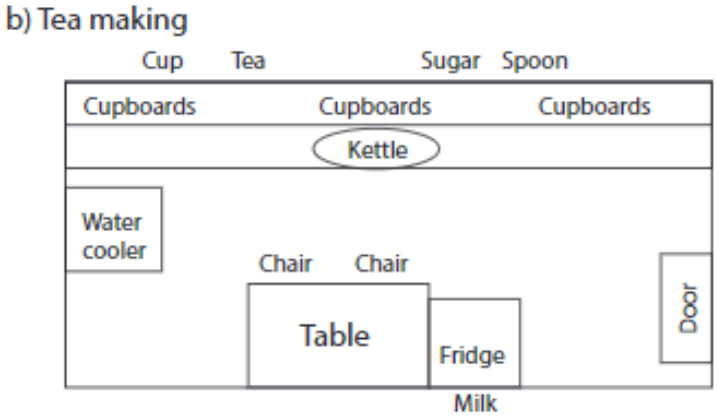
Based on the estimated time that would be needed for annotating the videos, a total of six videos were used for this study, considering the time limit. A fixation cross inside the videos shows eye fixations while performing this task (Figure 3).

Figure 3
Example of a video frame with the fixation cross while performing the tea-making task.



Annotating the data was necessary to perform further analysis of the data. Objects were annotated by using the same areas of interest (AOIs) as used Ioannidou et al. (2016). Figure 4 shows what AOIs will be annotated. The original annotation of the videos was not used since the goal of this research is different from the article of Ioannidou et al. (2016). They examined whether the fixation of the eyes in the middle of the visual field is related to viewing distances that are involved in tasks.

Figure 4
Plan of the kitchen while performing the task in Ioannidou et al. (2016).



The coding program *Solomon coder* was used to create useful data for the analysis. First, different objects were distinguished: cup, tea, sugar, cupboards, milk, kettle, spoon, table, fridge, water cooler, chair and the door. The videos show a fixation cross that indicates the direction of the gaze of the participants' eyes. If an object appeared in between the fixation cross and halfway of the horizontal or vertical midline, the object was annotated. In case there was more than one object present within that range, the object that appeared closest to the fixation cross was indicated. Next, different actions were distinguished: Finding tea, prepare the cups, finding sugar, pouring water, putting kettle on, finding spoon, placing sugar back, placing tea back, finding cup, finding milk, placing milk back.

The annotation of the data and the coordinates of the fixation cross inside the videos provide us with two sources of information: information of the eye movements and gaze when performing tasks and sequential information, which was retrieved through meaningful relationships in sequences of timeframes. For example, the different shifts from one object to another during an action.

3.3.2 Data cleaning/ pre-processing

The annotation of the data produced two CSV files with useful information; the data representing each activity and object during time frames and the coordinates that indicate the eye movements during the videos. Pre-processing is necessary for these two files in order to conduct further analysis. The first step in the data pre-processing was to combine the coded data from all participants into one file and adding a row indicating the participant's ID. The data was analysed in Rstudio and the following columns were distinguished: "Time", "Action", "Object" and 'file_name". The column file_name showed the participants' ID (su1, su2, su3, su4, su5, su6). This dataset, called *coded,* consisted of a total of 17,063 rows. Table 1 shows an example of the first 10 rows.

Table 1
Example of the annotated data acquired from Solomon coder.

Time	Object	Action	File_name
380.4	Door		su1
380.6	Cupboards		su1
380.8	Cupboards		su1
381.0	Cupboards		su1
381.2	Kettle		su1
381.4	Kettle		su1
381.6	Kettle		su1
381.8	Kettle		su1
382.0	Kettle		su1
382.2	Kettle		su1

Because each participant performed three activities (navigation, tea making and card sorting) and only the tea-making activity was needed for this study, only that specific part of the videos was annotated. All rows that concerned performance of the other tasks were removed from the data. Furthermore, all actions were converted to lowercase and an extra column was added showing the time difference between two rows that was needed for further analysis of the dataset (e.g. calculating the duration of actions). The commas in the time values were replaced with a point to keep values consistent over the dataset.

Additionally, the file containing the coordinates of eye movements needed pre-processing as well. In line with Land et al. (1999) fixations could be derived from these coordinates that could provide us with more information of eye movements during certain actions. These fixations could be detected using the *saccades* package with the formula **'detect.fixations'**. A second data frame was created with the fixation data consisting of 6498 rows and 11 columns (Table 2). Each row in this dataset represents one fixation that could take up a certain amount of time frames. That causes the two datasets to differ in the amount of rows.

Table 2
An example of the first six rows of the fixations data frame. Each row represents one fixation. It shows the start and end time of each fixation, the x and y coordinates and their corresponding standard deviations, the maximum coordinates of the eye movements (peak.vx and peak.vy), the duration of the fixation (dur) and the participant.

Trial	Start	End	X	Y	Sd.x	Sd.y	Peak.vx	Peak.vy	Dur	Participant
1	0.02	1.58	0.41	0.42	0.02	0.04	0.15	0.16	1.56	1
1	1.80	1.98	0.67	0.44	0.02	0.04	0.02	0.03	0.18	1
1	2.18	2.48	0.48	0.45	0.03	0.01	0.01	0.11	0.30	1
1	2.58	2.62	0.00	0.00	0.00	0.00	0.00	0.00	0.04	1
1	2.70	3.18	0.42	0.39	0.03	0.03	0.07	0.06	0.48	1
1	3.24	3.26	0.00	0.00	0.00	0.00	0.06	0.07	0.02	1

3.3.3 Descriptive analysis

Descriptive analysis was performed to explore the dataset and determine what features should be selected for the new dataset on which a classification method will be applied. Figure 5 presents an overview of the overall duration of performing the tea-making activity for each participant (Mean 266.47, *SD* 123.60). The figure shows there were considerable differences between participants.

Figure 5
An overview of the overall duration of performing the tea-making activity (all actions combined) for each participant.

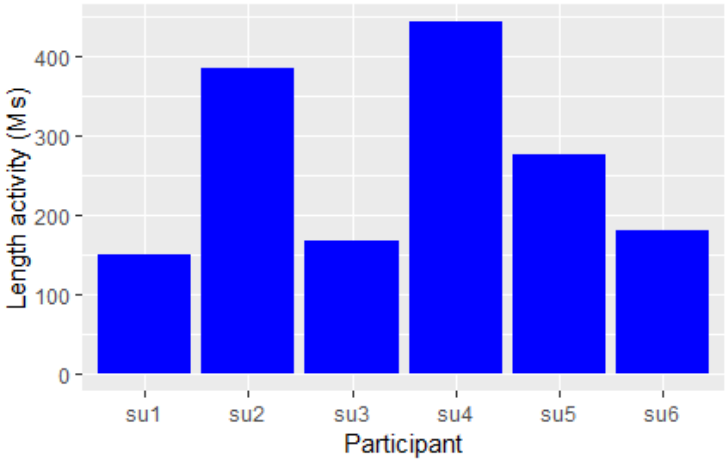
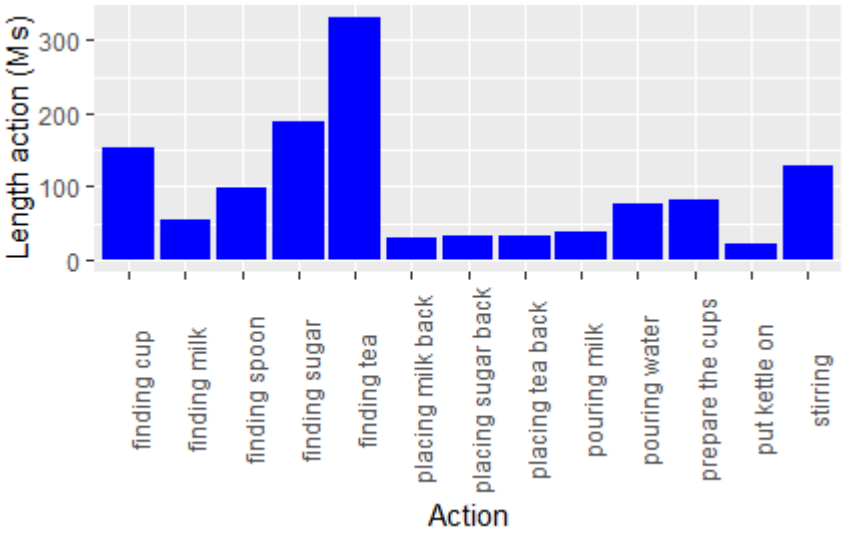


Figure 6 shows the mean duration of each action within the tea-making performance across all participants. It can be concluded that overall, the 'finding' actions took up the most time for participants, which could be explained by the fact that the

participants needed to search for these objects in the cupboards of the kitchen and the other actions could be completed without looking for objects. As shown in Figure 6, *finding the tea* was the action that took up most of the time and *putting the kettle on* was the action with the shortest duration. A possible explanation for this could be that the object 'Tea' was relatively hard to find for participants compared to other objects and *putting the kettle on* was a quick, easy to perform act.

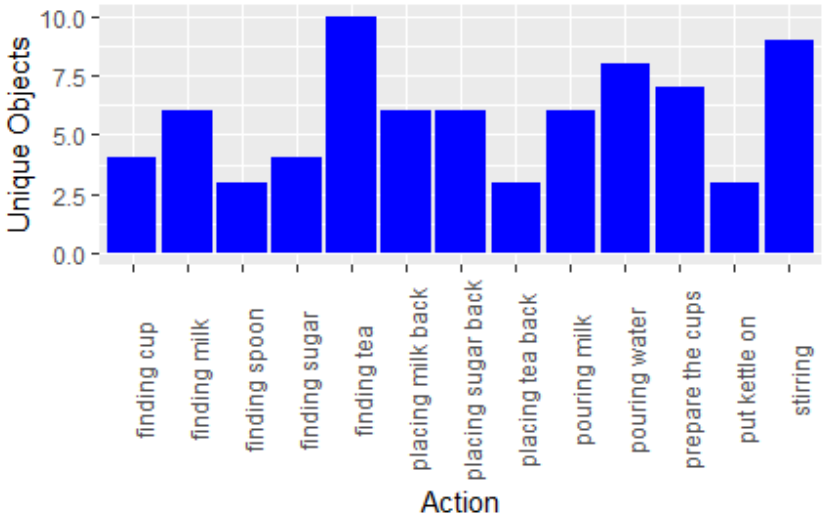
Figure 6

An overview of the mean duration of each action within the tea-making performance across all participants.



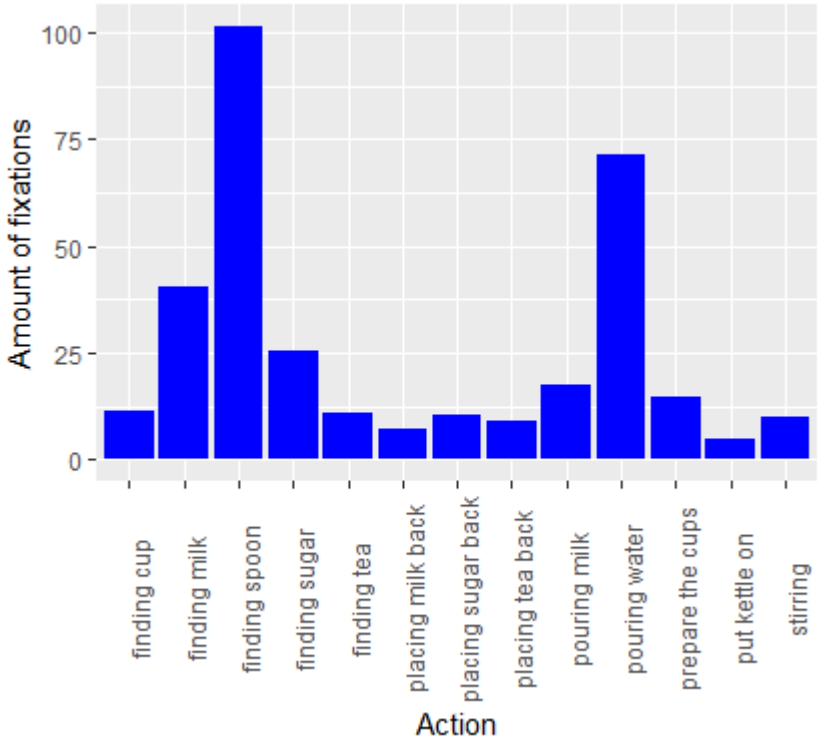
Next, the distribution of objects that were fixated on during the tea-making task was explored. Figure 7 shows the mean amount of unique objects that were fixated on during an action across all participants. What stands out is that the action *finding tea* was the action with most unique objects. A possible explanation for this could be that overall participants took longer to perform this action than other actions (as shown in Figure 6) and while doing this, they came across more objects. In addition, while performing the action *placing the tea back*, relatively few unique objects are looked at. It could be possible that in case it took participants more effort to find an object, they remember where to put it back more easily (Land et al. 1999).

Figure 7
An overview of the mean amount of unique objects that were fixated on during an action across all participants.



From the *fixation* data frame, the amount of fixations were analysed as shown in figure 8. It seemed that the most fixations occurred during the action 'finding spoon' and the least during the action 'putting the kettle on'. The fact that there are relatively few fixations during 'putting the kettle on' could be linked to the short duration of the action (Figure 6).

Figure 8
Presenting the amount of fixations per action across all participants.



3.3.4 Feature engineering

The following features were created to solve the classification problem as explained in section 3.2: the length of actions, the amount of unique objects, the duration of an object in sight, the transitions from one object to another, the time it took to find objects, the amount of fixations and the mean duration of fixations per action. Appendix 1 shows a short overview of the features. Since descriptive analysis showed significant differences per action, the features below were selected:

- *Length of actions* was calculated by taking the sum of the time frames on which each specific action took place, called Length_act.

- *The amount of unique objects* was calculated by counting the number of unique objects for each action and participant, called diff_object.

The duration of an object was determined by finding every unique object that was in sight during a certain action and calculating the duration that object was fixated on by the participant. This was done by taking the sum of the time frames on which a specific object was present during an action, called fix_dur. This feature consisted of 12 separate columns for each object and the corresponding duration it was fixated on. A short overview of the first six objects for the action finding cup is shown in Table 3.

Table 3
An overview of the feature fix_dur, presenting the first six objects for the action finding cup.

Action	Participant	Cup	Cupboards	Sugar	Tea	Spoon	Milk
Finding cup	Su1	0.67	11.53	0.20	0.00	0.00	0.00
Finding cup	Su2	6.80	33.00	0.00	0.00	0.00	0.00
Finding cup	Su3	2.40	1.68	0.70	2.20	0.00	0.00
Finding cup	Su4	3.00	0.80	0.00	0.00	0.00	0.00
Finding cup	Su5	1.40	3.00	0.00	0.00	0.00	0.00
Finding cup	Su6	1.60	6.10	0.00	0.00	0.00	0.00

The features that provide sequential information are: *Time.until.found* and *Transitions:*

- *Time.until.found*: Land et al. (1999) described *locating* as one of the functions of fixations. The feature presents the time it took to find the object that was needed to perform the action. This was done by taking the cumulative sum of the time frames before a specific object was present during an action.

- *Transitions*: this feature was selected with a view to sequential information of the data. In this case transitions mean the different shifts from one object to another during an action (e.g. cup- cupboards, cup-tea). This was done for every combination per object. This resulted in a total of 75 combinations of shifts. This feature was stored as transitions, which shows the number of times each combination occurred per action.

In order to analyse the *amount of fixations* and the *duration of fixations* during an action, the *fixations* data frame was used. Every row represents one fixation and shows a start and end time for it.

- *Amount of fixations* was calculated by looking at the range of time in which an action took place and counting the rows during that time, this feature was called Amount_fix.

- *The duration of fixations* was calculated by using the column called ‘dur’ in the *fixations* data frame, which shows the duration of a fixation. The mean duration per action was calculated, called Mean_duration.

3.3.5 Missing values

Before doing further analysis, missing values had to be handled. When inspecting the *Features* data frame, it turned out that the values that were missing were all derived from

the fix_dur feature. Since this feature represented duration and apparently a participant did not fixate on this object at all during this action, these missing values were replaced with the number 0.

3.3.6 Final dataset

All of the features described above were combined into a final data frame called *Features*. This data frame contains 93 columns which represent the features and 71 rows. This dataset was used for solving the classification problem. Table 4 represents the first 10 rows and first 6 columns of the final dataset.

Table 4
Representing the first 10 rows and first 6 columns of the final dataset.

Action	Length_ act	Unique_ objects	Cup_ Cupboards	Cup_ Door	Cup_ Kettle	Cup_ Milk
Finding cup	36.8	3	1	0	0	0
Finding cup	72.8	2	1	0	0	0
Finding cup	14.4	4	1	0	0	0
Finding cup	4.6	2	1	0	0	0
Finding cup	11.8	2	1	0	0	0
Finding cup	13.8	2	1	0	0	0
Finding milk	8.2	5	0	0	0	0
Finding milk	21.8	4	0	0	0	0
Finding milk	6.2	3	0	0	0	0
Finding milk	6.0	5	0	0	0	0

3.4 Method / Models

The following section will discuss what classification method was used in this study (section 3.4.1). Besides, the evaluation methods that were used will be explained in section 3.4.2.

3.4.1 Method

The features selected (section 3.3.4) above characterize actions in the tea-making task. For example, the action ‘pouring water’ could be characterized by a small amount of fixations and many unique objects during the action. To be able to accurately predict to what class (action) new input belongs to, the classification method ‘RandomForest’ was

chosen. This method was chosen because it produces clear output that shows the importance of features. This could be useful when finding out which features are good predictors of actions which could improve activity recognition (sub-question 1). Moreover, the Random Forest model reduces over-fitting of the data and is more accurate than decision trees (Gahukar, 2018). After running the Random forest model, important features can be extracted from the Mean decrease in Gini index. This index indicates the node impurity of an input feature. The higher the mean decrease in Gini index, the higher the importance of the feature.
To find out what features are good predictors when predicting actions, the original dataset (model 1) was split up in different subsets.

- The first subset, containing the twenty best performing features based on the Mean Decrease Gini index of the original model, called model 2.
- The second subset, consisting of the Length of actions, amount of unique objects and the duration of object in sight, called model 3.
- The third subset, consisting of the features providing sequential information; the time until the object was found and the transitions, called model 4.
- The fourth subset, consisting of the fixation features, called model 5.

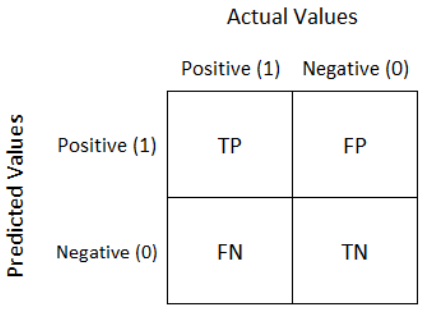
Before training the models, parameter grid search with Leave-one-out cross validation (LOOCV) was performed on the dataset to find the best parameters for the classification model. The model was repeatedly split into samples of 70, leaving 1 out for every row of the dataset. Because the model will be trained on the entire dataset, the bias will be reduced. A disadvantage of LOOCV could be the computational time it takes, but since the dataset was small it made sense to use LOOCV in our study. For the RandomForest method important parameters are: the number of features taken into account when conducting the optimal split (mtry) and the number of trees that need to be grown (ntree).The default values are 9 for mtry and 500 for ntree. These optimal parameters were used in the RF model.

3.4.2 Evaluation model

Since no prior research was conducted on activity recognition on this dataset, it was difficult to use prior work as benchmark for evaluation of the classification model. For that reason, the different models created in this study were compared to a majority baseline score in order to determine the model's predictive power. This shows the performance of the largest (majority) class before running any algorithm on the model and can be used as a reference point.

Confusion matrix
In order to evaluate the model, a confusion matrix (Figure 9) was created for all classes individually. A confusion matrix, as shown in table 4, shows the actual values and the values that were predicted by the model. This results in True Positives (TP), which show the positive values also predicted as such and the False Positives (FP) which represent the incorrectly predicted positive values. Moreover, the confusion matrix contains True Negatives (TN), that show the negative values also predicted as such and False Negatives (FN) which represent values that are incorrectly predicted as negative (Sunasra, 2017).

Figure 9
Example of a confusion matrix



Accuracy
Since multiple confusion matrices were created for all individual classes, the mean of these confusion matrices was taken in order to compute the evaluation metrics described below. To determine what part of prediction the classification model predicted right, the accuracy was measured. Accuracy can be measured by dividing the number of correct predictions by all predictions, as shown in Figure 10 (Sunasra, 2017). Furthermore, the accuracy of every action was evaluated in order to see whether some actions were easier to predict for the model than others (sub-question 2).

Figure 10
Formula of classification accuracy

$$Classification\ accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.5 Software used

The following section provides an overview of the software that was used to conduct this study.

As described earlier, for the annotation of the videos, the program *Solomon coder* was used (András Péter, http://solomoncoder.com). The R programming language in R studio was used to analyze the data (R Core Team, 2013). To create the *fixations* data frame the R package saccades was used (von der Malsburg, 2015). The packages dplyr (Wickham, François, Henry & Müller, 2019) and tidyr (Wickham & Henry (2019) were used for descriptive analysis of the data, the engineering of features and creating the subsets of the final dataset. These analyses were plotted using ggplot2 (Wickham, 2016). Last, to create the random forest model the package randomForest was used (Liaw & Wiener, 2002).

4. Results

This section will give insight into the results of the classification task performed to answer the research question:

To what extent can actions performed in ADLs be predicted, based on eye movements from egocentric view?

- *What features are good predictors for actions?*
- *To what extent do actions differ in predictability?*

Table 5 shows the results of the different models used in this study. Further details will be discussed below.

Table 5
Results of the five RandomForest models

Model	Accuracy
Majority baseline	0.21
Model 1	0.509
Model 2	0.535
Model 3	0.507
Model 4	0.507
Model 5	0.085

4.1 Random Forest

Sub-question 1
As shown in table 4 our first model including all 92 features, reaches an accuracy of 50.9 %, which means 50.9 % of the classes were predicted correctly. The confusion matrix of this model can be found in Table 6.

Table 6
Confusion matrix of model 1

		Actual												
		finding cup	finding milk	finding spoon	finding sugar	finding tea	placing milk back	placing sugar back	placing tea back	pouring milk	pouring water	prepare the cups	put kettle on	stirring
Predicted	finding cup	**15**	0	0	0	2	0	0	0	3	0	4	0	3
	finding milk	0	**8**	0	0	0	11	0	0	0	0	0	0	0
	finding spoon	0	0	**15**	0	0	0	2	0	0	0	0	0	3
	finding sugar	1	0	0	**20**	3	1	8	1	1	0	4	0	1
	finding tea	2	0	0	0	**6**	0	0	7	1	0	0	0	0
	placing milk back	0	10	0	0	0	**3**	0	0	0	0	0	0	0
	placing sugar back	0	0	0	1	0	0	**2**	2	0	0	0	0	2
	placing tea back	0	0	0	0	6	0	1	**2**	0	0	1	0	0
	pouring milk	0	0	0	0	0	0	0	0	**4**	3	0	0	0
	pouring water	0	0	0	0	0	0	1	0	2	**15**	0	0	1
	prepare the cups	0	0	0	0	0	0	0	0	0	0	**2**	0	4
	put kettle on	0	0	0	0	0	0	1	0	0	0	3	**18**	0
	stirring	0	0	0	0	1	0	0	0	1	0	4	0	**1**

Model 2 was created to decrease the complexity of the model and check whether the model would perform better when leaving out certain features. The twenty best performing features were included in this model (Table 7) .This resulted in an accuracy of 53.5%, which shows a slight improvement compared to the first model. The third model, consisting of the features: Length_act, Unique_objects and the duration an object was in sight, scored less on accuracy (50.7%) than the second model. Moreover, the fourth model, representing the 75 Transitions features and the feature Time until an object is found, scored 50.7% on accuracy, which is equal to the third model. Last, it was chosen to create a subset of the dataset with only the *fixation* features (Amount_fix and Mean_duration), called model 5. This model did not perform well with a 0.09% accuracy score. Compared to the majority baseline score, all models made an improvement in the predictability of actions, except for model 5. The confusion matrices of models 2-5 can be found in appendix 2-5.

Table 7
Variables included in Random forest model 2

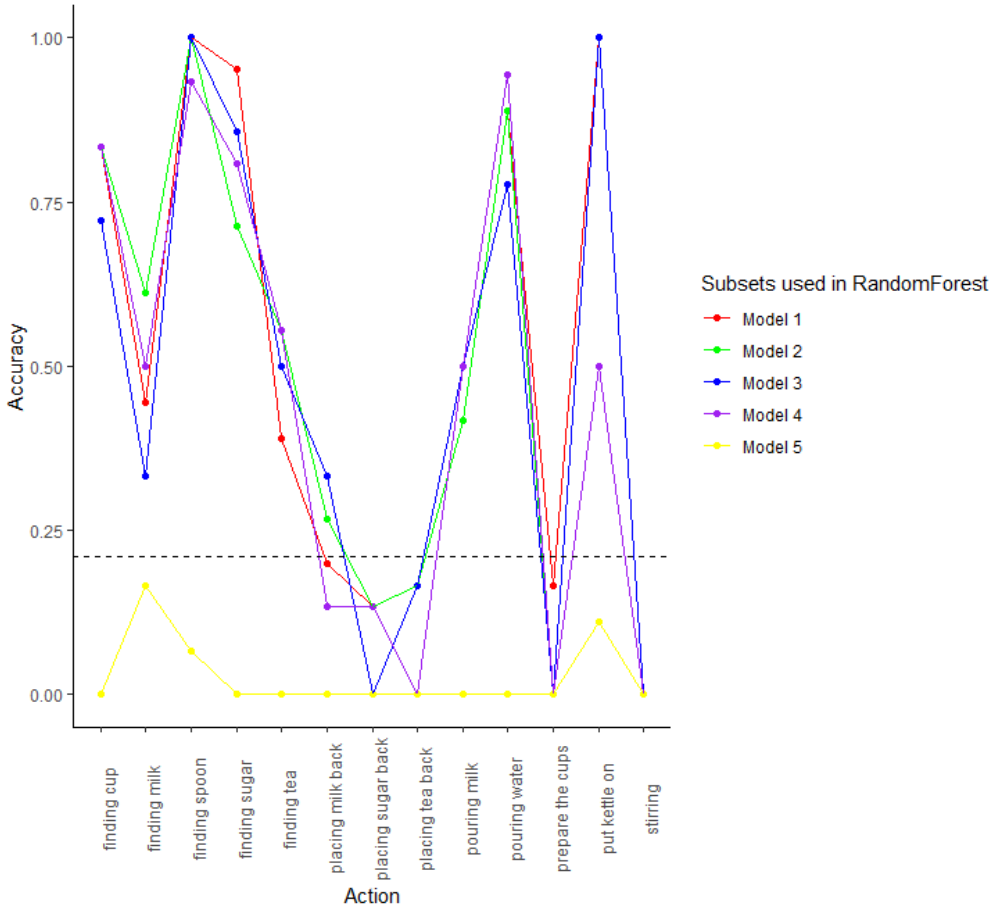
Variables			
Kettle	Length_act	Time.untill.found	Cupboards_Cup
Cup	Amount_fix	Milk	Cup_Cupboards
Fridge	Cupboards_Tea	Cup_Kettle	Unique_objects
Sugar	Spoon_Cupboards	Tea	Kettle_Cup
Cupboards	Kettle_NA	Spoon	Cupboards_NA

Sub-question 2

Figure 11 shows the accuracy scores of all actions separately for every model. Results show that all models perform relatively well when predicting the actions 'finding spoon' and 'pouring water' compared to other actions and the baseline score. Besides, the action 'put kettle on' scores high on accuracy as well, except for model 4 (containing sequential features). However, the actions 'prepare the cups' and all of the actions where an object is placed back are more often incorrectly predicted and often perform less than the baseline score of 0.21. What stands out is that the action 'stirring' is never predicted correctly in any of the models.

Analysing the confusion matrices of the models, it seems that the action 'stirring' is often predicted as 'prepare the cups' and vice versa. Moreover, the 'placing back' actions are often incorrectly predicted as a 'finding' action containing the same object.

Figure 11
Representing the accuracy score for all of the models per action.



5. Discussion

The goal of this study was to be able to classify data to determine to what types of actions it belongs to, based on eye-movements and objects in the video frames. Since no prior research was conducted on activity recognition on this dataset, it was difficult to use prior work as benchmark for evaluation of the classification model. For that reason, different models were created in this study and compared to one another by evaluation metrics and a majority baseline score.

To answer the first sub-question: '*What features are good predictors for actions?',* several models were created. The results of the first Random forest model, including the original dataset, show that the model scores 50.9% on accuracy. The second model, including the twenty most important features (accuracy 53.5%), performed slightly better than the first model. It could be concluded that model 2 is slightly better in predicting what actions were carried out, compared to model 1. Besides, model 3 (consisting of the 'object' features) and model 4 (consisting of the sequential features) did not show a major difference in performance compared to the first model, both with an accuracy of 50.7 %. However, when using only the 'fixations' features in model 5, the model performed poorly and even below the majority baseline with an accuracy of

0.09%. The poor performance of this model could possibly be explained to the fact that the subset only consisted of two features.

It could be concluded that there is no major difference in the performance of models when using different types of features. However, all models, except model 5, improve compared to the majority baseline score of 21%.This implicates that the models improve their predictive ability of actions compared to the accuracy of the majority class without a model.

In addition, analysis of the performance of the models per action was conducted to answer the second sub-question: '*To what extent do actions differ in predictability?'* It can be concluded there is a difference in the predictability of certain actions. The actions 'finding spoon', 'pouring water' and 'putting kettle on' were more often correctly predicted compared to other actions. These results could be explained by previous descriptive analysis of the dataset (section 3.3.3). This analysis showed that the action 'finding spoon' and 'pouring water' differed from the other actions, because it contained relatively many fixations. The action 'putting kettle on' contained relatively few fixations and had a short duration compared to other actions. These characteristics could make it easier for the models to predict the actions correctly, because they differentiate these actions from other actions.

However, the actions 'stirring', 'prepare the cups', 'placing milk back', 'placing sugar back' and 'placing tea back' were harder to predict for all models. What stands out is that the action 'stirring' is never predicted correctly in all of the models. Results show that the action 'stirring' seems to be predicted as 'prepare the cups' more often than other actions and vice versa. An explanation for this could be that, while performing both of these actions, participants look at many unique objects (section 3.3.3), which makes the actions similar. Moreover, the 'placing back' actions are often mistaken for 'finding' actions containing the same object. For example, the action 'placing milk back' is regularly predicted as 'finding milk'. A possible explanation for this could be that the model mistakes the 'placing back' action for the 'finding action' because the participant looks at the same object while performing both types of actions.

Previous research

Other research on ADLs, such as research of Ogaki, Kitani, Sugano & Sato (2012) studied activity recognition in an office environment, their highest scoring SVM model scored 57% on accuracy. The classification model of Fathi et al. (2011) scored 47.7 % on accuracy when studying the roles of objects and hands while performing daily tasks. Besides, the KNN model in Spriggs et al. (2009) scored 57.8 % on accuracy. However, it is hard to compare the models in this study to previous work. There are several factors that influence the guiding of eye movements; e.g. the type of stimulus, the spatial structure of the environment, performance of multiple actions (Triesch, Ballard, Hayhoe and Sullivan, 2003; in Xu and Nakayama, 2003; Aivar, Hayhoe, Chizk & Mruczek, 2005). Requirements for eye movements could differ for different tasks and produce different results. Since activities performed in previous studies differed from the tea-making task in this study, no reliable comparisons could be made.

Limitations

The current study has a number of limitations. First, due to time-constraints, the amount of videos annotated and analysed was limited to six. The study would be more representative if more videos were taken into account. Another limitation of this study

was that only one coder annotated the videos. If more coders would be assigned, the reliability of the annotations could increase.

As stated in Padilla-López, Chaaraou and Flórez-Revuelta (2015), action recognition can be divided in four levels: behaviour, activity, action and motion level. The current study was able to research action and motion level in order to characterize activity level. The tea-making task could be divided into smaller actions such as, 'put kettle on' and 'pouring water'. However, these actions could be divided into deeper sublevels. For example the action pouring water could include even smaller sub actions such as: 'lifting the kettle' and 'lifting the cup'.

Furthermore, not all factors that could influence activity recognition according to literature could be taken into account, again due to time-constraints. For example, as described in Land et al. (1999) saccades, fixations, relocating objects and objects not fixated on could be features for activity recognition in a tea-making task. During this study only fixations were taken into account. In addition, as stated in Fathi et al. (2011), changes in the state of objects support action recognition as well. This study did not go into that much detail.

Contributions and suggestions for future research

The first contribution to future work in the field of activity recognition regards the general demand for more fully annotated datasets. The annotation of the videos in this study contributes to more research on this subject possible. Besides, the current study contributes to the representativeness of the tea-making task in research on activity recognition in ADLs. By obtaining more knowledge on activity recognition on specific tasks, more generally validated conclusions could be drawn.

This study could contribute to research on activity recognition in healthcare. As stated in Ramzaoui, Faure, Spotorno (2018), studies on activity recognition in ADLs on healthy individuals is informative for research on diseases, because studying habits of healthy people contribute to understanding those with functional deficits. Moreover, this study contributes to more real-world research of activity recognition in healthcare.

Future research could take into account that this study has shown that some actions are easier to predict than others. For example, it could try to improve the prediction of the least performing actions by finding better predictors for these actions. Besides, including features that differentiated the actions that were easier to predict would make sense.

In addition, future research could explore more algorithms on this dataset to evaluate their results and see if the performance could be improved. Moreover, since this study only analysed six of the forty-two videos, the same research could be extended to all of the videos to improve the need for annotated and representative datasets of tasks. Last, since there is a strong need for more research on activity recognition, it would be interesting to use the approach of this study to other tasks as well.

6. Conclusion

Overall, the aim of this study was to find out to what extent actions performed in ADLs can be predicted, based on eye movements from egocentric view. This was examined by classifying actions, based on eye-tracking data. Participants in this study performed a tea-making task and that consisted of smaller sub-actions. It could be concluded that the models in this study, except for model 5, perform relatively well compared to the majority baseline score and contribute to the prediction of actions. In addition, this study indicates that some actions are easier to predict than others. However, more research is necessary to support the findings. It would be interesting to explore the performance of more algorithms on this dataset and to examine what features are good predictors for actions in further detail.

References

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Adams, N. M. (2010). Perspectives on data mining. International Journal of Market Research, 52(1), 11-19.

Aivar, M. P., Hayhoe, M. M., Chizk, C. L., & Mruczek, R. E. (2005). Spatial memory and saccadic targeting in a natural task. Journal of Vision, 5(3), 3-3.

Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. BioMed research international, 2015.

Bennewitz, M., Burgard, W., & Thrun, S. (2002, May). Learning motion patterns of persons for mobile service robots. In Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292) (Vol. 4, pp. 3601-3606). IEEE.

Betancourt, A., Morerio, P., Regazzoni, C. S., & Rauterberg, M. (2015). The evolution of first person vision methods: A survey. IEEE Transactions on Circuits and Systems for Video Technology, 25(5), 744-760.

Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., & Ertl, T. (2017, December). Visualization of eye tracking data: A taxonomy and survey. In Computer Graphics Forum (Vol. 36, No. 8, pp. 260-284).

Clarkson, B., Sawhney, N., & Pentland, A. (1998). Auditory context awareness via wearable computing. Energy, 400(600), 20.

Wilson, D. H., & Atkeson, C. (2005, May). Simultaneous tracking and activity recognition (STAR) using many anonymous, binary sensors. In International Conference on Pervasive Computing (pp. 62-79). Springer, Berlin, Heidelberg.

Fathi, A., Farhadi, A., & Rehg, J. M. (2011, November). Understanding egocentric activities. In 2011 International Conference on Computer Vision (pp. 407-414). IEEE.

Fathi, A., Li, Y., & Rehg, J. M. (2012, October). Learning to recognize daily actions using gaze. In European Conference on Computer Vision (pp. 314-327). Springer, Berlin, Heidelberg.

Fathi, A., Ren, X., & Rehg, J. M. (2011, June). Learning to recognize objects in egocentric activities. In CVPR 2011 (pp. 3281-3288). IEEE.

Gellersen, R. Want, and A. Schmidt, editors., vol. 3468. Springer; pp. 62-79, 2005.

Gerber, S. M., Müri, R. M., Mosimann, U. P., Nef, T., & Urwyler, P. (2018, July). Virtual reality for activities of daily living training in neurorehabilitation: a usability and feasibility study in healthy participants. In 2018 40th Annual International

Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1-4). IEEE.

Giannakeris, P., Avgerinakis, K., Vrochidis, S., & Kompatsiaris, I. (2018, September). Activity Recognition from Wearable Cameras. In 2018 International Conference on Content-Based Multimedia Indexing (CBMI) (pp. 1-6). IEEE.

González-Díaz, I., Buso, V., Benois-Pineau, J., Bourmaud, G., Usseglio, G., Mégret, R., ... & Dartigues, J. F. (2015). Recognition of instrumental activities of daily living in egocentric video for activity monitoring of patients with dementia. In Health Monitoring and Personalized Feedback using Multimedia Data (pp. 161-178). Springer, Cham.

Gandhi, R. (2018). Support Vector Machine—Introduction to Machine Learning Algorithms. Retrieved from: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Gulde, P., Hughes, C., Parekh, M., Russell, M., Ferre, M., Wing, A., ... & Hermsdörfer, J. (2014). Analysis of eye movements, kinematics and dynamic aspects of performance during activities of daily living in stroke patients. In Replace, Repair, Restore, Relieve–Bridging Clinical and Engineering Solutions in Neurorehabilitation (pp. 393-401). Springer, Cham.

H. Wickham (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Hadley Wickham and Lionel Henry (2019). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.3. Retrieved from: https://CRAN.R-project.org/package=tidyr

Wickham, H., & Wickham, M. H. (2016). Package 'plyr'. Obtenido de https://cran. rproject. org/web/packages/dplyr/dplyr. pdf.

Hand, D.J., Mannila, H. and Smyth, P. (2001) Principles of Data Mining. MIT press.

Intille, S. S., & Bobick, A. F. (1999). A framework for recognizing multi-agent action from visual evidence. AAAI/IAAI, 99(518–525).

Intille, S. S., Larson, K., Tapia, E. M., Beaudin, J. S., Kaushik, P., Nawyn, J., & Rockinson, R. (2006, May). Using a live-in laboratory for ubiquitous computing research. In International Conference on Pervasive Computing (pp. 349-365). Springer, Berlin, Heidelberg.

Ioannidou, F., Hermens, F., & Hodgson, T. (2016). The centrial bias in day-to-day viewing. Journal of Eye Movement Research, 9(6), 1-13.

Jug, M., Perš, J., Dežman, B., & Kovačič, S. (2003, April). Trajectory based assessment of coordinated human activity. In International Conference on Computer Vision Systems (pp. 534-543). Springer, Berlin, Heidelberg.

Asnaoui, K. E., Hamid, A., Brahim, A., & Mohammed, O. (2017, April). A survey of activity recognition in egocentric lifelogging datasets. In 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS) (pp. 1-8). IEEE.

Kang, W., & Han, Y. (2014). SmartPDR: Smartphone-based pedestrian dead reckoning for indoor localization. IEEE Sensors journal, 15(5), 2906-2916.

Kitani, K. M., Okabe, T., Sato, Y., & Sugimoto, A. (2011, June). Fast unsupervised ego-action learning for first-person sports videos. In CVPR 2011 (pp. 3241-3248). IEEE.

Kumar, K. S., & Bhavani, R. (2017). Human activity recognition in egocentric video using PNN, SVM, kNN and SVM+ kNN classifiers. Cluster Computing, 1-10.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities?. Vision research, 41(25-26), 3559-3565.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. Perception, 28(11), 1311-1328.

Lee, S. W., & Mase, K. (2002). Activity and location recognition using wearable sensors. IEEE pervasive computing, 1(3), 24-32.

Li, Y., Liu, M., & Rehg, J. M. (2018). In the eye of beholder: Joint learning of gaze and actions in first person video. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 619-635).

Ma, M., Fan, H., & Kitani, K. M. (2016). Going deeper into first-person activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1894-1903).

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). caret: Classification and Regression Training. R package version 6.0-81. https://CRAN.Rproject.org/package=caret

Mégret, R., Dovgalecs, V., Wannous, H., Karaman, S., Benois-Pineau, J., El Khoury, E., ... & Dartigues, J. F. (2010, October). The IMMED project: wearable video monitoring of people with age dementia. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1299-1302). ACM.

Meyer, D. (1997). Human gait classification based on hidden Markov models. In 3D Image Analysis and Synthesis (Vol. 97, pp. 139-146).

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. Jama, 309(13), 1351-1352.

Nguyen, T. H. C., Nebel, J. C., & Florez-Revuelta, F. (2016). Recognition of activities of daily living with egocentric vision: A review. Sensors, 16(1), 72.

Ogaki, K., Kitani, K. M., Sugano, Y., & Sato, Y. (2012, June). Coupling eye-motion and ego-motion features for first-person activity recognition. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (pp. 1-7). IEEE.

Oliver, N., Horvitz, E., & Garg, A. (2002, October). Layered representations for recognizing office activity. In Proceedings of the International Conference on Multimodal Interaction (ICMI 2002) (pp. 3-8).

Ordóñez, F. J., Iglesias, J. A., De Toledo, P., Ledezma, A., & Sanchis, A. (2013). Online activity recognition using evolving classifiers. Expert Systems with Applications, 40(4), 1248-1255.

Padilla-López, J. R., Chaaraoui, A. A., & Flórez-Revuelta, F. (2015). Visual privacy protection methods: A survey. Expert Systems with Applications, 42(9), 4177-4195.

Patterson, D. J., Liao, L., Fox, D., & Kautz, H. (2003, October). Inferring high-level behavior from low-level sensors. In International Conference on Ubiquitous Computing (pp. 73-89). Springer, Berlin, Heidelberg.

Pirsiavash, H., & Ramanan, D. (2012, June). Detecting activities of daily living in first-person camera views. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2847-2854). IEEE.

Poleg, Y., Arora, C., & Peleg, S. (2014). Temporal segmentation of egocentric videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2537-2544).

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1), 3.

Ramzaoui, H., Faure, S., & Spotorno, S. (2018). Alzheimer's Disease, Visual Search, and Instrumental Activities of Daily Living: A Review and a New Perspective on Attention and Eye Movements. Journal of Alzheimer's Disease, (Preprint), 1-25.

Ryoo, M. S., & Matthies, L. (2013). First-person activity recognition: What are they doing to me?. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2730-2737).

Sarkar, A. J., Lee, Y. K., & Lee, S. (2010). A smoothed naive bayes-based classifier for activity recognition. IETE Technical Review, 27(2), 107-119.

Shiga, Y., Toyama, T., Utsumi, Y., Kise, K., & Dengel, A. (2014, September). Daily activity recognition combining gaze motion and visual features. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (pp. 1103-1111). ACM.

Singh, S., Arora, C., & Jawahar, C. V. (2017). Trajectory aligned features for first person action recognition. Pattern Recognition, 62, 45-55.

Spriggs, E. H., De La Torre, F., & Hebert, M. (2009, June). Temporal segmentation and activity classification from first-person sensing. In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (pp. 17-24). IEEE.

Stewart, W. (2018). How do we accelerate data driven health care?

Taralova, E., De la Torre, F., & Hebert, M. (2011, November). Source constrained clustering. In 2011 International Conference on Computer Vision (pp. 1927-1934). IEEE.

Team, R. C. (2013). R: A language and environment for statistical computing.

Gahukar, G. (2018). Classification Algorithms in Machine Learning…Retrieved from: https://medium.com/datadriveninvestor/classification-algorithms-in-machine-learning-85c0ab65ff4

Sunasra, M. (2017). Performance Metrics for Classification problems in Machine Learning. Retrievedfrom:https://medium.com/thalus-ai/performance-metrics-for classification-problems-in-machine-learning-part-i-b085d432082b

Von der Malsburg, T. (2015). Saccades: Detection of fixations in eye-tracking data. R package version 0.1-1.

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. Journal of vision, 3(1), 9-9.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. science, 185(4157), 1124-1131.

Xu, Y., & Nakayama, K. (2003). Placing objects at different depths increases visual short-term memory capacity. Journal of Vision, 3(9), 27-27.

Yan, Y., Ricci, E., Liu, G., & Sebe, N. (2015). Egocentric daily activity recognition via multitask clustering. IEEE Transactions on Image Processing, 24(10), 2984-2995.

Yu, H., Jia, W., Li, Z., Gong, F., Yuan, D., Zhang, H., & Sun, M. (2019). A multisource fusion framework driven by user-defined knowledge for egocentric activity recognition. EURASIP journal on advances in signal processing, 2019(1), 14.

Zhou, X., & Curry, W. (2015). U.S. Patent Application No. 14/452,025.

Appendices

Appendix 1

Presenting an overview of the features used in this study.

Feature	Name	Data frame	Description
Length action	Length_act	Coded	*The length of each action per participant*
Amount of unique objects	Unique_objects	Coded	*The number of unique objects that is fixated on during each action per participant*
Duration of object in sight	Fix_dur	Coded	*The duration each unique object is fixated on during each action per participant, divided over 12 columns*
Transitions	Transitions	Coded	*The different combinations of shifts from one object to another during an action per participant, divided over 75 columns*
Time until object was found	Time.until.found	Coded	*The cumulative time before the object necessary to perform the action was found for each action per participant*
Amount of fixations	Amount_fix	Fixations	*The amount of fixations per action per participant (counting the rows)*
Mean duration of fixations	Mean_duration	Fixations	*The mean duration of fixations during each action per participant*

Appendix 2

Confusion matrix of model 2

		Actual												
		finding cup	finding milk	finding spoon	finding sugar	finding tea	placing milk back	placing sugar back	placing tea back	pouring milk	pouring water	prepare the cups	put kettle on	stirring
Predicted	finding cup	**13**	0	0	0	0	0	0	0	3	0	1	0	5
	finding milk	0	**14**	0	0	0	9	0	0	1	0	0	0	0
	finding spoon	0	0	**15**	0	1	0	1	0	0	0	0	0	3
	finding sugar	0	0	0	**19**	0	0	7	2	0	0	3	0	0
	finding tea	3	0	0	0	**8**	0	0	7	0	0	0	0	0
	Placing milk back	0	4	0	0	0	**6**	0	0	0	0	0	0	0
	placing sugar back	0	0	0	2	0	0	**2**	1	0	0	0	0	0
	placing tea back	0	0	0	0	7	0	0	**2**	0	0	2	0	0
	pouring milk	0	0	0	0	0	0	0	0	**5**	2	0	0	1
	pouring water	0	0	0	0	0	0	0	0	1	**16**	0	0	0
	prepare the cups	0	0	0	0	0	0	0	0	0	0	**6**	0	2
	put kettle on	0	0	0	0	0	0	3	0	0	0	3	**18**	3
	stirring	2	0	0	0	2	0	2	0	2	0	3	0	**1**

Appendix 3

Confusion matrix of model 3

Actual														
		finding cup	finding milk	finding spoon	finding sugar	finding tea	placing milk back	placing sugar back	placing tea back	pouring milk	pouring water	prepare the cups	put kettle on	stirring
Predicted	finding cup	**14**	0	0	0	0	0	0	0	1	0	0	0	3
	finding milk	0	**7**	0	0	0	10	0	0	0	0	0	0	0
	finding spoon	0	0	**15**	0	0	0	0	0	0	0	0	0	3
	finding sugar	0	0	0	**18**	0	0	9	2	0	0	2	0	0
	finding tea	0	0	0	0	**10**	0	0	6	0	0	0	0	0
	placing milk back	0	11	0	0	0	**5**	0	0	0	0	0	0	0
	placing sugar back	0	0	0	3	0	0	**0**	2	0	0	0	0	0
	placing tea back	0	0	0	0	5	0	2	**2**	0	0	2	0	0
	pouring milk	0	0	0	0	0	0	0	0	**7**	3	0	0	0
	pouring water	0	0	0	0	0	0	0	0	1	**15**	4	0	1
	prepare the cups	2	0	0	0	1	0	1	0	0	0	**1**	0	8
	put kettle on	0	0	0	0	0	0	1	0	0	0	4	**18**	0
	stirring	2	0	0	0	2	0	2	0	3	0	5	0	**0**

Appendix 4

Confusion matrix of model 4

		Actual												
		finding cup	finding milk	finding spoon	finding sugar	finding tea	placing milk back	placing sugar back	placing tea back	pouring milk	pouring water	prepare the cups	put kettle on	stirring
Predicted	finding cup	**15**	0	0	0	2	0	0	0	3	0	4	0	2
	finding milk	0	**9**	0	0	0	8	0	0	0	0	0	0	0
	finding spoon	0	0	**15**	0	0	0	2	0	0	0	0	0	2
	finding sugar	1	1	0	**17**	5	1	8	1	1	1	4	2	2
	finding tea	1	0	0	0	**10**	0	0	9	0	0	0	0	0
	placing milk back	0	6	0	0	0	**2**	0	0	0	0	0	0	0
	placing sugar back	0	1	0	4	0	0	**2**	1	0	0	0	0	2
	placing tea back	1	0	0	0	1	2	0	**0**	0	0	0	1	0
	pouring milk	0	0	0	0	0	0	0	0	**5**	0	0	0	1
	pouring water	0	0	0	0	0	0	1	0	3	**17**	0	1	1
	prepare the cups	0	0	0	0	0	0	0	0	0	0	**2**	4	4
	put kettle on	0	0	0	0	0	0	0	1	0	0	3	**10**	1
	stirring	0	1	0	0	0	2	2	0	0	0	5	0	**0**

Appendix 5

Confusion matrix of model 5

		Actual												
		finding cup	finding milk	finding spoon	finding sugar	finding tea	placing milk back	placing sugar back	placing tea back	pouring milk	pouring water	prepare the cups	put kettle on	stirring
Predicted	finding cup	**0**	1	0	0	0	2	1	0	0	0	0	1	0
	finding milk	0	**3**	0	0	1	0	0	0	0	1	1	0	0
	finding spoon	0	0	**1**	0	0	1	0	0	0	0	0	0	0
	finding sugar	0	1	1	**1**	1	0	0	0	2	0	2	0	0
	finding tea	0	1	1	1	**0**	0	2	0	1	1	0	1	0
	placing milk back	1	0	1	0	0	**0**	1	1	0	1	0	2	1
	placing sugar back	1	0	0	0	2	1	**0**	0	0	0	0	0	1
	placing tea back	1	0	0	0	0	0	0	**0**	0	1	1	0	1
	pouring milk	0	0	0	3	1	0	0	0	**0**	1	0	0	0
	pouring water	0	0	0	1	1	0	0	0	1	**0**	1	0	0
	prepare the cups	2	0	1	1	0	0	0	0	0	1	**0**	0	2
	put kettle on	1	0	0	0	0	1	1	1	0	0	0	**2**	0
	stirring	0	0	0	0	0	0	0	2	0	0	1	0	**0**