What we see is what we do: A peripheral vision based HMM framework for gaze guided human action recognition

Felix S. Wang[1*], Thomas Kreiner[1], Alexander Lutz[1], Quentin Lohmeyer[1], Mirko Meboldt[1]

[1] ETH Zurich
Product Development Group Zurich
Leonhardstrasse 21
CH-8092 Zurich

[*] wangfe@ethz.ch

Abstract

Video-based human action recognition models have shown remarkable performances, but rely heavily on a large quantity of training data. In specific applications with only limited data available, eye tracking may provide valuable additional sensory information to achieve accurate classification performances using only small datasets. We propose the new Peripheral Vision Based HMM (PVHMM) classification framework, which utilizes context-rich and object-related gaze features for the detection of sequential human actions. The gaze information is quantified using two gaze features, the object-of-interest hit and the object-gaze distance, which contains peripheral gaze information, and human action recognition is achieved by employing a hidden Markov model. The classification performance of the framework is tested and validated on a safety-critical medical device handling task with a specific intended use, using less than 50 eye tracking recordings. Our proposed framework achieves f1-scores of over 90% which surpasses previous gaze-based methods, while showing high robustness to adversarial noise, when using the object-gaze distance gaze feature. Consequently, this gaze guided human action recognition approach shows high potential to be implemented in real-world applications, such as surgical training, performance assessment, or automated use error analysis.

Keywords: Eye tracking, human action recognition, Hidden Markov Models, machine learning, automated task sequencing

1. Introduction

In the last two decades, accurate automated detection of human actions has attracted significant interest in a wide variety of fields. Due to its potential for improving workflow efficiency and quality assessment, human action recognition (HAR) has been used in fields such as health and surgical training [1]–[4], autonomous driving [5], [6] and human-robot interaction [7], [8]. Video-based HAR methods aim to achieve accurate detection of general human actions using only video information and have been able to show recognition accuracies of over 90% on the standard UCI HAR dataset [9] and over 70 % on the HMDB51 dataset [10], [11]. A major strength of video-based HAR is that general human actions can be accurately detected when trained with large quantities of training data [12].

However, in a variety of applications, the high effort required for training video-based machine learning (ML) algorithms often outweighs its benefits. In many applications, such as skill evaluation in surgical procedures [1] or the detection of manual handling tasks in usability studies, increasing the data labeling and algorithm training can increase classification accuracy [13], but are often cost-intensive and therefore unfeasible [14]. Consequently, researchers are often required to increase the complexity of the video-based HAR architectures to extract more information from the same video data, such as temporal relationships between video frames [15]–[17]. Instead of increasing architecture complexity, which generally correlates with an increase in computational cost [18], [19], others have suggested that better results may be achieved when including more features [20]. Including more features in our dataset can be achieved by using additional sensory input, such as hand, eye or object tracking.

In the past decade, mobile eye tracking (ET) has established itself as a valuable tool for the measurement of cognitive processes and visual attention, given the fact that the human gaze reveals important object and scene-related information during cognitively demanding manual tasks [21], [22]. In literature, ML models have successfully used gaze data to infer, detect and predict user actions [23] and thus shown that ML models could vastly benefit from exploiting human attention during human-object interaction [24]. Hidden Markov models (HMM) have been popular, due to their ability to create stochastic relations between observable parameters (gaze data) and hidden states (i.e. human actions), while being able to process the temporal relationship of actions within the input data [25], [26]. In mobile ET, contextual object-related gaze information is mainly obtained using tedious manual mapping procedures [27], [28] limiting the effectiveness and applicability for HAR. Consequently, a video-based HAR framework is needed that can combine automated object detection with gaze behavior information for accurate and context-specific action recognition using limited data.

In this paper, we propose a gaze-based HAR framework, the Peripheral Vision Based HMM (PVHMM) classifier, for the accurate recognition of a sequence of human actions in an exemplary manual handling task. The PVHMM uses Mask R-CNN for automated object of interest (OOI) detection in ET video recordings and uses context-aware gaze features as input for an HMM classifier. Two gaze features, the OOI Hits and the Object-Gaze Distance (OGD), which include information from the peripheral field of vision (Wang et al 2021), were investigated using 43 recordings. The gaze-guided action recognition is demonstrated on a mobile ET dataset of a medical device handling study. The performed handling task consisted of a specified safety-relevant sequence of seven distinct human-object interactions that follow the manufacturer's instructions for use. We conducted an ablation study to investigate the PVHMM's robustness to adversarial noise, using Gaussian noise and used leave-one-person out (LOPO) cross-validation for the investigation of the classification performance of new subjects. With our findings, we contribute a peripheral vision-based HAR framework that shows accurate results and thus offer a promising solution for the application of HAR in specific applications, using a small dataset.

2. Related work

Current research on human action recognition has focused on the use of video-based ML approaches, which are able to process general physical activities using large datasets. In specific applications where large training efforts and complex architectures outweigh the benefits, ET has been popular as an additional sensory tool to increase the informational density of the gathered data. Researchers have shown promising activity recognition results by combining gaze data with ML approaches such as support vector machines (SVM), random forests (RF) or hidden Markov models (HMM).

2.1. Video-based HAR

Over the years, many ML models, such as SVMs, RFs, HMMs, conditional random fields (CRF) and K-nearest neighbors (KNN), have been successfully applied for video-based HAR under strictly controlled environments [29]. More recently, neural network-based approaches, such as convolutional neural networks (CNN), recurrent neural networks (RNN) or long short-term memory (LSTM) have gained favorable attention, due to their remarkable performances in image and object classification on public HAR datasets. When the training and labeling efforts are limited, models such as the LSTM can increase informational density by making additional temporal connections. Ng et al. [15] employed an RNN that included Long Short-Term Memory (LSTM) cells and a 2D-ConvNet using images extracted from video recordings, as model input. Others have further increased their HAR model architectures and presented combinations of recurrent layers and 2D convolutional layers [17], 2D convolutional and recurrent layers with LSTM cell [30] or proposed 3D convolutional networks [31]–[35]. However, generally, the training efforts required of neural network-based HAR models for the recognition of very specific, task-related actions, can considerably limit the applicability of these state-of-the-art architectures.

2.2. Eye Tracking and Machine Learning

Eye movements have long been known to provide valuable insights into a person's thought processes [36]. The position and the duration of the human gaze can reveal crucial contextual information about the performed actions [37], while also being highly indicative of the cognitive processes during manual tasks [27], [38]. In recent years, more research efforts have been aimed towards answering the question of how gaze movement information can be exploited for the development of accurate HAR and workflow detection, using state of the art ML models. Predominantly, the evaluated gaze features have been either fixations, categorized as moments where the eyes are relatively still and visual information is extracted, or saccade, fast eye movements in between fixations, based. In their extensive review, Klaib et al. [23] presented a large variety of ET concepts, algorithms, methods and applications that used ML models over the last decades. Learning algorithms such as SVM, Naïve Bayes, RF, HMM and 2D and 3D CNN structures were used for the modeling of scan paths during visual inspection [39], the classification of ET data [25], [40]–[42], event detection [43] and disease detection [44].

Bader [45] developed a probabilistic model that categorizes fixations, into proactive and reactive, as well as goal- and object-related fixations. Using this approach they predicted users' intent to interact with a virtual object, achieving an average binary classification accuracy of 80.7%. Eivazi and Bednarik [46] applied a Support Vector Machine (SVM) based approach using seven combined fixation and saccade based gaze features and achieved an average accuracy of 53% during the classification of five different cognitive states. Similarly, Hu et al [42] have applied k-means clustering to the fixation coordinates of a screen-based graphical user interface study and achieved 64% classification accuracy using an SVM approach. Boisvert & Bruce [47] applied an RF classifier using fixation density features during static image viewing tasks, achieving an average classification accuracy of 43%. Going a step further, Huang et al. [22] have proposed a mutual context network (MCN) that can jointly achieve action-dependent gaze prediction and gaze-guided action recognition. The MCN achieved classification accuracies of 55.7% and 61.5% when tested on the EGTEA and GTEA Gaze + datasets, respectively.

2.3. Hidden Markov Models (HMMs) in Eye Tracking

Fixations do not have a fixed duration, which makes classical time-series classifiers less effective. Therefore, the sequential HMMs have been one of the most popular ML models for the classification of ET data. Great predictive performances have been achieved using HMMs in the recognition of computer tasks [48], control panel decision-making [49], face recognition tasks [50] and online e-learning activities [51]. HMMs that use gaze feature inputs were also applied for the detection of moving stimuli [52] and the classification of expert and novice operators during the visual inspection of front panels of a home appliance device [26]. Coutrot et al. [39] used a combined HMM and discriminant analysis model for the classification of tasks during the viewing of static natural scene images. Using the parameters of trained HMM models as gaze feature input, the suggested algorithm achieved an average classification accuracy of 55.9%.

2.4. Gaze Guided Human Activity Recognition (HAR)

Many of the above-mentioned approaches are concerned with user intention modeling or binary classification, yet the gaze guided classification of general human activities is addressed in only a few works [21]. Bulling et al. [53] performed the classification of five typical office activities using electrocardiography (ECG) signals on an SVM model, achieving an average of 76.1% precision and 70.5% recall. Compared to screen-based ET, mobile ET allows to capture human attention during interactions with tangible objects, but the dynamically changing scene increases the data evaluation effort manifold. However, to the best of our knowledge, only two studies have investigated HAR using mobile ET data. Kit & Sullivan [25] collected an ET data set of 5 natural tasks representative of daily living activities, such as navigating an office hallway, making a sandwich, or typing printed text into a word processor, and trained an HMM model on saccade amplitude and direction time series. The final model achieved an overall classification accuracy of 36% (with chance at 20%). Liao et al. [54] trained an RF classifier for classifying five real-world navigation tasks. Using five different types of gaze features, such as basic fixation and saccade based features and fixation density features, they were able to achieve an average accuracy of 67%.

Consequently, based on the promising results of the presented approaches, the use of context-rich gaze data has shown great potential for human action recognition, vindicating further exploration into the field of gaze guided HAR models. In this paper, we present a gaze guided HAR framework that exploits context-rich, object-related gaze data to explore the classification of coherent human action sequences using small datasets.

3. The proposed Peripheral Vision Based Hidden Markov Model framework (PVHMM)

The proposed HAR workflow of our Peripheral Vision Based Hidden Markov Model (PVHMM) is visualized in Fig 1. First, the data is collected using the Tobii Pro Glasses 2 eye tracker and fixations are detected using the default settings of Tobii Pro Labs (Version 1.162). Second, a Mask R-CNN algorithm is trained on the supervise.ly platform for automated object detection. We used the supervise.ly native image augmentation tool to increase the number of training images, including crop, flip, and gaussian blur. Third, gaze features are extracted and transformed for HMM input. In this paper, we investigated the HAR performance using two different gaze features, which were extracted as follows: The trained model and the gaze coordinates of each recording are concatenated using the cGOM (see Wolf et al. [55]) and the OGD (see Wang et al. [56]) algorithm. The cGOM algorithm matches the gaze point with the detected OOIs to create a list of OOI Hits. The OGD includes the wearer's peripheral gaze information by measuring the 2D Euclidean distance between the gaze center and the OOI in each frame of the ET recording. Subsequently, the peripheral gaze information is transformed into dictionary values, where each

Use Color in Print

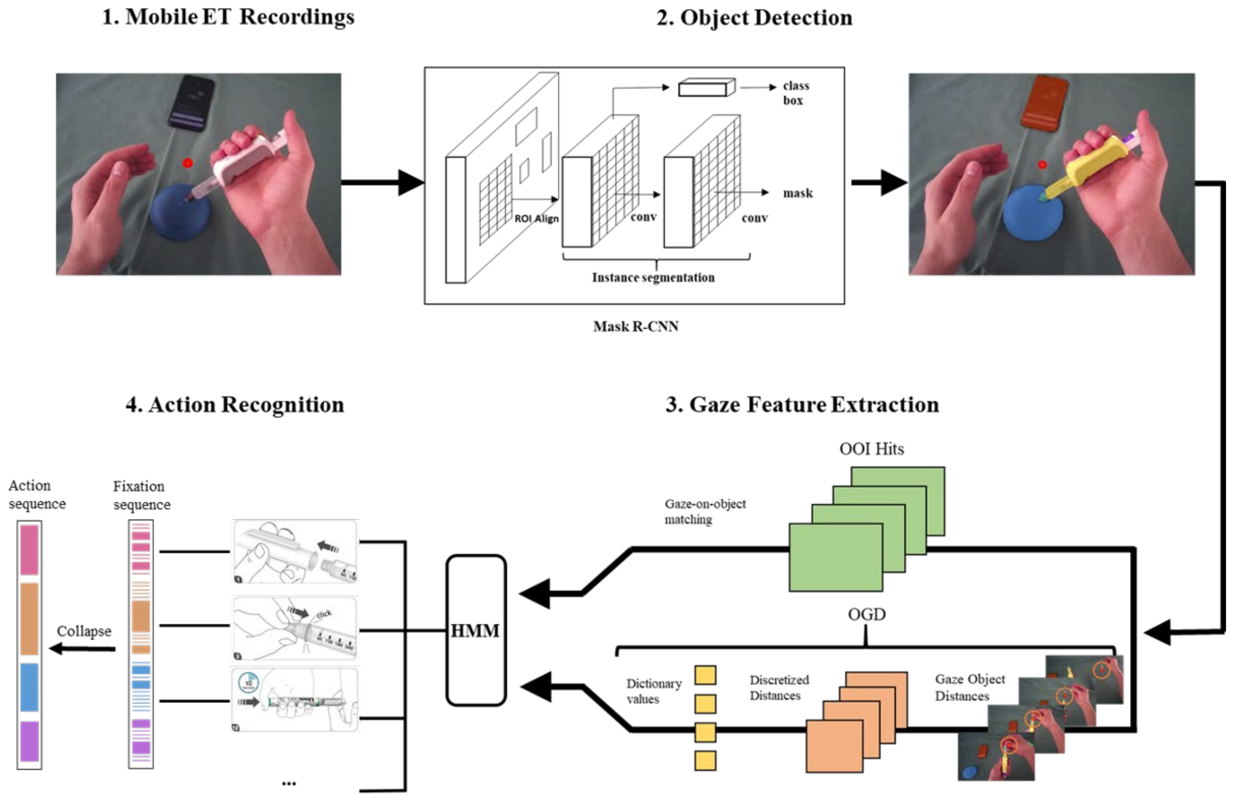


Fig. 1. The proposed Peripheral Vision Based HMM (PVHMM) framework consists of four main parts: (1) ET Data recording of the handling task. (2) Detection of objects of interest (OOI) using the MASK R-CNN algorithm. (3) Extraction and transformation of gaze features OOI Hits or object gaze distance (OGD). (4) Training and validation of the hidden Markov model (HMM) classifier for a fixation-by-fixation and action sequence classification.

entry contains a single numerical value. Fourth, action recognition is achieved using an HMM classifier and the investigated gaze features. The HMM allows the classification of actions on a fixation-by-fixation basis and, by collapsing successive fixations with the same action label into a single entry, a temporally connected sequence of the performed actions.

3.1. Mobile Eye Tracking Recording

The majority of publicly available HAR datasets, such as the UCI HAR [9], the HMDB51 (Yadav 2019) for the EGTEA Gaze + dataset [37] include only simple non-specific human actions, such as walking, drinking, or opening a fridge. Consequently, using these datasets can limit our knowledge of the models' performance, when applied to more specific scenarios involving complex actions, such as during product interaction, manufacturing, or surgery. Therefore, to validate the proposed approach a dataset was needed, which include human activities within a task-specific context with limited training data capabilities. To fulfil these criteria, a medical device handling procedure was chosen, which is defined by a safety-critical intended use, prescribed by the device's instruction for use (IFU). The gaze data was extracted as x-y gaze coordinates for each fixation, as calculated by the Tobii Pro Lab software and was used as input for the object segmentation (step 2 of the PVHMM), along with the video recordings.

3.1.1. Materials and Equipment

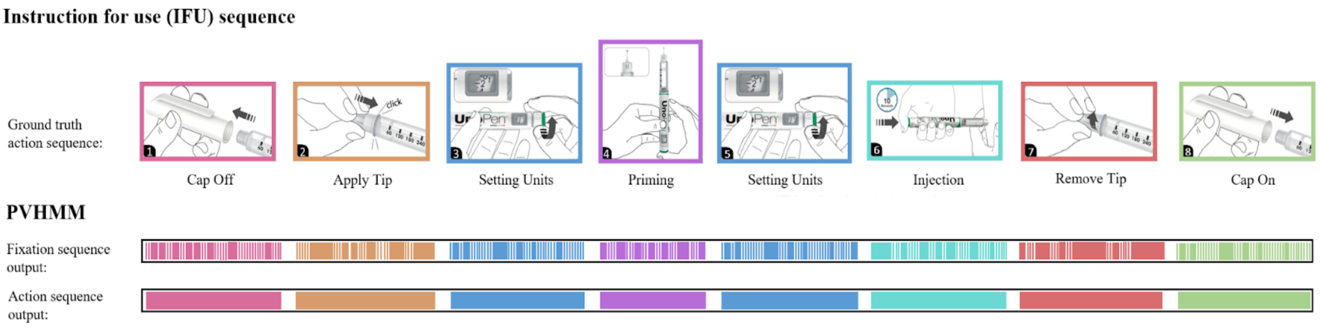
Data was recorded using the Tobii Pro Glasses 2 mobile ET glasses. The device has a reported accuracy of 0.6° at a distance of 1.5m. The front-facing camera has reported viewing angles of 82° horizontal and 52° vertical, recording with a resolution of 1920 x 1080 px at 25 frames per second and thus representing the user’s field of vision. The tracking percentage of the evaluated gaze samples was 85.2 7.2 %, as reported by the Tobii Pro Lab software (https://connect.tobiipro.com/s/article/Sample-percentage-calculated-in-Studio-Lab-and-Controller?language=en_US, WEB 28.01.2022). Algorithm training and testing were conducted on a GPU enabled workstation with the following characteristics: GPU NVIDIA GeForce RTX 2080 (8 Gb), working memory (16 Gb), CPU: AMD Ryzen 7 3700X 8-core processor.

3.1.2. Participants

Twenty subjects, university students and a PhD candidate, participated in the study (17 male, 3 female). All participants reported normal or corrected-to-normal vision and no neurological conditions. Due to the absence of task-native experts, seven subjects were trained to expert level on the chosen device handling task. The expert training was conducted prior to data recording and was carried out until subjects acquired the ability to finish the assembly repeatedly, without the use of the IFU and without making any mistakes. Each participant provided informed consent before testing and received a small monetary compensation.

3.1.3. Device Handling Task

To test the HAR system on a safety-relevant application that follows a specific set of sequential actions, a commercially available insulin injection device, the UnoPen, along with a novel connected add-on smart device prototype, the UnoPen Smartpilot, was chosen for the handling task. The Smartpilot add-on is connected to a smartphone application, which automatically counts the administered insulin units, displays instructions and displays the recommended holding time, during which the injection device is not to be removed from the injection site. Both devices were provided by the Swiss medical device manufacturer Ypsomed AG. For this task, participants were asked to carry out one successful injection into an injection pad (representative of the human body), using the UnoPen with the Smartpilot. The correct sequence of actions, given by the IFU, that are to be detected during the fourth step of the PVHMM framework is shown in Fig. 2. Additionally, the output of the HAR classifier of the PVHMM is depicted as both a fixation sequence and the resulting collapsed action sequence (step 4 of the PVHMM). Herein, a safe injection cycle, as given by the IFU, consists of a sequence of eight actions in the following order: 1) *Cap Off*, 2) *Apply Tip*, 3) *Setting Units*, 4) *Priming*, 5) *Setting Units*, 6) *Injection*, 7) *Remove Tip*, and 8) *Cap On*. Thus, the task consist of seven distinct actions, where the action *Setting Units* had to be carried out twice.



Use Color in Print

Fig. 2. The sequence of Human-Device actions as provided by the Instruction for Use (IFU) of the UnoPen (Ypsomed AG). A complete workflow consists of eight sequential actions, where Setting Units (3) is carried out twice, once before action priming (4) and once before action injection (6). A visualization of the output of the PVHMM is given for a fixation based HAR and for the resulting collapsed sequence of actions, where successive action labels are summarized into a single entry (step 4 of the PVHMM).

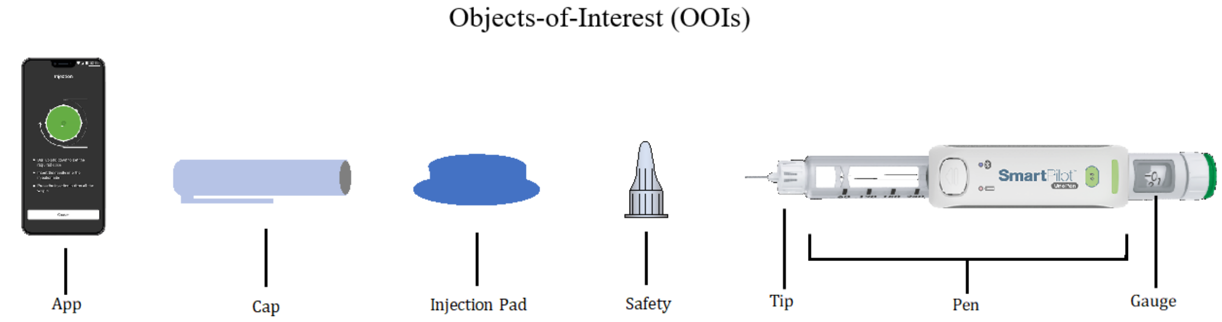


Fig. 3. The seven OOIs chosen for the object detection algorithm and the gaze feature transformation. The Smartpilot on top of the UnoPen is shown on the right and was separated into OOIs *Tip*, *Pen* and *Gauge*.

3.2. Object Segmentation

The object detection builds upon the Mask R-CNN (Region-Based Convolutional Neural Network) instance segmentation algorithm. Fig. 3. shows the seven OOIs which were chosen for the training of the algorithm, including both larger objects, such as the phone, the injection pad and the pen, as well as very small objects such as the unit gauge, the needle tip and the safety. OOIs such as *Gauge*, *Cap* and *safety* were chosen because of their distinct appearance in no more than two of the eight investigated actions. For example, users were expected to focus their visual attention on OOI Cap only during actions *Cap Off* and *Cap On*. Using this method for OOI selection and algorithm training, we expect a more accurate classification performance for the HMM classifier.

Images for training were extracted from several different participants recordings. 344 images were labeled for the training dataset. To increase the number of training images using a small dataset, and to avoid over-fitting we used image augmentation as a way of artificially proliferating the training images to boost the network's performance, using the supervise.ly platform. The following augmentations were included: multiplication, vertical flip, crop, rotation, Gaussian blur, contrast and brightness to obtain 55 728 augmented training images. The augmented images were grouped into a training and a validation set, with a split of 95:5. To increase the training efficiency of the neural network a pre-trained Mask R-CNN model was used. The model was trained using the transfer learning approach for an initial 15 epochs and a learning rate (LR) of 0.001, as well as a second training step using a LR of 0.0001 over 25 epochs to achieve the effect of LR decay. The quality of the masks was evaluated using the intersection over union (IoU) metric (see section 3.5.2).

We extracted the model weights from the epoch with the lowest loss value, which were used for object segmentation in all subsequent analyses.

3.3. Gaze Feature Extraction

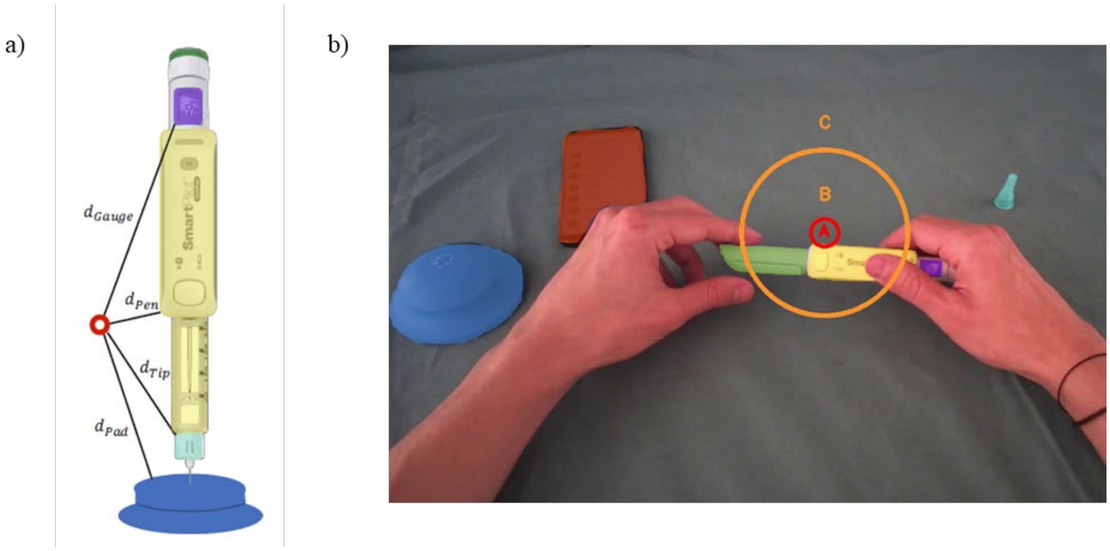
To examine the capability of the proposed framework to achieve accurate HAR using a small dataset and to determine the influence of information density of a gaze feature on the HAR performance, two fixation-based, object-related gaze features were investigated. Using the weights that were extracted from the object segmentation model training, the gaze coordinates of each trial are transformed into OOI hit [55] and the OGD [56] data format. The OOI hits feature stores the gaze information as a list of looked-at OOI, one entry for each fixation, which can be can be used to train the HMM model for action recognition as is. On the other hand, the OGD provides a 2D pixel (px) distance measurement between the gaze point and each segmented OOI, resulting in an x-by-seven sized matrix, where x is the number of fixations in a trial.

Since this format is not HMM compatible, we performed a gaze feature transformation, which is explained in the following sections.

3.3.1. Object Gaze Distance (OGD) Feature Transformation

In their recent work, Wang et al. [56] have presented a way to increase the informational value of gaze data by using automated, machine learning assisted computation of a 2D Euclidean distance between each OOI to the gaze point for each fixation. The calculation of distances to the gaze point allows for the simultaneous acquisition of time-series data for all OOIs, thus multiplying the object-related gaze data by the number of OOIs. Moreover, studies have also shown that near-peripheral vision is frequently used in the decision-making process [57] [58], allowing us to include more task-related information into our gaze features. The illustration in Fig. 4a. shows four detected OOIs (gauge, pen, tip, pad) and their matching distances to the gaze point (red circle). Fig. 4b. shows a visualization of the different fields of vision during the device handling study, displayed as red and orange rings around the gaze point. The recorded near-peripheral vision is restricted to a 52° degrees visual angle, which was given by the maximum measurable vertical viewing angle of the Tobii Pro Glasses 2 front camera.

As the HMM action recognition algorithm expects a sequence of single-dimensional observations as an input, for each fixation the OGD of multiple OOIs were transformed to obtain a single entry. First, the areas of each field of vision were assigned a letter from A to D (see Table 1). As the influence of the number of vision areas on the HMM classification performance was unknown at the time of our investigation, two different numbers of vision areas, three and four, were investigated. Then, the calculated distance of each OOI was retained as the letter of the matching vision area and strung together to form a word containing n_{OOI} letters, in alphabetical OOI order. Furthermore, a dictionary was created with each possible combination of OOI positions in a subject's visual field $n_c = n_P * n_{OOI}$, where n_P, which represents the number of chosen vision areas, was given a unique dictionary integer value within the range R:{ y | 1 ≤ $y < n_{OOI}$}. When using three vision areas ($n_P = 3$) and seven OOIs ($n_{OOI} = 7$), a dictionary is created with $n_c = 2187$ values, ranging from AAAAAAA→ $n_{dic} = 1$ to CCCCCCC→ $n_{dic} = 2187$. Similarly, using four vision areas and the same number of OOIs results in n_c= 16'384 dictionary values. The number of dictionary values is therefore directly influenced by the number OOIs



Use Color in Print

Fig. 4. In a) the Euclidean pixel distances from the gaze point (red) to each OOI is visualized. The output of the OGD consists of an array of 2D distances, such as OGD = [d_{Gauge} , d_{Pen} , d_{Tip} , d_{Pad}], with one entry for each OOI. In b) three fields of vision, foveal (circular area A), para- and perifoveal (circular area B) and peripheral vision (area C, rest of the image) are shown in a snapshot of the egocentric ET recording of the insulin injection task.

and the number of vision areas chosen for analysis. Noticeably, many dictionary values do not lie within the solution space, since many string combinations (such as AAAAAAA) are physically impossible. Due to the small area of the foveal field of vision, only a limited combination of OOIs exist that can lie within the area at the same time.

After OGD gaze transformation, a one-dimensional numerical vector with one entry per fixation is obtained, which still holds contextual information about each object and its positions within the peripheral field of vision. This vector is used as a gaze feature input for the HMM model for HAR. We expect that by discretizing the OGDs, we can reduce both the computational complexity of the classifier while achieving high classification accuracies using small datasets.

3.4 Hidden Markov Model (HMM)

Due to its ability to process sequential time-series data the PVHMM framework utilizes a HMM as the main HAR classifier. A HMM estimates the probability of the presence of unmeasurable hidden states X = $(x_1, \dots, x_t)$, through a sequence of measurable associated observations $O = (o_1, \dots, o_t)$. Observations can be any of those features (fixation durations, OOI Hits, transformed OGD, etc...) that convey information about the probability of occurrence of underlying hidden states. These observations $o_t \in K = (k_1, \dots, k_M)$ are contained in K, the set of possible observables. The set of possible hidden states is defined by S, $x_t \in S = (s_1, \dots, s_N)$ which in the presented case is the set of seven possible actions during the device handling task. HMMs are built upon the Markov assumption, which states that the probability of each event is solely dependent on the current observation and the previous state, which can be written as

$$\mathrm{P}\,(x_t \mid x_1, \dots, x_t - 1) = \mathrm{P}\,(x_t | x_t - 1) \tag{1}$$

Table 1. Vision areas used for the gaze features for the HMM, determined by biological accordance, converted into pixel ranges that accounts for the resolution of the Tobii ET glasses 2.

Assigned Letter	Biological Vision Area	Pixel Range
Three vision areas		
A	foveal	$d_{AOI} \leq 25\,px$
B	parafoveal + perifoveal	$25\,px \leq d_{AOI} \leq 150\,px$
C	near-peripheral	$150\,px \leq d_{AOI}$
Four vision areas		
A	foveal	$d_{AOI} \leq 25\,px$
B	parafoveal	$25\,px \leq d_{AOI} \leq 60\,px$
C	perifoveal	$60\,px \leq d_{AOI} \leq 150\,px$
D	near-peripheral	$150\,px \leq d_{AOI}$

Furthermore, HMMs use two probabilistic distributions P_A and P_B. P_A models the conditional transition probabilities from state s_i to s_j, a_{ij}=P_A $(x_t = s_j|x_{t-1} = s_i)$, where each transition is stored in a transition matrix a. The transition matrix is commonly initialized during training of the model and carries the transition probabilities between each pair of hidden states, which in our case is the transition probability between different actions such as *Setting Units* and *Remove Tip*. The transition probabilities can be adjusted manually, based on task-related information [48] in order to increase the model accuracy. PB models the conditional probability of observable emission k_l, given the state s_i, where each emission is displayed in the emission matrix $b_{i,l}$=P_B $(o_t = k_l \mid x_t = s_i)$. Thus, HMM is defined by a tuple with five elements $\mu = (S, K, P, a, b)$. For the actual prediction, the Viterbi algorithm was used to find the most likely sequence of hidden states X, given an observation sequence O (for example an OOI sequence). Thereby, the Viterbi algorithm seeks optimization for the globally complete sequence [59].

Within the PVHMM, the HMM takes a one-dimensional data vector of one trial, containing either OOI hits or the transformed OGD, and predicts one pre-trained action for each fixation. Thus, for the evaluation of the HAR performance, we obtain a vector with a fixation-by-fixation prediction of actions for each investigated handling trial.

3.5 Experimental Validation

The proposed PVHMM action recognition framework was evaluated in two parts: First, we tested the classification accuracy under laboratory conditions using data that was collected with two expert subjects. Both expert subjects (authors of this work) were familiar with the task and showed the ability to perfectly perform the task prior to the data recording. To prevent overfitting of the HMM model, we included trials with alternative order action sequences that did not follow the IFU sequence of eight actions, including repeated and omitted actions. Consequently, the dataset contains trials with actions sequences of various lengths. Using this set of 25 trials (16 normal order and 9 alternative order sequences), we conducted an ablation study, where we evaluated the influence of the number of vision areas on the performance of the classifier and the robustness of the system to Gaussian noise, using the leave-one-out cross-validation method (LOO). The goal of the ablation study was to choose the optimal parameters for our ML framework using a small dataset while retaining the larger dataset for the final evaluation. Second, we included 18 more trials from 5 new experts and 13 novices, resulting in a total of 43 recorded trials from 20 participants, with a mean of 142.58 ± 44.17 fixations and a mean of 8.23 ± 0.86 performed actions per trial. We applied a leave-one-person-out (LOPO) cross-validation method, changing the subject used for evaluation in each fold for cross-validation. This way we are able to conduct a subject-independent evaluation for the performance of the framework for new subjects [60], [61]. To determine the influence of the information density in the engineered gaze features on the HAR performance, all evaluations were performed on two gaze features, the OOI Hits and the OGD.

3.5.1. Ablation Study

To see the influence of different gaze features on the classification accuracy of the HMM, the algorithm performance was evaluated using two gaze features, the OOI Hits and the transformed OGD, for both the LOO and LOPO cross-validation. The two gaze features differ in the object-related informational density, since OOI hits carry binary gaze on OOI information, while the OGD contains information of the distance of the gaze center to each OOI and thus carries information of the near-peripheral area of vision. To gain a better understanding of the advantages and limitations of our proposed PVHMM classification framework, we evaluated the following range of parameters during a preliminary ablation study:

Number of vision areas. To evaluate the influence of the number of discretized fields of vision on the classification accuracy, we have evaluated a set of three and four vision areas, namely ABC (foveal, parafoveal & perifoveal and near-peripheral area of vision) and ABCD (foveal, parafoveal, perifoveal and near-peripheral area vision). The HAR performance was evaluated in the ablation study, using LOO cross-validation.

Gaussian Noise. To test the framework's robustness to noise, we apply three Gaussian noise levels on the raw gaze point coordinate data. Here, we chose the peak noise level according to the outer threshold of each field of vision, 25 px, 60 px and 150 px. The influence of noisy data on the classification accuracy was compared using both the OOI Hit and the OGD gaze feature and evaluated during the ablation study.

3.5.2. Evaluation Metrics

First, the performance of the object detection algorithm was evaluated, due to the dependency of the generated gaze features on the quality of the detected OOI masks. Therefore, the object detection algorithm was assessed using the widely used intersection over union (IoU) metric, which is calculated by dividing the area of overlap between the predicted and the manually labeled mask by their area of union. Additionally, the precision, recall and the f1-score were computed. The instance segmentation performance was evaluated on a set of 20 sample images from 5 study recordings. The sample images mostly contain scenes of human interaction with the OOIs, increasing the complexity of detection through occlusion. The sample images contained 91 ground truth OOIs. The widely used IoU with a confidence threshold of 0.5 and 0.7 sets a standard for the predicted masks to be evaluated, by only considering those masks for IoU calculation that were predicted with a confidence value of above 0.5 and 0.7, respectively.

Second, we evaluate the classifier accuracy on a fixation-by-fixation level, where each fixation contains a single classification made by the HMM model. The ability of the PVHMM to accurately detect the correct fixation sequence is evaluated using the f1-score. Due to the novelty of the intended application of the proposed framework, i.e. for workflow and performance analysis in HAR, traditional classifier evaluation metrics do not suffice for the evaluation of all relevant aspects. Therefore, for the evaluation of the performance of our PVHMM framework, we introduce three new action sequence based metrics. The HAR performance is evaluated using the *action sequence accuracy,* the *action sequence sensitivity* and the *action sequence precision* metrics. For this evaluation, consecutive fixations with the same classified action label are collapsed into a single entry (see Fig. 1.), transforming a fixation sequence into an actions sequence. The introduced metrics compare the list of actions recognized by the PVHMM to the manually labeled ground truth. A classification error was counted for each misclassified, undetected and/or wrongfully detected action. Here, the duration of the collapsed actions and their deviation from the ground truth was not considered.

We evaluated the classifier using the following four evaluation metrics:

f1 - score. Frequently used in ML classifier performance evaluations, the f1-score is calculated as the harmonic mean of the precision and recall and used to evaluate the fixation-by-fixation recognition accuracy. This value is expected to be higher than the action sequence evaluation metrics due to the larger number of fixations that influence the score (i.e. one wrongly classified fixation does not significantly influence the f1-score, while during the action sequence accuracy one wrongly classified fixation can signify a wrongly detected action that can have a big influence on the metrics described below).

Action sequence accuracy (1 - (no. of classification errors/ no. of classifications made))**.** The accuracy gives a sense of the overall ability of the classifier to achieve the ground truth action sequence. An action sequence accuracy of 1.0 signifies that the algorithm is able to classify the performed sequence of actions perfectly. If the number of errors exceeds the number of classifications, for example when the HMM classifies a sequence of two actions but the ground truth contains a sequence of eight actions, the value 0 is assigned.

Action sequence sensitivity (no. of correctly made classifications/total no. of sequences in ground truth)**.** The action sequence sensitivity allows us to quantify the number of actions that were correctly identified by the classifier, compared to the number of actions contained in the ground truth. A maximum value of 1.0 signifies a correct detection of all actions of the ground truth sequence.

Action sequence precision (no. of correctly made classifications/ total no. of made classifications)**.** The action sequence precision expresses the fraction of correctly classified actions, out of all classifications made by the algorithm. It gives an indication of how precise the classifier makes a prediction if a prediction is made. A precision value of 1.0 means that all classifications that were made were made correctly.

4. Results

In order to judge the validity of our proposed PVHMM framework, we evaluate and report the results of the object segmentation algorithm, the influence of Gaussian noise and choice of vision areas, as well as the main evaluation using a small dataset with many different subjects.

4.1. Object Segmentation

The object detection model shows excellent values for the classification metrics across all objects for confidence thresholds 0.5 and 0.7. The qualitative results showed that differences between predicted and ground-truth masks mostly occur when OOIs were close to one another (e.g. Tip and Pen), which can cause misdetections between OOIs. Averaged over all seven OOIs, the trained model achieved a mean IoU of 0.82 and 0.90 for thresholds 0.5 and 0.7 (see Table 2). For smaller OOIs, such as OOI safety and for OOI tip the mean IoUs showed lower values, indicating a less accurate segmentation performance. Imperfectly segmented object masks of small OOIs, therefore, have a non-neglectable effect on the IoU. The results suggest that the IoU is biased towards larger OOIs due to the higher ratio of circumference to the surface area, but that the algorithm can be used for gaze feature extraction (step 3 of the PVHMM) with high confidence.

4.2. Ablation Studies

Table 3. shows the results of the HMM classifier's performance within our ablation study, using the leave-one-out (LOO) cross-evaluation. When using three vision areas, without added noise in the raw gaze data, both the OOI Hit and the OGD gaze feature show very high f1-scores, actions sequence accuracies, action sequence sensitivities and action sequence precision. Specifically, during the fixation sequence

Table 1. Vision areas used for the gaze features for the HMM, determined by biological accordance, converted into pixel ranges that accounts for the resolution of the Tobii ET glasses 2.

		0.5 IoU			0.7 IoU		
OOI	Mean IoU	Precision	Recall	F_1- score	Precision	Recall	F_1- score
App	0.95	1.00	1.00	1.00	1.00	1.00	1.00
Cap	0.88	1.00	1.00	1.00	0.94	1.00	0.97
Gauge	0.89	1.00	1.00	1.00	1.00	1.00	1.00
Pad	0.92	1.00	1.00	1.00	1.00	1.00	1.00
Pen	0.85	1.00	1.00	1.00	0.95	0.95	0.95
Safety	0.78	0.94	0.94	0.94	0.82	0.82	0.82
Tip	0.51	0.80	0.80	0.80	0.60	0.60	0.60
OOI Average	0.82	0.96	0.96	0.96	0.90	0.90	0.90

classification, the use of the peripheral vision based OGD feature led to a statistically significant increase ($p = 0.05$) of the f1-score. The evaluation of the action sequence classification performance yields very similar results for both gaze features and no statistical significance was observed. After increasing the number of discretized vision areas from three to four during the OGD gaze feature transformation, the model shows slightly reduced performances in both the fixation sequence and action sequence metrics. The ablation study of the vision areas suggests that the use of more vision areas, and thus more dictionary values, seem to lead to a small, statistically non-significant decrease in the classification performance.

The classifier's robustness was investigated by evaluating the classification performance in response to noisy gaze data. Contrary to our initial expectations, the introduction of a small Gaussian noise of 25 px, which is the size of the foveal field of vision, resulted in a statistically non-significant increase in the mean action sequence accuracy for both the OOI Hits ($p = 0.282$) and the OGD feature ($p = 0.405$). With a further increase of the noise level to 60 px and 150 px the HMM's performance decreases noticeably, showing the lowest overall f1-score for both OOI Hits (0.439 ± 0.12) and OGD (0.791 ± 0.11) at the highest investigated noise level. However, even when subjected to an adversary noise of up to 150 px, the OGD based PVHMM showed a reasonably high classification performance, significantly outperforming the OOI hits based PVHMM in the f1-score ($p < 0.001$), actions sequence accuracy ($p < 0.001$) and action sequence sensitivity ($p < 0.001$). The results suggest that the implementation of peripheral gaze information can provide higher model stability and classification accuracy when dealing with noisy data.

The results of the ablation study suggest that both investigated gaze features for the PVHMM result in similar performances, even though the OGD based PVMM performs significantly better on a fixation-by-fixation level. Based on the findings of the ablation study, the subsequent LOPO evaluation of the OGD method was investigated using three vision areas (ABC) and without additional Gaussian noise.

Table 3. Results of the Leave-one-out (LOO) cross-validation ablation study. The evaluation metrics are shown for three (ABC) and four (ABCD) vision areas, as well as added Gaussian noise of 25 px, 60 px and 150 px.

Leave-one-out (LOO)	**No Noise (ABC)**	**No Noise + ABCD**	**25 px noise (ABC)**	**60 px noise (ABC)**	**150 px noise (ABC)**
f1 - score					
OOI Hits	0.908 ± 0.07*		0.888 ± 0.08*	0.746 ± 0.13*	0.439 ± 0.12*
OGD	0.952 ± 0.03*	0.935 ± 0.05 *	0.958 ± 0.03	0.949 ± 0.04	0.832 ± 0.10
Subtask accuracy					
OOI Hits	0.866 ± 0.15		0.909 ± 0.12	0.765 ± 0.27	0.281 ± 0.25*
OGD	0.871 ± 0.12	0.855 ± 0.14	0.900 ± 0.11	0.864 ± 0.14	0.807 ± 0.18*
Subtask Detection Sensitivity					
OOI Hits	0.995 ± 0.02		0.979 ± 0.06	0.888 ± 0.16*	0.593 ± 0.13*
OGD	0.992 ± 0.04	0.992 ± 0.04	1.00 ± 0.00	0.984 ± 0.05*	0.781 ± 0.20*
Subtask Detection Precision					
OOI Hits	0.874 ± 0.15		0.922 ± 0.11	0.961 ± 0.07*	0.885 ± 0.12
OGD	0.881 ± 0.12	0.865 ± 0.14	0.895 ± 0.14	0.893 ± 0.13*	0.839 ± 0.16

* differences between the gaze features OOI Hits and OGD are statistically significant ($p < 0.05$)

4.3. Main Evaluation

In the main evaluation of the algorithm's performance, the data was evaluated using a leave-one-person-out (LOPO) cross-validation method, where the trial of a single subject is used as test set for each iteration. Fig. 5. shows the confusion matrices for the classification of the seven investigated actions during the device handling task, based on the fixation-by-fixation classification. Here, we see that on average the HMM that was trained using the OGD gaze feature achieves higher true positive classification rates for all investigated actions. Especially for actions *Priming* and *Remove Tip,* both involving small OOIs, the rate of true positive classified fixations is increased by over 20 %. Moreover, the confusion matrices suggest that the HMM is capable of distinguishing actions that involve similar movements and identical OOIs, such as *Cap On/Cap Off* and *Apply Tip/Remove Tip,* in both observed instances. However, both HMMs show difficulties in differentiating actions *Setting Units, Priming* and *Injection.* Overall, most misclassified fixations are wrongfully recognized as either the preceding or succeeding action.

Table 4. shows the quantitative result of the LOPO cross-validation of the full insulin injection dataset. The HMM that was trained using the OGD gaze feature significantly outperforms the HMM trained with the OOI Hits feature, showing a more accurate classification on a fixation-by-fixation level, as well as on a collapsed action sequence. Between the two gaze features, the inclusion of peripheral vision significantly increased the f1-score (p = 0.009) and the action sequence sensitivity (p = 0.025), while achieving higher mean action sequence accuracy (p = 0.170) and action sequence precision (p = 0.24). Even when subjected to data sequences of subjects the HMM was not trained on, the high action sequence sensitivity (0.958 ± 0.10) shows that the model can accurately recognize the majority of the human actions of the ground truth action sequence when using the OGD feature. Compared to the LOO evaluation, the addition of 18 trials of 18 new subjects led to a decrease in the classification performance. However, the PVHMM using the OGD feature as input still shows the ability to classify fixation sequences and action sequences with reasonable accuracy, with an average f1-score of 0.849 ± 0.09 and an average action sequence accuracy of 0.840 ± 0.20.

Use Color in Print

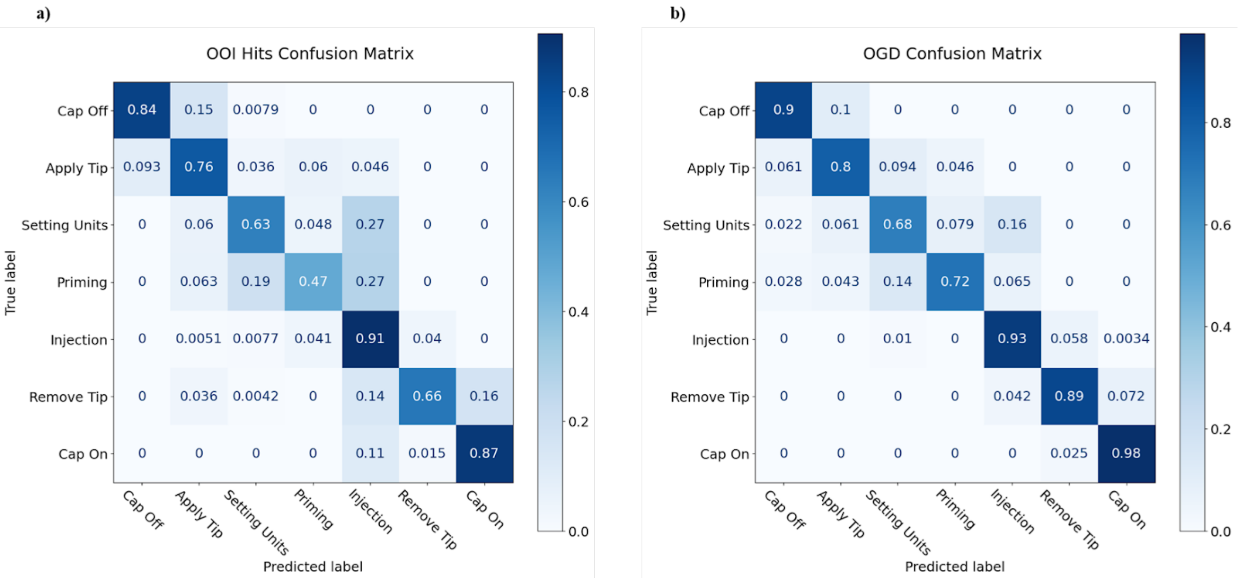


Fig. 5. Confusion matrix of the HAR classification results of the PVHMM trained using the OOI Hits (a) and OGD (b) gaze feature.

Table 5. Results of the Leave-one-person-out (LOPO) cross-validation of the main study.

Leave-one-person-out (LOPO) evaluation

	f1- score	Actio Sequence Accuracy (chance: 0.142)	Action Sequence Sensitivity	Action Sequence Precision
OOI Hits	0.697 ± 0.21*	0.713 ± 0.31*	0.822 ± 0.22	0.853 ± 0.18
OGD	**0.849 ± 0.09***	**0.840 ± 0.20***	**0.958 ± 0.10**	**0.871 ± 0.18**

* statistically significant difference between gaze features OOI Hits and OGD ($p < 0.05$), bold indicates highest values

5. Discussion

In this manuscript, we presented a new HAR framework that combines contextual task-related gaze information, video-based object segmentation and a hidden Markov model for the HAR of coherent human actions during physical device handling. The results show that the use of context-rich gaze features as input for an HMM model is highly effective for human action recognition during a medical device handling task using a small dataset of fewer than 45 data samples. The object segmentation algorithm was shown to work well within the PVHMM framework. Smaller objects were harder to detect and the overall performance might improve with higher IoUs for OOIs such as *safety* and *tip.* In the main data evaluation, both investigated gaze features, the OOI Hits and OGD, achieved remarkable overall classification performances with f1-scores of 70 % and 85 %, respectively. The presented study of the PVHMM framework has provided the following insights:

The introduction of noisy gaze data leads to a decrease of the PVHMM's classification performance at high noise levels, regardless of the chosen gaze feature input. Contrary to our expectations, small noise of up to 25 px has been shown to slightly increase the average algorithm performance. We assume that a small noise level can shift the gaze center in situations where it lies close to small OOIs, such as OOI needle*,* to a location where the gaze center coincides with said OOI. Consequently, in some cases, a small noise-induced gaze point offset can lead to an advantageous increase in the algorithm performance that should be investigated further, particularly, when using OOI Hits as a gaze feature. Higher noise distortions affect the ability of the HMM to accurately detect the correct action and lead to overall lower performance since it becomes more likely that an OOI Hit can be falsely detected as either a "no Hit", or as an OOI Hit on a different OOI within close proximity. Conversely, as the OGD feature does not rely on binary gaze-on-target logic, it shows remarkably higher robustness to noise, confirming suggestions by Wang et al. [56] that the quantification of peripheral gaze information can be used effectively for HAR.

High fixation-by-fixation classification performances during the main LOPO cross-validation show that human actions can be detected within the time window of single fixations, which commonly have a duration of 100 to 600 ms. Compared to other studies, where the data samples of each classifiable action range from 2 to 11 minutes [25], the proposed gaze-based features show the potential for the application in real-world scenarios.

In their work, Pirsiavash et al. [62] have highlighted the importance of classification and segmentation of continuous streams compared to the more widespread classification of pre-segmented action clips. The presented PVHMM has shown the capability of processing long video segments containing a sequence of coherent actions. Furthermore, using the engineered gaze features, the HMM has been able to differentiate effectively between human actions that involve the same objects, background and often very similar hand-object movements. The results have shown the HMMs tendency to mislabel some fixations as either the preceding or proceeding action. Courtemanche et al. [48] have shown that a stability factor for the transition properties can optimize the reactivity of the HMM, which could be included in future iterations.

The confusion matrices have shown that the OOI Hit trained HMM classifier experienced difficulties in differentiating between actions *Setting Units, Priming* and *Injection.* While subjects would primarily focus their attention on OOI *Tip* during action *Priming* and on OOI *Gauge* during *Setting Units*, the gaze point would often switch between *Gauge* and *Needle* when subjects were carrying out action *Injection.* This highlights a shortcoming for the OOI Hit gaze feature, which only carries information on the currently looked at object and can thus confuse the HMM classifier, which was trained to recognized both actions using the same gaze behavior. These misclassifications were reduced when using the OGD as gaze feature for the HMM. As the OGD feature incorporates the distance of the gaze center to all OOIs, the HMM can leverage the added peripheral gaze information to differentiate between fixations with a gaze point close to only OOI T*ip* (action *Priming*) and fixations with the gaze point close to both OOI *Tip* and *Pad* (action *injection*). This trend supports our theory that the same classifiers trained with gaze features that include more contextual information can improve the HAR performance. However, misclassifications between actions *Setting Units* and *Injection* still occur, since in both cases all other distinct OOIs lie outside of the foveal, parafoveal and perifoveal field of vision. Consequently, in cases where multiple actions contain identical gaze behavior and OGD dictionary values, a drop in classification performance has to be considered.

Both investigated gaze features have been shown to lead to great HAR performances within the PVHMM framework. The use of peripheral gaze information using the OGD has additionally led to improvements such as increased classification accuracies when subjected to inter-subject variability, as well as increased classification rates of actions involving small objects.

In real-world applications, where the collection of large data quantities is often unfeasible, such as complex surgery procedures, the PVHMM constitutes a valuable alternative to more advanced video-based ML models. The ability to achieve f1-scores of up to 95 % and action sequence accuracies of up 87 %, when using the data of expert individuals, can be especially useful for the assessment of applications with repetitive execution, such as the monitoring of assembly processes and the prevention of quality problems in a manufacturing assembly line [63], [64]. The increasing popularity of ready-to-use ET systems and the integration of ET in popular AR and VR head-mounted displays, such as the Hololens 2 and the HTC Vive Pro Eye, has made the technology more accessible to the public. Nevertheless, research using eye movement data for the classification of human actions via ML systems has been scarce [21]. We have shown that the PVHMM performs well in a specific manual task involving human device interaction, using a modern connected medical device.

We believe that the PVHMM provides a significant contribution to the advancement of ET based HAR, showing accurate results can be achieved during the classification of a manual handling task involving small objects, closely related and coherent actions. The use of context-rich gaze features as input for an HMM model, such as the OGD, has achieved high classification performances for HAR using only a small dataset of fewer than 45 data samples and

6. Limitations

Naturally, the presented evaluation possesses some limitations, which we would like to address. First of all, we have investigated HAR within a specific exemplary use case using only gaze-based data. Consequently, in our evaluation, other sensory information, such as tactile feedback or hand tracking was neglected and could be considered for future evaluations. We are planning to study the presented framework in the context of more use cases, to validate the presented findings. Furthermore, Wolf et al. [65] have shown that hand and eye movement data can be successfully combined for a human action prediction model. With head-mounted AR displays now capable of capturing even more contextual data, such as hand movement data, we expect that the framework's performance can be further improved. The main evaluation of the gaze data from novice individuals shows that for cases with a high number of inter-person handling variability, despite the reasonably high classification performance, the overall accuracy decreases. Therefore, future investigations could include an extensive study on the influence of more training data on the PVHMM performance.

7. Conclusion

In the present work, we introduced the Peripheral Vision Based HMM (PVHMM), a gaze-based HMM framework for the classification of task-specific human action sequences. Through the evaluation of a manual handling task of a medical device, we have shown that the PVHMM allows the classification of coherent human action sequences and shows accurate results, using a small mobile ET dataset of 43 samples. By investigating two object-related gaze features, evidence was given that using context-rich data leads to remarkable classification performances during the non-trivial classification of multiple, visually similar sequences of actions. Additionally, we were able to show that the application of a peripheral vision-based object gaze distance (OGD) gaze feature led to an increased classification performance, as well as increased robustness to noise, compared to a traditional gaze feature. The validation using an exemplary manual handling task showed the potential to achieve accurate HAR with limited data resources, creating an opportunity for applications in environments such as performance assessment and surgical workflow analysis in the future.

8. Acknowledgments

We would like to thank Berk Nergiz for his valuable support during the collection of the eye tracking data and Ypsomed AG for providing the UnoPen materials and the Smartpilot prototype for the study. This research was financially supported by Innosuisse, the Swiss Innovation Agency, under Project No. 43122.1.

9. References

[1] S. Eivazi, M. Slupina, W. Fuhl, H. Afkari, A. Hafez, and E. Kasneci, “Towards automatic skill evaluation in microsurgery,” *Int. Conf. Intell. User Interfaces, Proc. IUI*, pp. 73–76, Mar. 2017, doi: 10.1145/3030024.3040985.

[2] N. Padoy, “Machine and deep learning for workflow recognition during surgery,” *Minim. Invasive Ther. Allied Technol.*, vol. 0, no. 0, pp. 1–9, 2019, doi: 10.1080/13645706.2019.1584116.

[3] T. Czempiel *et al.*, “TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12263 LNCS, pp. 343–352, Oct. 2020, doi: 10.1007/978-3-030-59716-0_33.

[4] E. Zdravevski *et al.*, “Improving Activity Recognition Accuracy in Ambient-Assisted Living Systems by Automated Feature Engineering,” *IEEE Access*, vol. 5, pp. 5262–5280, 2017, doi: 10.1109/ACCESS.2017.2684913.

[5] T. Billah, S. M. M. Rahman, M. O. Ahmad, and M. N. S. Swamy, “Recognizing distractions for assistive driving by tracking body parts,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1048–1062, Apr. 2019, doi: 10.1109/TCSVT.2018.2818407.

[6] E. Ohn-Bar and M. M. Trivedi, “Looking at humans in the age of self-driving and highly automated vehicles,” *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 90–104, 2016, doi: 10.1109/TIV.2016.2571067.

[7] R. Mojarad, F. Attal, A. Chibani, S. R. Fiorini, and Y. Amirat, “Hybrid Approach for Human

Activity Recognition by Ubiquitous Robots," *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 5660–5665, Dec. 2018, doi: 10.1109/IROS.2018.8594173.

[8] L. Martínez-Villaseñor and H. Ponce, "A concise review on sensor signal acquisition and transformation applied to human activity recognition and human–robot interaction:," *https://doi.org/10.1177/1550147719853987*, vol. 15, no. 6, Jun. 2019, doi: 10.1177/1550147719853987.

[9] N. Sikder, M. S. Chowdhury, A. S. M. Arif, and A.-A. Nahid, "Human Activity Recognition Using Multichannel Convolutional Neural Network," *2019 5th Int. Conf. Adv. Electr. Eng. ICAEE 2019*, pp. 560–565, Jan. 2021, doi: 10.1109/icaee48663.2019.8975649.

[10] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human Action Recognition from Various Data Modalities: A Review," Dec. 2020, Accessed: Nov. 05, 2021. [Online]. Available: https://arxiv.org/abs/2012.11866v4.

[11] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Syst.*, vol. 223, p. 106970, Jul. 2021, doi: 10.1016/J.KNOSYS.2021.106970.

[12] X. Ji, J. Cheng, D. Tao, X. Wu, and W. Feng, "The spatial Laplacian and temporal energy pyramid representation for human action recognition using depth sequences," *Knowledge-Based Syst.*, vol. 122, pp. 64–74, Apr. 2017, doi: 10.1016/J.KNOSYS.2017.01.035.

[13] M. Stonebraker and E. K. Rezig, "Machine Learning and Big Data: What is Important?," 2019.

[14] J. Zhou, R. Cao, J. Kang, K. Guo, and Y. Xu, "An Efficient High-Quality Medical Lesion Image Data Labeling Method Based on Active Learning," *IEEE Access*, vol. 8, pp. 144331–144342, 2020, doi: 10.1109/ACCESS.2020.3014355.

[15] J. Yue *et al.*, "Beyond Short Snippets: Deep Networks for Video Classification." pp. 4694–4702, 2015.

[16] S. H. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Impact of fully connected layers on performance of convolutional neural networks for image classification," *Neurocomputing*, vol. 378, pp. 112–119, Feb. 2020, doi: 10.1016/J.NEUCOM.2019.10.008.

[17] C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention based LSTM networks," *Appl. Soft Comput. J.*, vol. 86, 2020, doi: 10.1016/j.asoc.2019.105820.

[18] H. M. Romero Ugalde, J. C. Carmona, J. Reyes-Reyes, V. M. Alvarado, and J. Mantilla, "Computational cost improvement of neural network models in black box nonlinear system identification," *Neurocomputing*, vol. 166, pp. 96–108, Oct. 2015, doi: 10.1016/J.NEUCOM.2015.04.022.

[19] K. Chen *et al.*, "77 Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges, and Opportunities," *ACM Comput. Surv*, vol. 54, 2021, doi: 10.1145/3447744.

[20] C. Jobanputra, J. Bavishi, and N. Doshi, "Human activity recognition: A survey," in *Procedia Computer Science*, Jan. 2019, vol. 155, pp. 698–703, doi: 10.1016/j.procs.2019.08.100.

[21] K. Lukander, M. Toivanen, and K. Puolamäki, "Inferring intent and action from gaze in naturalistic behavior: A review," *International Journal of Mobile Human Computer Interaction*, vol. 9, no. 4. IGI Global, pp. 41–57, Oct. 01, 2017, doi: 10.4018/IJMHCI.2017100104.

[22] Y. Huang, M. Cai, Z. Li, F. Lu, and Y. Sato, "Mutual Context Network for Jointly Estimating Egocentric Gaze and Action," *IEEE Trans. Image Process.*, vol. 29, pp. 7795–7806, 2020, doi: 10.1109/TIP.2020.3007841.

[23] A. F. Klaib, N. O. Alsrehin, W. Y. Melhem, H. O. Bashtawi, and A. A. Magableh, “Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies,” *Expert Systems with Applications*, vol. 166. Elsevier Ltd, p. 114037, Mar. 15, 2021, doi: 10.1016/j.eswa.2020.114037.

[24] Y. Rong, W. Xu, Z. Akata, and E. Kasneci, “Human Attention in Fine-grained Classification,” 2021, Accessed: Dec. 13, 2021. [Online]. Available: http://arxiv.org/abs/2111.01628.

[25] D. Kit and B. Sullivan, “Classifying mobile eye tracking data with hidden Markov models,” in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, MobileHCI 2016*, 2016, pp. 1037–1040, doi: 10.1145/2957265.2965014.

[26] B. H. Ulutas, N. F. Özkan, and R. Michalski, “Application of hidden Markov models to eye tracking data analysis of visual quality inspection operations,” *Cent. Eur. J. Oper. Res.*, vol. 28, no. 2, pp. 761–777, 2020, doi: 10.1007/s10100-019-00628-x.

[27] M. F. Land and M. Hayhoe, “In what ways do eye movements contribute to everyday activities?,” in *Vision Research*, Nov. 2001, vol. 41, no. 25–26, pp. 3559–3565, doi: 10.1016/S0042-6989(01)00102-X.

[28] B. Reimlinger, Q. Lohmeyer, R. Moryson, and M. Meboldt, “A comparison of how novice and experienced design engineers benefit from design guidelines,” *Des. Stud.*, vol. 63, pp. 204–223, Jul. 2019, doi: 10.1016/J.DESTUD.2019.04.004.

[29] L. Minh Dang, K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon, “Sensor-based and vision-based human activity recognition: A comprehensive survey,” *Pattern Recognit.*, vol. 108, p. 107561, Dec. 2020, doi: 10.1016/J.PATCOG.2020.107561.

[30] V. Hernandez, T. Suzuki, and G. Venture, “Convolutional and recurrent neural network for human activity recognition: Application on American sign language,” *PLoS One*, vol. 15, no. 2, p. e0228869, Feb. 2020, doi: 10.1371/journal.pone.0228869.

[31] Y. Wan, Z. Yu, Y. Wang, and X. Li, “Action Recognition Based on Two-Stream Convolutional Networks with Long-Short-Term Spatiotemporal Features,” *IEEE Access*, vol. 8, pp. 85284–85293, 2020, doi: 10.1109/ACCESS.2020.2993227.

[32] S. H. S. Basha, V. Pulabaigari, and S. Mukherjee, “An Information-rich Sampling Technique over Spatio-Temporal CNN for Classification of Human Actions in Videos,” Feb. 2020, Accessed: Jan. 25, 2022. [Online]. Available: https://arxiv.org/abs/2002.02100v2.

[33] N. Almaadeed, O. Elharrouss, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, “A Novel Approach for Robust Multi Human Action Recognition and Summarization based on 3D Convolutional Neural Networks,” 2019, Accessed: Jan. 25, 2022. [Online]. Available: https://www.researchgate.net/publication/334735494.

[34] S. N. Boualia and N. E. Ben Amara, “3D CNN for Human Action Recognition,” in *18th IEEE International Multi-Conference on Systems, Signals and Devices, SSD 2021*, Mar. 2021, pp. 276–282, doi: 10.1109/SSD52085.2021.9429429.

[35] Z. Shou, D. Wang, and S. F. Chang, “Temporal action localization in untrimmed videos via multi-stage CNNs,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 1049–1058, doi: 10.1109/CVPR.2016.119.

[36] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer, *Eye Tracking A Comprehensive Guide to Methods and Measures*, vol. 53, no. 9. 2013.

[37] Y. Li, M. Liu, and J. Rehg, “In the Eye of the Beholder: Gaze and Actions in First Person Video,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021, doi: 10.1109/TPAMI.2021.3051319.

[38] S. Fuchs, “Gaze-Based Intention Estimation for Shared Autonomy in Pick-and-Place Tasks,” *Front. Neurorobot.*, vol. 15, p. 647930, Apr. 2021, doi: 10.3389/fnbot.2021.647930.

[39] A. Coutrot, J. H. Hsiao, and A. B. Chan, “Scanpath modeling and classification with hidden Markov models,” *Behav. Res. Methods*, vol. 50, no. 1, pp. 362–379, Feb. 2018, doi: 10.3758/s13428-017-0876-8.

[40] Y. Yin, C. Juan, J. Chakraborty, and M. P. McGuire, “Classification of Eye Tracking Data Using a Convolutional Neural Network,” in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, Jan. 2019, pp. 530–535, doi: 10.1109/ICMLA.2018.00085.

[41] R. Bhattarai and M. Phothisonothai, “Eye-Tracking Based Visualizations and Metrics Analysis for Individual Eye Movement Patterns,” in *JCSSE 2019 - 16th International Joint Conference on Computer Science and Software Engineering: Knowledge Evolution Towards Singularity of Man-Machine Intelligence*, Jul. 2019, pp. 381–384, doi: 10.1109/JCSSE.2019.8864156.

[42] B. Hu, X. Liu, W. Wang, R. Cai, F. Li, and S. Yuan, “Prediction of interaction intention based on eye movement gaze feature,” in *Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC 2019*, May 2019, pp. 378–383, doi: 10.1109/ITAIC.2019.8785737.

[43] R. Zemblys, D. C. Niehorster, O. Komogortsev, and K. Holmqvist, “Using machine learning to detect events in eye-tracking data,” *Behav. Res. Methods*, vol. 50, no. 1, pp. 160–181, Feb. 2018, doi: 10.3758/s13428-017-0860-3.

[44] I. M. Pavisic *et al.*, “Eyetracking metrics in young onset alzheimer’s disease: A Window into cognitive visual functions,” *Front. Neurol.*, vol. 8, no. AUG, p. 1, Aug. 2017, doi: 10.3389/fneur.2017.00377.

[45] T. Bader, M. Vogelgesang, and E. Klaus, “Multimodal integration of natural gaze behavior for intention recognition during object manipulation,” *ICMI-MLMI’09 - Proc. Int. Conf. Multimodal Interfaces Work. Mach. Learn. Multimodal Interfaces*, pp. 199–206, 2009, doi: 10.1145/1647314.1647350.

[46] S. Eivazi and R. Bednarik, “Predicting Problem-solving Behavior and Performance Levels from Visual Attention Data,” *Proc. Work. Eye Gaze Intell. Hum. Mach. Interact. IUI*, pp. 9–16, 2011, Accessed: Nov. 10, 2021. [Online]. Available: http://tobii.se.

[47] J. F. G. Boisvert and N. D. B. Bruce, “Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features,” *Neurocomputing*, vol. 207, pp. 653–668, Sep. 2016, doi: 10.1016/j.neucom.2016.05.047.

[48] F. Courtemanche, E. Aïmeur, A. Dufresne, M. Najjar, and F. Mpondo, “Activity recognition using eye-gaze movements and traditional interactions,” *Interact. Comput.*, vol. 23, no. 3, pp. 202–213, May 2011, doi: 10.1016/j.intcom.2011.02.008.

[49] J. Grobelny and R. Michalski, “Applying hidden markov models to visual activity analysis for simple digital control panel operations,” in *Advances in Intelligent Systems and Computing*, 2017, vol. 523, pp. 3–14, doi: 10.1007/978-3-319-46589-0_1.

[50] T. Chuk, A. B. Chan, S. Shimojo, and J. H. Hsiao, “Eye movement analysis with switching hidden Markov models,” *Behav. Res. Methods*, vol. 52, no. 3, pp. 1026–1043, 2020, doi: 10.3758/s13428-019-01298-y.

[51] A. Elbahi, M. N. Omri, and M. A. Mahjoub, “Possibilistic reasoning effects on Hidden Markov Models effectiveness,” in *IEEE International Conference on Fuzzy Systems*, Nov. 2015, vol. 2015-Novem, doi: 10.1109/FUZZ-IEEE.2015.7338045.

[52] J. Kim, S. Singh, E. D. Thiessen, and A. V. Fisher, “A hidden Markov model for analyzing eye-tracking of moving objects: Case study in a sustained attention paradigm,” *Behav. Res. Methods*, pp. 1–19, Jan. 2020, doi: 10.3758/s13428-019-01313-2.

[53] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, “Eye Movement Analysis for Activity Recognition Using Electrooculography,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 741–753, Apr. 2011, doi: 10.1109/TPAMI.2010.86.

[54] H. Liao, W. Dong, H. Huang, G. Gartner, and H. Liu, “Inferring user tasks in pedestrian navigation from eye movement data in real-world environments,” *https://doi.org/10.1080/13658816.2018.1482554*, vol. 33, no. 4, pp. 739–763, Apr. 2018, doi: 10.1080/13658816.2018.1482554.

[55] J. ; Wolf, S. ; Hess, D. ; Bachmann, Q. ; Lohmeyer, and Meboldt, “Automating areas of interest analysis in mobile eye tracking experiments based on machine learning,” *J. Eye Mov. Res.*, 2018, doi: 10.3929/ethz-b-000309840.

[56] F. S. Wang, J. Wolf, M. Farshad, M. Meboldt, and Q. Lohmeyer, “Object-Gaze Distance: Quantifying Near- Peripheral Gaze Behavior In Real-World Application,” *J. Eye Mov. Res.*, vol. 14, no. 1, pp. 1–13, 2021, doi: 10.16910/jemr.14.1.5.

[57] E. M. Reingold and H. Sheridan, “Eye movements and visual expertise in chess and medicine,” in *The Oxford Handbook of Eye Movements*, Oxford University Press, 2012.

[58] E. A. Krupinski *et al.*, “Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience,” *Hum. Pathol.*, vol. 37, no. 12, pp. 1543–1556, Dec. 2006, doi: 10.1016/j.humpath.2006.08.024.

[59] A. Allahverdyan and A. Galstyan, “Comparative analysis of Viterbi Training and Maximum Likelihood estimation for HMMs,” Dec. 2011, Accessed: Jan. 26, 2022. [Online]. Available: https://arxiv.org/abs/1312.4551v1.

[60] D. Gholamiangonabadi, N. Kiselov, and K. Grolinger, “Deep Neural Networks for Human Activity Recognition with Wearable Sensors: Leave-One-Subject-Out Cross-Validation for Model Selection,” *IEEE Access*, vol. 8, pp. 133982–133994, 2020, doi: 10.1109/ACCESS.2020.3010715.

[61] H. Gunduz, “Deep Learning-Based Parkinson’s Disease Classification Using Vocal Feature Sets,” *IEEE Access*, vol. 7, pp. 115540–115551, 2019, doi: 10.1109/ACCESS.2019.2936564.

[62] H. Pirsiavash and D. Ramanan, “Parsing videos of actions with segmental grammars,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 612–619, doi: 10.1109/CVPR.2014.85.

[63] J. Chen *et al.*, “Current status of artificial intelligence applications in urology and their potential to influence clinical practice,” *BJU International*, vol. 124, no. 4. Blackwell Publishing Ltd, pp. 567–577, Oct. 01, 2019, doi: 10.1111/bju.14852.

[64] K. Bauters, J. Cottyn, D. Claeys, M. Slembrouck, P. Veelaert, and H. van Landeghem, “Automated work cycle classification and performance measurement for manual work stations,” *Robot. Comput. Integr. Manuf.*, vol. 51, pp. 139–157, Jun. 2018, doi: 10.1016/J.RCIM.2017.12.001.

[65] J. Wolf, Q. Lohmeyer, C. Holz, and M. Meboldt, “Gaze Comes in Handy: Predicting and Preventing Erroneous Hand Actions in AR-Supported Manual Tasks,” Nov. 2021, pp. 166–175,

doi: 10.1109/ismar52148.2021.00031.