International Journal of Neural Systems, Vol. 30, No. 7 (2020) 2050025 (13 pages)

DOI: 10.1142/S0129065720500252

World Scientific
www.worldscientific.com

Behavioral Activity Recognition Based on Gaze Ethograms

Javier De Lope
Department of Artificial Intelligence
Universidad Politécnica de Madrid (UPM)
Madrid, Spain
javier.delope@upm.es

Manuel Graña*
Computational Intelligence Group
University of the Basque Country (UPV/EHU)
San Sebastian, Spain
manuel.grana@ehu.eus

Accepted 5 March 2020
Published Online 9 June 2020

Noninvasive behavior observation techniques allow more natural human behavior assessment experiments with higher ecological validity. We propose the use of gaze ethograms in the context of user interaction with a computer display to characterize the user's behavioral activity. A gaze ethogram is a time sequence of the screen regions the user is looking at. It can be used for the behavioral modeling of the user. Given a rough partition of the display space, we are able to extract gaze ethograms that allow discrimination of three common user behavioral activities: reading a text, viewing a video clip, and writing a text. A gaze tracking system is used to build the gaze ethogram. User behavioral activity is modeled by a classifier of gaze ethograms able to recognize the user activity after training. Conventional commercial gaze tracking for research in the neurosciences and psychology science are expensive and intrusive, sometimes impose wearing uncomfortable appliances. For the purposes of our behavioral research, we have developed an open source gaze tracking system that runs on conventional laptop computers using their low quality cameras. Some of the gaze tracking pipeline elements have been borrowed from the open source community. However, we have developed innovative solutions to some of the key issues that arise in the gaze tracker. Specifically, we have proposed texture-based eye features that are quite robust to low quality images. These features are the input for a classifier predicting the screen target area, the user is looking at. We report comparative results of several classifier architectures carried out in order to select the classifier to be used to extract the gaze ethograms for our behavioral research. We perform another classifier selection at the level of ethogram classification. Finally, we report encouraging results of user behavioral activity recognition experiments carried out over an inhouse dataset.

Keywords: Neuroethology; activity recognition; gaze tracking; gaze ethogram; screen-based eye tracker; noninvasive eye tracker.

*Corresponding author.

1. Introduction

The computer-based recognition and subsequent analysis of the human and animal behavior falls in the broad field of Computational Ethology.[3] Human studies are often referred as human activity recognition (HAR).[61] Most HAR works reported in the literature use computer vision techniques[38] and/or wearable inertial sensors.[65] However, most studies are oriented to the identification of low level activities, such as abnormal behavioral situations in the elderly.[41] Hence, the studies do not deal with higher level behavior representations, i.e. ethograms. An ethogram is a time plot of the low level actions carried out by the subject under observation that provides a high level representation which has been used for animal phenotypical characterization.[1] Specifically, in this paper, we are interested in the characterization of behavioral states of a laptop computer user.

The underlying hypothesis of our work can be stated as follows: The subject's gaze fixation information allows to determine what kind of behaviors and activities the subject is engaged in Ref. 27 . The same hypothesis underlies the business of neuromarketing[33] companies that routinely analyze the visual behavior of users while visiting a web page, trying to asses the marketing value of the user's visual interaction. In our study, we want to assess the behavioral state of a laptop computer user on the basis of the sequence of gaze fixations on the computer display that we call *gaze ethograms*. There is a trade-off between eye tracking accuracy and invasiveness. The most accurate techniques for eye tracking are very invasive. Electrooculography (EOG) and video-oculography (VOG)[8] use a series of electrodes situated in the user's face to measure the eye movement, and a head-mounted mask that is equipped with small cameras, respectively. Glass frames with mounted infrared-based eye trackers are mildly invasive but very accurate. However, we aim to carry out observation and measurement with minimal or no interference to the natural behavior.

To carry out our experimental exploration of the behavioral activity recognition on the basis of gaze ethograms in a noninvasive way, we have developed an open source screen-based desktop eye tracker and gaze fixation area estimation system that uses conventional laptop web cameras without any additional hardware. Both conventional techniques[2,23,40,48] and artificial neural networks[5,47,51,60] have been used to develop systems for eye tracking. However, in order to detect the display target area, we need an additional model that maps the eye tracking information into display fixations. We solve the issue by applying additional machine learning models that have become ubiquitous in neuroscience and behavior studies.[29,44] In this paper, we report encouraging experimental results, showing that the sequence of gaze fixations on the screen detected by our system can be composed into an ethogram that allows the discrimination of the actual behavioral activity being carried out.

The rest of the paper is organized as follows. Section 2 reviews the background of behavioral activity recognition. Section 3 revisits some concepts concerning eye movements and gaze analysis. We provide a short view of the state-of-the-art to set the stage for our proposal. Section 4 details how user activities can be characterized by ethograms and how they can be compared and classified. We describe how we recognize the behavioral activities that a user carries out in front of a computer. Section 5 describes our gaze tracking system. Section 6 describes the recording procedure for (a) the gaze tracking system training data, and (b) the data for the activity classification based on the gaze ethogram. Section 7 provides the experimental results of both the gaze ethogram-based behavioral activity recognition, and the tuning of the gaze tracking system. Finally, Sec. 8 provides our conclusions and directions for future work.

2. Behavioral Activity Recognition

Machine learning-based recognition of behavioral activities belongs to the broad scientific field of Computational Neuroethology[16,22,30] that deals with the causal or correlation relationship between observable behavior and the neural pathways in the brain and central nervous system. However, as we are not including direct neural activity observation in our computations, this paper can be considered a Computational Ethology study.[3] A central tool of ethological research is the *ethogram*, i.e. the quantitative representation of the observed behavior as a time sequence of elementary actions that are recognized independently. High level reasoning about

the behavior can be carried out over the ethogram representation.[3]

Much effort on human activity recognition research is currently directed to the monitoring of the aging people,[34] and to the improvement of performance in sports.[4] Though monitoring elderly people is motivated by behavioral decline due to neurodegenerative diseases, no behavioral activity recognition is pursued, because the goal is to detect abnormal situations in order to raise alarms.[41] Often the kind of activity modeled is rather atomic and short timed, like sitting, walking, standing up, or falling. Fall detection and gait analysis in elderly people have attracted a lot of researches.[56] However, up to this date, we found no reference of works at the level of composing and recognizing human ethograms. There are two main modalities of the sensing systems: wearable and external. The wearable sensors are most of the time inertial measurement units,[63] which provide very limited information and suffer from drift, sensitivity to electromagnetic noise, and other artifacts.[62] Some works try to infer activity information from physiological sensors measuring heart rate,[9] but such indirect methods are very unreliable. External sensors are often cameras that can be complemented with depth information (RGB-D).[42] They are minimally invasive, but require the subject to be in the camera field of view. Such computer vision-based approaches have been extensively developed for animal behavior accurate measurement allowing to extract detailed ethograms for machine learning-based phenotyping.[35] However, we have not found works on ethogram-based HAR analysis previous to our own work reported here. The requirement of non-invasiveness is critical for ecologically valid behavior observation, hence we emphasize in this paper, the role played by our own developed gaze tracking system.

3. Gaze Tracking Background

3.1. *Eye movements*

Three basic types of eye movements can be detected by sensors which can be either attached to the face or remote: saccadic eye movements, fixations and blinking. *Saccades* are quick, simultaneous movements of both eyes between two or more phases of fixation in the same direction.[11,21] The brain does not retrieve information about the visual stimulus from these movements, because it suppresses visual sensation during saccades to avoid blur/motion. It has been found that the spatial distribution of eye movements remains optimal after losing central vision.[58] *Visual fixations* occur between saccades. A fixation is the sustained gaze during a time interval in a specific direction which falls upon a single location in the visual stimulus. The averaged duration of fixations is 200 ms. Fixation is needed because of the limited detailed visual angle (2°). Information retrieval is carried out by the brain during fixation periods, which take most of the viewing time. Fixations are not static positions of the eye, instead tremor, drift, and micro-saccades occur during fixation periods. They serve to stabilize the gaze on the point of interest and to prevent adaptation, which eventually produces the fading of the image from the perceived view. Tremor and micro-saccades are high frequency motions that can be confused with noise in high temporal resolution systems $\geq$300 Hz, or be undetectable by low resolution systems. Finally, *Blinking* is a semi-autonomic rapid closing of the eyelids. Generally, the rate of blinking is about 12–18 blinks per minute, although it may decrease to about 3–4 times per minute when the eyes are focused on an object for an extended period of time, such as when reading. The averaged duration of blinking is about 200–300 ms.

3.2. *Gaze detection state-of-the-art*

Gaze detection has been a research and application area for a long time.[19] Some early successful systems[67] were based on EOG, the recording of electrodes placed around the eye, and the use of scleral contact lenses rigged with a coil that allowed measurement of motion in an electromagnetic field. Such systems are very invasive. Optical-based systems use specific illumination systems (often infrared) that enhance the detection of eye features such as the pupil and the cornea for *point of regard* estimation. Such systems, are less invasive but still impose very stringent positioning of illumination sources and cameras. However, there is a need for much less invasive systems, that do not require the subject to have in hand and wear specific technology. Systems based on computer vision have recently been proposed based on the localization of the eyebrows,[24] the estimation of the 3D face motion from single camera,[50]

and deep learning architecture[66] for demanding environment in neuroscience studies.

Gaze information has been used for diagnostic and active interaction purposes.[20] Gaze interaction has been used for communicating with people suffering extreme disability,[6] and for games. Diagnostic applications have been widespread in areas such as diverse, neuroscience and marketing. Some recent reported examples: Determining how students' visual attention may influence school failure,[57] and evaluating the decision-making process during sports playing.[64] The entropy of gaze trajectory has been found to be a good detector of alcohol induced driving impairment,[53] gaze cuing (following the gaze of a partner in a social interaction) is not impaired in patients with Alzheimer's Disease (AD), though it impairs the distinction between direct and averted gaze.[36] Gaze detection systems have been used to confirm that impaired gaze leading abilities in subjects with autism spectrum conditions underly joint attention impairment[32] as well as diverse mechanisms for the detection of direct versus adverted gaze,[52] correlated gaze behavior and electroencephalography readings were found in a boredom study.[39] Gaze detection has been used for the analysis of facial expression exploration in subjects with social anxiety.[31] Finally, gaze detection and tracking contributes to the analysis of general cognitive processes.[28] discovering the mechanisms underlying visual memory, visual attention and learning.

4. Gaze Ethogram Modeling of Behavioral Activities

We recall that an *ethogram* is a representation of the sequence of actions that a subject is executing while engaged in an activity or a series of activities. From the ethogram, we can extract higher level information, such as the frequency or probability that an action is followed by another (either the same or a different one)[3] when carrying out a specific activity.

In our study, the atomic action is the area of the display that receives the user attention, which is determined by the gaze fixation. Hence, we consider *gaze ethograms*. An activity is what the user is doing in front of the computer for a period of time. Under the assumption that we can predict the screen area that receives the user attention, we want to use this information to identify the behavioral activity that the user is carrying out.

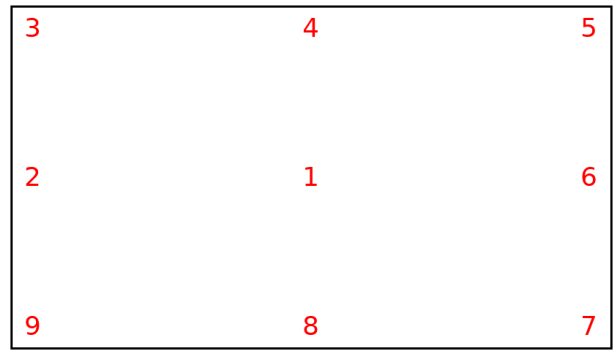

Fig. 1. Calibration template with the identification of the target areas for gaze localization. Note that the target order has been arbitrarily defined to reduce the user fatigue during the calibration.

The target screen areas are identified in Fig. 1. For gaze ethogram inference, it is not needed to receive the gaze destination coordinates on the screen with the resolution required, for example, in an HCI application, where gaze fixation coordinates may be used to choose an option or to gain access to a specific feature. Thus, the broad set of gaze fixation targets determine broad display areas that may receive the user's attention. Gaze detection system is implemented by a gaze fixation screen area predictor using the pupil and face pose information as features.

Figure 2 depicts a gaze ethogram that has been extracted from the fixations detected by the gaze tracking system while the user is carrying out the activity "reading a text". The activity has been video recorded for 200 s. The vertical axis corresponds to the target display area that receives the user's attention, the horizontal axis gives the time stamp when the sample images are taken. The gaze destination

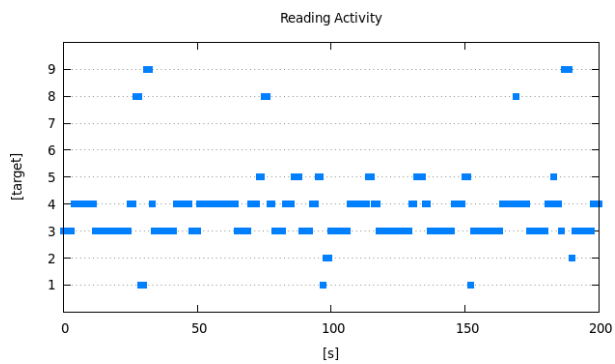

Fig. 2. Gaze ethogram of a reading activity.

falls mainly in the target corresponding to the top area of the display, namely targets 3, 4, and 5.

5. Gaze Tracking System

As discussed above, for the purposes of our research on gaze ethogram modeling of behavioral activities, we needed to develop an efficient gaze tracking working on–off the shelf laptop computers. In this section, we provide a description of this system emphasizing some innovative contributions.

5.1. *Overall description of the gaze tracking system*

The system hardware configuration is a laptop computer endowed with a web camera on top of the screen upon which a user is working. The distance of the face to the camera is roughly 0.5 m, and the camera view of the face is frontal, although the subject can move freely and change pose at will. The goal of our system is to identify the screen area where the user gaze is fixated while he is carrying out some behavioral task. We are using off the shelf web cameras that are factory installed in laptops, therefore robustness is a challenge and a limitation. The resolution of those cameras is limited and the quality of the image is quite low. Additional difficulties arise from the uncontrolled illumination conditions, and the user freedom of movement in front of the camera.

The gaze destination detection process is decomposed into a pipeline of low level tasks depicted in Fig. 3. Firstly, we localize the user's face in the image. Once the face position in the image is known, we extract the facial landmarks. Next, we compute the eye aspect ratio (EAR) that is used to determine if the eyes are closed in an image to detect eye blinks.

We test two approaches to estimate the gaze direction. The first approach is based on the image coordinates of the centers of the pupils. The second approach is based on novel local texture features. The system implementation in Python is published as free open source code.[15]

5.2. *Face localization*

Face localization can be be achieved in a number of ways, such as edge image matching to face edge templates.[25,26] In our system, we use a pre-trained detector based on histograms of oriented gradients

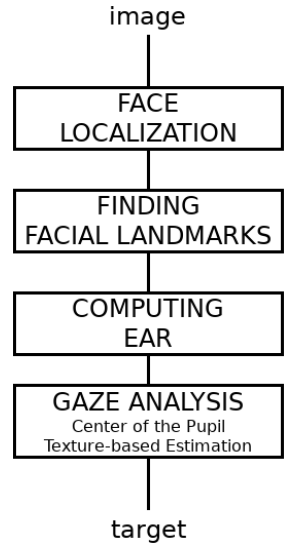

Fig. 3. Overall processing pipeline to estimate the destination of the gaze. EAR = eye aspect ratio.

(HOG)[43] as input features for classification by linear support vector machines (SVM).[12] The HOG-based detection has been successfully applied to human shape detection in pedestrian datasets with a large range of pose variations and backgrounds.[13] The HOG-based method has been already compared to Haar wavelets as descriptors for classification using polynomial kernel SVM[45,49] or AdaBoost.[59]

5.3. *Face alignment*

The face alignment problem has been addressed using several techniques, such as a cascade of regression functions,[10,18] consensus of exemplars,[7,17] conditional regression forests,[14] and nonparametric shape models.[54] We use an ensemble of regression trees to estimate the face landmark positions directly from a sparse subset of pixels intensities.[37] The method returns 68 2D points in the image that describe that can be used to localize the eyes, eyebrows, nose, mouth, and jawline. This approach allows almost real-time response. We have found trouble when the user is wearing some kind of glasses during the data capture.

5.4. *Eye aspect ratio*

The eye aspect ratio (EAR)[55] is defined as the proportion between height and width of an eye. It is used to determine if the eye is open (the value is mostly constant) or it is closed (the value will be getting

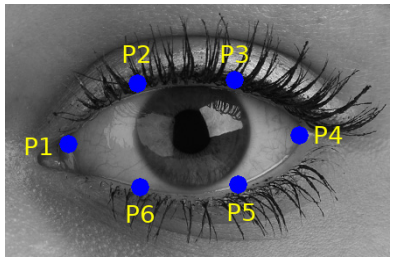

Fig. 4. Points used to compute the EAR.

near to zero). The ratio is computed as shown in (1).

$$\mathrm{EAR} = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\,\|p_1 - p_4\|}, \tag{1}$$

where $p_1, \ldots, p_6$ are the points that describe the eye position that have been detected in the previous step. p_1 and p_4 correspond to the left and right edges of the eye, p_2 and p_3 are two points above the eye about the intersection between the eyelid and the pupil, and p_6 and p_5 are the points below the eye relative to the two previous ones. The points are numbered clockwise starting from the leftmost one as shown in Fig. 4. The EAR is very helpful in order to determine if the gaze destination is the top or the bottom of the display. The greater the EAR value, the higher the gaze destination in the screen (EAR value is zero when the eye is closed).

We use EAR for the detection of blinking events, setting an empirically derived threshold. Specifically, when EAR ≥ 0.18 for both eyes, we consider that the blinking may be starting. The duration of blinking is in the range of 200–300 ms. Thus, we use this estimation to determine if the user is just blinking or if he or she almost closed the eyes.

5.5. *Center of the pupil*

One of the proposed methods for gaze fixation determination is based on the centers of the pupils of both eyes, which are estimated as follows: Following the line between the points p_1 and p_4 in Fig. 4, we find out the horizontal coordinates where the iris starts at both sides. Then, we compute the middle point, which is used as an approximation to the horizontal center of the pupil. This is the initial point from which the upper and lower eyelids are searched to determine the vertical coordinates. The final result of this process is shown in Fig. 5.

Once the approximate centers of the pupils are determined, they must be referred to a coordinate

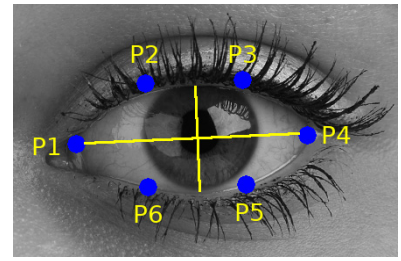

Fig. 5. The coordinates in the image of the center of the pupil are determined by computing the intersection of the orthogonal longest lines that cross the iris.

frame relative to the face in order to obtain invariance to user head motion. For this purpose, two face landmark points computed during the face alignment are used as the base reference. Specifically, these are the points between both eyebrows and the nose lower point.

5.6. *Texture-based gaze estimation*

The second set of features for gaze fixation estimation are computed from the area determined by four of the facial landmark points detected during the face alignment. Specifically, we use the upper eyelid landmarks p_2 and p_3, and the lower eyelid landmarks p_5 and p_6. Those points define a polygon whose centroid is computed. The centroid induces the partition of the polygon in four sub-polygons denoted A, B, C and D in Fig. 6. The averaged intensity of each subpolygon is then computed as well as the global intensity of the main polygon. The polygon average global intensity is used to normalize the intensity within each subpolygon. Thus, we detect which areas are darker or lighter than the global polygon. The iris and the pupil will fall in those darker areas, while

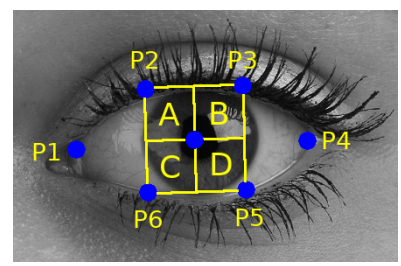

Fig. 6. The structure of subpolygons used to estimate the gaze destination obtained from landmark points p_2, p_3, p_5, and p_6 and the centroid of the polygon defined by them.

the rest of the eye image will correspond to lighter areas.

6. Data Capture and Experimental Design

Our experimental work has two phases. Firstly, we have to tune our gaze tracking implementation by training the display target area predictor. Secondly, we have done experiments on the behavioral activity recognition. We have separate datasets for these experiments. In this section, we describe how we collected the data and how it was used in the computational experiments.

6.1. *Gaze tracking system calibration data*

The gaze tracking system calibration data consists of a collection of pairs {face–eye image, display target area}. There is a special label for images where the user has closed eyes. For the experiments reported in this paper, this data has been obtained from two research collaborators that have quite different iris colors looking for some robustness in the subsystem for eye image processing. For general system deployment, fine tuning for specific users may be desirable. To obtain the calibration pairs, the volunteers were positioned in front of the laptop, looking to the display target areas in a controlled sequence and timing, so that the training images where the user gaze is fixated on a target area can be easily extracted from the calibration videos using the time stamp. The dataset for the training and validation experiments reported below contains more than 4000 images.

The calibration data recording sessions are carried out in an office during the workday with natural light and temperature conditions. The recorded subjects are seated in front of a conventional computer with a 15.6 inches display. The face distance to the computer display (and camera) ranges between 40–60 cm. The head movements are restricted to small angles ($\leq 10°$) during the recording sessions. The camera height is determined for each subject in order to optimize the EAR and position computation accuracy, in order to improve the sensibility of the whole system.

The gaze tracking system calibration data recording procedure is performed as follows: Target area numbers are displayed following its numeric order (see Fig. 1). The user fixes his gaze on the number in the screen. Timing of fixations is determined by the display program which is synchronized with the laptop camera. The order of presentations has been designed to minimize the fatigue and to reduce the eye movements between target areas. The system takes 60 face/eye image samples for each target area. As we set the video frame rate to 20 frames per second, the duration of each fixation is about 3 s.

For each face image sample, the center of the pupil coordinates, the texture-based gaze estimation and the EAR are computed. Note that blinking is also computed. These values are the input features of the gaze target area predictor to be tuned in order to be used for gaze ethogram recognition.

6.2. *Activity classification data*

We define three kinds of activities to be performed in front of a laptop computer display: reading a text, watching a movie, and typing a text. All of them are very common activities in an office-like environment.

For the experiments on behavioral activity recognition based on gaze ethograms, we have captured two datasets. The first was devoted to the selection of the most appropriate classification model by cross-validation experiments. The subjects for the capture were the two researchers that have contributed to the gaze tracking calibration data. The second was devoted to computational experiments on the generalization of the results, examining intra- and inter-subject performance of the selected classifier. The subjects in this case were volunteered undergraduate students ($n = 12$, average age 22 years, no. of female $= 4$) that are fluent in the use of the laptop computer (OS Linux). Students were compensated with credits for their collaboration. In both dataset recordings, the protocol was the same. Each subject recorded three sessions in three different days. Each session was composed of 20 activity blocks of 200 s. duration. We allowed a time to shift between activities of 5 s. A controller was indicating the next activity and keeping control of recording the actual start and end of the recorded activity. The schedule of the activities in each session was randomized, but repeated between subjects. As a result, we have recorded 60 activity blocks *per* subject. The final datasets are composed of the gaze ethograms extracted from each activity block by the

tuned gaze tracking system. In summary, we have 120 gaze ethograms for classifier model selection, and 360 for gaze ethogram classifier generalization results.

7. Results

In this section, we report results of the two aspects of the system. Firstly, we report on the accuracy of the gaze target area prediction achieved by several state-of-the-art classifiers, comparing the pupil centre localization with the texture-based gaze localization approach. Secondly, we report on the recognition of the user activity based on the gaze ethogram information. The classifier model implementations used are the Python-based scikit-learn environment (https://scikit-learn.org/stable/). In all validation experiments reported below, we have repeated a 100 times a 75% hold out validation, where we by randomly select a 75% of the dataset for training the classifiers and use the other 25% for test. We report the average accuracy of the 100 test results.

7.1. *Gaze target area prediction*

We have experimented with several classifiers to predict the target display area associated with the gaze destination. Table 1 summarizes the accuracy results corresponding to the k-NN classifiers using either the center of the pupil or the texture-based gaze estimation features. The first feature vector (center of the pupil method) is composed of the horizontal and vertical coordinates and the EAR of each eye, a total of six features. These results have been partially reported elsewhere.[46] The second feature vector (texture-based gaze estimation) is composed by the averaged intensity of each subpolygon and the EAR of each eye, a total of 10 features.

Table 1. Accuracy results of k-NN using either the center of the pupil or texture-based features for gaze estimation to determine the target area that the user is looking at (averaged values of a 100 repetitions).

k-NN	Center of pupil	Texture-based estimation
$k=1$	84.41%	89.10%
$k=3$	85.26%	90.20%
$k=5$	85.25%	89.43%
$k=7$	83.96%	88.34%

We obtain better performance with the texture-based features for gaze estimation. In particular, we get the best performance for $k=3$, which is about 90.20%. After reviewing the misclassified images, we conclude that the quality and resolution of the images that we are using do not allow accurate computation of the center of the pupil. It is hard to detect the facial landmark points in such kind of low resolution and poorly illuminated images. The texture-based gaze estimation seems to be more robust against these conditions.

Table 2 shows the results corresponding to texture-based gaze estimation using a set of classifiers: support vector machines (SVM) with several kernel functions, random forest with several numbers of decision tree estimators (RF), and multi-layer perceptrons with several different architectures. For the sake of clarity, we replicate the k-NN classifier results shown in Table 1.

The best accuracy performance is achieved by SVM with polynomial function: 98.63%. The misclassified images correspond to cases in which the user's eyes are closed. When treating a video sequence, these errors are usually recovered in the next image of the sequence. In this computational experiments, we did not apply the thresholding on the EAR value to decide if the user has the eyes closed. Multi-layer perceptrons with 2 hidden layers (80 and 40 neurons, respectively) and hyperbolic tangent activation function gets a top accuracy performance of 93.83%. The best values for the other types of classifier (the random forest computation time is prohibitive for our application) are: k-NN with $k=3$ achieves 90.20%, and random forest with

Table 2. Accuracy results of a set of classifiers using texture-based features for gaze estimation to determine the target area that the user is looking at (averaged values of a 100 repetitions).

k-NN	$k=1$	$k=3$	$k=5$	$k=7$
Perf.	89.10%	90.20%	89.43%	88.34%
SVM	RBF	Linear	Poly	Sigmoid
Perf.	94.33%	96.78%	98.63%	13.48%
RF	est=10	est=100	est=500	est=1000
Perf.	82.14%	83.51%	83.92%	84.43%
MLP	100/sig	100/tanh	60+40/tanh	80+40/tanh
Perf.	92.07%	93.91%	93.62%	93.83%

1000 estimators achieves 84.43%. Not included in Table 2, the Naïve Bayes classifier gets a performance of 77.29%. Finally, note that SVM with a sigmoid kernel function achieves a very poor performance, about 13.48%, perhaps due to the instability of the nonlinear kernel transformation. The SVM family of classifiers therefore provides both the best and the worst results.

7.2. *Behavioral activity classification*

Figure 7 shows sample gaze ethograms of an activity recording session during which the subject is reading

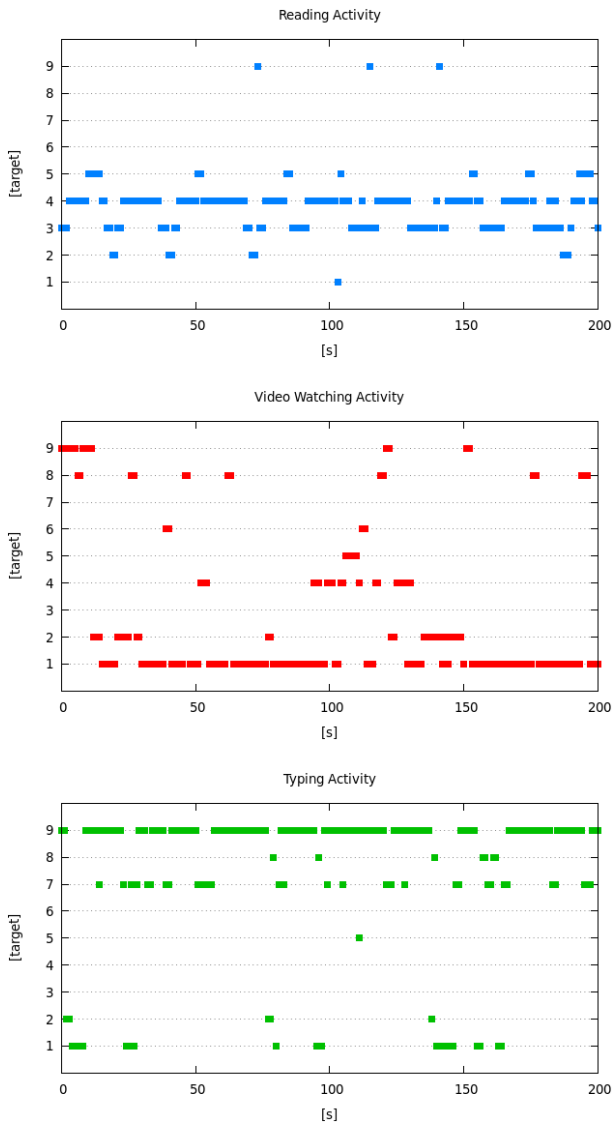

Fig. 7. Gaze ethograms associated with reading activity (top), watching a movie (center), and typing a text with the keyboard (bottom).

Table 3. Accuracy results of using gaze ethogram-based activity prediction (averaged values of 100 repetitions of 75% hold out validation).

k-NN	$k=1$	$k=3$	$k=5$	$k=7$
Acc.	94.52%	93.58%	93.43%	91.87%
SVM	RBF	Linear	Poly	Sigmoid
Acc.	97.12%	97.82%	98.32%	34.71%
RF	est=10	est=100	est=500	est=1000
Acc.	86.36%	87.71%	87.92%	88.14%
MLP	80/sig	80/tanh	60+20/tanh	40+40/tanh
Acc.	93.21%	95.17%	94.62%	93.75%

a text (top), watching a movie (center), and typing a text (bottom). The user's gaze fixates on a particular set of target areas depending on the task that is been performed. While reading a text (Fig. 7 (top)) the experimental subjects visits the upper display target areas 4 and 3 more frequently, due to the fact that the text starts in the top of the display and the subject scrolls the text. This behavior is obviously conditioned to the western reading convention. During the video watching activity (Fig. 7 (center)) the experimental subjects visit the display center target areas 1 and 8 more frequently, with some excursions to target area 4. Target 1 corresponds to the center of the display, so it is a highly expected gaze detection output. It is where the visual field is wider and higher in comparison to the other targets and where the subject can gather much more visual information from the screen. The gaze ethogram of the typing activity (Fig. 7 (bottom)) shows that the user visits more often the targets 7–9 at the bottom of the screen. This activity has a frequency of blinking events that is much lower than the reading and movie watching activities, probably because the gaze destination is mostly at the bottom of the display.

Table 4. Average confusion matrix of a SVM classifiers with a polynomial function over the model selection dataset.

	Reading	Video watching	Typing
Reading	32.7%	0.8%	0.0%
Video watching	0.5%	31.9%	0.8%
Typing	0.0%	1.1%	32.5%

Table 5. Intra and inter subject accuracy of behavioral activity SVM classifiers with a polynomial function over the generalization dataset.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S1	92.7%	83.4%	78.5%	84.7%	83.3%	83.9%	86.1%	84.6%	85.3%	81.8%	86.3%	83.4%
S2	79.7%	95.4%	77.5%	84.6%	81.7%	86.3%	83.7%	82.4%	81.6%	87.2%	88.1%	82.5%
S3	79.4%	78.1%	98.6%	81.8%	82.6%	87.2%	81.8%	83.1%	83.4%	85.0%	85.9%	79.4%
S4	82.7%	82.7%	82.3%	97.3%	84.6%	85.8%	79.2%	81.7%	82.8%	83.0%	86.4%	84.6%
S5	81.5%	80.3%	83.7%	82.9%	94.6%	83.8%	83.8%	85.6%	84.9%	80.7%	84.9%	83.3%
S6	83.1%	84.8%	85.7%	82.7%	84.0%	97.5%	82.6%	83.7%	84.5%	86.2%	87.8%	82.7%
S7	84.8%	83.1%	82.8%	80.3%	81.7%	81.8%	93.2%	81.6%	83.5%	85.8%	88.3%	84.8%
S8	83.7%	83.3%	83.4%	80.9%	86.4%	83.3%	83.0%	92.4%	82.5%	84.9%	87.7%	81.5%
S9	82.5%	82.4%	83.3%	83.9%	85.5%	83.9%	86.7%	83.7%	93.6%	81.6%	83.8%	82.4%
S10	79.6%	86.2%	82.8%	84.2%	82.6%	87.0%	87.3%	86.2%	83.0%	98.5%	86.5%	84.1%
S11	84.3%	89.0%	86.2%	87.1%	85.9%	84.9%	87.0%	87.2%	82.8%	87.2%	93.6%	84.8%
S12	82.7%	83.5%	81.5%	80.5%	82.9%	84.6%	83.5%	84.0%	82.3%	84.1%	84.4%	92.7%

The first computational experiment is devoted to the selection of the most appropriate classifier model using the first dataset of 120 gaze ethograms described above. We have explored the performance of k-NN classifier for several values of the parameter k, SVM with linear, radial basis function, polynomial, and sigmoid kernel functions, RF with several numbers of decision trees, and multi-layer perceptrons (MLP) with several different architectures. The averaged accuracy values of the results are presented in Table 3.

The best activity prediction results have been achieved by a SVM with a polynomial function, which is about 98.32%. We get a pretty good performance for most classifiers, though SVM with Sigmoid kernel is very bad. As a conclussion of this computational experiment, we select the SVM with polynomial kernel for ensuing experiments. Table 4 shows the average confusion matrix of the SVM with polynomial kernel, we can appreciate that the video activity can be confused with the other two. Reading and typing are not confused at all.

The second computational experiment tries to assess the generalization of the gaze ethogram classification approach to new subjects different from the ones that provided the data for the gaze tracking system tuning and the activity recognition model selection. Table 5 provides the average accuracy results of the hold out validation experiments carried out over the generalization dataset described above. The validation process was carried out as follows, for each subject, we repeated 100 times a hold out procedure, where 75% of its data was used for training an SVM with polynomial kernel. The trained model was applied to the 25% of the data hold out for test. The same model was applied to the data of the other subjects using their entire data as test, thus testing the ability to transfer the classifier trained with one subject to the remaining subjects. The average intra-subject classification accuracy achieved is 95.0%, corresponding to the average of the diagonal of Table 5. The average inter-subject accuracy is 83%, thus we appreciate a big decrease in performance when we try to do transfer learning.

8. Conclusions and Further Work

We propose a system for the recognition of behavioral activities carried out in front of a computer by means of gaze tracking information. The sequence of

gaze fixations is analyzed in order to create a gaze ethogram, which is the input for the behavioral activity classifier. In order to obtain noninvasive, thus highly ecologically valid, observations, we have developed our own open-source gaze tracking system that works on low resolution images captured by a conventional laptop computer web camera without any special equipment for illumination. The gaze tracking system innovations are (a) the definition of texture features for the eye direction characterization which are more robust than the center of the pupil, and (b) a machine learning approach to estimate the gaze fixation.

Based on the experimental results, we conclude that the system is able to classify activities usually performed in normal office-like conditions by several experimental subjects with different physical characteristics as color of the skin and eyes, shape of the face, and so on.

In the future, we have several lines of work to pursue. We want to explore other screen partition template designs with more targets and different target topology could help to improve the behavioral activity recognition results. We have to trade-off the low resolution of the employed equipment and the detection of gaze destination. We are also considering using directly an estimation of the display area of the gaze destination and to overcome the use of landmarks. We want to increase the number and diversity of activities. Although the selected ones for this work are quite representative, they are just some activities that a user performs in front of a computer. We also will be working on the improvement of the transfer of trained classifiers between subjects in order to generalize the system. We will be considering the recognition of activities with variable duration, which requires dynamic programming approaches in order to cope with varying gaze ethogram sizes, and on the determination of the minimum length of video that is needed to identify a particular activity.

Regarding the improvement of the gaze tracking system, future work is planned towards improving the image processing techniques that are used in order to detect the face and facial landmarks in the image. Currently, the head movements are restricted in order to improve the results. If a user turns the head several degrees, the face is not detected and no data is gathered.

Acknowledgments

This work has been supported by FEDER funds through MINECO project TIN2017-85827-P. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 777720. Additional support comes from grant IT1284-19 of the Basque Country Government.

References

1. J. H. F. Abeelen, Mouse mutants studied by means of ethological methods, *Genetica* **34** (1964) 79–94.
2. A. Al-Rahayfeh and M. Faezipour, Eye tracking and head movement detection: A state-of-art survey, *IEEE J. Transl. Eng. Health Med.* **1** (2013) 1–12.
3. D. J. Anderson and P. Perona, Toward a science of Computational Ethology, *Neuron* **84** (2014) 18–31.
4. G. Andrienko, N. Andrienko, G. Budziak, J. Dykes, G. Fuchs, T. von Landesberger and H. Weber, Visual analysis of pressure in football, *Data Min. Knowl. Disc.* **31**(6) (2017) 1793–1839.
5. S. Baluja and D. Pomerleau, Non-intrusive gaze tracking using artificial neural networks, Technical Report CMU-CS-94-102, Carnegie Mellon University, Pittsburgh, Pennsylvania (1994).
6. N. Barbara, T. A. Camilleri and K. P. Camilleri, EOG-based eye movement detection and gaze estimation for an asynchronous virtual keyboard, *Biomed. Signal Process. Control* **47** (2019) 159–167.
7. P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman and N. Kumar, Localizing parts of faces using a consensus of exemplars, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (Colorado Springs, USA, 2011), pp. 545–552.
8. B. W. Blakley and L. Chan, Methods considerations for nystagmography, *J. Otolaryngol. Head Neck Surg.* **44** (2015) 25.
9. M. Boukhechba, L. Cai, C. Wu and L. E. Barnes, Actippg: Using deep neural networks for activity recognition from wrist-worn photoplethysmography (PPG) sensors, *Smart Health* **14** (2019) 100082.
10. X. Cao, Y. Wei, F. Wen and J. Sun, Face alignment by explicit shape regression, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (Providence, USA, 2012), pp. 2887–2894.
11. B. Cassin and S. Solomon, *Dictionary of Eye Terminology* (Triad Publishing Company, Gainesville, Florida, 1990).
12. C. Cortes and V. N. Vapnik, Support-vector networks, *Mach. Learn.* **20**(3) (1995) 273–297.
13. N. Dalal and B. Triggs, Histogram of Oriented Gradients for Human Detection, *IEEE Conf. Comp.*

Vision and Pattern Recognition (CVPR) (San Diego, California, USA, 2005), pp. 886–893.
14. M. Danotne, J. Gall, G. Fanelli and L. V. Gool, Real-time facial feature detection using conditional regression forest, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (Providence, USA, 2012), pp. 2887–2894.
15. J. de Lope and M. Graña, Face feature detection for gaze estimation, *Zenodo*, http://doi.org/10.5281/zenodo.2565713.
16. S. R. Datta, D. J. Anderson, K. Branson, P. Perona and A. Leifer, Computational Neuroethology: A call to action, *Neuron* **104**(1) (2019) 11–24.
17. L. Ding and A. M. Martínez, Precise detailed detection of faces and facial features, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (Anchorage, Alaska, USA, 2008).
18. P. Dóllar, P. Welinder and P. Perona, Cascade pose regression, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (San Francisco, California, USA, 2010), pp. 1078–1085.
19. A. T. Duchowski, *Eye Tracking Methodology — Theory and Practice* (Springer, Cham, 2017).
20. A. T. Duchowski, Gaze-based interaction: A 30 year retrospective, *Comput. Graph.* **73** (2018) 59–69.
21. J. D. Enderle and D. A. Sierra, A new linear muscle fiber model for neural control of saccades, *Int. J. Neural Syst.* **23**(2) (2013) 1350002, PMID: 23578053.
22. J. P. Ewert, *Neuroethology: An Introduction to the Neurophysiological Fundamentals of Behavior* (Springer International Publishing, 1980).
23. O. Ferhat and F. Vilariño, Low cost eye tracking: The current panorama, *Comput. Intell. Neurosci.* **2016** (2016) 1–14.
24. L. Florea, C. Florea and C. Vertan, Recognition of the gaze direction: Anchoring with the eyebrows, *J. Vis. Commun. Image Rep.* **35** (2016) 67–77.
25. D. M. Gavrila and V. Philomin, Real-time object detection for smart vehicles, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (Ft. Collins, Colorado, USA, 1999), pp. 87–93.
26. D. M. Gavrila, J. Giebel and S. Munder, Vision-based pedestrian detection: The protector + system, *Proc. IEEE Intelligent Vehicles Symp. 2004*, Parma, Italy, 2004, pp. 13–18.
27. A. George, Image based Eye Gaze Tracking and its Applications, arXiv:1907.04325.
28. K. Gidlöf, A. Wallin, R. Dewhurst and K. Holmqvist, Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment, *J. Eye Movement Res.* **6**(1) (2013) 1–14.
29. J. M. Gorriz, J. Ramirez, F. Segovia, F. J. Martinez, M. C. Lai, M. V. Lombardo, S. Baron-Cohen and J. Suckling, A Machine Learning Approach to Reveal the NeuroPhenotypes of Autisms, *Int. J. Neural Syst.* **29**(7) (2019) 1850058.
30. M. Graña and J. de Lope, A Short Review of Some Aspects of Computational Neuroethology, in *Understanding the Brain Function and Emotions*, LNCS Vol. 11486, eds. J. Ferrández Vicente, J. Álvarez-Sánchez, F. de la Paz López, J. Toledo Moreo, H. Adeli (Springer Nature, Cham, Switzerland, 2019), pp. 275–283.
31. A. Gutierrez-Garcia, A. Fernandez-Martin, M. Del Libano and M. G. Calvo, Selective gaze direction and interpretation of facial expressions in social anxiety, *Pers. Individ. Differ.* **147** (2019) 297–305.
32. O. Grynszpan, J. Bouteiller, S. Grynszpan, F. Le Barillier, J. C. Martin and J. Nadel, Altered sense of gaze leading in autism, *Res. Autism Spect. Disord.* **67** (2019) 101441.
33. R. Hof, How do you Google? New eye tracking study reveals huge changes, *Forbes Online* (2015), https://www.forbes.com/sites/roberthof/2015/03/03/how-do-you-google-new-eye-tracking-study-reveals-huge-changes/.
34. R. G. Hussain, M. A. Ghazanfar, M. A. Azam, U. Naeem and S. U. Rehman, A performance comparison of machine learning classification approaches for robust activity of daily living recognition, *Artif. Intell. Rev.* **52**(1) (2019) 357–379.
35. E. Itskovits, A. Levine, E. Cohen and A. Zaslaver, A multi- animal tracker for studying complex behaviors, *BMC Biol.* **15**(1) (2017) 29.
36. P. M. Insch, G, Slessor, J. Warrington and L. H. Phillips, Gaze detection and gaze cuing in Alzheimer's Disease, *Brain Cogn.* **116** (2017) 47–53.
37. V. Kazemi and J. Sullivan, One millisecond face alignment with an ensemble of regression trees, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (Columbus, Ohio, USA, 2014), pp. 1867–1874.
38. S.-R. Ke, H. Le Uyen Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo and K.-H. Choi, A review on video-based human activity recognition, *Computers* **2**(2) (2013) 88–131.
39. J. Kim, J. Seo and T. H. Laine, Detecting boredom from eye gaze and EEG, *Biomed. Signal Process. Control* **46** (2018) 302–313.
40. K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik and A. Torralba, Eye Tracking for Everyone, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, Nevada, USA, 2016), pp. 2176–2184.
41. A. Lentzas and D. Vrakas, Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review, *Artif. Intell. Rev.* **53** (2020) 1975–2021.
42. Q. Li, W. Lin and J. Li, Human activity recognition using dynamic representation and matching of skeleton feature sequences from RGB-D images, *Signal Process. Image Commun.* **68** (2018) 265–272.

43. D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Computer Vision* **60**(2) (2004) 91–110.
44. O. Martinez Manzanera, S. K. Meles, K. L. Leenders, R. J. Renken, M. Pagani, D. Arnaldi, F. Nobili, J. Obeso, M. Rodriguez Oroz, S. Morbelli and N. M. Maurits, Scaled subprofile modeling and convolutional neural networks for the identification of Parkinson's disease in 3d nuclear imaging data, *Int. J. Neural Syst.* **29**(9) (2019) 1950010.
45. A. Mohan, C. Papageorgiou and T. Poggio, Example-based object detection in images by components, *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(4) (2001) 349–361.
46. S. Moraleda, J. de Lope and M. Graña, Recognizing cognitive activities through eye traking, in *Understanding the Brain Function and Emotions*, LNCS, Vol. 11486, eds. J. Ferrández Vicente, J. Álvarez-Sánchez, F. de la Paz López, J. Toledo Moreo, H. Adeli (Springer Nature, Cham, Switzerland, 2019), pp. 291–300.
47. R. A. Naqvi, M. Arsalan, G. Batchuluun, H. S. Yoon and K. R. Park, Deep learning-based gaze detection systems for automobile drivers using a NIR camera sensor, *Sensors* **18** (2018) 1–34.
48. U. Obaidellah, M. Al Haek and P. C.-H. Cheng, A survey on the usage of eye-tracking in computer programming, *ACM Comput. Surv.* **51**(1) (2018) 1–58.
49. C. Papageorgiou and T. Poggio, A trainable system for object detection, *Int. J. Comput. Vis.* **38**(1) (2000) 15–33.
50. K. R. Park, J. J. Lee and J. Kim, Gaze position detection by computing the three dimensional facial positions and motions, *Pattern Recogn.* **35**(11) (2002) 2559–2569.
51. W. Sewell and O. Komogortsev, Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network, *ACM CHI Conf. Human Factors in Computing Systems* (Atlanta, Georgia, 2010), pp. 3739–3744.
52. A. Senju, Y. Kikuchi, T. Hasegawa, Y. Tojo and H. Osanai, Is anyone looking at me? Direct gaze detection in children with and without autism, *Brain Cogn.* **67**(2) (2008) 127–139.
53. B. A. Shiferaw, D. P. Crewther and L. A. Downey, Gaze entropy measures detect alcohol-induced driver impairment, *Drug Alcohol Depend.* **204** (2019) 107519.
54. B. M. Smith and L. Zhang, Joint face alignment with non-parametric shape models, *European Conf. Computer Vision (ECCV 2012)*, LNCS, Vol. 7574, eds. A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, Schmid (Springer, Berlin, Heidelberg, Florence, Italy, 2012), pp. 43–56.
55. T. Soukupová and J. Čech, Real-time eye blink detection using facial landmarks, in *21st Computer Vision Winter Workshop*, eds. L. Cehovin, R. Mandeljc and V. Struc, Rimske Toplice, Slovenia, 3–5 February 2016.
56. E. E. Stone and M. Skubic, Unobtrusive, continuous, in-home gait measurement using the Microsoft Kinect, *IEEE Trans. Biomed. Eng.* **60**(10) (2013) 2925–2932.
57. M.-J. Tsai, H.-T. Hou, M.-L. Lai, W.-Y. Liu and F.-Y. Yang, Visual attention for solving multiple-choice science problem: An eye-tracking analysis, *Comput. Educ.* **58** (2012) 375–385.
58. A. Vasilyev and M. Hansard, Spatial distribution of eye-movements after central vision loss is consistent with an optimal visual search strategy, *Int. J. Neural Syst.* **29**(10) (2019) 1950026.
59. P. Viola, M. J. Jones and D. Snow, Detecting pedestrians using patterns of motion and appearance, *IEEE Int. Conf. Computer Vision (ICCV 2003)*, Vol. 2 (Nice, France, 2003), pp. 734–741.
60. S. Vora, A. Rangesh and M. M. Trivedi, On generalizing driver gaze zone estimation using convolutional neural networks, *IEEE Intelligent Vehicles Symp.* (Redondo Beach, California, 2017), pp. 849–854.
61. M. Vrigkas, C. Nikou and I. Kakadiaris, A review of human activity recognition methods, *Front. Robot. Artif. Intell.* **2** (2015) 11.
62. J. Wang, Y. Chen, S. Hao, X. Peng and L. Hu, Deep learning for sensor-based activity recognition: A survey, *Pattern Recogn. Lett.* **119** (2019) 3–11.
63. Y. Wang, S. Cang and H. Yu, A survey on wearable sensor modality centred human activity recognition in health care, *Expert Syst. Appl.* **137** (2019) 167–190.
64. P. Weigel, M. Raab and R. Wollny, Tactical decision making in team sports: A model of cognitive processes, *Int. J. Sports Sci.* **5**(49) (2015) 128–138.
65. J. Y. Yang, J. S. Wang and Y. P. Chen, Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers, *Pattern Recogn. Lett.* **29**(16) (2008) 2213–2220.
66. Y.-H. Yiu, M. Aboulatta, T. Raiser, L. Ophey, V. L. Flanagin, P. Zu Eulenburg and S.-A. Ahmadi, Deepvog: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning, *J. Neurosci. Methods* **324** (2019) 108307.
67. L. R. Young and D. Sheena, Survey of eye movement recording methods, *Behav. Res. Methods Instrum.* **7**(5) (1975) 397–439.