Analysis of the Usefulness of Mobile Eyetracker for the Recognition of Physical Activities

Peter Hevesi*, Jamie A. Ward*†, Orkhan Amiraslanov*, Gerald Pirkl* and Paul Lukowicz*‡
* German Research Center for Artificial Intelligence, Kaiserslautern, Germany
†University College London
‡University of Kaiserslautern
* peter.hevesi@dfki.de, † jamie@jamieward.net,
‡ orkhan.amiraslanov@dfki.de, § gerald.pirkl@dfki.de,
¶ paul.lukowicz@dfki.de

***Abstract*—We investigate the usefulness of information from a wearable eyetracker to detect physical activities during assembly and construction tasks. Large physical activities, like carrying heavy items and walking, are analysed alongside more precise, hand-tool activities like using a screwdriver. Statistical analysis of eye based features like fixation length and frequency of fixations show significant correlations for precise activities. Using this finding, we selected 10, calibration-free eye features to train a classifier for recognising up to 6 different activities. Frame-by-frame and event based results are presented using data from an 8-person dataset containing over 600 activity events. We also evaluate the recognition performance when gaze features are combined with data from wearable accelerometers and microphones. Our initial results show a duration-weighted event precision and recall of up to 0.69 & 0.84 for independently trained recognition on precise activities using gaze. This indicates that gaze is suitable for spotting subtle precise activities and can be a useful source for more sophisticated classifier fusion.**

Keywords–eyetracker; activity recognition; sensor fusion

I. Introduction

Understanding complex, self-organising, physical labour intensive work processes involving multiple persons is a key competence in applications like production line optimisation or construction management. Our research goal is to lay the foundation of a system, that is capable to analyse and understand such processes and provide hints for possible improvements. The first step towards that goal is to evaluate different unobstrusive sensing modalities that can be used in real word scenarios to detect physical activities of single group members. As part of this research, we investigated the usefulness of mobile eyetrackers for detecting physical activities.

Eyetracking is known to provide important insights about one's attention. Attention in turn is known to provide important indications of one's activity. As a consequence, as unobtrusive, affordable mobile eye trackers have started to emerge, there has been an increasing interest in using them for activity recognition [1]. To date the vast majority of such research had concentrated on activities directly related to visual attention and cognition (reading, watching TV, etc.) [2][3]. In this paper, we investigate the use of gaze information for the recognition of physical activities, specifically activities related to an assembly/construction task. In doing so, we focus on the following questions:

1) Can gaze information help spot subtle precise activities such as for example screw driving in a stream of heterogeneous physical activity data? Such spotting is a well known, hard problem in wearable activity recognition.
   One of the reasons, it is so hard, is the variability associated with the motions involved in many such activities (e.g., tightening a screw can be done with one hand, with two hands, with a screwdriver, or with a drill – all using a variety of grips). Another is the fact that many of the involved motions occur spuriously, for example during walking or random gesticulation. We hypothesize that the need to fix the gaze in a certain pattern during many such precise activities can help overcome those problems.
2) How fine grained is the discriminative power of gaze information for physical activities? Can it be used to distinguish activities on a fine grade or it is restricted to broad categories characterized by the need for an increased attention or focus level.
3) How strongly user and setting dependent is the gaze information?
4) How does gaze information compare to standard wearable activity recognition sensors such as accelerometers and sound? Can it complement such information?

We investigate these questions in an experiment where four participants have to build up a large TV wall (described previously in [4]). Note that the purpose of this work is not to present a highly optimized ready-to-use solution with the best possible recognition rate in the above specific application. Instead, it is to provide an initial analysis of the suitability of gaze tracking for physical activity recognition with respect to the questions above.

In Section II, we provide a brief overview of state-of-the-art methods in our research field. In Section III, we describe our experimental setup, used sensors and generated datasets. Our evaluation methodology including the proposed feature sets is described in Section IV and finally our results are presented and discussed in Section V.

II. Related work

Research into activity recognition using wearable sensing has continued to grow in recent years. Many studies deploy distributed body-worn or mobile inertial sensors to recognise a wide-range of physical activities (see [5] for an overview).

A common sensing modality is sound. In [6], Lu et al. introduce a mobile-phone based system for classifying ambient sound, voices and music. Previous works use multiple streams of audio to recognise social situations [7][8], or to infer collocation and social network information [9].

Combined sound and movement data obtained from the mobiles of groups was recently used to analyse pedestrian congestion at busy thoroughfares, making use of changes in people's step-intervals and ambient audio [10]. Wrist-worn microphones and accelerometers were first used together to detect hand-tool activities in a wood workshop scenario [11]. More recently, these sensors were used to recognise physical collocation and collaboration of co-workers performing a group task [4].

A. *Eye-based activity recognition*

Eye tracking is a widely used technique in human computer interaction (HCI), for example in assistive technologies for people with limited motor skills [12], and is used in a growing body of research in Ubicomp (e.g., on attention [2]). Typically, researchers are interested in the object of a user's gaze – what it is that the user is looking at – however, another approach is to analyse the patterns created by eye movement in different situations. Patterns of eye fixation and saccadic movement recorded from changes in the eye's electrical activity (electrooculography, or EOG), were first used in a wearable setting to detect reading activities while walking [1]. This work was then developed to detect activities such as writing, reading, watching a video, etc. [13]. An advantage of a pattern-based approach is that no calibration is needed with a worldview video. Platforms like Google Glass include the ability to record blink rate, which when combined with head movement can be an effective method for recognising activities [3].

Vidal et al. introduced a calibration-free, gaze interaction method based on tracking the smooth pursuit movements that occur when the eye follows a moving target [14]. And in [15] a commercial, wearable EOG system, the Jiins Meme, was used as a novel gestural input device based on a similar approach.

Closest to our research is the work in [16]. The authors proposed a system based on eyetracker and first person videos to recognise daily activities. However, this work focuses on activities directly related to gaze (e.g., reading, video watching) compared to our approach, where we want to detect rather physical activities (e.g., screw driving), where a direct gaze contact is not an essential part of the activity.

III. Experiment

We designed an experiment as a benchmark to evaluate different sensors and algorithms for group activity recognition.

A. *Scenario*

In the experiment, four participants collaborate to build a 2.5 meter high TV wall consisting of 8 large LCD screens, 3 base panels, 18 screen spacers, and more than 50 screws. The parts are stored in containers at a storage area which is separated by a ca. 25 meter long hallway from the assembly area.

The building phase included the following main steps: 1.) Unload screens (each screen weights 8 kg) and other TV parts from the containers, 2.) Carry items to the assembly area, 3.) Assemble and place base items, 4.) Lift screens onto the wall, 5.) Fix screens on the wall by tightening the screws. After the build phase and a short break the participants perform the reversed process: 6.) removing the screws, 7.) taking down the screens and other parts carefully, 8.) carrying back to the storage area, 9.) put them back into the containers.

Generally, the participants had the freedom to organise and execute the tasks as they thought it's best. The overall task takes usually from 40 minutes up to 1 hour.

B. *Wearable sensors*

As shown on Figure 1, the participants were equipped with a mobile eyetracker, a sound recording device with two separate microphones and three inertial measurement units.

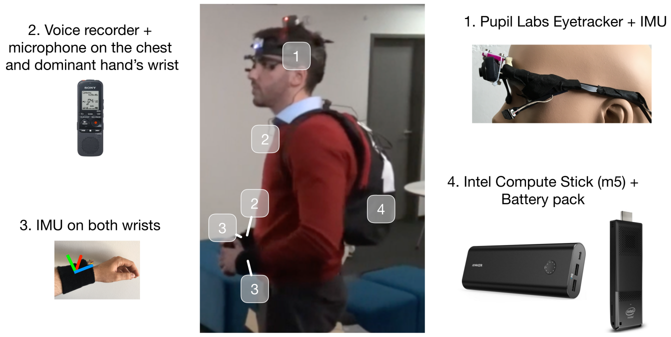
2. Voice recorder + microphone on the chest and dominant hand's wrist
1. Pupil Labs Eyetracker + IMU
1
2
2
3
4
3
4. Intel Compute Stick (m5) + Battery pack
3. IMU on both wrists

Figure 1. Recording setup for each participant includes an eyetracker connected to a small recording computer. Additional sensors: IMU on both arms and head, microphone on the wrist and at the chest.

a) Mobile eyetracker: The eyetracker setup consists of a head worn device from Pupil Labs [17] connected to an Intel Compute Stick with an m5 1.6 GHz processor as recording device (running Ubuntu 16.10). Both devices were powered by a portable 20100 mAh battery. The recording itself was done using Pupil Capture (v0.82) software. We implemented scripts to remotely control and monitor the recordings. The overall cost of this eyetracker setup is around 1600 Euros, which is significantly lower then many other commercially available mobile eyetracker solutions. This makes the setup better scalable for real world applications.

b) Inertial measurement unit (IMU): For tracking movements of the participants, they wore IMUs on both wrists and one on the head. The IMU devices record 3-axis acceleration, gyro, and magnetic field as well as 3D orientation with approximately 40 Hz.

c) Sound recorder: Each participant wore two microphones: one on the dominant hand's wrist and a second one attached on the chest. The microphones were connected to a voice recorder capable of recording stereo sound and were saved as the two channels of the sound file.

C. *Datasets and labels*

We created two datasets (referred to as dataset A and dataset B) by recording the above described experiment performed by two different group of participants. Each dataset includes IMU data, sound recording and eyetracker recordings (world camera video, eye camera video, eye and eye movement data, fixation events). To help the annotation process, four additional stationary cameras recorded the scene. Two cameras were recording the assembly area, one the storage area and one the hallway.

We created two label sets to analyse the discriminative power of the features, whether the system can distinguish single activities or rather just specific type of an activity.

a) Six class problem: The detailed label set includes six classes as follows:

1) Adjust: during these activities the subject is interacting (placing, taking or adjusting) with screws without any tool.
2) Screwdriver: subject tightens or loosens screws using screwdriver.
3) Drill: events when a participant tightens or loosens screws using a powerdrill with screwdriver attachment.
4) Carry: the times when one or two participants carry the heavy TV screens to or from the assembly area.
5) Screen placement: segments where screens are taken out of or placed back into the container or put on or taken off the TV wall.
6) Walk: person moves between assembly area and storage area (without carrying heavy objects).

b) One class problem: The second set looks only at the single class of Precise activities:

1) Precise Activity: mostly consists of small and precise movements. Typically it requires increased attention of the subjects. This includes the above labeled instances of screwdriver, adjust and drill events.

The ground truth labels for both sets were annotated using mainly the first person view (world camera) of the eyetracker recordings for each participant. The degree of freedom to organise and perform the experiment resulted often in unexpected event flows with lot of short interruptions and activity changes. This proved to be a real challenge for the labeling making low level event annotation nearly impossible, on the other hand this makes the data realistic. By keeping this in mind, we consider each ground truth label as a rough description what a participant is mainly doing in a given time interval of a few seconds up to a minute. Short interruptions (e.g., person taking additional screw from the desk or interacting with other participants) are not represented in this ground truth.

In total, we labeled 606 events with an overall length of ca. 260 minutes.

IV. Activity Recognition System

A. Features

For further analysis, we extracted features on the time series data of each participant using a centered sliding window of 30 seconds. The label for each sample is defined as the current ground truth event at the center of the window, or null if there are no active events for the current person.

a) Eyetracker features: One important eye movement feature is fixation (looking at something for a time period), because it could be an indicator of increased attention. The recording software already provides extracted fixation events described by start time and duration. A feature vector is easily calculated by sliding a window across this output and taking the sum of the fixation durations inside each window.

Figure 2 shows the correlation between the fixation length feature and a subject's activities. The temporal changes in the fixation length values are synchronous to the currently performed task (color on the top represent different activities). Statistical analysis confirms the relationship between an activity and fixation length values during it. The average fixation length for "drill" and "screwdriver" events is significantly higher than for any other activity which indicates that it might be a strong feature (see box-plot on the right side of Figure 2).

A similarly interesting feature can be extracted using the gap or duration between two fixations. A lower gap duration means that there are frequent fixations over a certain time, meanwhile a higher duration represent times when the participant is not looking at anything for a long time (scanning the environment).

Information about the pupil size, could help to distinguish dark and bright environments. Accommodation (change of viewing distance) can also cause changes of pupil size.

For this study, we calculated 10 eye-derived features for each sliding window: 1,2) average and median of the durations of fixation events starting or ending in the window, 3,4) average and median of the fixation gap values, 5,6) average of the pupil position in spheric coordinates (ϕ and θ) 7,8) standard deviation of the pupil position in spheric coordinates, 9) average pupil size, 10) standard deviation of the pupil size.

The above features are calibration-free meaning that the device displacement (or wrong calibration) does not influence the results.

b) Acceleration features: The 3-axis accelerometer signals (x,y,z) are combined to give a single orientation-invariant reading, $a = \sqrt{x^2 + y^2 + z^2}$, for each of the head, left, and right-wrist IMUs (a_h,a_l, and a_r). For each of these readings four standard features are calculated across a 1 second rolling window, these are: mean (μ), standard deviation (σ), short-term energy (E), and zero-crossing rate (ZC). ZC, a simple measure of dominant signal frequency, is calculated by counting the zero-crossings on each window after subtracting μ.

c) Sound features: Sound signals from each participant's dominant wrist, s_r (all were right-handed), and head, s_h, are downsampled from the recording rate of $44.1kHz$ to $8kHz$ (16 bit). Two features are extracted for each of these across a rolling window of 1 second: short-term energy, E, and zero-crossing rate, ZC. These features were chosen because of their widespread use in low-cost speech and audio analysis [18]. We also included an intensity analysis feature, calculated as $ia = \frac{E_r}{E_h} - \frac{E_h}{E_r}$, where E_r and E_h are the short term sound energies of the right-hand and head-recordings respectively. This measure can be used to distinguish sounds made closer to one microphone or another from sounds made further away from both [19]. For a sound made close to the hand, $ia > 0$, and for further away, $ia \approx 0$.

d) Feature sets: For each participant in each dataset, we created three feature matrices where each row represents the following features:

1) As baseline, we use the feature vector including the 12 acceleration and 5 sound features, since this is

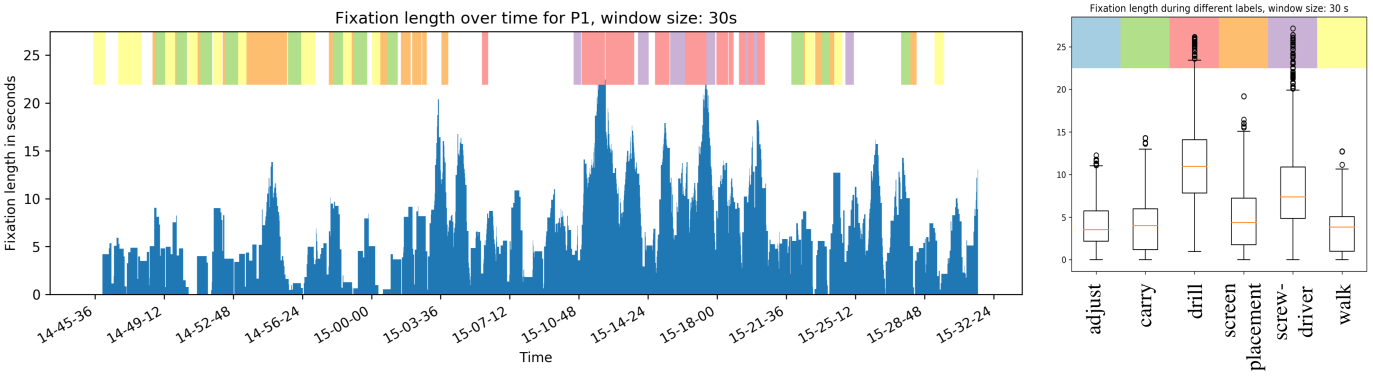

Figure 2. Left: sum of the lengths of each fixation event inside a 30 s window over time, the color on the top indicates the current activity of the subject. Right: statistical box plot about fixation length feature values during specific activities (see x axis) - same color scheme on top.

the most widespread approach. (Later referred as acc+snd)

2) The feature vector of the 10 eyetracker based features as described above is the topic of the main investigation in this paper (Later referred as eye)
3) For a preview, how well a combination of sensors performs, we combine simply the 12 acceleration and 5 sound features with the 10 eye-derived information to create a single feature vector for each time frame. (Later referred as all)

B. Evaluation methods

For evaluation, the features matrices are split up to training and test sets depending on the evaluation method. The training set is used then to train a Naive Bayes classifier. We applied several different standard classifiers, but found Naive Bayes to be sufficient for the purposes of the current work. Training and testing was implemented on Python using the scikit-learn toolkit [20].

In the testing phase, for each new samples (one row of the feature matrix) the classifier predicts an activity label. This is referred to as a frame based result. If sequential rows receive the same activity class predictions, these are merged together into an event. These predicted events are then used for event based evaluation.

a) Experiment dependent evaluation: The experiment dependent evaluations were performed on the primary dataset (dataset A). Ideally, the leave-one-person out would be the preferred method for training and evaluation. This approach however doesn't work well in this case because there are some activities performed almost exclusively by one participant. This leads to insufficient data for training.

Instead features are divided into six smaller sets, while trying to keep an equal distribution of samples for each label. A purely random selection of features for each split can result in a misleadingly high accuracy when samples from the same event are used for training and test. To avoid this the training and test samples are always strictly separated by the events they belong to. On the splits a six-fold cross validation was performed and the average scores were calculated over the iterations.

b) Experiment independent evaluation: In this evaluation, the performance of the recognition is tested on completely unseen data (not used for training in any way). The system is trained on all samples of a dataset B (including every person). The test is performed then on the extracted features of dataset A, which were not used for training at all in this case.

This indicates how well the system can generalize the results and handle later datasets without any additional training effort.

c) Frame based evaluation: Standard precision and recall values are calculated for the frame-based evaluation. Each predicted label is considered as true positive if it's equal to the sample's ground truth label or as false positive otherwise. A ground truth label is a false negative if the predicted label for the same sample is different.

d) Event based evaluation: In many cases it's more important to detect activity events rather than detecting each frame on an activity. For example, the information that a subject performed an activity is sufficient and the exact timings are less relevant. To get comparable results to the frame-based analysis, event based precision and recall values are calculated.

In the event based evaluation, we compare detection events with the ground truth. A detected event is considered as a true positive (TP_{det}) if it has an overlap with a ground truth event of the same activity (for the same participant) or as a false positive (FP_{det}) otherwise. Similarly ground truth events are labeled as true positives (TP_{gt}) if they are detected at least once otherwise as false negatives (FN_{gt}).

Analogous to the standard frame definitions the event-based metrics are calculated as:

$$precision = \frac{TP_{det}}{TP_{det} + FP_{det}} \tag{1}$$

and,

$$recall = \frac{TP_{gt}}{TP_{gt} + FN_{gt}} \tag{2}$$

V. Results and Discussion

The precision/recall results for different sensor combinations are given for the six class problem in Figure 3 and for the one class problem in Figure 4.

Before we go into discussion of individual issues, it is important to point out a general observation related to all

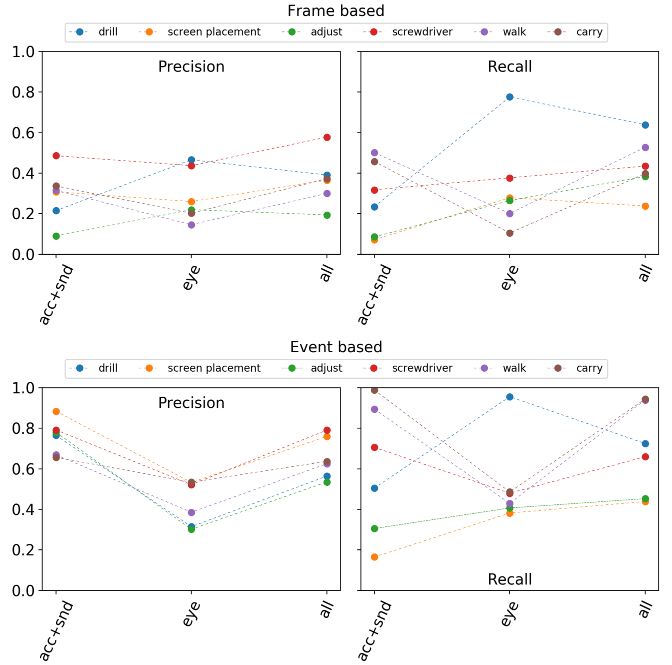

Figure 3. Six class, dataset dependent evaluation using different feature sets (acc+snd: acceleration and sound features, eye: eye features only, all: sound, acceleration, and eye). Top plots show frame, bottom plots show event-based precision and recall results.

results that involve gaze features: whereas acceleration and sound only results always show a big improvement when going from frame by frame to event based results (which is to be expected), this improvement is much smaller when gaze features are involved. Where in many cases eye features are better on frame level, acceleration/sound win on event level. The explanation of this fact is related to the way people's visual attention works. In very few cases, we keep our attention 100% on a single task. Instead, while focusing mainly on the main task, we tend to glance at other things (e.g., someone we speak to while tightening a screw). Since such distractions tend to be short, on frame level, they do not have much impact. However on event level, they fragment the result. Thus where there is in reality a single event, the system detects several separated by short breaks.

In our event evaluation, this means that what is a single insertion in the accelerometer data becomes several insertions in the eye related data. This is nicely illustrated by the duration-weighted normalized event results shown in Figure 4. The values are calculated as follows:

$$precision_{weighted} = \frac{\sum len(TP^i_{det})}{\sum len(TP^i_{det}) + \sum len(FP^i_{det})} \quad (3)$$

and,

$$recall_{weighted} = \frac{\sum len(TP^i_{gt})}{\sum len(TP^i_{gt}) + \sum len(FN^i_{gt})} \quad (4)$$

where $len(TP^i)$ means the length of the i-th true positive event (for false positive and false negative events analog). With this measure, small errors (short false positive or if a short event isn't recognized) are less significant. It can be seen that for the duration-weighted case, eye based features (1) significantly improve when moving to event based recognition and (2) are better than acceleration+sound case. On the other hand the non-weighted results get worse for event based evaluation and are lower than acceleration+sound. In future work, we will investigate more sophisticated temporal smoothing for eye based features to address this problem.

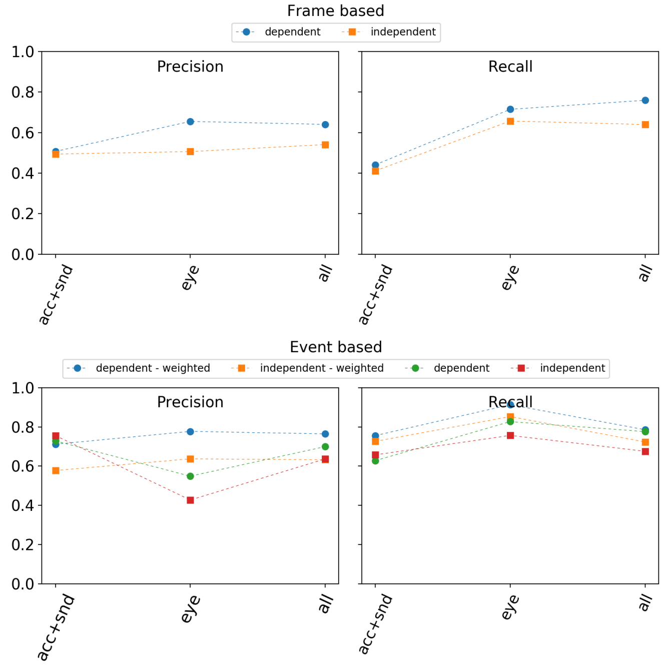

Figure 4. One class results for both dataset-dependent and independent evaluation. Frame (top) and event-based (bottom) precision and recall as well as event length weighted values are also represented. (acc+snd: acceleration and sound features, eye: eye features only, all: sound, acceleration, and eye)

With respect to the four questions raised in the introduction, the following can be said:

1) Gaze features seem clearly suitable for spotting subtle precises activities. Thus, for example, the weighted precision and recall for the one class problem in Figure 4 are 0.78 and 0.91 using the eye feature set (dataset dependent training).
2) As shown by the 6 class case (Figure 3), gaze based features allow a finer distinction then just a broad "Precise" activities class. Results for the screen placement, adjusting, and the individual types of screw driving are not perfect but well above random.
3) As a comparison of the experiment dependent and experiment independent results in all the graphs shows gaze based features are fairly robust with respect to different users (we have different subjects in the two experiments so that experiment (in)dependent means subject (in)dependent. Indeed the best performing combination (eye) in Figure 4 achieves a weighted event precision and recall of 0.69 & 0.84.
4) As can be seen in Figure 3, screen placement, adjusting and the individual types of screwdriving are resolved much better by gaze features then by acceleration+sound. Whenever acceleration and sound are not too bad, combining them provides further

improvement, which means that they do contain complementary information.

From the above, the focus of our future work will be on individualized recognition chains for each type of events (with subsequent plausibility like fusion), temporal smoothing for the gaze features on event level and combining eye gaze with image recognition methods for detecting visual context.

Acknowledgment

The work is funded by the German Federal Ministry of Education and Research (BMBF).

References

[1] A. Bulling, J. A. Ward, H.-W. Gellersen, and G. Tröster, "Robust recognition of reading activity in transit using wearable electrooculography," in PERVASIVE, May 2008, pp. 19–37.

[2] M. Vidal, D. H. Nguyen, and K. Lyons, "Looking at or through?: Using eye tracking to infer attention location for wearable transparent displays," in Proceedings of the 2014 ACM International Symposium on Wearable Computers, ser. ISWC '14. New York, NY, USA: ACM, 2014, pp. 87–90. [Online]. Available: http://doi.acm.org/10.1145/2634317.2634344

[3] S. Ishimaru, K. Kunze, K. Kise, J. Weppner, A. Dengel, P. Lukowicz, and A. Bulling, "In the blink of an eye: Combining head motion and eye blink frequency for activity recognition with google glass," in Proceedings of the 5th Augmented Human International Conference, ser. AH '14. New York, NY, USA: ACM, 2014, pp. 15:1–15:4. [Online]. Available: http://doi.acm.org/10.1145/2582051.2582066

[4] J. A. Ward, P. Lukowicz, G. Pirkl, and P. Hevesi, "Detecting physical collaborations in a group task using Body-Worn microphones and accelerometers," in 13th Workshop on Context and Activity Modeling and Recognition (CoMoRea'17), Big Island, USA, Mar. 2017, pp. 268–273.

[5] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," ACM Computing Surveys (CSUR), vol. 46, no. 3, 2014, p. 33.

[6] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "Soundsense: Scalable sound sensing for people-centric applications on mobile phones," in Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services, ser. MobiSys '09. New York, NY, USA: ACM, 2009, pp. 165–178. [Online]. Available: http://doi.acm.org/10.1145/1555816.1555834

[7] Y. Yang, B. Guo, Z. Yu, and H. He, "Social activity recognition and recommendation based on mobile sound sensing," in 2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing, Dec 2013, pp. 103–110.

[8] N. Eagle and A. S. Pentland, "Social network computing," in International Conference on Ubiquitous Computing. Springer, 2003, pp. 289–296.

[9] D. Wyatt, T. Choudhury, J. Bilmes, and J. A. Kitts, "Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science," ACM Trans. Intell. Syst. Technol., vol. 2, no. 1, Jan. 2011, pp. 7:1–7:41.

[10] T. Nishimura, T. Higuchi, H. Yamaguchi, and T. Higashino, "Detecting smoothness of pedestrian flows by participatory sensing with mobile phones," in Proceedings of the 2014 ACM International Symposium on Wearable Computers, ser. ISWC '14. New York, NY, USA: ACM, 2014, pp. 15–18. [Online]. Available: http://doi.acm.org/10.1145/2634317.2642869

[11] J. A. Ward, P. Lukowicz, G. Tröster, and T. E. Starner, "Activity recognition of assembly tasks using body-worn microphones and accelerometers," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 10, October 2006b, pp. 1553–1567.

[12] R. Barea, L. Boquete, M. Mazo, and E. Lopez, "System for assisted mobility using eye movements based on electrooculography," Trans. on Rehabilitation Engineering, vol. 10, no. 4, December 2002, pp. 209–218.

[13] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, "Eye movement analysis for activity recognition," in (Ubicomp) Proceedings of the 11th international conference on Ubiquitous computing. New York, NY, USA: ACM, 2009, pp. 41–50.

[14] M. Vidal, A. Bulling, and H. Gellersen, "Pursuits: spontaneous interaction with displays based on smooth pursuit eye movement and moving targets," in Proc. UbiComp. ACM, 2013, pp. 439–448.

[15] M. Dhuliawala, J. Lee, J. Shimizu, A. Bulling, K. Kunze, T. Starner, and W. Woo, "Smooth eye movement interaction using eog glasses," in Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016, pp. 307–311.

[16] Y. Shiga, T. Toyama, Y. Utsumi, K. Kise, and A. Dengel, "Daily activity recognition combining gaze motion and visual features," in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. ACM, 2014, pp. 1103–1111.

[17] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ser. UbiComp '14 Adjunct. New York, NY, USA: ACM, 2014, pp. 1151–1160. [Online]. Available: http://doi.acm.org/10.1145/2638728.2641695

[18] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in American Society for Engineering Education (ASEE) Zone Conference Proceedings, 2008, pp. 1–7.

[19] J. A. Ward, P. Lukowicz, and G. Tröster, "Roc analysis of partitioning method for activity recognition using two microphones," in Adjunct Proc. of the 3rd Int. Conf. on Pervasive Comp., vol. 191, May. 8-13 2005a.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.