Vision and Acceleration Modalities: Partners for Recognizing Complex Activities

Alexander Diete
Data and Web Science Group
University of Mannheim
Mannheim, Germany
alex@informatik.uni-mannheim.de

Timo Sztyler
Data and Web Science Group
University of Mannheim
Mannheim, Germany
timo@informatik.uni-mannheim.de

Heiner Stuckenschmidt
Data and Web Science Group
University of Mannheim
Mannheim, Germany
heiner@informatik.uni-mannheim.de

***Abstract*—Wearable devices have been used widely for human activity recognition in the field of pervasive computing. One big area of in this research is the recognition of activities of daily living where especially inertial and interaction sensors like RFID tags and scanners have been used. An issue that may arise when using interaction sensors is a lack of certainty. A positive signal from an interaction sensor is not necessarily caused by a performed activity e.g. when an object is only touched but no interaction occurred afterwards. In our work, we aim to overcome this limitation and present a multi-modal egocentric-based activity recognition approach which is able to recognize the critical activities by looking at movement and object information at the same time. We present our results of combining inertial and video features to recognize human activities on different types of scenarios where we achieve a F_1-measure up to 79.6%.**

***Index Terms*—activity recognition, machine learning, computer vision**

I. Introduction

Human Activity Recognition is an active field of research in pervasive computing [1]–[3]. One popular task of this field is the recognition of so called activities of daily living [4]. Especially in the field of health care and nursing, recognizing such activities becomes increasingly important as the cost for care increases [5], [6]. The detection of these activities poses a difficult problem and proposed solutions often rely on smart homes with many sensors in the environment. We propose the usage of off-the-shelf smart-devices to recognize such activities, where we rely on inertial sensors and an ego-centric camera. Several studies already investigate activity recognition, be it low-level [2], [7] or high-level activities [3], [8]. Usually, the former comprises actions like *walking* where the latter refers to context-enriched actions such as *preparing food*. Their results show on one hand that object-based activity recognition is the most promising vision-based approach [9] but on the other hand that the object recognition itself is error-prone and crucial in respect of the recognition quality [8]. In contrast, inertial-based activity recognition approaches perform poorly for high-level tasks, but are reliable for low-level activities which also include the tracking of the users arm [7]. For that reason, researchers started to shift to the idea of fusing this information. Approaches for fusing inertial and vision sensor have been made by other researchers [10]. However, most of the work focuses on the fusion of sensor streams that belong to the same on-body position [3], [11]. In this paper, we present a multi-modal ego-centric activity recognition approach that relies on smart-watches and smart-glasses to recognize high-level activities like activities of daily living. Particularly, we consider the inertial data of our smart-watch to derive the movement pattern of the forearm where in turn the egocentric video from smart-glasses provides information about objects. In this context, we aim to investigate to what extend vision information can improve the recognition of activities that are hard to recognize purely through motion sensing. We present our results of a multi-modal activity recognition approach based on manually annotated recordings as well as on a similar public dataset. Ideas for this paper were presented in a previous Work-In-Progress paper [12] and extended and implemented in this work. Our contributions in this publications are:

1) We collected a new dataset with two subjects performing a set of activities in two different environments with a focus on activities that are hard to distinguish as they involve similar motions (e.g. eating and drinking) and are often interleaved.
2) We present a new method for multi-modal activity recognition, utilizing deep learning models for object detection and evaluating this method on our presented dataset, achieving a F_1-measure of 79.6%.

II. Related Work

There are several methodologies from the domains of image and video processing targeting sub-problems of our research question. These approaches have shown to perform well in their respective applications. In the following, we summarize methods that can be used to support multi-modal activity recognition, namely separate methods for vision and inertial data and solutions for combining them.

A. Image object detection

Recently, there have been advances in deep and neural network based object detection where especially the TensorFlow Object Detection API[1] achieves promising results. Many different neural network architectures are available and have their separate advantages, offering trade-offs between

[1]https://github.com/tensorflow/models/tree/master/object_detection

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

performance and run-time. In our case, we rely on a ResNet FPN model as described in [13], as the reported performance of 35% mAP is still among the best offered, while having the advantage of a significantly lower run-time compared to the state-of-the-art network (1833ms vs. 76ms).

B. Activity recognition based on objects

Researchers have been using object information for activity recognition, especially when considering complex activities like cooking [14], [15]. For this purpose, the occurrence of objects and possibly the interaction with them is used to recognize an activity. Similarly, Lei et al. [16] build their system on a RGB-D camera system, detecting activities in a kitchen environment, focusing on the recognition of actions as well as objects with tracking and detection methods. Adding a camera to a wrist-worn sensor is another approach for detecting activities and was analyzed by Lei et al. [17]. A wrist worn camera has the benefit of having interactions with objects always in frame in addition to having a camera movement that corresponds to the hand movement of a test subject. One drawback from image based recognition is a limited field of view. When an activity occurs that is not fully captured within the video, the information is lost to a system. Therefore, we also look at inertial data, which is another modality that has been analyzed by many.

C. Activity recognition based on inertial data

Especially with the rise of smart-devices, researchers focused greatly on using inertial sensors for recognizing activities (in this case inertial data refers to acceleration, gyration and magnetic field data). Sliding windows in combination with acceleration data is a typical method to predict activities and has been analyzed by many researchers before [2], [18]. Especially activities like walking, jogging, and climbing stairs have been predicted successfully. Features that are calculated from these windows are often from the time and frequency domain and may contain statistical values like mean and variance but also more computationally expensive features like energy [2]. Apart from cyclic activities, there is also research involving inertial data that is focused on detecting short activities or events like falling [19], [20]. Falling however, is an activity with a unique motion that is hard to mix up with other activities of everyday living, thus these methods may not fully work in our scenario. Finally, for classification, classifiers that are commonly used are Decision Trees [18], Hidden Markov Models [21], and kNN [22] and recently Neural Networks [23]. In our work, we rely on a sliding window approach, similar to [2]. But in contrast to low-level activities, where the window size can be fairly long, we rely on short windows with greater overlap between consecutive windows, capturing the short nature of the activities. For our final goal of fusing together modalities, we also look at methods for multi-modal activity recognition.

D. Multi-modal activity recognition

Previous work has combined multiple sensors to create and analyze multi-modal datasets [14], [24]. Scenarios recorded in the datasets vary greatly and involve activities such as office work [24], sport activities [25], and cooking [14]. On top of these datasets, researchers tested different methods to recognize activities. One problem that is central in dealing with multi-modal datasets is the fusion of sensors with different sampling rates where a prominent example is the fusion of vision data with inertial data. Inertial data is usually sampled at a higher rate than video data, especially when using off the shelf sensors. Spriggs et al. [10] solved this problem by downsampling the inertial data to the capture rate of the video, thus having a one to one mapping of frames to single inertial measurements. When dealing with windowed feature, some of these problems can be mitigated. If windows are defined by timespans rather than by number of instances, merging them together can be simpler, when for example the same window size has been chosen. Another issue when dealing with multi-modal data is the fusion method. Song et al. [24] published their ego-centric multi-modal dataset which contains video data from smart-glasses along with inertial data from the same device. To recognize life-logging activities, they developed and presented a fusion method for inertial and video data based in Fisher Kernels. In this work, we rely on a windowing approach based on timespans for both of our modalities which allows us to fuse data both early and late.

III. Dataset

In our work, we consider two different datasets for evaluating our methods. The first dataset was collected by us and contains a set of activities that are hard to distinguish due to similar motions and often very short duration. The second dataset (CMU-MMAC) contains a wider variety of activities with more test subjects.

A. ADL dataset

To evaluate our methods, we recorded the ego-centric view of two subjects in a common home by smart-glasses and a chest mounted tablet as well as a third person camera recording the whole scenario. The subjects were also equipped with smart-watches and smart-phones to capture the movement of their arms and thigh, i.e., we recorded for all devices acceleration, gyration, and magnetic field data simultaneously. The subjects performed common and interleaved activities which include *drinking* (A_1), *eating* (A_2), *taking medicine* (A_3), *preparing meal* (A_4), *taking snack* (A_5), and *wipe mouth* (A_6). The sequence of activities was predefined and performed twice, once in a natural fashion and another time with artificial breaks between activities. Overall, we recorded six sessions per subject which reflects 30 minutes of activities.

The required data was collected using customary smart-devices[2] (see Figure 1) which were attached to the head (P_1), the left (P_2) and right (P_3) wrist, the chest (P_4), and also to the left (P_5) and right (P_6) thigh. Video and inertial data was recorded with a resolution of 1920x1080 (25fps) and 50Hz, respectively.

[2]"Vuzix M100" (Glasses), "LG G Watch R" (Watch), "Tango" (Tablet), "Samsung Galaxy S4" (Phone)

Fig. 1: Sensor placement. The subject wears the wearable devices on the head, chest, forearm, and thigh (top down).

Subsequently, we annotated the video recordings manually, i.e., we labeled on an activity level the start and stop times of the performed activities using third person video recordings and Boris [26]. On an object level, we drew the required bounding boxes around the visible objects within the egocentric video of the smart-glasses, annotating 14 objects including *bread*, *napkin*, *glass*, *knife*, *pillbox*, and both hands with Vatic [27].

Our labeled dataset is publicly available including a detailed description and images of each subject and the environment[3].

B. CMU-MMAC - Quality of Life dataset

The Quality of Life dataset [14] was created by the Carnegie Mellon University and contains a fairly large set of test subjects, cooking a variety of recipes. Recordings consist of different modalities such as first person overhead video, inertial measurement units that record acceleration, gyration, and magnetic field data on different body positions, audio from five different microphones, and in some cases even motion capturing data.

For our analysis, we focused on a subset of recipes, the brownie recipe, as labels for these recordings are provided on the website. One challenge within the dataset is the complexity of the labels. These are given in the form of *verb-object1-preposition-object2*, with the brownie recipe consisting of 14 different verbs, 34 different objects and 6 different prepositions. Overall, we counted 43 different labels in the subset we considered. Given the dataset size, building a multi-class model for 43 labels is not feasible. Therefore, we consider only the verb part of the activity as our target class, reducing the amount of classes to 14.

[3]https://sensor.informatik.uni-mannheim.de/#dataset_egocentric

In total we looked at 13 different subjects, considering the overhead camera frames and the acceleration data on both arms in our analysis.

IV. Methods

A. Acceleration data

For the acceleration information we considered the data from both smart-watches, as we aim to use a minimal amount of sensors to recognize activities. Initially, we planned to only consider data form the dominant hand of the test subjects, but as activities were often performed with a mix of both hands, we decided to use both. We transform the raw data into window features both from the time and frequency domain that can be seen in Table I. The windows have a length of 1000ms and an overlap of 75% and are created for each watch separately. With an overlap of this size, we make sure that the image windows and the inertial windows can be mapped. Inertial data on its own may be sufficient to recognize motions like raising an arm, but to properly detect the different activities, we also have to consider the visual information.

TABLE I: Set of features from acceleration data. Features are in the time and frequency domain.

Time domain	Frequency domain
Mean, Median, Standard Deviation, Variance, Inter Quantil Range, MAD, Kurtosis, Correlation Coefficient, Gravity, Orientation, Entropy	Energy, Entropy, MeanDC

B. Video

All features in our model are based around object information within the frames. A main factor for detecting activities, is the interaction of the test subject with objects. We assume that interaction with different objects is a good indicator for an activity and were able to see this in initial experiments. Thus, we try to estimate interactions by looking at objects overlap with a detected hand using a pre-trained neural network. We first pre-filter the frames and only consider those, that contain a positive detection for a hand (person class in the pre-trained network). Within these frames, we calculate the overlap of each detected object's bounding box with the hand's bounding box. The result is a vector of overlap percentages for all detectable classes with the rest of the frames receiving a vector of the same size with all values set to negative one.

To transform these vectors to windows, we calculate the average overlap of each object with the hand within each window where the window size is ten and the stride is five. By applying this approach, we try to work around movements of the arm within a sequence that generate an overlap with objects for a short period of time (e.g. when the arm passes over an object to get to another one). The whole process of extracting vision features is described in Figure 2.

To evaluate the approach further, we ran the experiments on learned image features as well as on the annotation ground truth data.

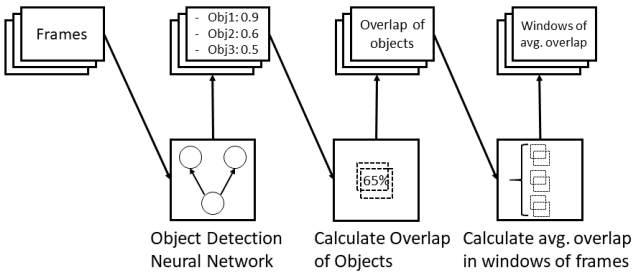

Fig. 2: Pipeline for the image feature generation

C. Combining both modalities

With both modalities we can estimate the overall sequence of activities. For that purpose, we define a method to combine both results into one classification. Before we combine the data, we first have to align both modalities as, at least in our dataset, the timestamps are not synched. After each of the three modalities was aligned, we consider the biggest overlap in time of the three sensors as our training and testing data. From the trimmed data we calculate our features like described before. In an initial approach we test early fusion, by concatenating the feature vectors and learning a model. Then, we also apply late fusion learning. First we assign each vision window the corresponding IMU windows that occurred at the same period of time. We merge both IMU windows into one feature vector and apply a learning algorithm. Simultaneously, we learn a model for the image window. For both windows we return the class probabilities and append them to the feature vectors. We then use another learning algorithm on top of both vectors concatenated together. For more insights, we evaluated the combination as well as both sensors separately in our experiments section.

V. Experiments

A. ADL dataset

For the experiments, we consider each subject separately and test our model with a 10-fold cross-validation. To test for stability, we run each cross validation 100 times with different folds and check for similar results. We tested a set of different configurations in respect to their performance to evaluate the influence of each modality and how they behave on their own. Configuration parameters include the classifier that is used for the late fusion learning, which modalities are used and which vision is assumed. These vision options are the ground truth object data or the detection results of the pre-trained neural network. For classification, we use Random Forest and Logistic Regression algorithms. When we consider all modalities, the classifiers used for the separate sensors are Random Forest for acceleration data and Logistic Regression for vision data. This way we keep the single modalities fixed and only change the fusion learning algorithm, reporting its performance.

TABLE II: Different configurations for our learning method. Values are reported as an average over all classes and for both subjects. RF = Random Forest, LR = Logistic Regression, ALL = both modalities were used, VIS = only vision features, IMU = only acceleration features, GT = ground truth vision, LEARN = vision features that have been detected by our neural network.

Config	Precision	Recall	F_1-measure
RF_ALL_GT	0.843	0.754	0.796
LR_ALL_GT	0.897	0.753	0.819
RF_ALL_LEARN	0.816	0.709	0.758
LR_ALL_LEARN	0.880	0.722	**0.793**
RF_IMU	0.673	0.556	0.609
LR_IMU	0.516	0.392	0.446
RF_VIS_GT	0.872	0.622	0.726
LR_VIS_GT	0.855	0.590	0.698
RF_VIS_LEARN	0.506	0.367	0.425
LR_VIS_LEARN	0.721	0.337	0.460

Using a sliding window approach with overlap poses the problem that two consecutive windows may end up in the training and in the testing set respectively. To avoid this, we sampled our data, depending on which modalities we evaluate, making sure that no data is present in training and testing simultaneously. In our vision and combined approach, the main point of reference is the image window. As it has an overlap of 50%, we consider every other vision-window and the attached IMU window as our dataset. When considering only acceleration data, the overlap of windows is 75%, thus we consider every fourth window in the experiments. We used a five-fold validation in these scenarios due to the amount of data available. Results are reported as an average of both test subjects.

In Table II, we can see that the best configuration uses all modalities and Logistic Regression as the fusion learning algorithm, yielding a F_1-measure of 79.3%. It can be seen, that results for vision improve greatly when assuming a perfect vision algorithm. The gap in performance is therefore attributed to the current state of object detection algorithms. When looking at the results of inertial data classification, the difference in performance among the learning algorithms is more emphasized. Overall, the results of the classification tend to prefer a high precision at the cost of recall which is beneficial in our scenario. Next, we examine the separate classes and the performance for each class using the best parameters of the previous experiment.

In Table III, the results for all classes are broken down for each class separately. Great performance can be achieved for the bread preparation class, with a F_1-measure of 90.9%. This makes sense, as for both modalities this class offers unique features. In the case of inertial data, the motion of buttering a bread is distinctively different than the other motions which all involve some sort of grabbing and lifting motion. For the video data, this scenario also offers unique views, as the test subjects were looking down, focusing on their plate.

TABLE III: An closer look at the results for our best configuration for each activity separately. Both vision and acceleration features were used in combination with Logistic Regression.

Class	Precision	Recall	F_1-measure
none	0.928	0.986	0.956
drink_water	0.886	0.62	0.729
eat_banana	0.868	0.511	0.643
eat_bread	0.867	0.749	0.804
prepare_bread	0.891	0.929	0.909
take_meds	0.894	0.676	0.769
wipe_mouth	0.837	0.585	0.688

TABLE IV: Results for CMU-MMAC dataset. Here we used the same method as above to evaluate our method. As we do not have bounding-box ground truth data, we can only learn on the output of our neural network.

Config	Precision	Recall	F_1-measure
RF_ALL	0.748	0.436	0.551
LR_ALL	0.738	0.482	**0.584**
RF_IMU	0.727	0.440	0.548
LR_IMU	0.230	0.115	0.153
RF_VIS	0.400	0.269	0.321
LR_VIS	0.395	0.236	0.295

Eating a piece of banana and wiping the mouth after the scenario were the worst performing activities, yielding F_1-measures of 64.3% and 68.8% respectively. There are separate reasons for both classes. In the case of eating a piece of banana, the shortness of the activity is the main problem. Test subjects were eating just one piece which was readily available on the table, thus few unique features are available to be learned. Wiping the mouth has the issue of hard to detect objects, as napkins are often hidden underneath the plate and hard to distinguish from the environment.

B. CMU-MMAC dataset

For the experiment of the CMU-MMAC dataset, we evaluated the whole dataset, among all subjects to see how well a model can be applied among a set of subjects instead of learning per single subject. Here we also ran the experiment 100 time and calculate the average precision and recall and the resulting average F_1-measure.

Given the harder task of the CMU-MMAC dataset, the lower F_1-measure of 58.4% (see Table IV) is not surprising. What contributes to this fact, is the larger amount of subjects that perform a greater set of activities, both of which adds more variation to the data. The bad performance using the vision features is also striking, with the performance going down to 32.1%. This can be attributed to our reduction of the labels to just the verb of the label. Thus, activities like *open-brownie_box* and *open-cupboard_top_left* are assigned the same label, even though they are performed on very different objects and in different situation. Vision features in this context are relying on the objects visible in frame and thus do not properly differentiate the different activities. When looking at the acceleration data though, the results are fairly

TABLE V: A closer look at our best performing configuration for the classes in the CMU-MMAC dataset. The model was learned in a 10-fold cross-validation among all subjects.

Class	Precision	Recall	F_1-measure
close	0.516	0.062	0.111
crack	0.757	0.389	0.514
none	0.674	0.783	0.724
open	0.690	0.481	0.567
pour	0.601	0.613	0.607
put	0.752	0.460	0.571
read	0.834	0.551	0.664
spray	0.890	0.726	0.800
stir	0.744	0.811	0.776
switch_on	0.859	0.630	0.727
take	0.708	0.648	0.677
twist_off	0.824	0.188	0.306
twist_on	0.793	0.196	0.314
walk	0.695	0.215	0.328

good. This is in line with results in [28] where it was shown that hierarchical clustering of the activities tends to favor activities with the same verb. Therefore, acceleration data is able to represent similar activities in a similar fashion. We could already see that Logistic Regression performs worse on our dataset when applied on acceleration data. This effect is even stronger in the CMU-MMAC dataset, most likely because of the bigger set of labels that have to be recognized. Random Forest behaves similar in both cases and yields good results which is in line with previous research [2].

To look deeper into the classification results, we consider the IMU classification results on their own and show the performance for each class.

Table V shows our findings. Good performance can be seen in classes like pouring and stirring with a F_1-score of 60.7% and 77.6% respectively, while generic classes like reading or closing are not recognized very well. This is in line with our assumption that the acceleration data is able to distinguish specific activities (i.e. stirring involves a motion that is very unusual compared to the others) and has problems distinguishing verbs that are very generic.

We can see that the combination of inertial and video data yields a better result than each sensor on its own. Depending on the activity that should be recognized, modalities perform differently as they are relying on the variation within the data. Inertial data for example, may not be as expressive when the activities that are performed are very similar in motion. Thus, it makes sense to consider the combination of both modalities to predict high level activities.

We also considered other sensors like infrared and depth cameras to mitigate issues like privacy concerns with the setup. These however, are not readily available in smart devices and can sometimes even divulge more information than regular cameras. Also, they are restricted in respect of their usage. Depth cameras for instance have a minimum working distance which impairs the usage in many scenarios. We believe that privacy concerns can be reduced when an application is only used in specific contexts like specific rooms. Here the usage of

smart devices can really help as they can be turned off based on such a context switches.

VI. Conclusion and Future Work

In this paper, we present a new multi-modal dataset that includes activities of daily living. Then, we present a method for recognizing activities by using window features and fusion, based on acceleration- and ego-centric video data. This way we were able to achieve a F_1-measure of 79.6% on our presented dataset and 58.4% on the CMU Multi-Modal Activity dataset. Both scenarios pose different challenges for our approach. For our dataset the similarity of the activities is challenging, while the CMU-MMAC dataset contains a wider variety of activities by a greater number of subjects. We can show that our approach is promising for the recognition of activities in a multi-modal setting, including the usage of off-the-shelf sensors build into smart-devices. Future work in this field can be done in multiple directions. It is obvious that gyration and magnetic field information may also be included for inertial analysis, as inertial measurement units often record these modalities simultaneously. Extending the vision features is another direction. More complex methods like object tracking can be employed, thus containing more in depth information about the motion. Finally, the method for fusing the modalities can be analyzed in more depth e.g. using different window lengths and methods of voting and boosting to learn a final model

References

[1] T.-H.-C. Nguyen, J.-C. Nebel, and F. Florez-Revuelta, "Recognition of activities of daily living with egocentric vision: A review," *Sensors*, vol. 16, no. 1, p. 72, 2016.

[2] T. Sztyler and H. Stuckenschmidt, "On-body localization of wearable devices: An investigation of position-aware activity recognition," in *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE Computer Society, 2016, pp. 1–9.

[3] S. Song, V. Chandrasekhar, B. Mandal, L. Li, J.-H. Lim, G. S. Babu, P. San, and N.-M. Cheung, "Multimodal multi-stream deep learning for egocentric activity recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE Computer Society, 2016, pp. 378–385.

[4] M. P. Lawton and E. M. Brody, "Assessment of older people: Self-maintaining and instrumental activities of daily living," *The gerontologist*, vol. 9, no. 3_Part_1, pp. 179–186, 1969.

[5] T. Hori, Y. Nishida, and S. Murakami, "Pervasive sensor system for evidence-based nursing care support," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 1680–1685.

[6] D. H. Wilson, *Assistive intelligent environments for automatic health monitoring*. Carnegie Mellon University, 2005.

[7] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber, "Smartwatch-based activity recognition: A machine learning approach," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE Computer Society, 2016, pp. 426–429.

[8] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2012, pp. 2847–2854.

[9] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 744–760, 2015.

[10] E. H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference On*. IEEE, 2009, pp. 17–24.

[11] J. Windau and L. Itti, "Situation awareness via sensor-equipped eyeglasses," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE Computer Society, 2013, pp. 5674–5679.

[12] A. Diete, T. Sztyler, L. Weiland, and H. Stuckenschmidt, "Improving motion-based activity recognition with ego-centric vision," in *2018 IEEE International Conference on Pervasive Computing and Communications : PerCom 2018, Athens, Greece, March 19-23, 2018 : PerCom Workshops proceedings*. IEEE Computer Society, 2018.

[13] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in *CVPR*, vol. 1, no. 2, 2017, p. 4.

[14] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," *Robotics Institute*, p. 135, 2008.

[15] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.

[16] J. Lei, X. Ren, and D. Fox, "Fine-grained kitchen activity recognition using rgb-d," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 208–211.

[17] T. Maekawa, Y. Yanagisawa, Y. Kishino, K. Ishiguro, K. Kamei, Y. Sakurai, and T. Okadome, "Object-based activity recognition with heterogeneous sensors on wrist," in *International Conference on Pervasive Computing*. Springer, 2010, pp. 246–264.

[18] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[19] Y. S. Delahoz and M. A. Labrador, "Survey on fall detection and fall prevention using wearable and external sensors," *Sensors*, vol. 14, no. 10, pp. 19 806–19 842, 2014.

[20] C. Krupitzer, T. Sztyler, J. Edinger, M. Breitbach, H. Stuckenschmidt, and C. Becker, "Hips do lie! A position-aware mobile fall detection system," in *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2018, pp. 1–10.

[21] R. San-Segundo, J. M. Montero, R. Barra-Chicote, F. Fernández, and J. M. Pardo, "Feature extraction from smartphone inertial signals for human activity segmentation," *Signal Processing*, vol. 120, pp. 359–372, 2016.

[22] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, and D. Howard, "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 871–879, 2009.

[23] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[24] S. Song, N. M. Cheung, V. Chandrasekhar, B. Mandal, and J. Liri, "Egocentric activity recognition with multimodal fisher vector," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Computer Society, 2016, pp. 2717–2721.

[25] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 168–172.

[26] O. Friard and M. Gamba, "Boris: a free, versatile open-source event-logging software for video/audio coding and live observations," *Methods in Ecology and Evolution*, vol. 7, no. 11, pp. 1325–1330, 2016.

[27] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, vol. 101, no. 1, pp. 184–204, 2013.

[28] A. Diete, T. Sztyler, and H. Stuckenschmidt, "Exploring semi-supervised methods for labeling support in multimodal datasets," *Sensors*, vol. 18, no. 8, 2018. [Online]. Available: http://www.mdpi.com/1424-8220/18/8/2639