A Smart Data Annotation Tool for Multi-Sensor Activity Recognition

Alexander Diete
University of Mannheim
Mannheim, Germany
alex@informatik.uni-mannheim.de

Timo Sztyler
University of Mannheim
Mannheim, Germany
timo@informatik.uni-mannheim.de

Heiner Stuckenschmidt
University of Mannheim
Mannheim, Germany
heiner@informatik.uni-mannheim.de

***Abstract*—Annotation of multimodal data sets is often a time consuming and a challenging task as many approaches require an accurate labeling. This includes in particular video recordings as often labeling exact to a frame is required. For that purpose, we created an annotation tool that enables to annotate data sets of video and inertial sensor data. However, in contrast to the most existing approaches, we focus on semi-supervised labeling support to infer labels for the whole dataset. More precisely, after labeling a small set of instances our system is able to provide labeling recommendations and in turn it makes learning of image features more feasible by speeding up the labeling time for single frames. We aim to rely on the inertial sensors of our wristband to support the labeling of video recordings. For that purpose, we apply template matching in context of dynamic time warping to identify time intervals of certain actions. To investigate the feasibility of our approach we focus on a real world scenario, i.e., we gathered a data set which describes an order picking scenario of a logistic company. In this context, we focus on the picking process as the selection of the correct items can be prone to errors. Preliminary results show that we are able to identify 69% of the grabbing motion periods of time.**

I. Introduction

Automated annotation of sensor and video data is of great interest especially due to the recent developments in wearable devices. Many successful approaches rely on them and enable, e.g., to recognize activities of daily living. However, typically the manually annotation of large amount of sensor data goes along with them. Already existing annotation tools often only provide visual support to annotate such data but rarely the needed semi-supervised support. Even then, semi-supervised labeling support is usually limited. Thus, typically the tool learns from the already labeled data to provide labeling recommendations or perform autocomplete but without considering the information source, i.e., the kind of sensor that in turn is important concerning suitable feature extraction techniques.

Therefore, we focus on an annotation tool for multi-sensor activity recognition that is able after a short phase of training to label comparable data sets automatically. To investigate the deployability, we consider a real world scenario, thus, we consider an order picking task in a logistic company. As part of a project, we were able to examine their environment and also to record sensor and video data of a grabbing task. Indeed, a lot of processes in the industry can be augmented to improve productivity. Egocentric video data as well as sensors on different body parts of a worker can gather information on the task at hand.

In this context, picking is the action of selecting items from boxes in shelves that make up an order in a warehouse environment. As of now identifying the location of the item and validating that the correct item is selected are done manually. We believe that an augmented system can improve this process significantly and therefore are currently working on a project that aims to deliver a full augmented picking system that can be used in warehouses without relying on specific barcodes or RFID techniques [1].

For that purpose, we use a custom wristband and smart-glasses which should support each other. Hence, the sensors of the wristband should be used to support the labeling, e.g., of the recorded video but can also provide clues if the worker is actually looking at the item which is grabbed. This enables to decide automatically if specific video frames should be ignored for training. As a first step, we primarily focus on the grabbing detection and the corresponding interpretation and labeling.

In this work, we present our method of using inertial sensors to recognize activities and create labels to be used for annotating video data. For that purpose, we focus on template matching in context of dynamic time warping as preceding works already presented promising results [2]. In contrast, a sliding window approach seems to be less promising concerning the recognition across subjects [3].

This paper is structured as follows: In Section II, we present the related work. Subsequently, we are going to describe the data set we collected and used for our system. In Section IV, we outline our methods that are used for automatic labeling and the corresponding tool for annotation. Section V summarizes our preliminary results. Finally, Section VI describes our conclusion and future work.

II. Related Work

In many respects, especially the automatic annotation of video data is challenging [4]. For that purpose, we focus on the feasibility to transfer automatic recognized labels of inertial sensor data to corresponding video recordings. In this context, the precision plays an important role as many scenarios require accurate labeling concerning the individual frames. For that purpose, we focus on template extraction and matching of certain motions in context of dynamic time warping.

978-1-5090-4338-5/17/$31.00 ©2017 IEEE

Indeed, Margarito et al. [2] already showed that templates that were extracted from a wrist worn accelerometer enable to recognize certain sports activities across different people. However, the performance concerning a broader set of activities is still unclear. Further, they pointed out that combining different template-matching metrics in context of statistical classifiers could be also promising.

In contrast to template matching, Spriggs et al. [3] investigated a multi-modal based classification approach considering also inertial sensors but as well first-person video data. They focused on daily kitchen actions and performed a frame based classification by relying on features that were extracted from the inertial sensor and video data. However, they clearly state that their approach do not generalize well across people.

Relying only on inertial and force sensors, Morganti et al. [5] stated that already different wrist shapes and muscles configurations across people can affect the recognition procedure. Further, they point out that especially the force sensors enable to recognize specific gestures that could not reliable be recognized by inertial sensors. However, they presented only preliminary results, hence, the performance in context of a user-independent scenario is unclear.

Focusing on sensor data annotation tools, several researchers already presented powerful and promising annotation tools. However, only few of them provide support concerning labeling recommendation or automated labeling. Palotai et al. [6] presented a labeling framework that relies on common machine learning approaches but was only designed for domain experts. In addition, it is unclear how their approach performance concerning different level of activity types or how different sensors are supported concerning their introduced learning approach (e.g. feature extraction). Indeed, Barz et al. [7] highlight that most data acquisition and annotation tools are mostly limited to a particular sensor. This can be attributed to the fact that it seems to be necessary to consider different techniques or feature sets for different kinds of sensors. Especially the combination of these concerning learning approaches in context of automated labeling seems to be challenging.

III. Data Set

In our work, we investigate the feasibility to apply template matching of simple motions represented by inertial sensor data across different people to support semi-supervised labeling of video data. For that purpose, we created our own data set which describes the procedure of a worker that prepares an order in a warehouse. The data recording followed a predefined protocol that contains a sequence of actions, i.e., *walking to shelf*, *locating the correct box*, and *grabbing from the box*. In this context, several scenarios were recorded including picking from different boxes on different rows and from different shelves.

The test environment consists of two shelves located next to each other where each shelf has three rows of boxes with three to five boxes per row. Thus, the boxes were placed on different heights and were spread horizontally among two shelves (see Figure 1). The environment setup is based on a real warehouse that we could inspect in regard of a company project. A problem with recognizing grabbing motions is the variation of that activity, thus, a grabbing motion can produce highly different sensor outputs depending on the location of the object to be grabbed. In contrast, activities like walking or running do not have this degree of variation. Therefore our data set contains multiple different cases of grabbing within the shelf to cover the space of different motions.

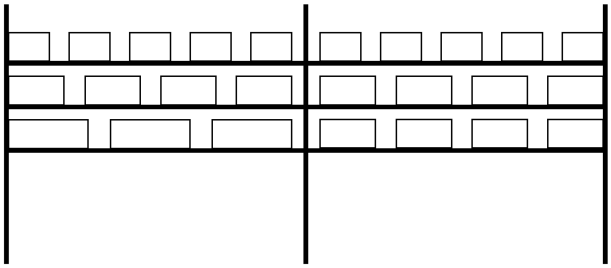


Fig. 1. Schematic of the shelves in the test environment.

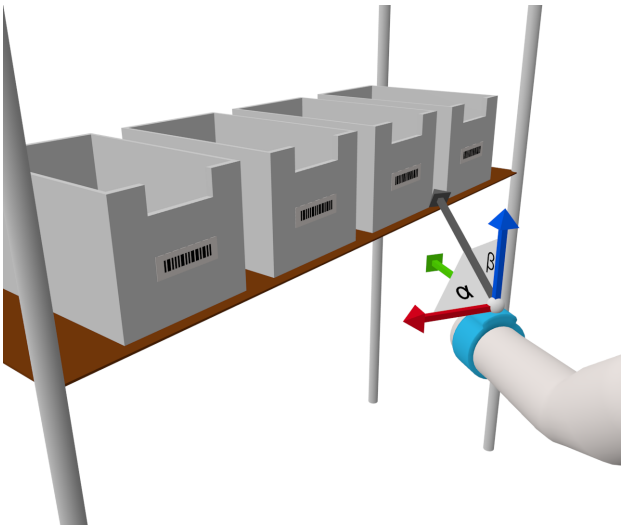


Fig. 2. Angle features from the wristband used for matching.

The required data for our data set was collected using smart-glasses[1] and a custom wristband. Both devices recorded acceleration, gyration, and magnetic field data while the smart-glasses also recorded video information of its front camera. The inertial sensors of the wristband are all read at the same point in time while the sensors of the smart-glasses cannot recorded synchronously[2]. Further, as the wristband and the smart-glasses are not connected, the recorded timestamp of the data has to be synchronized afterwards. For that purpose, the subjects were instructed to stand idle for a period of time before and after the performance of the activities. Data on the wristband was collected at 40 Hz for all the sensors while

[1] Vuzix M100

[2] Under Android, sensors are not queried but the system pushes new values. Hence, there is no guaranty concerning the specified frequency.

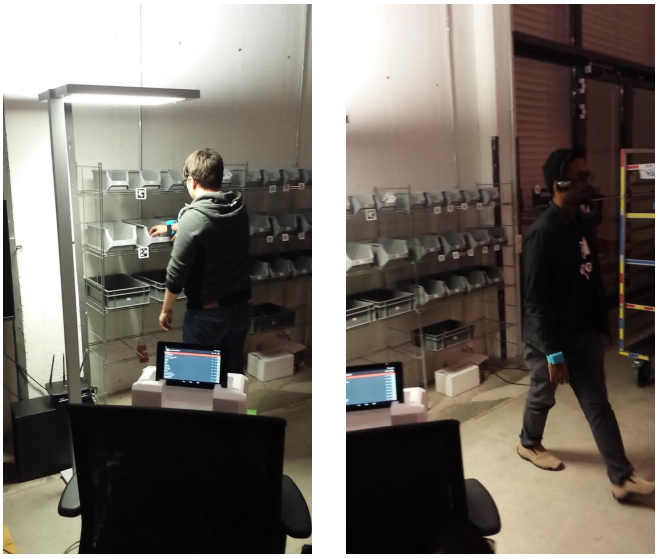


Fig. 3. Two subjects in picking data set. On the left: grabbing action in place. On the right: walking away from shelf as indicator for end of grabbing action.

the smart-glasses recorded the sensors at 50 Hz and 25 fps respectively. Hence, we recorded with the highest possible frequency. For better interpretation, each recording session was also filmed from a third person perspective (see Figure 3) using a depth sensor enabled tablet[3]. This enables also to collect depth information of the recorded images in form of point clouds.

For the recording, we relied on a self-developed application. Thus, we enhanced an Android application of a previous work [8] where especially the support of smart-glasses was added. The recorded data of the smart-glasses is stored locally on the device. However, the custom wristband does not have enough storage to store the data locally, hence, we had to send the recorded data directly over Wi-Fi to a server.

As an additional reference, we also considered the data set *Quality of Life* which was published by F. Torre et al. [9]. In contrast to our own data set this one is far more complex in multiple ways. Hence, the data set contains more different arm motions. For instance, getting a cup from a shelf includes opening the shelf and grabbing the cup. The motions themselves are also not as homogeneous as they are in our data set since the setting of a kitchen leads to many different arm movements for retrieving objects. Besides, they also recorded the movement of both arms instead of just one.

In our experiments, we consider both cases, i.e. simple and complex, to clarify the feasibility and performance of our approach. In this context, we consider a subset of the *Quality of Life* data set. In particular we only looked at one recipe, i.e. the brownie recipe, for a subset of all the participants because it was the only one which was completely labeled.

IV. Method

We developed a web-based application which provides support concerning the alignment, analysis, and labeling of inertial sensor and video data (see Figure 4). The tool maps the inertial sensor data onto the video by visualization which in turn enables to adjust the data regarding a possible offset but also to define a label for a certain period of time. The user defined labels are utilized to learn patterns so that labeling recommendations can be provided. This reduces the labeling effort but also enables to gather labels for video recordings by inspecting the inertial sensor data. Besides, the visualization and recommendations may also allow to assign the labeling task to a non-domain expert.

As a first step, we align our recorded data in respect of the timestamp where we considered zerolines at the beginning of each recording that allow us to pinpoint the starting time of an action. More precisely, we used the peaks of walking motions in accelerometer plots to align the data as those were easily identifiable as the first action. Subsequently, we labeled, e.g., the grabbing action by analyzing the inertial sensor data that represents the motion and countercheck against the video data that describes the same time period. This allows to label all sensor recordings simultaneously. Once the boundaries of an action are defined, the web application replays the corresponding part of the video that were recorded from the data glasses. After the confirmation of the correctness, the corresponding inertial sensor data is extracted for creating a template of this action where a template is represented by start and end timestamp, the corresponding inertial sensor data, and a label.

For now, we focus on the acceleration data because preliminary experiments have shown that the angles relative to the three axes are promising concerning the characterizing of the grabbing motion in context of a wristband. After a certain number of templates of the same action are available, we apply dynamic time warping [10] to identify possible matches. We assume that the same motions produce similar outputs which only differ in respect of their length due to the varying speed the action was performed. Thus we chose

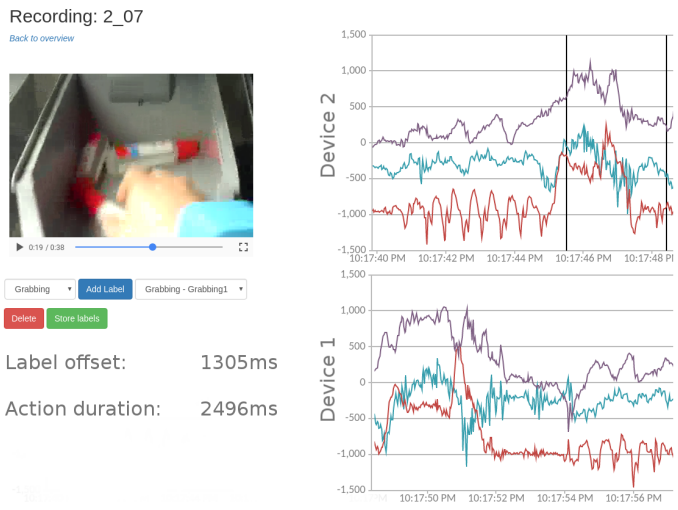


Fig. 4. Labeling tool aligning wristband and video data.

[3]Google Tango

dynamic time warping as it allows us to match time series of different lengths. Dynamic time warping works by finding a path between two time series that have the smallest distance. The minimal distance is found by first initializing the distance from every point in series A to the first point in series B to $infinity$ and vice versa. Afterwards the algorithms iterates over the combination of all points in both series and calculates their distance by using a $cost$ function (in our case euclidean distance). The function compares single points and the cost of the path leading to the previous points (recursive):

$$d(i,j) = cost(i,j) + min(d(i,j-1), d(i-1,j), d(i-1,j-1))$$

By considering three preceding options that lead to i, j, the algorithm can cope with different lengths of series. The extracted templates slide over an unlabeled data set to detect the time when an action occurs. In this context, we try to find the position of the template with the smallest deviation while assuming that at least one action occurs in the unseen data. Future work would use a threshold value for the distance to determine if an action we want to match occurs in the data.

Focusing on the *Quality of Life* data set, due to its complexity, we have to consider additional steps. Hence, in contrast to our own data set, people switch between the left and right hand which means that it is also necessary to identify which hand was used for the current activity. Therefore, we unify the data of the same sensor type of both hands so that the current activity is represented by a single vector. Considering the corresponding labels, it stands out that the described activities cover several motions, e.g., grabbing is only a sub-activity. Therefore, still focusing on acceleration data and considering the corresponding gold standard to extract the templates, we segment the data of a template into small windows to compute features that have a stronger expression concerning more complex activities. This includes the used energy (Fourier transform) and median absolute deviation. Due to these high-level labels, several different activities may cover common sub-activities, e.g., taking a pot or turning on the stove includes grabbing. Therefore, we also investigate if the extracted templates have a label independent correlation. We assume that the extracted templates could be grouped to actions that are specific in their motion and not in their semantic. For that purpose, we apply agglomerative clustering to group the templates where the distances between the clusters are the result of the dynamic time warping. Detecting the motion similarity between certain activities, may allow to generalize activity labels but also to construct more robust templates due to the varying executions which in turn supports to avoid overfitting.

For the experiments, we perform leave-one-out cross validation. Thus, extracted templates from n-1 data sets, and applied them on the remaining one.

V. Preliminary Experiments

As this paper is a work in progress, we have yet to evaluate the whole labeling tool in regards to usability for manual annotation of datasets. We rather focus on the performance of our labeling support tool to see if it is a feasible approach to be used in greater scale. Afterwards we plan to evaluate the whole labeling tool using standard agreement measures between annotators and suggested labels like Cohen's kappa or weighted kappa.

TABLE I
Recognition performance of template matching. The overlap (avg. 69%) is excluding outliers and represents only the best match within a data set.

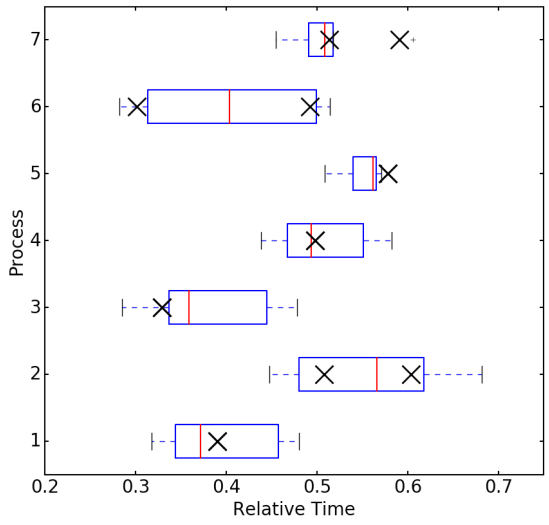
Data set	1	2	3	4	5	6	7
Overlap	0.43	0.67	0.78	0.52	0.72	0.74	0.99
Motion [s]	5.02	2.49 2.22	2.55	4.23	2.86	2.43 4.11	2.04 2.60
Δ Start [s]	1.41	1.89 1.81	0.91	0.86	0.71	2.88 2.61	0.65 2.91
Δ Duration [s]	1.65	0.74 0.68	1.40	1.46	0.63	0.68 1.52	1.99 1.43

In our initial experiments, we only focused on the grabbing action in context of the inertial sensor data that correspond to the wristband. Thus, we want to investigate the feasibility to apply template matching across different people to identify certain activities where in turn the result should be used to provide recommendations concerning the labeling of the video recordings.

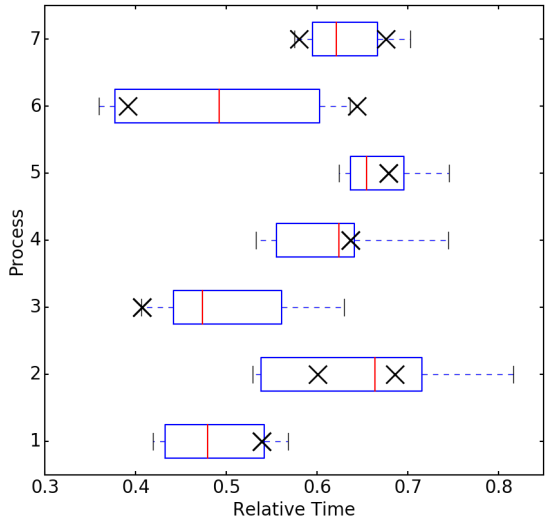
For that purpose, we applied our introduced approach on our own data set. Hence, we extracted the grabbing motion templates from all except one data set where each set covers a complete picking process. We measured the temporal overlap of the estimated and the actual grabbing motions. For the average overlap per data set we took the best match (i.e. the match with the least distance) for each template. Afterwards, we selected the most promising subsets of matches and used them to calculate the average overlap for each test data set. The most promising subset of matches is determined by evaluating all the subsets of the match results with k elements and then selecting the one with greatest overlap among itself. Testing showed that a value of k=6 yielded the best results in our data set. Table I summarizes the results and points out that we were able to detect nearly all grabbing motions but have an issue concerning the accuracy of the start and stop boundaries. Indeed, inspecting the individual results strikes that the assumption that the searched action has to have the same length as the considered template leads to an inaccuracy.

Figure 5 describes in detail the recognition and distribution results for all start and stop timings. We want to emphasize that the x-axis does not represent the recognition rate but the relative duration of the whole process. Hence, the box plot represents the time interval were we suppose the start point, respectively the stop point, for the action that should be recognized. Every box represents the best match for the templates where the x markers show the actual point in time of the grabbing motion. As there can be two grabbing motions in a data set we plotted both positions. The boxes provide an interesting insight concerning the reliability, i.e., most of the

extracted templates were able to identify the correct area of a certain action across different recordings of the same process.



(a) Result Start



(b) Result End

Fig. 5. Overall estimate of grabbing start and end point. Crosses represent correct start or end points.

Considering the *Quality of Life* data set, our first results were misleading because different activities covered similar arm movement. For instance, the extracted templates of the activity *take oil* also recognized *put oil into cupboard*. Thus, we tried to cluster the actions based on their similarity to get an insight regarding their meaning. Figure 6 illustrates the clustering result of one sample set. It is striking that some actions that use items within a similar location are ending up in the same cluster fairly consistently. For instance, we can observe that motions like taking the big and small measuring cup are very similar. However, in contrast the fork and the scissors for instance are both located in a drawer but end up in the same cluster fairly late. We believe that this is most likely due to fact that the actions are more variable in length than they are in our own data set. Even though dynamic time warping is able to handle different lengths of timeseries it is still very likely that the distances of short templates are generally smaller and thus end up faster in clusters than the longer actions like for instance taking the baking pan from the oven.

Fig. 6. Dendogram of the clustering of the templates. Marked boxes are actions using the same item.

Further experiments will deal with the main problem of our approach: better identifying the correct start and end points of the action we recognized. As of now our approach finds the general position of the action but struggles to pinpoint the borders correctly as template length varies greatly between single instances. We see two possible solutions that can be used to tackle the problem. Hence, on one hand it would be possible to create more templates for other actions and try to create a coherent time series of actions. By using the distance values for all the template matches we can readjust the start and end points accordingly. On the other hand, after we initially guessed the position of a certain action, we also want to test readjusting the window size to maximize their similarity.

VI. Conclusion

In this work, we investigated the possibility of a smart data annotation tool that provides labeling recommendations based on the already labeled inertial sensor data. We aim to reduce the labeling effort but also want to determine to which extent the recognized labels from the inertial sensor data could be used to label and process video recordings. For that purpose, we performed preliminary experiments to investigate the feasibility of applying template matching in context of dynamic time warping to recognize certain actions

across different processes and people. In this context, we focused on acceleration data of a wristband to recognize certain actions. It has emerged that it depends on the granularity of the considered activity labels which recognition technique is promising. Hence, activities that actually consist of several sub-activities may have to be considered initially separately. In this context, we also showed that clustering existing templates from a labeled data set allows to infer similarities in motion from semantically different actions. This can be considered as a starting point to construct more robust templates while the clustering results also yield more information for a specific motion which in turn reduces the need to perform a certain activity more frequently to get enough characterizing information. Besides, in contrast to other approaches [3], [4], we need significantly less data to guess the correct time frame of a certain action.

In our next steps, we want to focus on the problems which came up during our investigations. This includes the recognition quality of the boundaries of actions due to the limitation of a predefined template length but also that we considered so far only acceleration data to extract labeling recommendations. Thus, considering further sensors may also increase the recognition accuracy. For that purpose, we want enhance our on data set concerning the number of instances but also regarding the considered activities since it turned out that the considered activity level is essential. For improving the performance of the overall system we could also implement active learning. By showing an annotator instances with very uncertain classification ratings, a machine learning algorithm could further tune its parameters.

VII. Acknowledgement

We would like to thank the Ubimax GmbH for providing the hardware and environment for generating our dataset as well as helping with collecting the data.

References

[1] A. Diete, L. Weiland, T. Sztyler, and H. Stuckenschmidt, "Exploring a multi-sensor picking process in the future warehouse," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 1755–1758.

[2] J. Margarito, R. Helaoui, A. M. Bianchi, F. Sartor, and A. G. Bonomi, "User-independent recognition of sports activities from a single wrist-worn accelerometer: A template-matching-based approach," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 788–796, 2016.

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

[4] T. Ishihara, K. M. Kitani, W. C. Ma, H. Takagi, and C. Asakawa, "Recognizing hand-object interactions in wearable camera videos," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 1349–1353.

[5] E. Morganti, L. Angelini, A. Adami, D. Lalanne, L. Lorenzelli, and E. Mugellini, "A smart watch with embedded sensors to recognize objects, grasps and forearm gestures," *Procedia Engineering*, vol. 41, pp. 1169–1175, 2012.

[6] Z. Palotai, M. Lng, A. Srkny, Z. Tsr, D. Sonntag, T. Toyama, and A. Lrincz, "Labelmovie: Semi-supervised machine annotation tool with quality assurance and crowd-sourcing options for videos," in *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2014, pp. 1–4.

[7] M. Barz, M. M. Moniri, M. Weber, and D. Sonntag, "Multimodal multisensor activity annotation tool," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. New York, NY, USA: ACM, 2016, pp. 17–20.

[8] 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.

[9] F. De la Torre Frade, J. K. Hodgins , A. W. Bargteil, X. Martin Artal, J. C. Macey, A. Collado I Castells, and J. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-08-22, April 2008.

[10] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.