# *Activity recognition in manufacturing: the roles of motion capture and sEMG+inertial wearables in detecting fine vs gross motion*

# Activity recognition in manufacturing: The roles of motion capture and sEMG+inertial wearables in detecting fine vs. gross motion

Alyssa Kubota[1], Tariq Iqbal[2], Julie A. Shah[2], Laurel D. Riek[1]

***Abstract*— In safety-critical environments, robots need to reliably recognize human activity to be effective and trustworthy partners. Since most human activity recognition (HAR) approaches rely on unimodal sensor data (e.g. motion capture or wearable sensors), it is unclear how the relationship between the sensor modality and motion granularity (e.g. gross or fine) of the activities impacts classification accuracy. To our knowledge, we are the first to investigate the efficacy of using motion capture as compared to wearable sensor data for recognizing human motion in manufacturing settings. We introduce the UCSD-MIT Human Motion dataset, composed of two assembly tasks that entail either gross or fine-grained motion. For both tasks, we compared the accuracy of a Vicon motion capture system to a Myo armband using three widely used HAR algorithms. We found that motion capture yielded higher accuracy than the wearable sensor for gross motion recognition (up to 36.95%), while the wearable sensor yielded higher accuracy for fine-grained motion (up to 28.06%). These results suggest that these sensor modalities are complementary, and that robots may benefit from systems that utilize multiple modalities to simultaneously, but independently, detect gross and fine-grained motion. Our findings will help guide researchers in numerous fields of robotics including learning from demonstration and grasping to effectively choose sensor modalities that are most suitable for their applications.**

## I. Introduction

Robots demonstrate great potential for decreasing physical and cognitive workload, improving safety conditions, and enhancing work efficiency for their human teammates in a variety of areas including hospitals and manufacturing environments [1], [3], [27], [35]. Particularly in safety-critical environments, robots need the ability to automatically and accurately infer human activity. This will allow them to operate either autonomously or with minimal user input to avoid distracting their human teammates.

Robots can learn valuable information about the activities of their human partners from their motion [4], [12], [14], [20]. Gross motion detection (e.g. movement of the arms, legs, or torso) is the primary area of focus for most human activity recognition (HAR) approaches, traditionally using RGB cameras, depth sensors, or motion capture systems [4], [20]. Thus, robots can recognize gross motion daily living activities, such as walking or lifting items, with accuracies of as high as 99% [9], [13], [15], [33].

Research reported in this paper is supported by the National Science Foundation under Grant Nos. IIS-1724982 and IIS-1734482.

[1]Computer Science and Engineering, UC San Diego, La Jolla, CA 92093, USA (email: {akubota, lriek}@eng.ucsd.edu)

[2]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA (email: tiqbal@mit.edu, julie_a_shah@csail.mit.edu)

However, recognizing fine-grained motion (e.g. movement of hands or fingers) is imperative for enabling robots to accurately understand human intention in safety-critical industrial environments. For example, in order to infer how a person is using a screwdriver or placing an item, the robot needs to perceive their hand and wrist motion. However, most conventional sensors do not provide adequate information to accurately detect these movements, so fine-grained activity recognition is unreliable using traditional HAR approaches.

To recognize these minute movements, one approach researchers have employed is hand-centric motion capture [18], [21]. However, motion capture often requires expensive equipment and a cumbersome installation procedure [20]. Furthermore, these sensors are easily occluded in dynamic environments, resulting in reduced recognition accuracy [20].

Thus, many researchers instead employ wearable sensors such as accelerometers, gyroscopes, or surface electromyography (sEMG) sensors for fine-grained motion detection [20], [26]. Recent examples include automatically recognizing American Sign Language, identifying gestures to interface with technology, and detecting different types of grasps to control robotic arms [2], [23], [41].

Both motion capture and wearable sensors have proved effective for HAR when recognizing different granularities of motion. However, especially in the context of robotics, their relative efficacy for detecting gross and fine-grained motion is unclear. If their relative capabilities were known in this context, then it may be possible to combine multiple sensor modalities in a complementary fashion to more accurately detect a wider variety of activity.

To our knowledge, we are the first to directly compare the efficacy of motion capture and wearable sensors for recognizing gross and fine-grained motion in the context of assembly manufacturing. We employed three common classification algorithms for HAR (support vector machine (SVM), linear discriminant analysis (LDA), *k*-nearest neighbors (KNN)). We chose these classifiers due to their success in recognizing activities using motion capture or wearable sensor data [4], [19], [20]. To evaluate these modalities on both granularities of motion, we introduce the new UCSD-MIT Human Motion dataset. We used a Vicon motion capture system and a Myo armband to record participants completing two assembly tasks. The first is an automotive assembly task consisting of primarily gross motor movements. The second is a block assembly task which required fine grasping movements.

Our empirical evaluations on the two tasks suggest that these two sensor modalities are complementary: motion capture yields better accuracy for gross motion, while wearable

sensors yield better accuracy for fine-grained motion. When classifying gross motion, using the motion capture system achieved up to 36.95% higher accuracy than the wearable sensor. On the other hand, the wearable sensor data was up to 28.06% more accurate for fine-grained motion recognition.

Our findings will help roboticists understand how motion capture and wearable sensors compare when classifying activities of different motion granularities. In turn, this will unveil which sensors are best suited for detecting activities that are relevant in a given context. Thus, robots can better infer the person's task by utilizing a multimodal system to simultaneously detect gross and fine-grained motion.

This work will also help the robotic learning from demonstration and grasping communities choose sensor modalities to recognize motion specific to their needs. Learning from demonstration researchers may choose to use motion capture data to teach a robot gross motion activities, and utilize wearable sensors for fine-grained manipulation training. Similarly, as grasping is a very fine-grained activity, our results suggest that the robotic grasping community can garner important insights by using wearable sensors to complement the traditional RGB-D and motion capture systems.

## II. Background

Many activities that occur in everyday life (e.g. walking, climbing stairs, lifting objects) primarily entail gross motion. Thus, the majority of HAR algorithms are designed to recognize these activities, typically using data gathered via external sensors such as RGB-D or motion capture systems.

In particular, motion capture has many applications. It can help robots track people and objects in an environment, generally using mounted cameras. For example, unmanned aerial vehicles rely on motion capture data to guide them and prevent collisions while in autopilot [11], [31]. It is also widely used for tracking human activity for applications such as security in public spaces and entertainment [25], [34].

Many researchers have explored using motion capture data to help robots predict gross motion in manufacturing, a safety-critical environment. For example, Unhelkar et al. [39] used a Kinect to create human-aware robots that can safely deliver parts to human workers in an automotive assembly environment. Similarly, Hayes and Shah [9] classified automotive assembly activities using 3D joint locations of people and objects from a Vicon system. Mainprice et al. [24] captured single-arm reaching movements of two people to help robots predict activities in collaborative environments.

However, in many settings, robots need to be able to recognize pertinent activities that involve fine-grained motion, such as grasping. Reliable classification of fine motion is particularly difficult due to the small, ambiguous movements that human hands are capable of [20], [40].

One approach to fine activity classification is using visual data to track hands and objects in the environment. For instance, Lei et al. [22] achieved high classification accuracy of seven kitchen activities by using RGB-D data to track hands interacting with 27 different objects. However, using visual data is not necessarily viable in all settings, especially

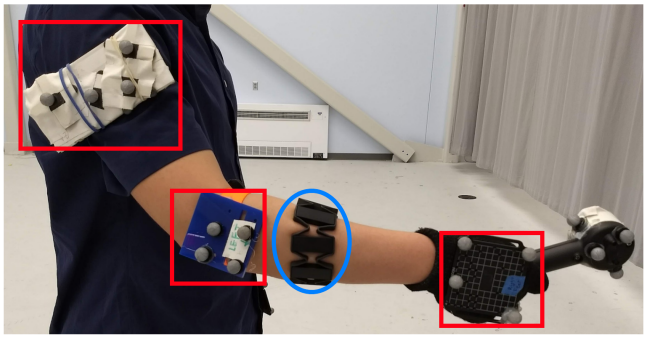

Fig. 1. Arrangement of sensors on a participant's arm. Vicon markers are in red boxes. Myo is circled in blue.

dynamic and chaotic environments where cameras can often be occluded. Additionally, cameras for motion capture and visual sensing are expensive to install, and their field of view is limited to a constrained physical space.

On the other hand, wearable sensors are mobile and thus can be used to recognize activity anywhere. Thus, body-worn non-visual sensors are another common approach to fine-grained activity recognition. For example, Zhu et al. [42] used data from an inertial measurement unit (IMU) worn on the finger to recognize five different hand gestures. Batzianoulis et al. [2] used arm muscle activity data from sEMG sensors in tandem with finger joint locations to recognize five different types of grasping motions.

A commonly used wearable sensor in recent studies is the Myo armband which measures sEMG and inertial data. Researchers have used it to recognize a wide variety of activities such as daily living, gym exercises, and wandering behavior in the elderly [19], [37], [38].

All of the aforementioned work used either motion capture or a wearable sensor to recognize gross or fine-grained motion. However, it is unclear whether they could have achieved higher accuracy for their activity set had they used a different sensing modality. Accurate recognition of both gross and fine motion is especially crucial for robots in safety-critical spaces where an error could result in harm to a human partner. To this end, we investigate whether there is an advantage to using one sensor modality over another for recognizing different granularities of motion.

## III. Methodology

In this work, we compared the efficacy of motion capture and wearable sensors for recognizing gross and fine-grained motion. We collected the UCSD-MIT Human Motion dataset, comprised of two tasks. The first task is an automotive assembly task entailing gross motion, and the second is a block assembly task consisting of fine grasping motion (see Section III-A.2). The automotive task contains four activity classes, and the block task has five. Five participants (two female, three male) performed both tasks. We trained three widely used machine learning algorithms with these data, and used F1 scores as our evaluation metric (see Section IV). In this section, we describe the data collection procedure, labeling method, and classification algorithms.

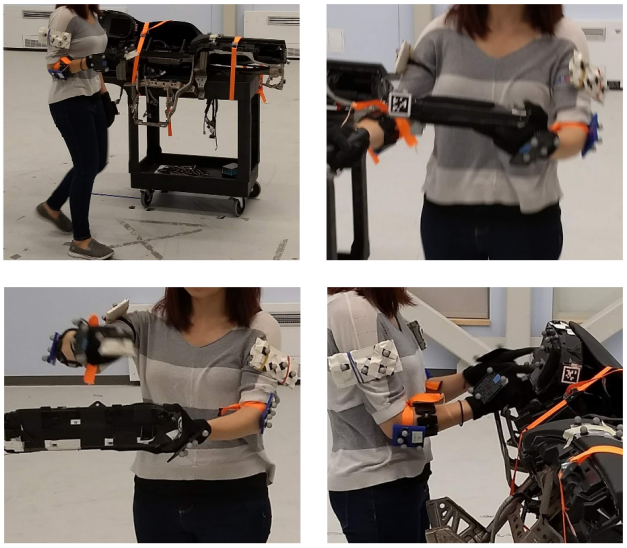

Fig. 2. Activities from the automotive assembly task. From left to right, top to bottom: *Walking*, *Receiving Part*, *Scanning Part*, *Attaching Part*.

### A. Data Collection

*1) Sensors:* We collected data using a Vicon motion-capture system and Myo armband simultaneously (See Fig. 1). We placed Vicon markers on the shoulders, elbow, and back of the hand on each of the participants' arms. Participants wore the Myo on the forearm of their dominant arm. We solely tracked participants' arm movements to avoid burdening them with excessive Vicon markers, while still capturing relevant information as they completed activities.

We connected the sensors to a single machine (Intel i7-6820HQ CPU, 16GB of RAM) to ensure a consistent timestamp across all data. We used the Robot Operating System (ROS) (version Indigo) to save time-synchronized data in *rosbag* format. The Vicon has a sampling rate of 120Hz, while the Myo has a sampling rate of 50Hz for IMU data and 200Hz for sEMG data. In accordance with other real-time activity recognition systems, we sampled our data at a consistent rate of 30Hz in order to reduce the computation required by the systems [8], [9], [30].

*2) Dataset Creation:* To evaluate the efficacy of these sensors on different granularities of motion, we constructed the automotive and block assembly tasks to have activities composed of either gross or fine-grained motion respectively.

The automotive assembly task, inspired by the Dynamic-AutoFA dataset, consists of four gross motion activities [9]. As such, no actions in this task depend on dexterous hand or finger movements. The four main activities are *Walking*, *Receiving Part*, *Scanning Part*, and *Attaching Part* (see Fig. 2, Table I). There are between two and four instances, or occurrences, of each activity throughout the task. Each participant completed five trials (i.e. repetitions of the task) yielding a total of 50 to 100 instances of each activity.

The block assembly task consists of five fine grasping motions. Participants received a box with one flat base block and four rectangular blocks. In order to simulate different dexterous hand movements, we asked participants to grab and affix each block to the structure in a distinct manner. The activities in this dataset are *Palmar Grab*, *Thumb-3 Fingers*, *Thumb-2 Fingers*, *Pincer Grab*, and *Ulnar Pinch Grab* (see Fig. 3, Table II). These grasps are similar to those used in other grasp recognition studies [2], [42]. Each participant completed five trials, performing each grasp once per trial, which yielded a total of 25 instances of each grasp.

TABLE I

SEQUENCE OF ACTIVITIES PERFORMED IN THE AUTOMOTIVE TASK.

| Class | Description |
|---|---|
| *Walk* | to dashboard |
| *Scan* | dashboard |
| *Walk* | to left side of dashboard |
| *Receive* | speedometer |
| *Scan* | speedometer |
| *Attach* | speedometer |
| *Walk* | to right side of dashboard |
| *Receive* | navigation unit |
| *Scan* | navigation unit |
| *Attach* | navigation unit |
| *Walk* | to exit |

We collected data from five participants who engaged in both the automotive and block assembly tasks. Participants were between the ages of 26 and 34, with a mean age of 28.2 years. Two of the five participants were female, and three were male. Four of the participants were right-handed, and one was left-handed.

### B. Data Processing and Labeling

*1) Feature Selection:* For both the Vicon and Myo data, we use low-level, raw data features in the temporal domain. This is to assess the baseline capabilities of these sensor modalities without the influence of high-level feature selection, which can drastically impact a classifier's accuracy [16]. Data are partitioned using a sliding window technique, with window size of 1 second with 50% overlap.

The Vicon markers provided the 3D position ($x$-, $y$-, $z$-coordinates) of the selected joints with respect to the Vicon's internal coordinate system. Since there were six joints (three on each arm), there were a total of 18 of these features in the dataset. We chose to track these joints because they are similar to the arm joints tracked in the Carnegie Mellon University Motion Capture Database [5].

For the Myo data, we collected the linear acceleration, angular velocity, and muscle activity data of each participants' dominant arm. This included $x$-, $y$-, and $z$- linear acceleration, $x$-, $y$-, and $z$- angular velocity, and the eight channels of sEMG data, yielding a total of 14 features. We chose these features to help detect arm position and orientation relative to the wearer, as the Myo does not sense movements relative to the global environment. Additionally, the sEMG signals can help detect differences in hand motion.

*2) Data Labeling:* Two annotators manually labeled the data by reviewing recorded video played back from a *rosbag* file. Annotators used a script to record the start and end time of each activity. In order to ensure consistency in our class labels, we conducted inter-rater reliability analysis by

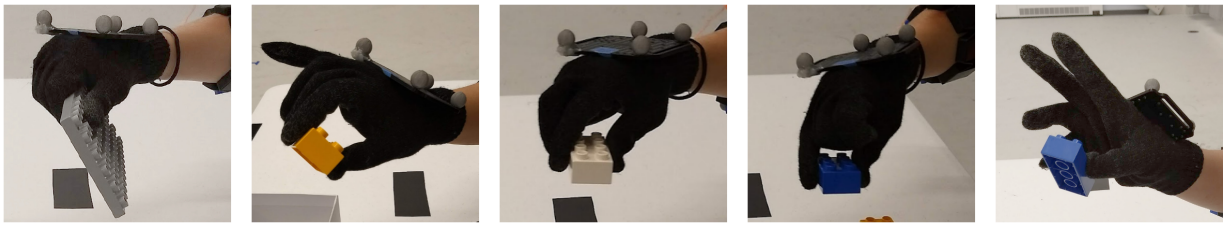

Fig. 3. Grasping activities in the block assembly task. From left to right: *Palmar*, *Thumb-3 Fingers*, *Thumb-2 Fingers*, *Pincer*, *Ulnar pinch*.

TABLE II
GRASP TYPES USED IN THE BLOCK TASK. EACH GRASP USED A DIFFERENT COMBINATION OF FINGERS.

| **Grasp Name** | **Fingers Involved** |
|---|---|
| Palmar | All |
| Thumb-3 Fingers | Thumb, Index, Middle, Fourth |
| Thumb-2 Fingers | Thumb, Index, Middle |
| Pincer | Thumb, Index |
| Ulnar Pinch | Thumb, Pinky |

computing the two-way mixed intraclass correlation (ICC) for our labeled data. ICC is a measure of similarity between class labels, in our case the similarity of the start and end times of the activities between annotators [6]. Thus, we normalized the timestamps to the start of each trial.

We found Cronbach's $\alpha = .81$ which indicates that our labels were consistent between annotators [36].

### C. Classification algorithms

We trained three machine learning classifiers on both datasets to determine which sensor modality is better suited for recognizing gross and fine-grained motion. We used a support vector machine (SVM) with a linear kernel function ($C = 1$), linear discriminant analysis (LDA), and *k*-nearest neighbors (KNN) ($k = 5$) [32]. We chose these classifiers as they have proven successful in HAR and other applications [4], [20]. As our goal was not to compare the classifiers against each other, we used standard values for additional parameters (e.g. $C$ for SVM, $k$ for KNN) to simplify the selection process.

SVMs are widely used for pattern recognition, classification, and regression [10]. They use kernel functions to calculate hyperplanes with which to divide training instances into proposed classes. These are then used to classify new instances. They have shown success in high dimensional spaces while producing interpretable results [10], [20].

LDA models training instances parametrically as multivariate means then uses linear decision boundaries to separate them into classes [19]. They inherently handle multiclass data such as ours and do not require hyperparameter tuning.

KNNs are a type of instance-based learning that classifies new samples as the most prevalent class of the *k* most similar training instances [20]. We chose $k = 5$ to maintain distinct classification boundaries between classes.

### D. Evaluation

To evaluate the relative efficacies of motion capture and wearable sensors, we performed leave-one-out cross-validation for each task (i.e. we tested each individual trial by training the classifier on all other trials of that task and then classified the original trial). In the case where we fused the Vicon and Myo data, we employed early fusion techniques, or combined the features before classification, which showed success in our prior work [29].

We calculated the mean F1 score to evaluate the classification efficacy across all trials of each participant for both tasks (see Table III). As such, the training set is not subject specific, but does contain data from that participant. The F1 score of a class is the average of its classification precision and recall. Its value lies in the range of 0 to 1, where values closer to 1 indicate higher precision and recall. We chose to use the mean F1 score over raw accuracy as our performance measure as it is a better indicator of performance, especially when class distributions are imbalanced [17]. This was the case for the automotive task, since each trial contained up to twice as many more instances of some activities than others.

To determine the significance of our independent variables on our dependent variable (F1 score), we performed a three-way repeated-measures analysis of variance (ANOVA) test. The independent variables we tested were motion granularity (gross or fine), sensor modality (Myo or Vicon), and classifier (SVM, LDA, or KNN) (see Table IV).

## IV. RESULTS

Mauchly's Test of Sphericity indicated that all combinations of motion granularity, sensor modality, and classifier violated the assumption of sphericity, i.e. the variances of differences between data of the same participant were not equal. The exceptions to this were Sensor Modality * Classifier), $\chi^2(2) = 1.37$, $p = .504$, and Motion Granularity * Sensor Modality * Classifier, $\chi^2(2) = 0.68$, $p = .712$. Thus, we corrected the degrees of freedom for all other combinations using Greenhouse-Geisser estimates of sphericity. We corrected family-wise error rate in post hoc comparisons using Bonferroni correction.

*1) Motion Granularity:* Motion granularity had a significant main effect on F1 score. Regardless of the sensor modality or classifier used, the type of motion being classified significantly impacted the F1 score, $F(1, 19) = 532.76$, $p < .001$, $r = 0.98$.

TABLE III

MEAN F1 SCORES OBTAINED FOR EACH DATA MODALITY ON EACH DATASET USING DIFFERENT CLASSIFIERS. ACROSS THE DATASETS AND SENSORS, WE AVERAGED THE F1 SCORES FROM EVERY TRIAL. A HIGHER F1 SCORE IS BETTER.

| | **SVM** | | | **LDA** | | | **KNN** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Vicon | Myo | Vicon+Myo | Vicon | Myo | Vicon+Myo | Vicon | Myo | Vicon+Myo |
| **Automotive** (Gross motion) | **.79** | .42 | .43 | **.76** | .48 | .49 | **.88** | .58 | .59 |
| **Block** (Fine-grained motion) | .09 | **.37** | .36 | .23 | **.39** | .36 | .32 | **.43** | .43 |

TABLE IV

F-TESTS OF FACTORS. $p \leq .05$ INDICATES A SIGNIFICANT EFFECT ON F1 SCORE FOR INDIVIDUAL VARIABLES, AND SIGNIFICANT INTERACTION BETWEEN VARIABLES FOR MULTIPLE. CONFIDENCE FOR ALL $p$-VALUES IS 95%. $r$-VALUE IS EFFECT SIZE.

| **Source** | $p$ | $r$ |
|---|---|---|
| Motion Granularity | $< .001$ | 0.98 |
| Sensor Modality | $< .001$ | 0.69 |
| Classifier | $< .001$ | 0.75 |
| Motion Granularity * Sensor Modality | $< .001$ | 0.96 |
| Sensor Modality * Classifier | $> .05$ | 0.21 |
| Motion Granularity * Classifier | $> .05$ | 0.28 |
| Motion Granularity * Sensor Modality * Classifier | $< .001$ | 0.54 |

*2) Sensor Modality:* The sensor modality also had a significant main effect on F1 score. Regardless of the motion granularity or classifier, the sensor significantly impacted F1 score, $F(1, 19) = 17.08$, $p < .001$, $r = 0.69$.

*3) Classifier:* We also found that the main effect of the classifier was significant, $F(1.54, 29.24) = 38.07$, $p < .001$, $r = 0.75$. Contrasts between each classifier found that the KNN achieved higher F1 scores than the SVM, $F(1, 19) = 52.22$, $p < .001$, $r = 0.86$, as well as the LDA, $F(1, 19) = 29.00$, $p < .001$, $r = 0.78$. The LDA also outperformed the SVM, $F(1, 19) = 17.53$, $p < .001$, $r = 0.693$.

*4) Motion Granularity * Sensor Modality:* There was a significant interaction between the motion type and sensor type, $F(4, 19) = 219.39$, $p < .001$. This indicates that the sensor had significantly different effects on the F1 score depending on the motion granularity being recognized, and vice-versa. Contrasts revealed that the Vicon yielded higher accuracy than the Myo for gross motion, but lower accuracy for fine. Conversely, the Myo yielded higher accuracy than the Vicon for fine-grained motion, but lower for gross, $F(1, 19) = 219.39$, $p < .001$, $r = 0.96$ (see Fig. 4a).

*5) Sensor Modality * Classifier:* There was no significant interaction between the sensor modality and classifier, $F(2, 38) = 1.83$, $p > .05$, $r = 0.21$. The interaction graph supports this finding (see Fig. 4b).

*6) Motion Granularity * Classifier:* There was also no significant interaction between the granularity of motion being classified and the classifier, $F(1.53, 29.10) = 3.27$, $p > .05$, $r = 0.28$. The interaction graph supports this finding (see Fig. 4c).

*7) Motion Granularity * Sensor Modality * Classifier:* Finally, the was significant interaction between all three of the independent variables, $F(2, 38) = 15.34$, $p < .001$, $r = 0.54$. This indicates that F1 score was significantly different for each combination of motion granularity, sensor modality, and classifier. This is reflected in the interaction graphs as the difference in F1 score is consistently greatest between the SVM and KNN (see Fig. 4b,c).

## V. DISCUSSION

Our results suggest that motion capture and wearable sensors offer complementary strengths for HAR. Motion capture is more accurate for detecting gross motion, while wearable sensors are more accurate for recognizing fine-grained motion. Our results also indicate that both sensor modalities yielded significantly more accurate recognition of gross motion than fine-grained which suggests that fine-grained motion is more difficult to classify than gross.

For gross motion recognition, we found that motion capture data yielded significantly higher accuracy than the wearable sensor data. This may be because the Vicon utilizes 3D position in the environment, so the relative position of the person may help the classifiers more accurately recognize gross motion. For example, the *Receiving Part* and *Attaching Part* activities occur in consistent, but different, locations in the environment. Thus, the classifiers can use the consistent arm positions to distinguish between these two activity classes. On the other hand, the Myo only obtains data relative to the user, so it does cannot distinguish activities in the same way. Arm movement may not be enough information to accurately detect gross body motion.

For recognizing fine-grained motion, we found that the wearable sensor data yielded significantly higher accuracy than motion capture data. The Myo can detect the muscle activity generated by the minute motion variations of each grasp to help the classifiers differentiate between them. In contrast, the Vicon tracks the position of the hands as opposed to the fingers, so the 3D motion it captures is similar between these fine-grained finger activities. Moreover, joints were often occluded from view, resulting in lower accuracy, a known problem when working with visual sensors [7].

Our results also indicate that fine-grained motion is more difficult to classify than gross. Across all classifiers, both the Myo and Vicon yielded lower accuracy on the block assembly task than on the automotive one. This may be because the movements between the grasps were similar (overhand, using some number of fingers) which led to ambiguities in the data. Fine-grained hand motion, as seen in our dataset, can be difficult to discern as it often entails analogous arm motion and muscle activity. In future work, we will explore higher level features and combinations of sensors to more accurately recognize these activities.

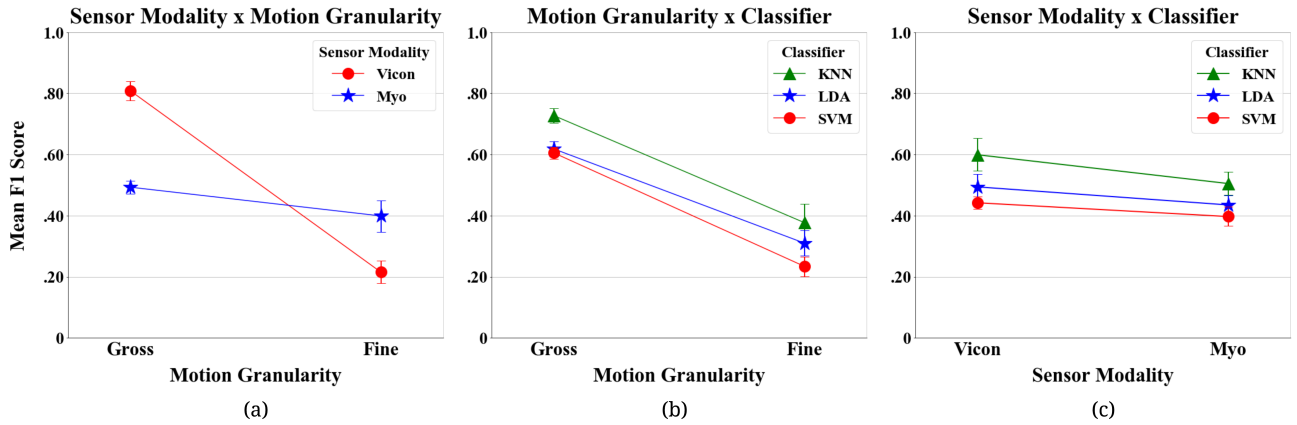


Fig. 4. Interactions between each pair of variables. The y-axis represents mean F1 score over all trials. Similar slopes between lines indicate insignificant interaction between variables. There was significant interaction between sensor modality and motion granularity, and insignificant interaction of classifier with both motion granularity and sensor modality.

Our results also suggest that multimodal sensor fusion resulted in lower classification accuracy than when using a single sensor for both tasks. Prior work in other recognition tasks showed that using similar multimodal approaches can improve classification accuracy, so we expected a similar result here [28], [29]. However, it is possible that the additional modalities contributed more noise than meaningful information, resulting in lower accuracy. In future work, we may be able to mitigate this by performing higher level feature extraction (e.g. mean absolute value for sEMG data, frequency domain features for inertial data), training a deep learning model to extract more significant information, or exploring alternate fusion techniques [29].

Depending on the types of relevant activities in the space, robots may need different kinds of sensor data in order to accurately recognize the intentions of their human counterparts. Our findings can help the robotics community make more informed decisions regarding which sensor modalities would be most beneficial for their specific tasks. This decision depends considerably on which activities are important for robots to recognize as well as the motion granularity of these activities. For instance, if the robot needs to know that a person is lifting a heavy object and may need help, motion capture systems are reliable. On the other hand, wearable sensors would better help a robot to determine which tool to fetch next depending on whether the person is currently assembling a part with a hammer versus a screwdriver.

A limitation of this work is that we only recorded the arm motion of the participants. In many HAR scenarios, movement of other body parts and environmental features can improve activity detection [9]. While it is possible that motion capture would have performed better with more markers, recognizing precise finger movements would still be a challenge due to their close proximity. Therefore, it is unlikely that using more markers would have increased accuracy of fine-grained motion, and improvements in accuracy of gross motion would further support our findings. Additionally, motion capture is not always viable for small tools and parts (e.g. screwdrivers for assembling small electronics). Thus, we subject both the Vicon and Myo to the difficult scenario where only human arm movements are measured.

Our findings suggest promising avenues for improving HAR of complex tasks in safety-critical settings. However, a limitation that should be addressed to improve the robustness of such systems is that we assume the classifier is trained on previous data from each participant, which may not always be the case in real-world scenarios. Additionally, as the amount of training data increases, so does the computational complexity of these classifiers. This is not ideal for a robot that must react quickly in dynamic settings. Therefore, as more data is collected, approaches that can handle larger datasets such as deep learning may be more suitable.

As we continue research in this area, we plan to develop a multimodal system that can leverage the complementary nature of these sensor modalities to recognize both gross and fine-grained motion so robots can better infer human activity. We will also extend our dataset in order to create a more reliable unimodal activity recognition system. Once we have a classifier that can reliably detect human activity, we plan to explore how robots can improve safety conditions for human workers in safety-critical settings.

Our findings can help the robotics community to understand which sensors work best for certain activities. These insights will enable researchers to design algorithms for robots that incorporate complementary multimodal approaches to better recognize activities that entail both motion types. These findings can also help guide both the robotic learning from demonstration and grasping communities as they choose sensor modalities best suited for their contexts. Our findings will help robots infer human intention regardless of the nature of the activities and environment. With the means to accurately distinguish particular activities, they can better support people and improve safety conditions in more specialized, safety-critical settings.

## REFERENCES


[1] P. Akella, M. Peshkin, E. Colgate, W. Wannasuphoprasit, N. Nagesh, J. Wells, S. Holland, T. Pearson, and B. Peacock. Cobots for the automobile assembly line. In *Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on*, volume 1, pages 728–733. IEEE, 1999.

[2] I. Batzianoulis, S. El-Khoury, E. Pirondini, M. Coscia, S. Micera, and A. Billard. Emg-based decoding of grasp gestures in reaching-to-grasping motions. *Robotics and Autonomous Systems*, 91:59–70, 2017.

[3] A. Cherubini, R. Passama, A. Crosnier, A. Lasnier, and P. Fraisse. Collaborative manufacturing with physical human–robot interaction. *Robotics and Computer-Integrated Manufacturing*, 40:1–13, 2016.

[4] M. Cornacchia, K. Ozcan, Y. Zheng, and S. Velipasalar. A survey on activity detection and classification using wearable sensors. *IEEE Sensors Journal*, 17(2):386–403, 2017.

[5] 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*, page 135, 2008.

[6] K. A. Hallgren. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23, 2012.

[7] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool. Tracking a hand manipulating an object. In *Computer Vision, 2009 IEEE 12th International Conference On*, pages 1475–1482. IEEE, 2009.

[8] N. Y. Hammerla, S. Halloran, and T. Ploetz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[9] B. Hayes and J. A. Shah. Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 6586–6593. IEEE, 2017.

[10] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[11] J. P. How, B. Behihke, A. Frank, D. Dale, and J. Vian. Real-time indoor autonomous vehicle test environment. *IEEE control systems*, 28(2):51–64, 2008.

[12] T. Iqbal, S. Rack, and L. D. Riek. Movement coordination in human–robot teams: a dynamical systems approach. *IEEE Transactions on Robotics*, 32(4):909–919, 2016.

[13] T. Iqbal and L. D. Riek. A method for automatic detection of psychomotor entrainment. *IEEE Transactions on affective computing*, 7(1):3–16, 2015.

[14] T. Iqbal and L. D. Riek. Coordination dynamics in multi-human multi-robot teams. *IEEE Robotics and Automation Letters (RA-L)*, 2017.

[15] T. Iqbal and L. D. Riek. Human-robot teaming: Approaches from joint action and dynamical systems. *Humanoid Robotics: A Reference*, pages 1–20, 2018.

[16] A. Janecek, W. Gansterer, M. Demel, and G. Ecker. On the relationship between feature selection and classification accuracy. In *New challenges for feature selection in data mining and knowledge discovery*, pages 90–105, 2008.

[17] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245–251. IEEE, 2013.

[18] Z. Ju and H. Liu. A unified fuzzy framework for human-hand motion recognition. *IEEE Transactions on Fuzzy Systems*, 19(5):901–913, 2011.

[19] H. Koskimäki, P. Siirtola, and J. Röning. Myogym: introducing an open gym data set for activity recognition collected using myo armband. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 537–546. ACM, 2017.

[20] O. D. Lara, M. A. Labrador, et al. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3):1192–1209, 2013.

[21] P. A. Lasota, G. F. Rossano, and J. A. Shah. Toward safe close-proximity human-robot interaction with standard industrial robots. *2014 IEEE International Conference on Automation Science and Engineering (CASE)*, 2014.

[22] 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*, pages 208–211. ACM, 2012.

[23] Z. Lu, X. Chen, Q. Li, X. Zhang, and P. Zhou. A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices. *IEEE Trans. Human-Machine Systems*, 44(2):293–299, 2014.

[24] J. Mainprice, R. Hayne, and D. Berenson. Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 885–892. IEEE, 2015.

[25] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126, 2006.

[26] S. C. Mukhopadhyay. Wearable sensors for human activity monitoring: A review. *IEEE sensors journal*, 15(3):1321–1330, 2015.

[27] B. Mutlu and J. Forlizzi. Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 287–294. ACM, 2008.

[28] A. Nigam and L. D. Riek. Social context perception for mobile robots. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 3621–3627. IEEE, 2015.

[29] M. F. O'Connor and L. D. Riek. Detecting social context: A method for social event classification using naturalistic multimodal data. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 3, pages 1–7. IEEE, 2015.

[30] F. J. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[31] M. Orsag, C. Korpela, and P. Oh. Modeling and control of mm-uav: Mobile manipulating unmanned aerial vehicle. *Journal of Intelligent & Robotic Systems*, 69(1-4):227–240, 2013.

[32] 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*, 12:2825–2830, 2011.

[33] N. Pham and T. Abdelzaher. Robust dynamic human activity recognition based on relative energy allocation. In *International Conference on Distributed Computing in Sensor Systems*, pages 525–530. Springer, 2008.

[34] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[35] L. D. Riek. Healthcare robotics. *Communications of the ACM*, 60(11):68–78, 2017.

[36] M. Tavakol and R. Dennick. Making sense of cronbach's alpha. *International journal of medical education*, 2:53, 2011.

[37] M. S. Totty and E. Wade. Muscle activation and inertial motion data for noninvasive classification of activities of daily living. *IEEE Transactions on Biomedical Engineering*, 65(5):1069–1076, 2018.

[38] T. Toutountzi, C. Collander, S. Phan, and F. Makedon. Eyeon: An activity recognition system using myo armband. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, page 82. ACM, 2016.

[39] V. V. Unhelkar, P. A. Lasota, Q. Tyroller, R.-D. Buhai, L. Marceau, B. Deml, and J. A. Shah. Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time. *IEEE Robotics and Automation Letters*, 3(3):2394–2401, 2018.

[40] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1553–1567, 2006.

[41] J. Wu, Z. Tian, L. Sun, L. Estevez, and R. Jafari. Real-time american sign language recognition using wrist-worn motion and surface emg sensors. In *Wearable and Implantable Body Sensor Networks (BSN), 2015 IEEE 12th International Conference on*, pages 1–6. IEEE, 2015.

[42] C. Zhu and W. Sheng. Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(3):569–573, 2011.