Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Machine Learning for Human Movement Understanding

Taizo Yoshikawa, Viktor Losing

2020

Preprint:

This is an accepted article published in Advanced Robotics. The final authenticated version is available online at:
https://doi.org/10.1080/01691864.2020.1786724

Machine Learning for Human Movement Understanding

Taizo Yoshikawa, Viktor Losing, and Emel Demircan

Abstract*— *Main purpose of this project is to develop fundamental technology for assist robots to recover and maintain human motor skill and to extend scope of human activity. Our goal is to provide a system that adapts to its user's personal behavior patterns in real-time. We aim to develop a continuous collaboration system between the physical assist robots and the user where both alternatively adjust to each other to maximize the system's utility. To understand human movement, we recorded motion sequence of several tasks for different subjects using motion capture system. Through analysis of human motion data, we extracted a general model by rule-base approach. On the other hand, since such tasks are not feasible with static models, we investigate the potential benefit of supervised online learning in the task of online action classification and Deep Learning in the task of acquiring human motion. Finally, developed system was extended to show its potential effect in ergonomics and in assist robotics.

1 INTRODUCTION

It is estimated that Japan's workforce will decline year by year due to population aging, population decline and low birth rate. According to the report of Ministry of Health, the number of the employees in 2012 was 62.7 million, on the other hand, 60 million are anticipated when economic growth and labor participation will assume the case which does not advance appropriately in future when they interrupt it. Furthermore, in 2030, it is simulated that the number of the employees decreases to 54.49 million. The low birthrate and aging are said to be more in future when one of 2.5 people of the nation becomes the elderly person 65 years or older in progress in 2060. A new approach to future human resources utilization becomes important to each company. In order to counteract the diminishing workforce, it is crucial not

only to increase the current proportion of foreigners but also to enable more females, elderly and handicapped persons to participate in physically demanding tasks. To utilize a variety of human resources, it is demanded to develop an environment where an employee can easily work and brings about result. It is necessary to actively supplement physical exercise work in the production environment to improve work efficiency and safety.

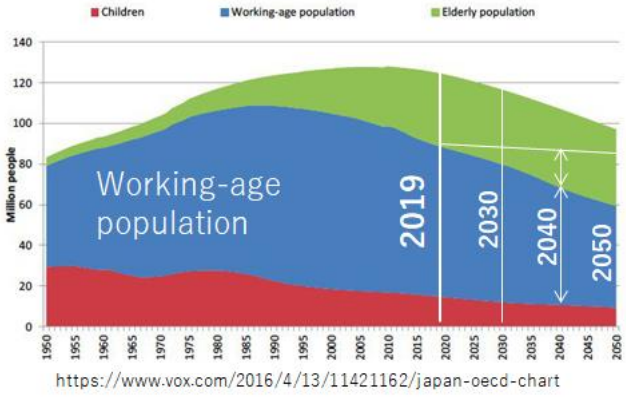


Fig. 1: Japanese population and age structure

To do so, it is required to evaluate the workload of standard work by ergonomics, reduce high-load posture work, improve the process, and arrange the process in the optimal process that matches the ability of the worker. Kinematic and kinetic models that apply robot technology to human work movement measurement data have been constructed, and more faithful kinetic and dynamic analysis based on musculoskeletal models has been incorporated [1]. Physiological and psychological models are incorporated into the workload as needed, and it has become possible to perform workload evaluation by ergonomics. In recent years, with the further progress of computerization and the spread of new technologies such as AI and robots, it has become possible to replace tasks that were difficult to mechanize with machines and systems. In order to construct a variety of production systems adapted to the various specialties of workers, introduction of robot technology that physically assists the movements of workers has attracted attention.

It is expected that the introduction of assisting devices constitutes a promising way to augment the diverse capabilities of workers according to the demands of a given task and secure a stable future labor force. In recent years, the introduction of assist robot technology for reducing the burden on the waist and technology for assisting walking have been advanced [2][3][4]. CYBERDYNE Co., Ltd [5]. has realized a technology to reduce the load on the waist when lifting and carrying objects by detecting the bioelectric

potential signal. The technology of CYBERDYNE Corporation has been put to practical use. ATOUN [6] developed It a POWERED WEAR, a wearable robot that alleviates burden to the waist when lifting heavy objects, through powerful motors. The technology of ATOUN also has been put to practical use.

The Honda Walking Assist Device [7] adopts cooperative control technology that was developed based on Honda's cumulative study of human walking. The Walking Assist Device features a function to influence the user to achieve efficient walking based on the inverted pendulum model, which is a theory of bipedal walking, and is designed as a device to be used in the training of walking. Honda also developed Walking Assist Device with Bodyweight Support System. The exoskeleton was first introduced in 2008 and since then Honda Motor Co. kept improving it. The Honda walking device was developed for manual workers who stand, walk and stay in a semi-crouching position during the day. The device is aimed to reduce the load on leg muscles and provide natural assistance in many motions and postures.



Fig. 2: Body-weight assist for factory workers

In this article, we focus on the following two aspects that are crucial for adapting the working processes and environments to individuals.

(1) Evaluating the work load by ergonomics to reduce high-load posture work.

(2) Introducing robot technology that physically assists the movement of the worker.

In both cases, we believe that the core technology is to estimate human motion in real time. Generally, it is difficult to estimate a status of human motion with a simple rule-based algorithm due to the diversity of movement. In this project, we applied machine learning to time-series human motion data, measured by motion capture system, in order to estimate the status of human motion by incorporating the characteristics of each person's motion.

2 ANALYSIS OF HUMAN MOTION BY RULE-BASE APPROACH

2.1 Motion Types

In an automobile factory, hundreds to thousands of cars are made a day, so it is a prerequisite that workers are able to handle as many products as possible. In order to assess and understand the typical motions of a factory worker, we executed pilot study to understand movement of worker. We focused on three typical assembling processes in an automobile factory to perform our analysis.

1. Engine assembling process
2. Shaft assembling process
3. Pipe assembling process

We created the same laboratory environment as the selected assembling process and measured the motion of the workers by performing the same operation as the factory workers do in the factory. In the laboratory environment, real equivalent parts were placed in the same location as the factory, and the subjects could use the same equipment as the factory to perform actual work. Four different subjects performed movement sequences consisting of several single actions. These sequences were repeated 20 times a day for 2 days for 3 selected processes. In total, 120 sequence classes are present in the data.

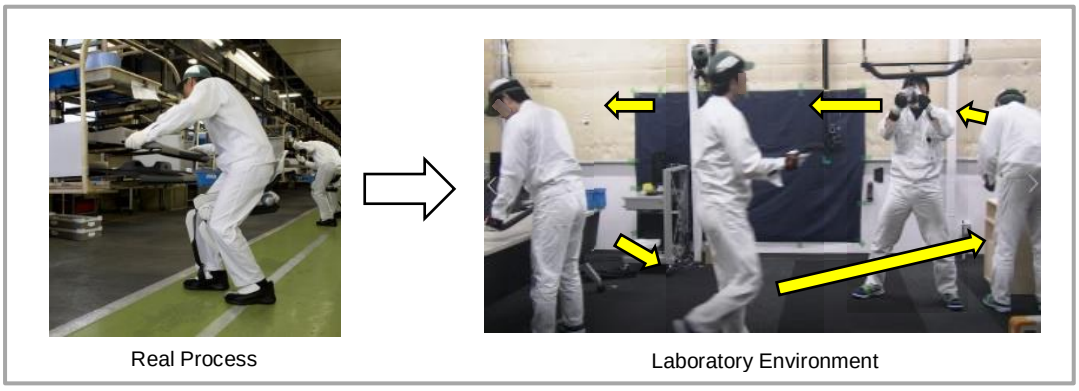
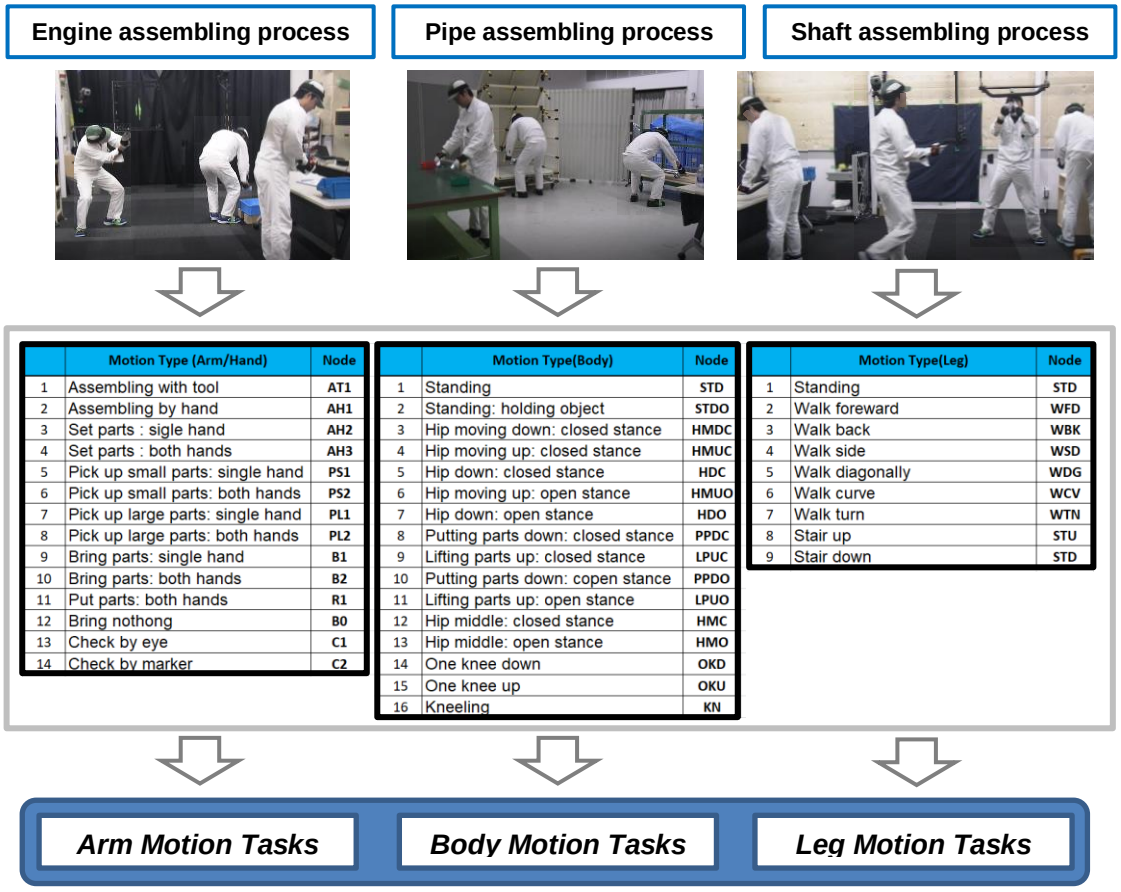


Fig. 3: Virtual environment assuming factory processes

Next, the operation of the three measured assembly processes was analyzed. The assembly work at the factory, selected this time, involves many complex operations using the whole body, such as lifting parts, moving parts, attaching parts, working with own equipment, and checking work after installation, etc. In order to simplify the complicated motion using the whole body, the motion estimation was divided into the motion of the arm, the body, and the leg, and each motion estimation is individually performed. The motion of the whole body was estimated based on a combination of the motion estimation results of each

part. The work process of each step was classified in detail, and the movement was performed on the arm, body, and foot parts. As a result, 14 movement types with regard to arm movements, 11 movement types with regard to body movements and 9 movement types with regard to leg movements were defined in Table 2. As a result of analyzing the measured three process operations, the classified operation type is considered to be the main operation in the selected line process. We manually labeled all the motion data recorded as ground-truth labels.

We consider that lifting and placing the load on the floor increases the burden on the operator in many processes. In Table 1, to focus more on the motion which increases the burden on the operator, we added several motion types such as, squat up, squat down, one knee down, and one knee up with closed leg or open leg, and standing, walking, sitting, etc. Four different subjects performed several movement sequences consisting of several single actions. These sequences were repeated 10-20 times. Altogether, the dataset encodes 2755 actions represented by 329021 single instances.



	Motion Type (Arm/Hand)	Node
1	Assembling with tool	AT1
2	Assembling by hand	AH1
3	Set parts : sigle hand	AH2
4	Set parts : both hands	AH3
5	Pick up small parts: single hand	PS1
6	Pick up small parts: both hands	PS2
7	Pick up large parts: single hand	PL1
8	Pick up large parts: both hands	PL2
9	Bring parts: single hand	B1
10	Bring parts: both hands	B2
11	Put parts: both hands	R1
12	Bring nothong	B0
13	Check by eye	C1
14	Check by marker	C2

	Motion Type(Body)	Node
1	Standing	STD
2	Standing: holding object	STDO
3	Hip moving down: closed stance	HMDC
4	Hip moving up: closed stance	HMUC
5	Hip down: closed stance	HDC
6	Hip moving up: open stance	HMUO
7	Hip down: open stance	HDO
8	Putting parts down: closed stance	PPDC
9	Lifting parts up: closed stance	LPUC
10	Putting parts down: copen stance	PPDO
11	Lifting parts up: open stance	LPUO
12	Hip middle: closed stance	HMC
13	Hip middle: open stance	HMO
14	One knee down	OKD
15	One knee up	OKU
16	Kneeling	KN

	Motion Type(Leg)	Node
1	Standing	STD
2	Walk foreward	WFD
3	Walk back	WBK
4	Walk side	WSD
5	Walk diagonally	WDG
6	Walk curve	WCV
7	Walk turn	WTN
8	Stair up	STU
9	Stair down	STD

Fig. 4: Common motion types categorized from 3 factory processes

Table 1: Defined common motion types from 3 factory processes

	Motion Type (Arm/Hand)	Node
1	Assembling with tool	AT1
2	Assembling by hand	AH1
3	Set parts : sigle hand	AH2
4	Set parts : both hands	AH3
5	Pick up small parts: single hand	PS1
6	Pick up small parts: both hands	PS2
7	Pick up large parts: single hand	PL1
8	Pick up large parts: both hands	PL2
9	Bring parts: single hand	B1
10	Bring parts: both hands	B2
11	Put parts: both hands	R1
12	Bring nothong	B0
13	Check by eye	C1
14	Check by marker	C2

	Motion Type(Body)	Node
1	Standing	STD
2	Standing: holding object	STDO
3	Hip moving down: closed stance	HMDC
4	Hip moving up: closed stance	HMUC
5	Hip down: closed stance	HDC
6	Hip moving up: open stance	HMUO
7	Hip down: open stance	HDO
8	Putting parts down: closed stance	PPDC
9	Lifting parts up: closed stance	LPUC
10	Putting parts down: copen stance	PPDO
11	Lifting parts up: open stance	LPUO
12	Hip middle: closed stance	HMC
13	Hip middle: open stance	HMO
14	One knee down	OKD
15	One knee up	OKU
16	Kneeling	KN

	Motion Type(Leg)	Node
1	Standing	STD
2	Walk foreward	WFD
3	Walk back	WBK
4	Walk side	WSD
5	Walk diagonally	WDG
6	Walk curve	WCV
7	Walk turn	WTN
8	Stair up	STU
9	Stair down	STD

Table 2: 799 dimension feature vector from Xsens

	Sensor Type	Node	DOF	Total Node
1	sensor acceleration	17	3	51
2	sensor angular velocity	17	3	51
3	sensor magnetic field	17	3	51
4	sensor orientation	17	4	68
5	orientation	23	4	92
6	position	23	3	69
7	velocity	23	3	69
8	acceleration	23	3	69
9	angular velocity	23	3	69
10	angular acceleration	23	3	69
11	joint angle	22	3	66
12	joint angle XZY	22	3	66
13	center of mass	3	3	9
Total				**799**

2.2 Motion Measurement

We used Xsens [8][9][10] motion capture system with seventeen IMUs, measuring linear and angular motions with a triad of gyroscopes and accelerometers, distributed on different body locations. The sample rate of 120Hz is sufficient to track even highly dynamic motions as common in sports [11]. We used a sub-sampled rate of 60Hz. The Xsens body suit offers various feature types of sensor information for 17 IMU sensors such as 3DOF gyro sensor information, 3DOF acceleration sensor information and 3DOF magnetic field sensor information. It also offers various feature types of kinematic information for 23

segments of whole body such as 3DOF vector of position, velocity, acceleration and 4DOF vector of orientation by quaternion and for 22 joints of whole body such as joint angle, angular velocity, angular acceleration. Consequently, Xsens system offers 799 dimension feature vector for each time series data (Table 2).

2.3 Analysis of Motion

In this study, we considered that lifting and placing the load on the floor increases the burden on the operator in many processes and we analyzed the lifting and placing operation mainly. In order to reduce the load on the waist and knees during the lifting and placing operation, it is an issue to estimate the timing at which the state transitions from standing to the placing and from lifting to standing. We analyzed 799 feature vectors of Xsens data sets manually by checking graph to understand particular timing at which the state changes. As a result, it was found that, when transitioning from the standing position to the placing operation, there was a particular movement in which the hip moves backward in the horizontal plane earlier than the timing at which the hip begin to fall. Analyzing the measurement data for four subjects, it was similarly verified that the hips moves backward in the horizontal plane earlier than the timing at which the hips begin to go down. This is a motion particular to a human and can be modeled by rule-base algorithm. On the other hand, the timing at which the movement occurs, the amount of movement, and the speed of movement differ depending on the size of the person, age, condition, etc. To identify common rule to capture particular timing of motion for variety of motion types, our rule-base algorithm becomes heuristic consequently. In this point of view, machine learning is effective to respond to variety of heuristic data and extract reasonable result. In this project, we applied rule-base algorithm to acquire common rule and machine learning to estimate personalized human motion situation by incorporating the characteristics of each person's motion (Table 3).

Table 3: Pros and cons of rule-based method in comparison to machine learning approach

	Rule-Base	Machine Learning
Cons	◆Know-how and skill from humanoid robotics ◆Possible to account for dynamics and kinematics ◆Easy to comprehend ◆Easy to debug	◆Trainable ◆Adapts automatically ◆Effective for variety of data (Number of data and type of data)
Pros	◆Heuristic ◆Requires experts and manual inspection ◆Not useful for variety of data	◆Difficult to account for dynamics and kinematics ◆Requires labeled data ◆Requires retraining for new data

3 ANALYSIS OF HUMAN MOTION BY SUPERVISED MACHINE LEARNING

Commonly, online action classification has been tackled with offline machine learning methods. A static model is generated by a large amount of labeled examples and then applied to a hold-out set [12]–[16]. However, we aim for a system that adapts to the personal behavior patterns of its user in real-time and on the fly based on inertial measurements. Ideally, a continuous collaboration between the system and user is triggered where both alternatively adjust to each other and maximize the utility of the system. A static model is simply not suitable in such a scenario. Hence, we apply online machine learning algorithms which continuously adapt to the incoming data stream. Since the data is processed one-by-one, these algorithms are able to handle infinite streams and thereby guarantee a low time and space complexity. Therefore, even limited computational hardware is able to process the stream locally without access to cloud services, allowing an application independent from internet connectivity and offering complete data privacy at the same time. Personalization denotes the modification of a system towards the characteristics of an individual user. Two different modes of personalization have been distinguished [17]: i) Active customization by the user, e.g. by making selections and setting parameters and ii) adaptive systems where the usage history is employed to estimate user preferences and situation statistics to adjust parameters and behavior. The general idea behind personalized learning is that the focus on one person drastically reduces the variance within the data, enabling a better performance with a smaller amount of data. Furthermore, the major problem of inter-person generalization is completely avoided, facilitating the task as well as the computational complexity, since post-processing steps such as normalization or temporal integration can often be omitted. The potential of personalization in the context of motion recognition was mostly analyzed with static models. Weiss and Lockhart investigated personalization in an activity recognition task on the basis of accelerometer-data obtained from smart phones [18]. In their experiments, personalized models were more accurate with a clearly smaller amount of training data. Medrano et al. used personalized models for the task of human fall-detection and achieved a higher performance in comparison to average user models [19]. Our contribution differs from the mentioned work, since we combine personalization with online learning to classify fine-grained motion classes on the basis of a large data foundation obtained with multiple IMUs. We first perform a greedy feature selection to determine a small sub-set of IMUs which are the most valuable ones in terms of classification performance. Subsequently, we thoroughly compare average offline machine learning models with online personalized ones. Our experiments show that online models require a small amount of data to outperform average user systems, particularly if only the raw sensor data is used, often a necessity for applications with strictly limited computational resources. The performance advantage of the online modes increases further with more personal training data.

3.1 Framework

Human movement classification and prediction can be done based on different modalities and different stages of sensor processing. The usage of skeletal data has been established as an efficient representation that can be obtained through machine learning. While most prior art has concentrated on offline action recognition, recently we see some new approaches for fast online action detection based on skeletal data[13][20][21]. We focus on extending prior art especially with respect to personalized models for movement prediction and incremental learning approaches to the problem. Our challenge in this approach is as follows.

1) Based on skeletal 3D motion and training data, to develop an algorithm capable of personalized offline movement prediction which can classify the movement and evaluate the movements.
2) Personalized offline movement prediction, to develop an algorithm capable of personalized online movement prediction for one subject, which can
    - Classify the current movement
    - Predict the next movement
3) To provide a real-time capable implementation of the movement classification and prediction algorithm.

Our focus is the evaluation of offline and online [22] learning models in the online action classification task. The main idea is that even though the classification in retrospective may often still be challenging, there are various tasks where it is certainly easier than doing it online. Therefore, we propose to pre-train an offline model that exploits this advantage and performs the classification in retrospective, yielding a higher performance than a model that performs the classification online [23][24][25]. In the following, we introduce the classification task and describe the characteristics of both learning schemes.

3.2 Offline Learning

Supervised learning requires the input data, in our case sensory data such as the joint positions / velocities / accelerations etc., and the corresponding ground-truth labels. One movement sequence consists of a set of data points (sensory data measured at specific times and additional motion data calculated during the movement) and corresponding ground-truth labels. The ground-truth labels of one movement sequence are equal. Here, the sensory data is recorded and manually labeled. The sensor state vector contains of the raw sensory data such as joint positions/velocities etc. and additionally higher level features which would facilitate the classification and prediction. In the offline learning setting, an algorithm generates a model function h: $R^n \stackrel{\text{def}}{=} \{1, \dots, c\}$ based on a training set $\mathrm{D}_{train} = \{(\mathrm{x}_i, y_i) | i \in \{1, \dots, \mathrm{j}\}\}$. In the subsequent test

phase, the model is applied on another set $\mathrm{D}_{test} = \{(\mathrm{x}_i, y_i) | i \in \{\mathrm{j}+1, \ldots, \mathrm{k}\}\}$, whose label are kept hidden. The model provides a label $\hat{y}_i = \mathrm{h}(\mathrm{x}_i)$ for every point $\mathrm{x}_i \in \mathrm{D}_{test}$ and the 0-1 loss $\mathcal{L}(\hat{y}_i, y_i) = 1(\hat{y}_i \neq y_i)$ is calculated. The test error is as follows.

$$E(\mathrm{D}_{test}) = \frac{1}{k}\sum_{i=j+1}^{k} \mathcal{L}(\hat{y}_i, y_i) \qquad (1)$$

In our case, an offline average user model is generated by using the data of all but one subjects for training and tested to classify the actions of the hold-out subject.

3.3 Online learning

In online learning, each time learning data is input, learning is performed sequentially using only newly input data. A major difference from offline learning is that when learning with new data, instead of recreating the model from scratch, the current model parameters are updated as needed [22]. Formally, a potentially infinite sequence $\mathrm{S}_t = (\mathrm{s}_1, s_2, \ldots, s_t)$ of tuples $\mathrm{S}_i = (\mathrm{x}_i, y_i)$ arrives one after another. In contrast to the offline setting, a model function is generated after each tuple. As t represents the current time stamp, the classification $\hat{y}_i = \mathrm{h}_{t-1}(\mathrm{x}_i)$ is done according to the previously learned model h_{t-1}. After the true label y_t is revealed, the applied learning algorithm generates a new model $\mathrm{h}_t = \mathrm{train}(\mathrm{h}_{t-1}, \mathrm{s}_t)$ on the basis of the current tuple s_t and the previous model h_{t-1}. Usually, the interleaved-test-train error is used for performance evaluation and is defined as follows.

$$\hat{E}(\mathrm{s}_t) = \frac{1}{t}\sum_{i=1}^{t} \mathcal{L}(\hat{y}_i, y_i) \qquad (2)$$

We train the online personalized models iteratively from the scratch [21]. Precisely, each action of one subject is first classified by the model and subsequently used for training. Our goal is to evaluate the potential of online learning schemes for model individualization as just formalized in the context of online action classification based on IMUs. Thereby, we will rely on state-of-the-art online learning schemes as exist in the literature, in particular the Online Random Forest algorithm.

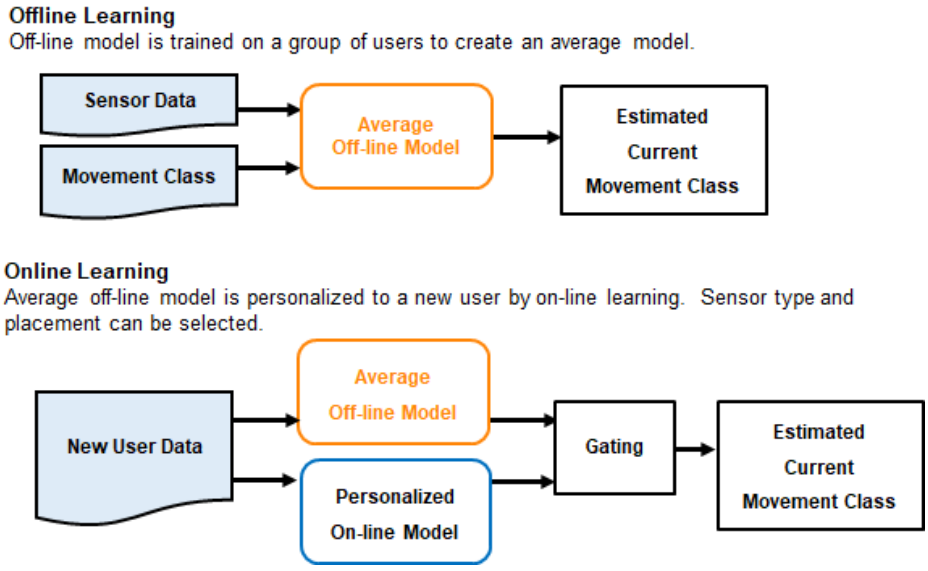
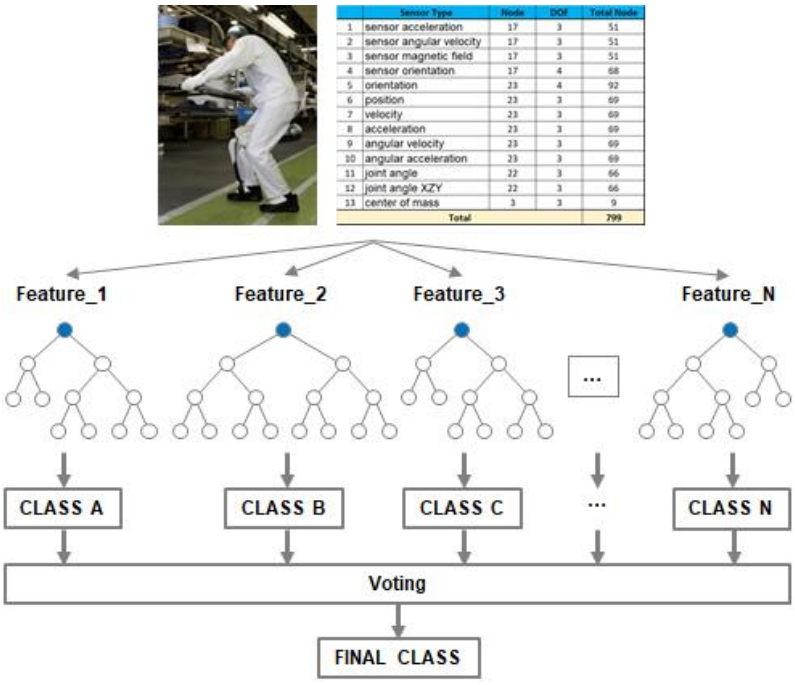


Fig. 5: Online and offline learning system



	Sensor Type	Node	DOF	Total Node
1	sensor acceleration	17	3	51
2	sensor angular velocity	17	3	51
3	sensor magnetic field	17	3	51
4	sensor orientation	17	4	68
5	orientation	23	4	92
6	position	23	3	69
7	velocity	23	3	69
8	acceleration	23	3	69
9	angular velocity	23	3	69
10	angular acceleration	23	3	69
11	joint angle	22	3	66
12	joint angle XZY	22	3	66
13	center of mass	3	3	9
	Total			799

Fig. 6: Classifier by Random Forest

3.4 Classifier

Random Forest [26][27] is used for offline and online learning. Random Forest uses a method called bootstrap sampling to select data. This is a method that allows sampling from one sample set with duplication and creates a new sample set. Allowing duplication from m data and taking m as learning data. A decision tree is constructed for each of a large number of training data sets obtained by repeatedly performing such sampling. Generally, single decision trees are known for overfitting. The average of the Random Forest is a model which is very robust against overfitting since it uses randomness and is based on an ensemble of different predictors. The Random Forest is designed with the goal to reduce its sensitivity to overfitting. This is one of the main reasons it uses a randomized ensemble of trees.

3.5 System Design

The main idea of our approach is based on the fact that even though classifying a sequence in retrospective, in our case a human body motion may still be difficult, it is often substantially easier than classifying it during the execution. Therefore, it can be expected that the retrospective model achieves a higher performance. We leverage this performance gap to adapt the personalized model during run-time. Concretely, we propose to construct a retrospective ground truth model that classifies the actions with delay. The model classifies the action of time t with a delay of m time units at time t + m:

$$\hat{y}_t = \underset{y_t \in \{1,\dots C\}}{\operatorname{argmax}} (y_t | (\mathrm{x}_{t-l-1}, \dots, \mathrm{x}_{t+m})) \qquad (3)$$

The delay allows the model to leverage features after time t which facilitates the task yielding a higher performance. In the case of action classification, this is quite intuitive as it is easier to classify the current motion sequence if information is given about how the motion progresses after the query time t. The retrospective model is an average user offline model which is trained in advance for the specific potentially utilizing a large datasets covering many different users. During the application, this model is applied to feed its classification outputs as pseudo-ground truth into the personalized online model for adaptation. The personalized model is performing online classification for the time t, meaning it classifies the current immediately without any delay. However, it is adapted at least with a delay of m time units, such that at time t+m it can use the tuple $(\mathrm{x}_t, \hat{y}_t)$ for adaptation. Figure 7 gives an overview of the overall architecture.

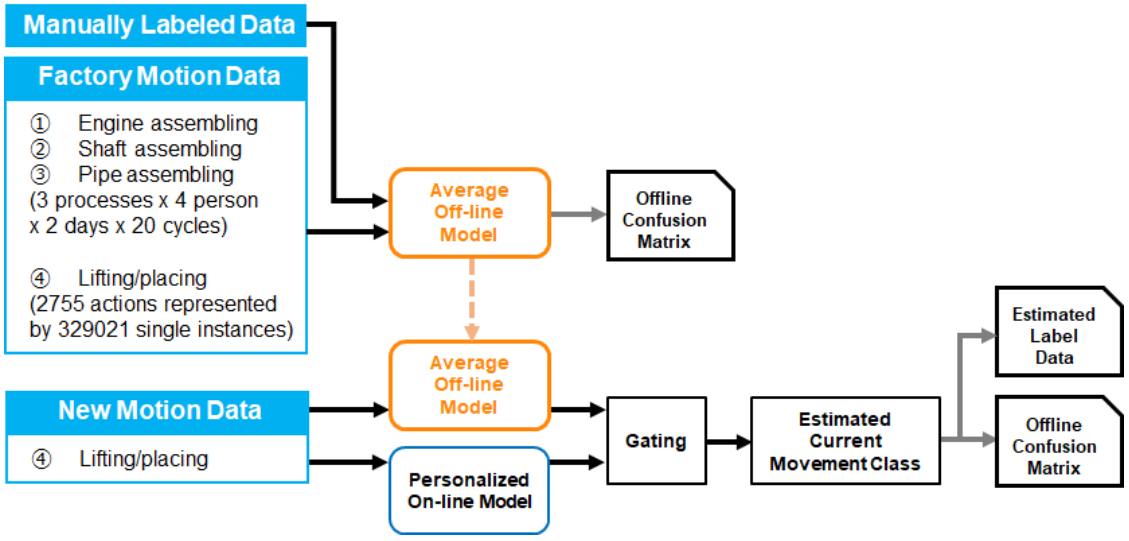


Fig. 7: Online motion estimation system

3.6 Learning schemes

The evaluation was performed with three different learning schemes. The comparison of these allows a detailed analysis regarding various aspects such as inter- and intra-subject generalization. The partitioning of the data among training and testing data is done on motion level, meaning all instances of a single motions are either all in the training or in the testing data.

(A) kFold cross-validation (kFold)

The complete data is divided in k subparts. Each time k-1 subparts are used for training and one for testing. This is repeated k times until each subpart is used for testing. This evaluation allows to estimate how well the motion classes can be discriminated.

(B) 5.3.2 Leave-one-subject-out (LOSO)

This scheme is performed to analyze the generalization across different subjects. All data of one subject is used for testing whereas the data of the other subjects is used for training. This is repeated 4 times until each subject-specific data is used for testing.

(C) Online personalized (PERS)

The data of one subject is used in an online fashion for testing and training. This is repeated for all subjects. Each personalized model is trained exclusively with the data of one subject. The

capabilities of personalization and incremental learning are investigated using this learning scheme.

Using all learning schemes we evaluate the resulting models in terms of average 0-1 accuracy. Table 4 shows the results. The high performance of the kFold evaluation shows that the motion classes can be discriminate with a high accuracy. The average accuracy of LOSO and PERS are similar, however, the average accuracy of PERS measured in the second half of the data indicates that it is clearly better after it has seen enough training data. With additional data it is expected to converge against the performance of kFold. The performance gap between the kFold and LOSO shows that the data of each subject is quite different. However, it is not clear whether this is exclusively due to personalized motions or rather specific characteristics during each recording, since the data of each subject was recorded in one session. The performance is significantly increased when the input contains not only the data of the current instance but those of the last 30 time steps (corresponds to the data obtained in 0.5 seconds). The disadvantage is the drastically increased number of dimensions. However, instead of the plain input data, statistical features such as means, extremes or discrete Fourier or Cosine coefficients could be used to reduce the number of dimensions again while maintaining the performance advantage.

Table 4: All learning schemes and result

number of past instances	kFold	LOSO	PERS	PERS 2. half
1	**94.61**	83.31	84.87	91.61
30	**94.80**	85.83	86.98	94.01

Figure 8 shows the learning curves of the models. Since personalized models have only access to the data of one subject they get less overall training data. The accuracy of the personalized model is depicted as the average accuracy (cumulative) as well as the current accuracy. The estimation of the current accuracy is simply an averaging over the last 1000 instances. In general, the average accuracy is lower, since it includes the initial learning phase in which the model has a low performance because it has seen very few data. All models are quickly reaching a high accuracy. Particularly, the personalized model would profit from additional data, since it seems to be not converged, yet.

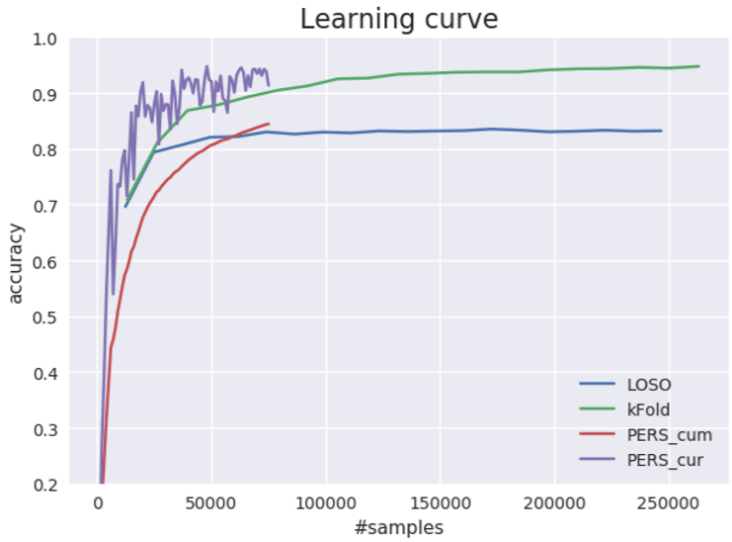


Figure 8: The learning curves of the different learning schemes. The accuracy of the personalized model is depicted as the average accuracy (cumulative) as well as the current accuracy, averaging the accuracy over the last 1000 instances.

3.7 Result of Motion Classification

We compare an offline average user model against an online personalized user model. The average model is trained in leave-one-subject-out scheme. Precisely, it is tested with the data of one specific subject, whereas data of the remaining three subjects is used for training. This is done repeatedly such that each subject is used for testing once. Personalized models are evaluated in the online learning setting. The model classifies the label of one sample and uses it afterward for model adaption. This is done for all samples in the dataset. However, the order of each instance within one performed action is predefined by the recording time, and therefore, there is a high degree of label auto-correlation, since each action consists of a number of samples with the same class. In this case, the ordinary online scheme is misleading because a naive classifier, simply using the previously seen label for classification, achieves a very low error rate without learning anything. Therefore, we perform an action-wise evaluation. Precisely, the model classifies all samples of one action, before the corresponding labels get revealed. The personalized models are trained from scratch, one online model for every subject in single pass. Consequently, online models are utilized without any form of pre-training and only access the data of one subject. We calculate both errors (off- and online) using the same data for testing, but the online algorithms continuously adapt their model to the test subject.

Confusion matrix of offline learning is shown in Fig. 9 and result of online learning is shown in Fig. 10. In statistical classification, typically a supervised learning, to evaluate the accuracy of classification, it is common to create a confusion matrix and compare the number of correctly identified and the number of incorrectly identified ones. In the confusion matrix, each column of the confusion matrix represents the instances in an actual class and each row of the confusion matrix represents the instances in a predicted class. The diagonal elements of the confusion matrix represent the accuracy of the actual class (generated motion type input) as estimated by the classifier, and the other elements represent the error rate. Accuracy of offline classification (Fig. 9) was 88%-97% and online (Fig. 10) was 72%-93%. As expected, the accuracy of online learning was lower than offline learning. A major difference from offline learning is that when learning with new data, instead of recreating the model from scratch, the current model parameters are updated as needed. On the other hand, online learning is susceptible to noise such as outliers and susceptible to the latest data.

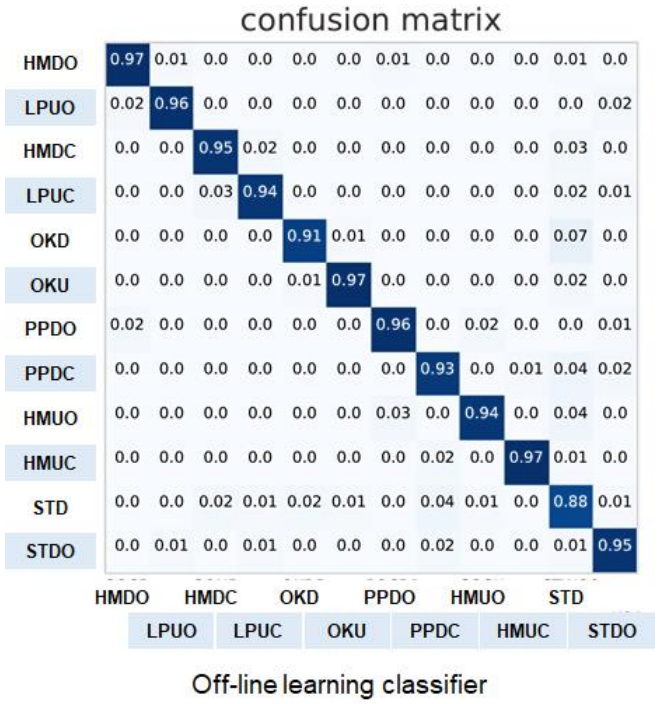


Fig. 9: Confusion matrix of offline motion classification

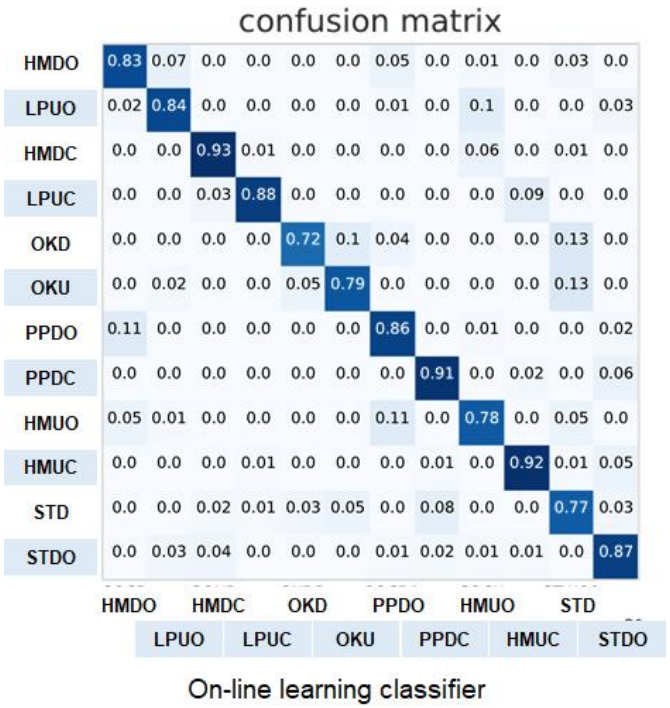


Fig. 10: Confusion matrix of online motion classification

3.8 Feature and Segment Selection

The Xsens body suit offers various feature types for each segment. To reduce the number of input dimensions, we performed for each model feature selection, determining the most valuable dimensions in two steps. We use forward feature selection to minimize the classification error. Concretely, we start with an empty set and iteratively add the feature type / segment which minimize the error the most until all of them are used. In the first step, we extract the most valuable feature types, thereby using all segments and in the second we determine the most valuable segments using only the previously extracted feature types. The normalized feature types were not offered to the personalized model, since they have no effect on person-specific and scale-invariant models. Figure 11 and Figure 12 shows the error rate depending on the number of feature types / segments. It can be seen that only few feature types and segments are necessary to reach a high performance. Using all feature types is even harmful which is probably caused by over fitting. The personalized model relies on less data and is therefore more affected. Based on these results, we decide to use the three best feature types together with the five most valuable segments. The chosen types by the average model are mostly temporally integrated and normalized, highlighting that the person-specific normalization improves the inter-person generalization

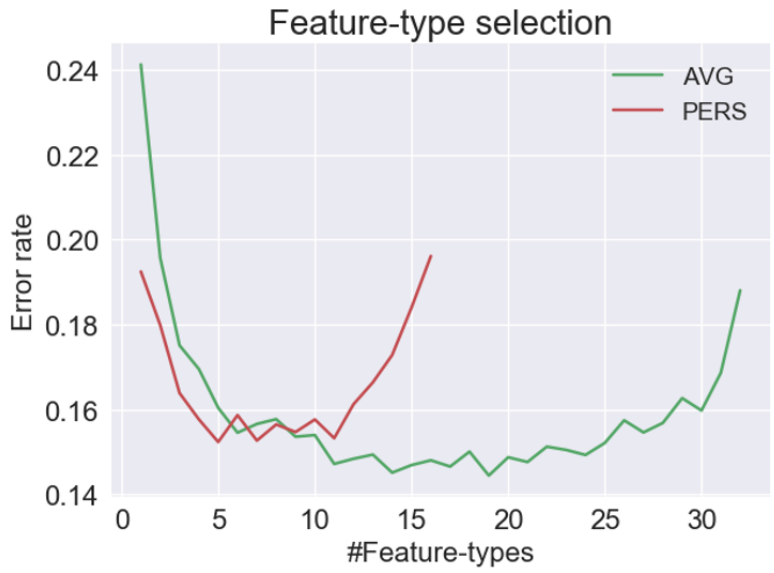


Fig. 11: The error rate depending on the number of feature types. It is denoted that average is AVG and personalized is PERS.

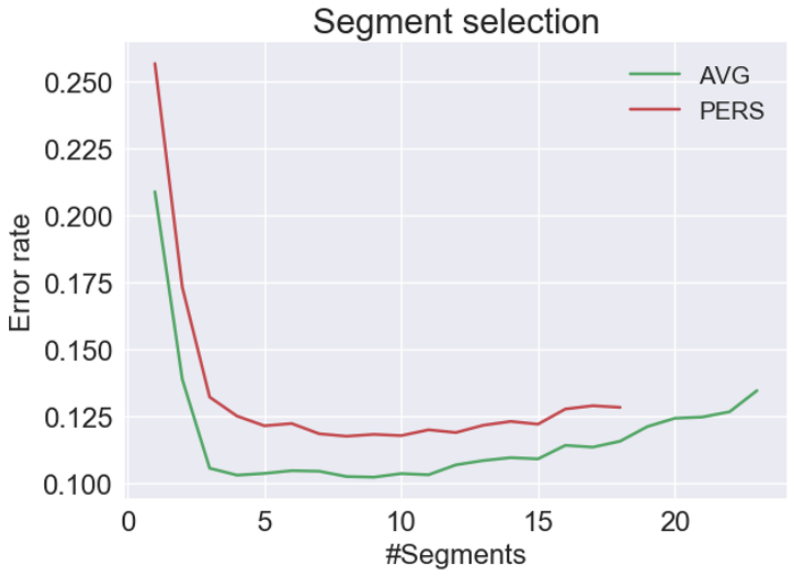


Fig. 12: The error rate depending on the number of segments.

3.9 Evaluation in Ergonomics

Ergonomics is the study of designing and researching from an engineering approach so that human can use all kinds of hardware and software comfortably and with less fatigue. At the factories, ergonomics is used to create an environment where workers can work efficiently using natural movements. If workers can work more comfortably, matching human natural movements, they should be more efficient and work more rewarding (Table 5). The philosophy of ergonomics in industry is not to "match people to work" but to "match work to people". The goal of ergonomic evaluation is to "adapt people to equipment, devices, work environments, and organizations." In this research, we have constructed a system to identify the work movements of factory workers using a rule-based approach and a machine learning approach. When a certain number of repetitive tasks can be measured, by applying the ergonomic evaluation index, it is possible to determine whether or not the worker can work efficiently with reasonable natural movement. An ergonomic index defined in the factory was applied to the engine assembly process among the three measured operations. Fig. 13 shows a three-dimensional visualization of the ergonomic evaluation during the execution of an arbitrary operation. Fig. 14 shows a visualization of the evaluation of the entire work operation after the work is completed. From these display results, it can be seen that even a work manager who has no knowledge of the ergonomics evaluation can easily determine whether or not the worker can work efficiently with reasonable natural movements.

Table 5: Purpose of Factory Ergonomics

	Topic	Purpose
1	**Safety**	To design efficient environment where workers can work safely
		To reduce physical load
		To design safe physical work
2	**Keeping up with workforce**	Effective use of elderly workers
		Effective use of the female workforce
		Effective use of the young workforce
3	**Management of workforce**	Efficient line design
		Efficient staffing
		Reduce employee health care costs

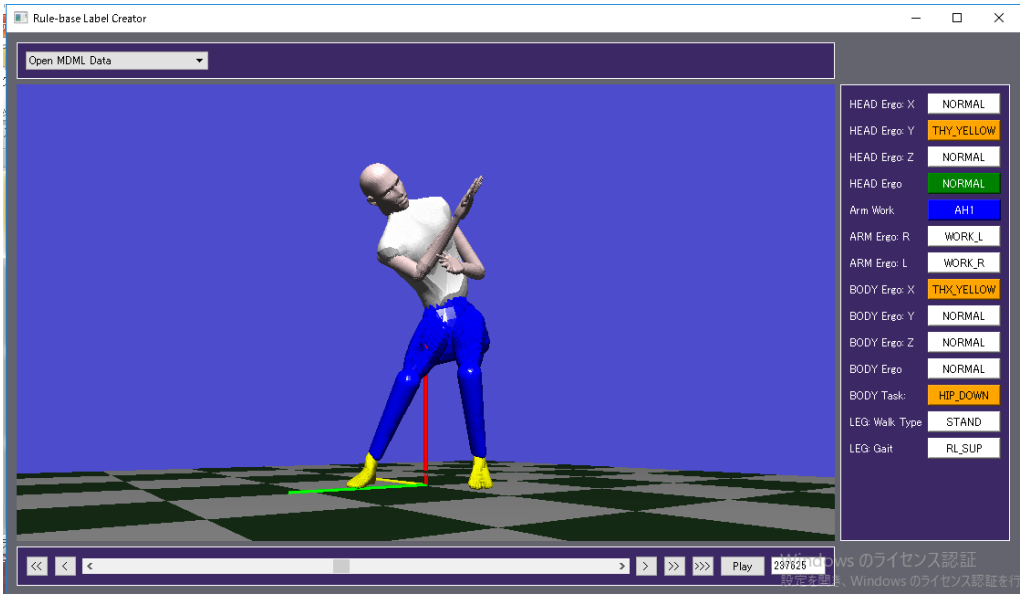


Fig. 13: A three-dimensional visualization of the ergonomic evaluation during the execution of an arbitrary operation. Right boxes show evaluated posture according to index of ergonomics. By moving scroll bar, evaluated posture is displayed in the boxes.

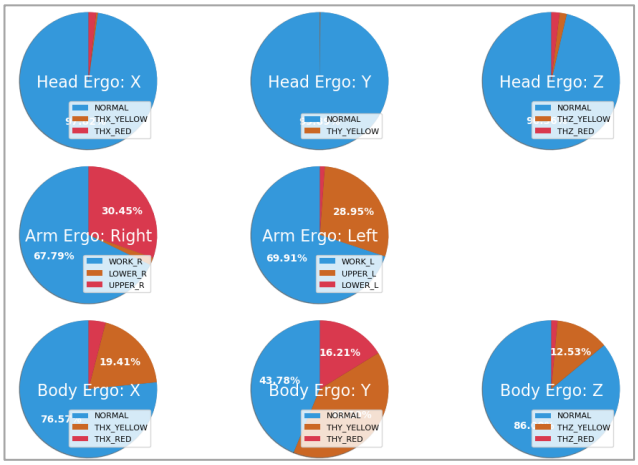


Fig. 14: A visualization of the evaluation of the entire work operation after the work is completed. Each circle shows ratio of good posture, warning posture and bad posture for each joints in XYZ axis.

4 ANALYSIS OF HUMAN MOTION BY DEEP LEARNING

In the previous chapter, we constructed a technology to identify motion types of factory worker by supervised learning (using Online Random Forest). The objective of this research was to develop the decision-making capability in assistive robots to execute a task with human intuition in dynamic situations. We used machine learning algorithms CNN and k-NN for real-time material identification. Here, we assumed a work movement of the upper body of the factory worker, focused on the movement of grasping an object by hand and the movement of picking up with tools, and constructed a technology to estimate the work target from human movement data. We selected the human activity of scooping different materials because it involves physiological manipulation skills [30] that are natural to humans but extremely difficult for robots to learn. We introduce capability for robot technology that physically assist the movement of the worker using a convolutional neural network (CNN).

4.1 Human Musculoskeletal Model

The human arm is a highly complex system. Thus, to understand the motion of scooping, we observed the Right Arm Flexion, Right Arm Adduction, Right Arm Rotation, Right Elbow Flexion, Right Wrist Flexion, Right Wrist Deviation, Right Wrist Pronation and Supination joint angles of the human arm (Fig. 15). We performed human motion capture to identify the natural human posture while executing the scooping activity. Human motion capture was performed using Optitrack system [35]. This system requires the human subject to place passive motion-capture markers on the body-part whose motion is under concern. In our study, subject's dominant arm which performed the scooping action was tracked. The markers were placed on the right arm of our subject such that the resulting motion of the arm would produce smooth and continuous trajectories in the captured data. We obtained motion data by placing the markers on all fingers, back of the hand, wrist, forearm, elbow, biceps, triceps and the shoulder. We also monitored the biceps and ECRL muscles which are primarily responsible for the action of scooping. The data files are then imported in Opensim and inverse kinematics is applied on the data to convert the motion data into joint-angles of the subject's arm. We used OpenSim's upper extremity kinematic model [31], which includes 15 DOFs, representing the shoulder, elbow, forearm, wrist, thumb, and index finger, and 50 muscle compartments crossing these joints. The scooping motion was reconstructed by placing virtual markers on the model exactly as they were positioned on the human subject. Alongside motion capture data, human subject's wrist linear acceleration data was also captured during the experiment and used for robot motion planning.

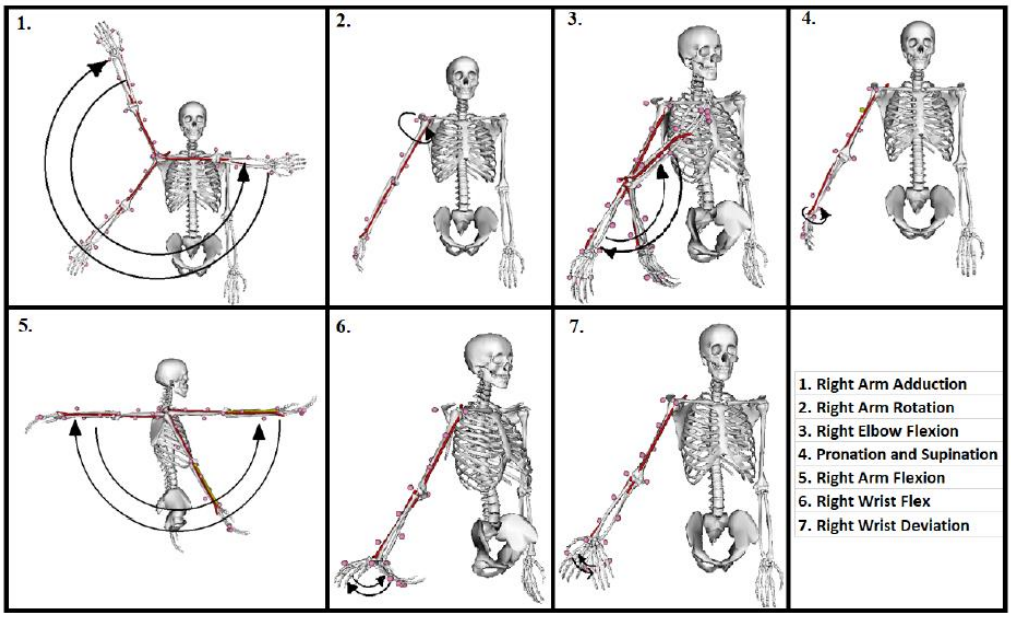


Fig. 15: Degrees of Freedom of Human Right Arm Considered for the Experiment

4.2 Experimental Setup

The experiment was performed for 8 different materials-cereal, honey, rock, pebble, sand, soil, stone, and water. We covered materials that are significantly different in appearance, nature, and force feedback upon being scooped. The Optitrack motion tracking system with five Prime-41 cameras was used for this experiment. Each camera had 170 LEDs and 4.1MP resolution. For tracking the motion, a 120Hz sampling frequency was used. From the captured marker-data, we reconstructed the human motion in real-time through consistent tracking of marker trajectories [32]. In addition to motion tracking, we also collected data on linear acceleration of the wrist and EMG signal activation in the biceps and ECRL using the EEGO [33] system. EMG data were post processed using high-pass filtering at 20 Hz to remove motion artifacts, full-wave rectifying, and then low-pass filtering at 5 Hz to generate a linear envelope. These data were then normalized to a maximum isometric muscle contraction. Out of the ten trials for each material, the one with least muscular effort [34] was selected as the optimal posture, and was investigated further. To perform tasks in real-time, probabilistic approach is considered where there is need for fast algorithms for motion planning of a high DOF robot. For this research, a Kinova JACO robotic arm with 6 DOFs was used. The robot was equipped with a generic USB camera for computer vision ability. The executable code to manipulate and control the robot was developed with the API exposed by Jaco SDK and related libraries.

4.3 Training

The very first stage of the image identification module pipeline is the convolutional neural network, which represents the input (image of the material to be scooped) in the program in an understandable format. For image classification we have selected to use transfer learning to extract image features from pre -trained convolutional neural network and build classifier on small sample of the training data. Transfer learning is a technique in which the already trained model is used for solving similar problems [36]. For representing the material images in the proposed approach, we used the transfer learning method [37] and a pre-trained neural network model with VGG architecture [38]. With transfer learning, the model can be extended to a greater number of materials with very small sample for each material to train the model. We have collected enough data to make the background information like bowl colors invariant. Using transfer learning, the model can understand more features other than visual representation which could be difficult to identify when features are hand crafted. We used the Visual Geometry Group (VGG) convolutional neural network model [33, 34]. The model achieved 92.7% top-5 test accuracy in ImageNet [35, 36], which is a dataset of over 14 million images belonging to 1000 classes. The last layer off the VGG16 network is SoftMax function, which classifies the input in one of the 1000 ImageNet classes depending upon the 4096 features learned by the previous layer (fully connected + ReLu layer output). For this research, the bottleneck features were captured after the fully connected + ReLu layer and the SoftMax function was replaced by the model trained using k-NN algorithm [43]. For each material, 15 images were taken by the USB camera and resized to an image of size 224x224x3. The feature vector representation from VGG network was used to build the k-NN classifier using ‘knn’ package of R programming language for value of k = 3.

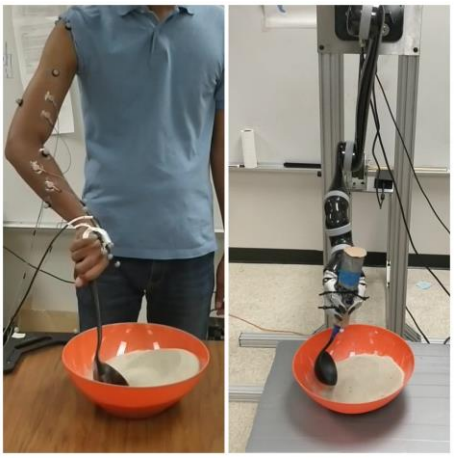


Fig. 16: Initial and Final Optimal Posture (i.e Minimum Muscular Effort) of the Human Subject and the Robot)

4.4 Results

The objective of the research was to develop the decision-making capability in assistive robots to execute a task with human intuition in dynamic situations. We used machine learning algorithms CNN and k-NN for real-time material identification. We achieved 100% image classification accuracy for material image classification algorithm for k = 3. The performance of image classification algorithm for different values of k is shown in the Table 6.

Table 6: k-NN Model Performance for Different Values of k

K-value	Classification Accuracy
1	78%
2	82%
3	100%
4	87%
5	92%
6	88%

We achieved 100% image classification accuracy for material image classification algorithm for k = 3. The performance of image classification algorithm for different values of k is shown in the Table II. By performing human motion analysis using musculoskeletal modeling, we observed how every joint angle changed according to the nature of the material. The joint angles of human arm for two extremely different materials, water and rock, can be observed in Fig. 17. The robot was programmed and could execute the scooping action for each material with high efficiency and was comparatively efficient than human subject in terms of execution time. The scooping execution time for the human and the robot was measured manually and is compared in the graph represented by Fig. 18. The success ratio while examining the scooping motion over all trials was found to be 85%.

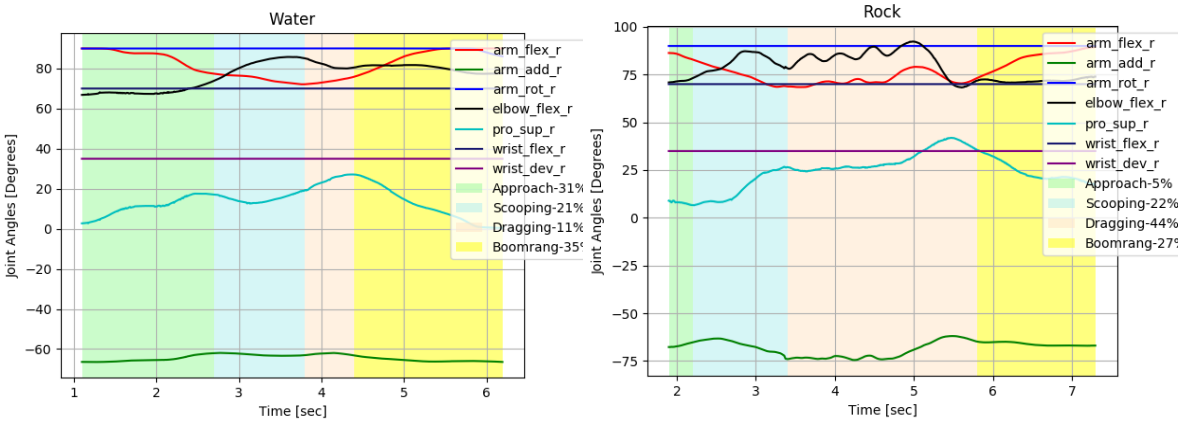


Fig. 17: Joint Angles in Different Motion Primitives for Scooping Water and Rock

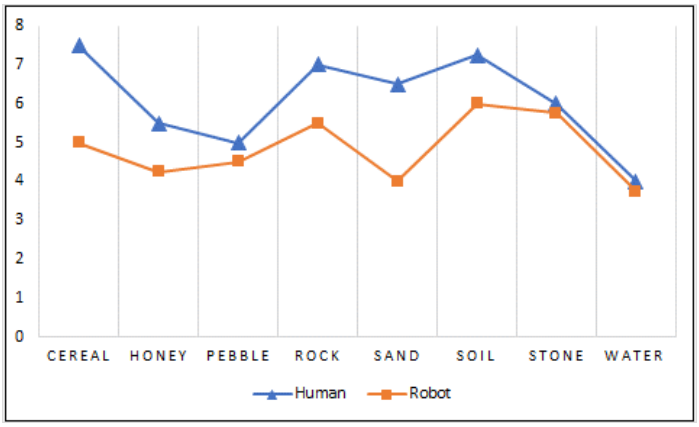


Fig. 18: Scooping Activity Time (in Seconds) by Human Subject and Trained Robot.

VI. Conclusion

1) It is expected that the introduction of assisting devices constitutes a promising way to augment the diverse capabilities of workers according to the demands of a given task and secure a stable future labor force. In this paper, we focused on the following two aspects that are crucial for adapting the working processes and environments to individuals.

   (1) Evaluating the work load by ergonomics to reduce high-load posture work.

   (2) Introducing robot technology that physically assists the movement of the worker.

   To respond to diverse capabilities of workers, we applied machine learning to time-series human motion data, measured by motion capture system, in order to estimate the status of human motion by incorporating the characteristics of each person's motion.

2) In valuating the work load by ergonomics, we selected (i) engine assembling process, (ii) shaft assembling process and (iii) pipe assembling process, created the same laboratory environment and measured the motion data of the workers by performing the same operation as the factory workers.

3) We proposed to pre-train an offline model that exploits this advantage and performs the classification in retrospective, yielding a higher performance than a model that performs the classification online. Using state-of-the-art machine learning algorithms, we compared personalized online models against average-user offline models. Accuracy of off line classification was 88%-97% and online was 72%-93%. We performed a forward feature selection, which showed that only a few IMUs are necessary

to obtain a high performance. Results were visualized to see the result of worker's motion and evaluate the motion in ergonomics.

4) Introducing robot technology, we recorded the dynamics of human motion for different situations by using techniques of motion capture, EMG signal monitoring and musculoskeletal modeling. We investigated the motion by simulating a human musculoskeletal model and deriving the essential information for robot motion planning. The robot was programmed to perform the desired tasks by using an angular control method, which provides more accuracy. Moreover, we provided computer vision capability to the robot, which enabled it to decide which program to execute based on the materials that it identified. As a result of these innovations, the resulting robotic motion was highly improved as the robot had human intuition (computer vision) and skills (learned from human motion).

VII. FUTURE WORK

A. Extension of typical motions of a factory worker

In this paper, we focused on evaluating the work load by ergonomics to reduce high-load posture work. In order to assess and understand the typical motions of a factory worker, we executed pilot study to understand movement of worker by selecting (i) engine assembling process, (ii) shaft assembling process and (iii) pipe assembling process. In an automobile factory, hundreds to thousands of cars are made a day through hundreds of different assembling processes. We extend the framework of motion estimation by machine learning to other processes. Then, we will consider whether to apply the constructed framework of motion estimation to individual processes or generalize it as a common framework for motion estimation. We suppose to apply this system to factory system to reduce high-load posture work and manage work force.

B. Implementation in online system

It is expected that the introduction of assisting devices constitutes a promising way to augment the diverse capabilities of workers according to the demands of a given task and secure a stable future labor force. In this paper, our focus was the evaluation of offline and online learning models in the online action classification task. We proposed to pre-train an offline model that exploits this advantage and performs the classification in retrospective, yielding a higher performance than a model that performs the classification online. We have constructed a system to identify the work movements of factory workers using a rule-based approach and a machine learning approach. Our

next goal is to run this ergonomics evaluation system on an online system and to establish a physical assist system for estimating motion while acquiring motion data.

REFERENCES

[1] https://opensim.stanford.edu/

[2] K. Kong and D. Jeon, "Design and control of an exoskeleton for the elderly and patients," IEEE/ASME Transactions on mechatronics, vol. 11, no. 4, pp. 428–432, 2006.

[3] J. E. Pratt, B. T. Krupp, C. J. Morse, and S. H. Collins, "The roboknee: an exoskeleton for enhancing strength and endurance during walking," in IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, vol. 3, April 2004, pp. 2430–2435 Vol.3.

[4] Y. Muramatsu, H. Kobayashi, Y. Sato, H. Jiaou, and T. Hashimoto, "Quantitative performance analysis of exoskeleton augmenting devices - muscle suit - for manual worker," International Journal of Automation Technology, vol. 5, pp. 559–567, 07 2011.

[5] https://www.cyberdyne.jp/

[6] http://atoun.co.jp/

[7] Honda. [Online]. Available: http://asimo.honda.com/innovations/ feature/body-weight-support-assist/

[8] D. Roetenberg, H. Luinge, and P. Slycke, "Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors," vol. 3, 01 2009.

[9] Xsens. [Online]. Available: https://www.xsens.com/research/ published-papers/

[10] R. G. J. Damgrave and D. Lutters, "The drift of the xsens moven motion capturing suit during common movements in a working environment, "in Proceedings of the 19th CIRP Design Conference– Competitive Design. Cranfield University Press, 2009.

[11] B. Tessendorf, F. Gravenhorst, B. Arnrich, and G. Trster, "An imu-based sensor network to continuously monitor rowing technique on the water," in 2011 Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Dec 2011, pp. 253–258.

[12] G. Garcia-Hernando and T.-K. Kim, "Transition forests: Learning discriminativetemporaltransitionsforactionrecognitionanddetection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 407–415, 2017.

[13] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 203–220.

[14] A. Gaidon, Z. Harchaoui, and C. Schmid, "Actom sequence models for efficient action detection," in CVPR 2011, June 2011, pp. 3201–3208.

[15] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in 2013 IEEE International Conference on Computer Vision, Dec 2013, pp. 3551–3558.
[16] M. Jain, J. v. Gemert, H. Jgou, P. Bouthemy, and C. G. M. Snoek, “Action localization with tubelets from motion,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 2014, pp. 740–747
[17] G. Fischer, “User modeling in human-computer interaction,” User Modeling and User-Adapted Interaction, vol. 11, no. 1-2, pp. 65–86, 2001.
[18] G. M. Weiss and J. W. Lockhart, “The impact of personalization on smartphone-based activity recognition,” in AAAI Workshop on Activity Context Representation: Techniques and Languages, 2012, pp. 98–104.
[19] C. Medrano, I. Plaza, R. Igual, ´A. S´anchez, and M. Castro, “The effect of personalization on smartphone-based fall detectors,” Sensors, vol. 16, no. 1, p. 117, 2016.
[20] G. Garcia-Hernando, T-K. Kim, Transition Forests: Learning Discriminative Temporal Transitions for Action Recognition and Detection, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2017.
[21] M. Hasenjger and H. Wersing, “Personalization in advanced driver assistance systems and autonomous vehicles: A review,” in 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Oct 2017, pp. 1–7.
[22] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, “Online human action detection using joint classification-regression recurrent neural networks,” in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 203–220.
[23] V. Losing, B. Hammer, and H. Wersing, “KNN Classifier with Self Adjusting Memory for Heterogeneous Concept Drift,”in16thInternational Conference on Data Mining (ICDM). IEEE, 2016.
[24] V. Losing, B. Hammer, and H. Wersing, “Personalized maneuver prediction at intersections,” in 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Oct 2017, pp. 1–6.
[25] V. Losing, B. Hammer, and H. Wersing, “Choosing the Best Algorithm for an Incremental Learning Task,” in European Symposium on Artificial Neural Networks (ESANN), 2016.
[26] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof, “On-line random forests,” in Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 1393– 1400.
[27] U. M. Fayyad and K. B. Irani, “The attribute selection problem in decision tree generation,” in AAAI, 1992, pp. 104–110.

[28] V. Losing, M. Hasenjger, T. Yoshikawa, B. Hammer and H. Wersing, “Personalized Online Learning of Whole Body Motions using Multiple Inertial Measurement Units,” in (ICRA), 2019.
[29] T. Yoshikawa, “Machine Learning for Human Movement Understanding” in (IROS), 2019., Cutting Edge Forum.
[30] E. Demircan, T. Besier, S. Menon, and O. Khatib, “Human Motion Reconstruction and Synthesis of Human Skills,” in Advances in Robot Kinematics, J. Lenarcic and M.M. Stanisic, Eds., Berlin, Heidelberg, Germany: Springer, pp. 283-292, 2010.
[31] K.R. Holzbaur, W.M. Murray, and S.L. Delp, ”A Model of the Upper Extremity for Simulating Musculoskeletal Surgery and Analyzing Neuromuscular Control.” Annals of Biomedical Engineering, 2005
[32] E. Demircan, L. Sentis, V. De Sapio, and O. Khatib, “Human motion reconstruction by direct control of marker trajectories,” in Advances in Robot Kinematics, J. Lenarcic,
[33] EEGO system: https://www.ant-neuro.com/products/eego sports
[34] O. Khatib, E. Demircan, V. De Sapio, L. Sentis, T. Besier, S. Delp, Robotics-based synthesis of human motion, Journal of Physiology Paris, vol. 103, pp. 211-219, 2009
[35] Optitrack system: https://optitrack.com/
[36] S. J. Pan, Q. Yang,’A Survey on Transfer Learning’, in IEEE Trans. Knowl. Data Eng. 2010, 22, 1345-1359.
[37] M. Oquab, L. Bottou, I. Laptev, J. Sivic,’ Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks’, in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1717-1724.
[38] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: A system for largescale machine learning,” Google Brain. 12th USENIX Symposium on OSDI., 2016.
[39] K.Simonyan, A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition.” arXiv:1409.1556v6 [cs.CV] 10 Apr 2015.
[40] Caffe Model available at: https://gist.github.com/ksimonyan/211839e770f7b538e2d8#file-readme-md
[41] Imagenet: http://image-net.org/
[42] A. Krizhevsky, I. Sutskever and G. Hinton. ImageNet classification with deep convolutional neural networks.In Advances in Neural Information Processing Systems (NIPS), 2012.
[43] K-nearest neighbor available at: https://en.wikipedia.org/wiki/K-nearest neighbors algorithm