Measuring motor actions and psychophysiology for task difficulty estimation in human-robot interaction

Domen Novak, Matjaž Mihelj, Jaka Ziherl, Andrej Olenšek and Marko Munih
University of Ljubljana, Faculty of Electrical Engineering
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
domen.novak@robo.fe.uni-lj.si
+386 41 969 753 (mobile), +386 1 4768 196 (office)

ABSTRACT
In this paper, we describe a method for estimating task difficulty in human-robot interaction using a combination of motor actions and psychophysiology. A number of variables are calculated from kinematics, dynamics, heart rate, skin conductance, respiration and skin temperature. Discriminant analysis of the variables is used to determine whether the user finds the task too easy or too hard. The discriminant function is recursively updated with Kalman filtering in order to better adapt to the current user. The method was tested offline in a task with 20 subjects. In cross-validation, nonadaptive discriminant analysis yielded a classification accuracy of 80.2 % while adaptive discriminant analysis yielded a classification accuracy of 84.3 %.

Author Keywords
Human-robot interaction, virtual reality, sensor fusion, psychophysiology, human factors.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. Only enter keywords in this section if you wish your extended abstract to be included in the ACM Digital Library.

INTRODUCTION
In the last few decades, robots have been developed that interact with humans in many different environments. For instance, humanoid robots provide entertainment and companionship while haptic robots assist patients in motor rehabilitation. However, studies have shown that it is difficult to fully evaluate and interpret the interaction between a robot and a human [1]. While most robots are equipped with sensitive force and position sensors that can measure the user's motor actions, such sensors cannot reveal information about the user's subjective feelings: stress, engagement etc. Such information could be obtained through the use of so-called psychophysiological measurements. Defined as measurements of physiological responses to psychological stimuli, these have been extensively used for user state estimation in various situations. For instance, users' emotional responses to computer games are reflected in their heart rate, skin conductance and skin temperature [5].

Once motor actions and psychophysiology are measured, they can be used in a feedback loop: while the user interacts with the robot in order to perform a task, the task difficulty is adjusted in order to avoid frustration (task is too hard) or boredom (task is too easy). The principle of such a feedback loop is illustrated in Figure 1.

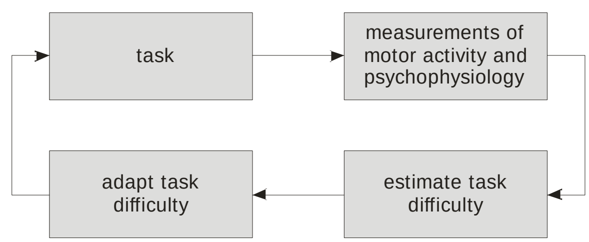

Figure 1: Basic principle of task difficulty adaptation.

Such a feedback loop was successfully created using only psychophysiological measurements in a flight simulator [2]. However, in that system, task difficulty was estimated somewhat arbitrarily: if physiological signals exceeded a manually set threshold, the task was considered too hard. For practical use, a more advanced method of task difficulty estimation is required. One possible method is discriminant analysis, a statistical method for classification of multidimensional data into two or more classes. Two classes ('too easy' and 'too hard') should be sufficient for basic task difficulty estimation.

Since psychophysiological responses exhibit large inter-individual differences, an optimal task difficulty estimation method would be able to adapt to a user as the task progresses and the system obtains data about that particular user. A variant of discriminant analysis, Kalman adaptive discriminant analysis [6], can be used here. With this method, the classifier is created offline from training data, recursively updated online after every new data point. The system can thus adapt to a particular user. The principle of such an approach is shown in Figure 2.

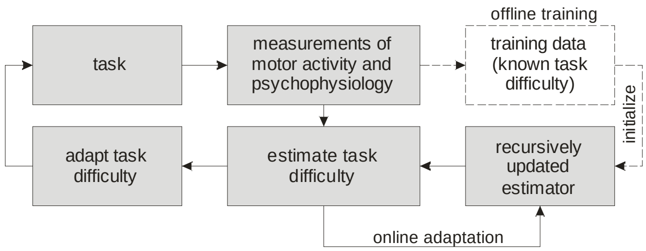

Figure 2: Task difficulty adaptation with task difficulty estimator updating.

We examined the suitability of using adaptive discriminant analysis of motor actions and psychophysiology for task difficulty estimation. Our first hypothesis was that adaptive methods would be more accurate than nonadaptive ones. Our second hypothesis was that classification based on both motor actions and psychophysiology would be more accurate than classification based on a single type of data.

MATERIALS AND METHODS

Task

For our task, we used a scenario previously used for reaching and grasping exercise in rehabilitation robotics [3]. A photo of a subject performing the task is shown in Figure 3. In the centre of the screen, there is a table sloped toward the subject. At the beginning of the task, a ball appears at the top of the slope and starts rolling downward. The subject's goal is to catch the ball using a haptic robot (the HapticMaster, manufactured by Moog FCS) before it reaches the lower end of the table. Once the ball is grasped, the subject must then hold the ball and place it in a basket above the table. Once the ball is dropped into the basket or falls off the table, another ball appears at the top of the table and the task continues. The haptic robot's grasping device allows the subject to feel each virtual item. Seven different difficulty levels were implemented, with each higher difficulty level featuring smaller and faster balls so that they were harder to catch.

Experiment procedure

Twenty students and staff members (16 males, 4 females) from the University of Ljubljana participated in the study. Mean age was 27.3 years, standard deviation 4.6 years.

Upon arrival, the task was explained and demonstrated to the subject. The physiological measurement equipment was attached. The subject rested for two minutes, then performed the task for six two-minute periods (12 minutes total). Within each period, the task difficulty was constant. At the end of a period, the subject was asked whether he or she would prefer the difficulty of the task to increase or decrease. The difficulty of the task then changed randomly by one or two levels in the selected direction. This randomness was introduced in order to expose subjects to a wider range of difficulty levels.

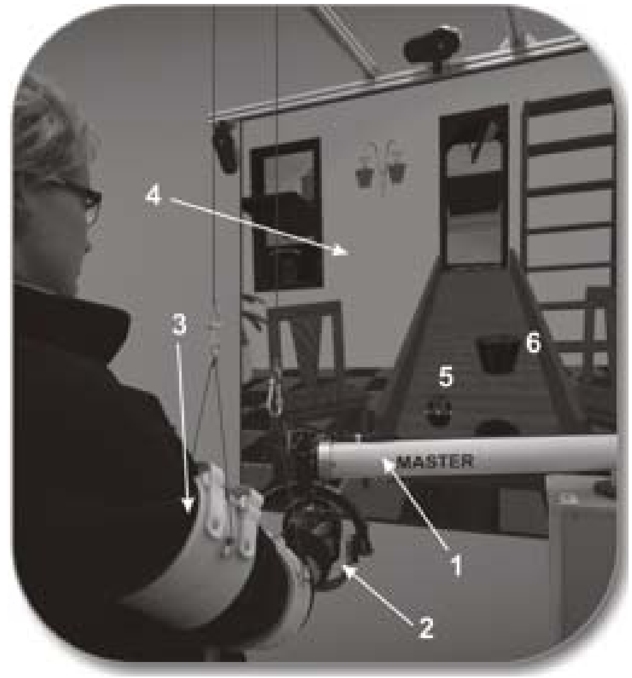

Figure 3: A subject performing the task using a haptic robot (1) and grasping device (2) while his/her arm is supported by cuffs (3). The screen (4) shows a sloped table, a ball (5) and a basket (6).

Input variables for classification

The variables used for classification were divided into two groups: motor variables and psychophysiological variables. A feature vector was defined as the vector of all variables from a single time period from a single subject.

Motor variables describe how well subjects did and how they moved during a particular time period. Measured using the robot's force and position sensors, they include variables such as the percentage of balls caught by the subject, the subject's mean velocity in different directions, and the mean forces exerted by the subject.

Physiological signals were sampled at 600 Hz using a g.USBamp amplifier (g.tec Medical Engineering GmbH). The electrocardiogram was recorded using disposable surface electrodes placed on the torso. Skin conductance was measured using a g.GSR sensor (g.tec). The electrodes were placed on the second and third fingers of the nondominant hand. Respiratory rate was obtained using a thermistor-based SleepSense Flow sensor placed beneath the nose. Peripheral skin temperature was measured using a g.TEMP sensor (g.tec) attached to the fifth finger of the nondominant hand.

Thirteen variables were extracted from the physiological signals. From the electrocardiogram, mean heart rate as well as seven time- and frequency-domain measures of heart rate variability were extracted. Detailed information about these variables is available in an extensive paper [4]. From the skin conductance signal, three variables were extracted: mean skin conductance, skin conductance response frequency and mean skin conductance response amplitude. Skin conductance responses are defined as transient increases in skin conductance whose amplitude exceeds 0.05 μS. From the respiration signal, two variables were extracted: mean respiratory rate and standard deviation of respiratory rate. From the skin temperature signal, mean temperature was extracted.

Linear discriminant analysis

Linear discriminant analysis (LDA) is a well-known method for feature extraction and classification, used to find a linear discriminant function that best separates data points into two or more classes. The discriminant function is built using a set of feature vectors (training data), each of which has a known class label assigned to it. This discriminant function is then used to determine the best class label for new feature vectors.

As previously noted, the goal of the discriminant analysis was to classify a measurement as either too easy or too hard. Thus, there were two possible classes. Two methods of discriminant analysis were used. The first was the classic, nonadaptive LDA, which is well-described in statistical literature. The second method was Kalman adaptive linear discriminant analysis (KALDA), an adaptive version of the LDA in which the weights of the discriminant function are recursively estimated online using a Kalman filter as new data becomes available. The Kalman gain varies the update coefficient and changes the adaptation speed depending on the properties of the data. Detailed equations can be found in [6].

Classifier fusion

While it is possible to use discriminant analysis to build a single, multivariate classifier using all the input variables at once, another option is to build a separate univariate classifier for each input variable. While the accuracy of any individual univariate classifier would be low, fusing the large number of classifiers may result in high accuracy. For classifier fusion, the result of each classifier was weighed according to its estimated accuracy, which was estimated from previously obtained training data. For instance, a classifier that was able to correctly classify 100% of the training data would be weighted with a factor of 1, a classifier that was able to correctly classify 75% of the training data would be weighted with a factor of 0.5, and a classifier that was able to correctly classify 50% or less of the training data would be weighted with a factor of 0. Classifier fusion is then done using the following formula:

$$C = \sum k_i \cdot c_i$$

where c_i is the class assigned to the feature vector by classifier i, k_i is the weighting factor of classifier i, and C is the final assigned class. If c_i can either be -1 (class 1) or +1 (class 2), the feature vector is assigned to class 1 if C is equal to or less than zero and assigned to class 2 if C is greater than zero.

Both multivariate classifiers with no classifier fusion and weighted vote fusion of univariate classifiers were tested.

Cross-validation

To test the accuracy of our classifiers, we used leave-one-out cross-validation. For a classifier, the entire data set was split into the test data (all data from one subject) and the training data (all data from all other subjects). The classifier was built using the training data, then validated using the test data. This procedure was repeated as many times as there were subjects, with each subject's data used as the test data exactly once. The accuracy rate of a classifier was calculated as the number of correctly classified feature vectors divided by the number of all feature vectors. A feature vector was considered to be correctly classified if the classifier's estimate (too easy or too hard) was the same as the choice that the subject had made.

Since the data was available offline, adaptive classifiers were tested as follows. The first feature vector from each subject (i.e. from the first time period of a session) was classified using the initial classifier obtained from the training data. Then, the classifier was recursively updated using this feature vector and the choice that the subject had actually made. The updated classifier was tested on the second feature vector from each subject, once again updated and so on. The weakness of such an approach is that, in online task difficulty estimation, the choice that the subject had made would not be available. This problem is further explored in the Discussion section.

RESULTS

Table 1 shows the accuracy rates for different classifier types and different input data (motor actions, psychophysiology or both). In the multivariate classifiers, a single classifier is made using all available inputs. In the weighted univariate classifiers, one classifier is made for each input variable and the classifiers are then fused as described in the 'Classifier fusion' section.

	motor.	**psychophys.**	**both**
multivar. LDA	79.3 %	60.3 %	71.1 %
weighted univar. LDA	79.3 %	61.6 %	80.2 %
multivar. KALDA	79.3 %	66.9 %	72.7 %
weighted univar. KALDA	**81.0 %**	**76.0 %**	**84.3 %**

Table 1: Classification accuracy rates for different classification methods.

DISCUSSION

Comparison of different classifiers

In all cases, KALDA yields higher classification accuracy than LDA. This confirms the hypothesis that adaptive methods improve classification accuracy. In weighted fusion of univariate classifiers, using both motor actions and psychophysiology yields higher classification accuracy than using only one data source. This is in agreement with our second hypothesis that multiple data sources improve classification accuracy. Interestingly, in the case of multivariate classifiers, nonadaptive classification using both data sources produces a worse result than using only motor actions. This may be because the data dimensionality is large, so it is difficult to find a single robust discriminant function. Weighted vote fusion of univariate classifiers should be more robust since each individual discriminant function covers only one dimension.

While nonadaptive task difficulty estimation using motor actions already gives a classification accuracy of almost 80 %, nonadaptive classification using psychophysiology yields an accuracy of less than 62 %. However, using adaptive methods can greatly increase accuracy even over a short time period. In our case, the task was only performed for a total of 12 minutes, and increasing the length of the task may allow even greater improvement in classification accuracy when using psychophysiological measurements. Nonetheless, results suggest that, at least in haptic interaction, motor actions should be used as a primary data source with psychophysiology providing supplementary information.

Naturally, perfect classification accuracy should not be expected. Several subjects occasionally expressed the desire to stay at the same difficulty level, but this option was not available. Additionally, it is not certain whether subjective choices are always perfectly reflected in measurable responses and thus whether completely accurate classification is even theoretically possible.

Use in online task difficulty adaptation

Since our classifiers can be used to determine whether a task is too easy or too hard, they can be directly used for online task difficulty adaptation. If the task is too hard, the system should decrease the difficulty. If the task is too easy, the system should increase the difficulty. Methods of increasing or decreasing difficulty must be defined in advance and can range from simple to very complex.

In our implementation, the user’s direct input (task is too easy / too hard) was used to update the KALDA classifiers. However, in a real-world application, this information would not be available and the update process would need to use its own estimate of the current class rather than the actual class. This would need to be done carefully since such an approach can also amplify classification errors. If an incorrect class estimate is used to update the classifier, the classifier will become worse. One way to address this would be to generate a measure of how ‘reliable’ the estimate is. The system would then only update the classifier if the estimate was sufficiently reliable. A simple variant of this has already been tested and resulted in a classification accuracy of 72.1 % with only psychophysiological inputs (compared to 76.0 % when the user’s actual input is available). Another possibility would be for the system to explicitly ask the user for input if certain potentially erroneous trends are detected (e.g. if the classifier repeatedly estimates that the task is too easy even though the user has reached a very high difficulty level).

CONCLUSION

We have demonstrated a classification method that can be used to estimate task difficulty in human-robot interaction based on motor actions and psychophysiology. The classifier can be recursively updated as new data becomes available, allowing it to gradually adapt to a particular user.

ACKNOWLEDGEMENTS

This work was funded by the EU Information and Communication Technologies Collaborative Project MIMICS grant 215756 and additionally supported by the Slovenian Research Agency. Moog FCS graciously loaned one of the two HapticMaster robots used in our research.

REFERENCES

1. Bethel, C. L., Salomon, K., Murphy, R. R. and Burke, J. L. Survey of psychophysiology measurements applied to human-robot interaction. *16th IEEE International Conference on Robot and Human Interactive Communication* (2007), 732-737.
2. Haarmann, A., Boucsein, W. and Schaefer, F. Combining electrodermal responses and cardiovascular measures for probing adaptive automation during simulated flight. *Applied Ergonomics*, *40* (2009), 1026-1040.
3. Novak, D., Ziherl, J., Olenšek, A., Milavec, M., Podobnik, J., Mihelj, M. and Munih, M. Psychophysiological responses to robotic rehabilitation tasks in stroke. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 18* (2010), 351-361.
4. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, *17* (1996), 354-381.
5. van Reekum, C. M. and Johnstone, T. Psychophysiological responses to appraisal dimensions in a computer game. *Cognition and Emotion, 18* (2004), 663-688.
6. Vidaurre, C., Schlögl, A., Cabeza, R., Scherer, R. and Pfurtscheller, G. Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces. *IEEE Transactions on Biomedical Engineering*, *54* (2007), 550-556.