On Gestures and Postural Behavior as a Modality in Ensemble Methods

Heinke Hihn, Sascha Meudt(✉), and Friedhelm Schwenker

Institute for Neural Information Processing, Ulm University, 89069 Ulm, Germany
{heinke.hihn,sascha.meudt,friedhelm.schwenker}@uni-ulm.de

Abstract. Knowledge about the users emotional state is important to achieve human like, natural HCI in modern technical systems. Humans rely on body gestures and posture when communicating. We investigate the relation between gestures and human emotion, specifically when completing tasks. The main focus of this work lies on discriminating between mental overload and mental underload, which can e.g. be useful in an e-tutorial system. Mental underload is a new term used to describe the state a person is in when completing a dull or boring task. It will be shown how to select suited features, such as gestures, movement and postural behavior. Furthermore those features will be investigated regarding their discriminative power. After features are selected, a multiple classifier system will be designed, trained and evaluated.

1 Introduction

A fundamental part of human communication is noticing a change in the affective state of the conversational partner. Affective state refers to the experience of feelings or emotions. To elaborate on this more, consider the following scenario: A person is telling another about a rather complex topic, e.g. in an teacher-student setting. During this conversation the student starts to look a bit overwhelmed by all the new information. In this case one would expect the teacher to change his pace as the student obviously can't follow up. Let that state the student is experiencing henceforth be referred to as *mental overload*. This term is meant to describe the state one is in when being confronted with a very complex task, e.g. understanding something completely new. The opposite, i.e. completing an easy task or listening to a teacher talking about a already well known topic, shall be called *mental underload*. In terms of the student-teacher example one can consider a electronic tutorial platform which controls its pace depending on the student's behavior. A user centered system should offer possibilities for the user to express their emotions [10,16]. Based on human interaction one can imagine two ways: verbal and non-verbal. Verbal communication focuses on information retrieved from speech. These can be loudness and pitch or the words being sad. There has been a lot of research in this area [2,9,11]. Non-verbal ways of expressing feelings can be facial expressions and gestures, to name a few. While facial expressions have been researched very thoroughly in the past [6,12,13,15], the same doesn't quite hold for gestures. Even tough they play a crucial part

F. Schwenker et al. (Eds.): ANNPR 2016, LNAI 9896, pp. 312–323, 2016.
DOI: 10.1007/978-3-319-46182-3_26

in human-to-human communication they have been only used little compared to other modalities in Affective Computing [17], e.g. in [7,8]. This work aims to close the gap and develop a method to employ postural behavior and gestures as a powerful additional modality. Specifically to distinguish between mental overload and underload, as described earlier.

2 Related Work

Kapur et al. [7] conducted a study in 2005 on gesture based affective computing. Their goal was to train a classifier to distinguish between the four basic emotions *Sad, Joy, Anger* and *Fear*. The authors equipped five actors with markers of a motion capture system and asked them to "perform" given emotions. After collecting the data 10 participants were asked to identify each emotion only by watching the moving points, i.e. the position of the markers. In a next step the authors compute mean of velocity and acceleration and the standard deviations of positions. The resulting data is then used to train and evaluate several classifiers. The participants achieved an average recognition rate of 93 %, while the classifiers achieved between 66.2 % and 91.8 %.

Kipp and Martin [8] investigate four basic gestural features and their respective relation to the emotional state. Their main goal was to create embodied conversational agents, i.e. defining a set of gestures to discriminate emotions. They introduced *lexemes* to describe gestures by a set of constraints on *handedness*, *hand shape*, *palm orientation* and *motion direction*. Data was gathered from the movie *Death of a Salesman* (1966 DS-1 and 1985 DS-2). In order to estimate the correlation between emotion category and gestures, the authors computed pairwise χ^2 values. Results suggested a highly significant correlation between emotion category and handedness (χ^2 = 40.14; p < .001, in film DS-1, χ^2 = 35.37; p < .001, in film DS-2). They also found for the film DS-2 a correlation between emotion category and palm orientation (χ^2 = 42.50; p < .05).

Bianchi-Berthouze et al. conducted a study on posture and gesture and immersion [1]. They focused on two things: is there a relationship between postural behavior and immersion and the importance of full-body control to improve user experience. High immersion occurs when the participant has the perception of being physically present in a virtual reality. Twenty participants were randomly assigned to two groups: a simple point and click game and a first person shooter. The authors hypothesized that the players in group two experienced a higher lever of immersion. After playing 10 min the players were asked to complete a *Immersion Questionnaire* [5] to quantify the level of immersion. Group 1 returned rather low immersion scores (mean 47.1, $\sigma^2 = 16.64$). They also showed many shifts in sitting position, e.g. from a very relaxed to a very attentive pose. Group 2 showed significantly higher scores (mean 68.11, $\sigma^2 = 11.95$). Movement in this group occurred fewer, with players scoring lower in immersion showing more movement. The authors argue that the results suggest that higher immersion causes fewer unnecessary movements. They also infer that the observed reduction in movement is caused by the higher engagement, i.e. players in group 2 are more focused.

3 Experimental Setting

The dataset is based on an experiment conducted within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems". Participants were asked to play a series of games based on the interaction paradigm of Schüssel et al. [14]. The task of each game sequence was to identify the singleton element, i.e. the one item that is unique in shape and color(see Fig. 1). The participants interacted with the system by speech. The difficulty was set by adjusting the number of shapes and the time to answer. If the given answer was incorrect, the player received no reward for that particular round. After a introduction each participant completed four game sequences of decreasing difficulty. The first sequence was designed to induce overload (6 × 6 board, 6 s to answer, see Fig. 1), the second was 5 × 5 with 10 s, the third was set to 3 × 3 with 100 s, sequence four was a 3 × 3 mode with 100 s (underload). The last sequence induces frustration, e.g. by purposely logging in a wrong answer. As the sequences 1 and 4 are explicitly designed to cause over- and underload, we focused only on those two. After each sequence the participants answered a self assessment questionnaire (SAM). The aim of those questions was to determine valence, arousal and dominance experienced in the particular sequence. A total of 52 participants were recorded. Of those were 26 male and 26 female. Their age spanned from 17 to 27 (mean 21.66, $\sigma^2 \approx 2.7$). During the experiment participants were monitored by several sensors. This work focuses on the depth data provided by a Kinect sensor to compute body movements and postural behavior. The skeleton is extracted by the Kinect itself. We do not employ any extraction algorithm.

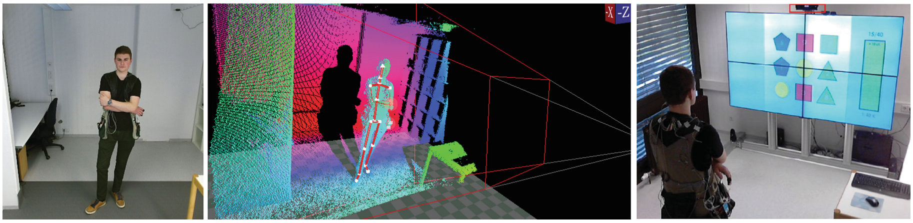

Fig. 1. Experimental setting. Left: Front view. **Middle**: 3D projection of the data provided by the Kinect sensor. The depth is color coded, such that green indicates near objects and rad objects further away. **Right**: Rear view. The red rectangle indicates the Kinect. (Color figure online)

4 Feature Engineering

By watching the recordings a couple of mentionable static gestures were found. For each of those a set of constraints was defined:

Arms crossed: the right hand is near the left elbow and vice versa.
Hands behind back: both hands are not visible to the Kinect.
Resting hand on hips: the right hand is near the right hip and vice versa.
Crossed feet: the left foot is right of right foot or vice versa.
Feet in front of another: the left foot is closer to the Kinect then the right one or vice versa.

The occurrences are almost identically distributed and therefore bare only little discriminative power. To overcome this, they have to be combined with further information. We chose to enhance the features by adding the duration. One drawback of this approach is that suitable thresholds for the constraints have to be defined. Assuming a threshold for a given person is found, the threshold doesn't necessarily apply to other persons just as well. To avoid setting thresholds the mean distance of the respective joints over the set of frames is computed. Figure 4 gives an example. Again, by watching the video material and observing the participants' behavior two main linear movements have been identified: moving both hands away from the torso and scratching the head or face. The latter occurred about evenly in both sequences. "Moving both hands away from the torso" is mostly done in combination with confused facial expression as Fig. 2 shows (this might be very useful in fusion paradigms, i.e. in combination with a facial expression detector). It seems to be a rather clear indicator whether a

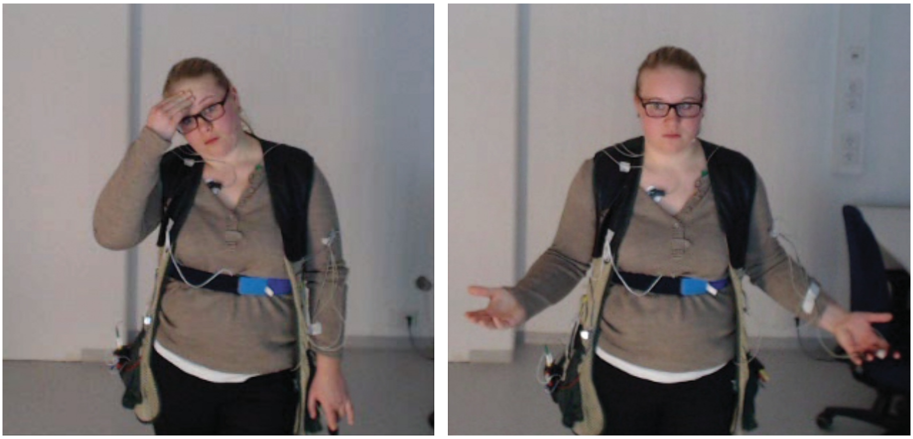

Fig. 2. Examples of participant behavior. Left: The participant is scratching their head. This occurred about evenly during overload and underload. **Right**: The participant is moving their hand away. This occurred far more often during overload. During underload it only occurred when the participant gave a wrong answer.

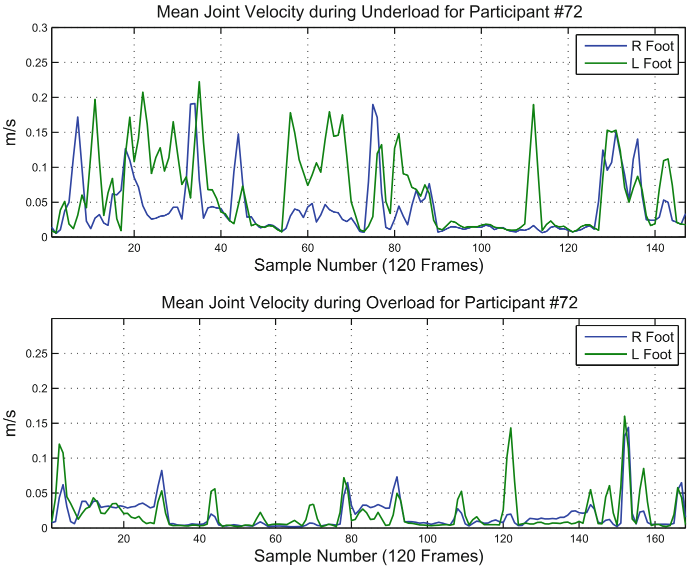

Fig. 3. Extraced features. Top: Mean joint velocities during underload. **Bottom**: Mean joint velocities during Overload. It can be seen that participant #72 had a rather active feet movement during underload while their feet were mostly standing still during overload.

person is experiencing over- or underload, because it occurs mostly during overload. We captured this gesture by computing mean joint distances (e.g. between both hands, hand and hip), velocity, and acceleration. For each joint the values were computed within a window of 120 frames. Figure 3 gives an example. Joint velocity is computed as the first derivative of the joint positions (x, y, z) w.r.t. time: $\bar{v} = \frac{\Delta \mathbf{s}}{\Delta t}$. Joint acceleration is computed by approximating the derivative of the joint velocities, i.e. $\bar{a} = \frac{\Delta \mathbf{v}}{\Delta t}$. To account for varying movement within the frame the standard deviation of velocity and acceleration are also computed. Additionally, most participants showed a rather highly active head movement. The Kinect sensor measures the rotation angles of the head in yaw, pitch, and roll notation. The yaw angle can be used to detect whether a participant is looking at the camera or not. Pitch angle indicates movement towards the floor or the ceiling and roll angle is measuring head tilt. To capture those movements a threshold for each angle is defined.

5 Results

We focus on person dependent classification. This type of classification refers to training a classifier such that it fits well to a given person. A Random Forest [3,4] (RF) of 200 trees was trained and evaluated with a 10-fold cross validation for each participant. We choose RFs because they can be trained and evaluated easily and can handle large amounts of data well. Furthermore they do not

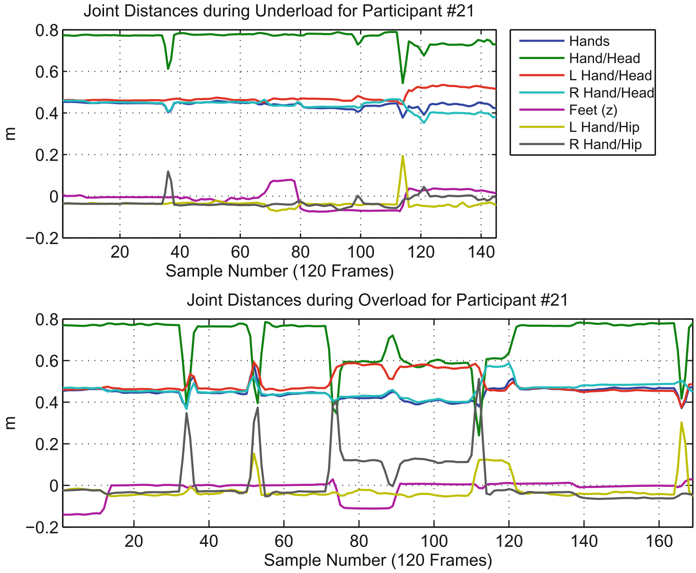

Fig. 4. Examples for postural behavior. Mean distances (window 120 frames, 60 frames overlap) between selected joints during underload (top) and overload (bottom). Negative values e.g. indicate that the left hand is below the left hip.

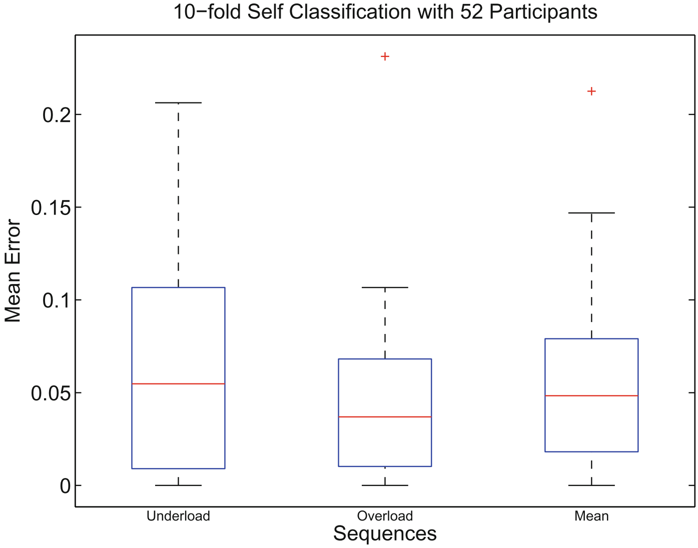

Fig. 5. Person dependent classification results. Each group contains 52 samples, i.e. the number of participants. Each sample represents the mean classification accuracy obtained by a 10-fold cross validation.

require many parameters. As the features were extracted with overlap, there is a high correlation between neighboring samples. To achieve an unbiased result the features containing overlap from test and training set were removed. Results are shown in Fig. 5. Due to the slightly more expressive behaviors during mental overload this class resulted in a smaller classification error. Overall an error rate of about 5 % was achieved. Further experiments of training a single classifier matching all participant behaviors at once (leave one subject out) yielded an error of about 38 % in overload recognition and up to 47 % overall. This is due to the highly individual character of human behavior and could possibly be overcome by grouping similar participants.

5.1 Approximating Feature Importance

The importance of the different features was investigated by randomly permuting the values of a feature and classifying based on that [4]. The resulting values indicate how much the mean classification error changed after permutation (*delta error*), i.e. high values represent important features. This was done for each of the features and for each participant. Figure 6 gives an example for postural behavior, i.e. selected joint distances.

Moving direction of the joints seems to be important for some participants and for others it's not. Tests without this feature did not yield significantly higher accuracies. Head positions yielded rather low importance measurements. In fact, the values obtained were the lowest of all features and for some participants the delta error was negative. Negative values indicate that the classification accuracy could be improved when leaving those features out. This was done and a new set of tests was run, but classification error did not improve significantly. This could be explained by interactions between features.

5.2 Approximating OOB Error

Recall, that the RF algorithm employs bagging during training phase. Bagging is a technique where samples are divided into subsets by drawing randomly with replacement [3] such that one subset is created per tree. The *Out Of Bag Error* is computed by running the samples that haven't been used for training through the classification tree and evaluating the results. Figure 7 shows the mean, maximal and minimal cumulative OOB error over all 52 participants. The low OOB error indicates the classification trees can learn the underlying distribution rather well. This is also backed by the high classification accuracies obtained when evaluating self classification.

5.3 Approximating Outlier Measurements

Outliers in RFs can be found by first computing the proximity of the data and then averaging by the number of trees. Proximity between two observations is defined as the fraction of trees in the ensemble for which these two observations

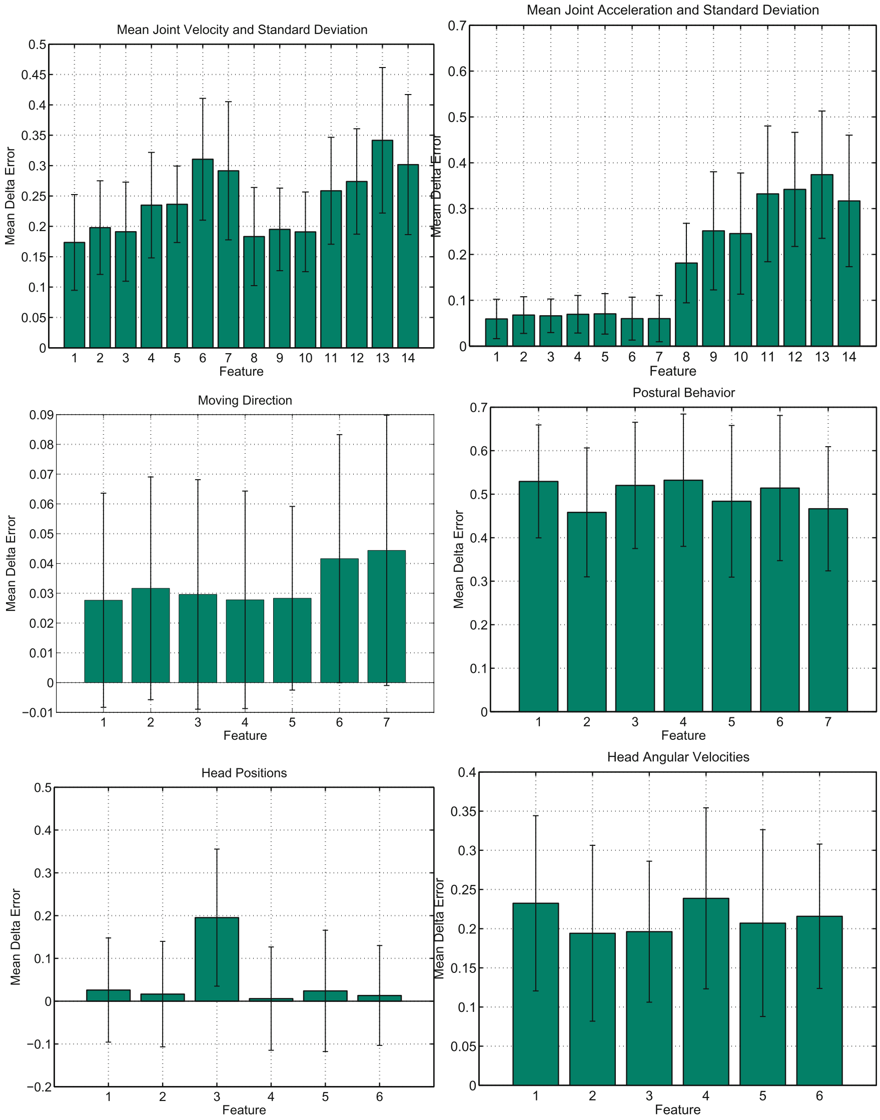

Fig. 6. Mean feature importance and respective standard deviation. Delta error refers to the change in classification error made when permuting values of a given feature. High values indicate important features and low values less important features.

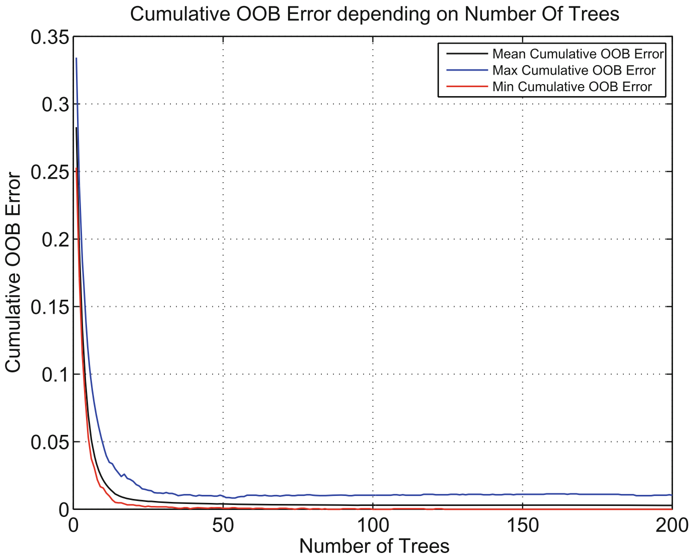

Fig. 7. Out-Of-Bag Errors. As the figure shows, the Out-Of-Bag Error doesn't improve further after about 50 trees.

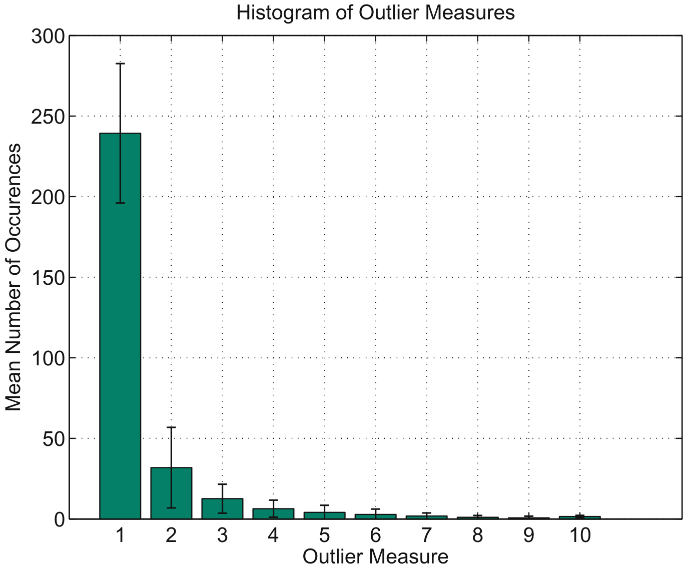

Fig. 8. Outlier measurement and respective standard deviation. The samples within one participant contain only little outliers. Note that the outlier value has been omitted and only the bin numbers are shown.

land on the same leaf [4]. Outliers can then be found by taking the squared inverse proximity of a given sample and compare that value with the squared inverse proximity of the remaining samples. A high value indicates this sample is an outlier. Figure 8 shows the histograms of outlier measurements over all 52 participants and the corresponding standard deviations. As most of the samples have a low measurement (first bin), the samples within each participant are very similar.

5.4 Comparing Different Ensemble Members

The previous section evaluated RFs. To get an idea of how well that classifier compares to others and how they influence classification accuracy, several tests were run using different classifiers. Figure 9 shows the results. A 3-NN ensemble, a Linear Discriminant Analysis (LDA) ensemble, and a mix of both were trained using the random subspace method, which is a generalization of the RF algorithm. It operates on a random subset of the feature space. In this case it was set to $m = \lfloor log_2(M) \rfloor$, as suggested by Breiman in is original paper. We chose these two classifiers because they are trained similarly to RF and don't require many parameters.

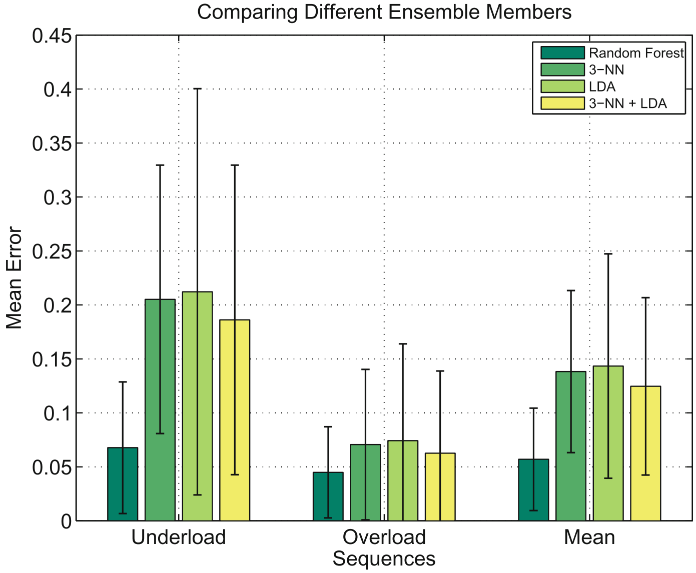

Fig. 9. Comparing ensemble members. As the figure shows, the RF performs well compared to the others.

6 Conclusion

It was shown that there is indeed a relation between postural behavior and mental over- and underload. Specifically, the findings suggested mental overload is in most cases accompanied by a rather high physical arousal, i.e. a lot of movement. This was then used as a basis to identify suitable features. These were mean joint velocity, acceleration, and distances between several selected joints. Additionally, the respective motion direction is important. Head movement has been captured by computing rotational velocities on each axis and six predefined positions. To prove the usefulness of the feature an extensive analysis was conducted. In particular, the training error (OOB error), the sample outlier measurement and the feature importance have been investigated. The analysis of the OOB error and the outlier measurement showed the features do indeed separate the samples well into mental overload and underload for each participant. For each of the features their respective discriminative power was also approximated. The results indicated mean joint velocities and accelerations bare the most information and head position and moving direction the least. In a last step the model itself has been evaluated to prove the RF algorithm is indeed the best fitting choice for this task. To achieve this the classification results were compared to ensembles of 3-NN, LDA, and a mix of both classifiers. Comparing the results revealed the RF outperformed the others. The overall results are promising in terms of HCI systems which are adaptable to the users bearings. Such a systems would interact with a single user over a longer period and learn to understand the users behavior. Remembering the initial described tutorial system one could imagine a systems which assists a student over at least a whole term by adapting the teaching pace based on the presented approaches. The theoretical findings of this work also used to successfully design and implement a live system. This system is able to record a given participant using a Kinect sensor, extract features and classify those features. Classification is achieved by training with the data from the participants used in the theoretical analysis.

For future works it could be of interest to investigate if other affective states can also be classified based on postural behavior. The data we used was from rather young participants (mean 21.66 years). This could have biased the gestures we found and analyzed, because elderly people may show different (less expressive) gestures. It would be worthwhile to investigate this further. Another drawback of our method is the rather long training phase, because we assumed a person dependent. This could be overcome by finding a suitable participant grouping and training classifier systems for each group.

Acknowledgments. The authors of this paper are partially funded by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Bianchi-Berthouze, N., Cairns, P., Cox, A., Jennett, C., Kim, W.W.: On posture as a modality for expressing and recognizing emotions. In: Emotion and HCI Workshop at BCS HCI London (2006)
2. Breazeal, C., Aryananda, L.: Recognition of affective communicative intent in robot-directed speech. Auton. Robots **12**(1), 83–104 (2002)
3. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Cairns, P., Cox, A., Berthouze, N., Jennett, C., Dhoparee, S.: Quantifying the experience of immersion in games (2006)
6. Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., Karpouzis, K.: Modeling naturalistic affective states via facial and vocal expressions recognition. In: Proceedings of the 8th International Conference on Multimodal Interfaces, pp. 146–154. ACM (2006)
7. Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., Driessen, P.F.: Gesture-based affective computing on motion capture data. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 1–7. Springer, Heidelberg (2005)
8. Kipp, M., Martin, J.-C.: Gesture and emotion: can basic gestural form features discriminate emotions? In: 3rd International Conference on Affective Computing and Intelligent Interaction Workshops, pp. 1–8. IEEE (2009)
9. Meudt, S., Zharkov, D., Kächele, M., Schwenker, F.: Multi classifier systems and forward backward feature selection algorithms to classify emotional coloured speech. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 551–556. ACM (2013)
10. Picard, R.W.: Affective Computing, vol. 252. MIT Press, Cambridge (1997)
11. Plesa-Skwerer, D., Faja, S., Schofield, C., Verbalis, A., Tager-Flusberg, H., Dykens, E.M.: Perceiving facial and vocal expressions of emotion in individuals with williams syndrome. Am. J. Ment. Retard. **111**(1), 15–26 (2006)
12. Russell, J.A., Bachorowski, J.-A., Fernández-Dols, J.-M.: Facial and vocal expressions of emotion. Annu. Rev. Psychol. **54**(1), 329–349 (2003)
13. Schels, M., Glodek, M., Meudt, S., Scherer, S., Schmidt, M., Layher, G., Tschechne, S., Brosch, T., Hrabal, D., Walter, S., et al.: Multi-modal classifier-fusion for the recognition of emotions. In: Coverbal Synchrony in Human-Machine Interaction (2013)
14. Schüssel, F., Honold, F., Bubalo, N., Huckauf, A., Traue, H., Hazer-Rau, D.: In-depth analysis of multimodal interaction: an explorative paradigm. In: Kurosu, M. (ed.) HCI 2016. LNCS, vol. 9732, pp. 233–240. Springer, Heidelberg (2016). doi:10.1007/978-3-319-39516-6_22
15. Shan, C., Gong, S., McOwan, P.W.: Robust facial expression recognition using local binary patterns. In: IEEE International Conference on Image Processing, ICIP, vol. 2, pp. II-370. IEEE (2005)
16. Wendemuth, A., Biundo, S.: A companion technology for cognitive technical systems. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) COST 2102. LNCS, vol. 7403, pp. 89–103. Springer, Heidelberg (2012)
17. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)