An Analytical Approach for Head Gesture Recognition with Motion Sensors

Nina Rudigkeit and Marion Gebhard
Sensors and Actuators in Medicine (SAM)
Westphalian University of Applied Sciences
45877 Gelsenkirchen, Germany
Email: {nina.rudigkeit, marion.gebhard}@w-hs.de

Axel Gräser
Institute of Automation (IAT)
University of Bremen
28359 Bremen, Germany
Email: ag@iat.uni-bremen.de

***Abstract*—Within this work, an analytical approach for robust recognition of four different head gestures in a continuous data stream is presented. Analytical solutions are more robust against signal variations than pure signal-oriented approaches. Furthermore, they enable user-independent gesture recognition.**

The proposed model integrates information about sensor placement and ideal shape of gestures. Furthermore, activity-based windowing was used to increase computational efficiency. Model parameter values were obtained empirically. For evaluation, data were collected from ten subjects using a 9-axis MEMS motion sensor system. The subjects were instructed to repeat each of the defined gestures five times. In addition a total number of 25 motion patterns slightly different to the defined gestures were recorded for each subject. Applying user-specific parameters an average classification rate of $93.56\% \pm 4.96\%$ was achieved. User independent parameters led to an average classification rate of $87.56\% \pm 8.90\%$. It is likely that the performance using user independent parameters can be further increased when giving the user meaningful feedback about how to adjust their movements. However, future research will cover real-time performance of the model in a natural environment.

I. Introduction

A. Previous Work

People with tetraplegia cannot move any of their limbs and are therefore limited in their capabilities of interacting with their environment. Nevertheless, many of them can still move their heads which can be utilized for the development of assistive devices. In [1] we already presented and tested a control structure for direct control of a robotic arm with head motion. Head motions were recorded using a 9-axis Inertial Measurement Unit (IMU) including on-board sensor fusion. Head orientation in terms of Euler angles was used to generate three independent signals to control a robot arm. It was proposed to control the 3D position of the robot gripper in world coordinates and the gripper 3D orientation in device coordinates. Moreover, an additional signal was necessary to open and close the gripper. As a result, at least 7 degrees of freedom (DOFs) had to be controlled while head movements provided 3 DOFs only. To map the 3 DOFs of the head onto the 7 DOFs of the robot, the robot movements were decomposed into motion groups with maximum 3 DOFs. For switching between the motion groups additional control commands were needed.

The generation and reliable recognition of these so called switching commands is proposed within this work. Switching is performed by a well-defined head movement that can be recognized with high reliability and low latency. Furthermore, interference with the head motions which are used for direct control of the robot arm has to be avoided.

B. Related Work

The presented research problem belongs to the field of pattern recognition, more precisely, Human Activity Recognition (HAR). Until now, most research in this topic addresses computer vision-based approaches as surveyed in [2]. A major drawback of vision-based approaches is that they require at least one camera that faces and captures the user. As a consequence, they can only be used in constrained environments. In addition, many people do not want to be filmed which generally decreases user acceptance of vision-based systems.

In contrast, motion sensors are self-contained. That means, they do not require any modification of the environment in order to record motion. Moreover, motion sensors based on MEMS[1] technology are small, low-cost and energy-efficient. Hence, approaches using body-worn motion sensors have attracted grown interest in HAR research of recent years. In [3] the authors give an extensive introduction into this topic.

A general purpose framework for designing Activity Recognition Chains (ARC) consists of the following elements:

1) *Preprocessing (optional):* Filters are applied to remove artifacts from raw sensor signals.
2) *Windowing:* A continuous data-stream is split into segments which contain data for further analysis.
3) *Feature Calculation:* Reduces the signal into features which may be discriminative for the classification problem.
4) *Dimensionality Reduction (optional):* Features of interest are selected or the feature set is projected onto a lower dimensional subspace.
5) *Classification:* A classifier assigns each data segment to a certain class based on its features.

There are mainly two distinct approaches to address pattern recognition problems, i.e, signal-oriented [4], [5] and analytical ones [6]. Whenever the relationship between input data and output is unknown or too complex to implement, the signal-oriented approach is a good choice. An example application

[1] Micro-Electro-Mechanical System

978-1-4799-6314-0/15/$31.00 ©2015 IEEE

is handwriting recognition using wearable sensors [7]. In this case the input data corresponds to the sensor data while the output corresponds to the performed gesture. Signal-oriented ARCs to address this type of problem are based on statistical signal properties, e.g., mean and variance, and statistical models. Common statistical models for classification are k-Nearest-Neighbor, Support Vector Machines, Artificial Neural Networks and Hidden Markov Models [8]. A major advantage of ARCs based on the signal-oriented approach is that they can be adapted to any pattern recognition problem. Furthermore, implementations of statistical models are often already provided as part of Integrated Development Environments or freely available online.

In contrast, analytical approaches consider prior knowledge in order to develop customized ARCs. Therefore, they are often not pursued even though they are more robust against signal variation [9]. Within this work, we present a new analytically derived algorithm to recognize head gestures. The algorithm is adapted to the boundary conditions of the research problem discussed in section I-A. That means, the proposed method considers kinematic knowledge about the system, such as sensor placement and ideal shape of the gestures. Furthermore, specifications were made for the consistent integration into the already existing control structure to meet the computational requirements for real-time gesture recognition.

II. Analytical Head Gesture Recognition

A. Sensor Placement

A 9-axis IMU is used to measure the user's head motion. Outputs are the raw sensor data from three accelerometers, three gyroscopes and three magnetometers, as well as sensor orientation obtained from these raw data. The 9-axis IMU sensor is placed on the user's head [10]. We have shown that the sensor is moved on a spherical surface when the sensor yaw axis coincides with the approximated yaw axis of the user's cervical spine. Given this sensor placement, changes in head and sensor orientation are identical and a transformation of sensor orientation to head orientation is not needed. Nonetheless, an offset calibration remains a necessary preprocessing step. Furthermore, the challenge is to avoid inaccurate sensor placement, to reduce sensor drift as well as coupled motion due to the complex kinematic of the cervical spine. However, these deviations are expected to be small. Fig. 1 shows the sensor placement and the coordinate systems. At this point, one should note that every head motion apart from rotation around the yaw-axis results in additional linear sensor movement as the sensor is not rotated around its own center.

B. Head Gesture and Feature Selection

The previously mentioned linear sensor movement provides important information for head gesture recognition. Given the presented sensor placement, the onset of pitch or roll motion of the head results in noticeable linear sensor acceleration. Moreover, during pure pitch or roll motion the other DOFs, which we call non-dominating DOFs, are expected to be activated to a small extent only, e.g., due to coupled motion or imperfect sensor placement [10]. Using this fact, we define four gestures which can be discriminated easily (Fig. 2):

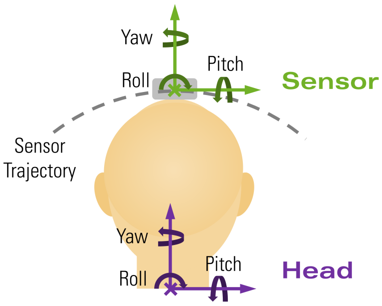


Fig. 1. Coordinate systems of the head (purple, seen from the back of the head) and of the sensor (green). Both head and sensor coordinate system have three rotational degrees of freedom (DOFs), i.e., roll φ, pitch ϑ and yaw ψ.

- *Nodding down:* Motion along the pitch DOF, ϑ, starting from the center in negative direction and back
- *Nodding up:* Motion along the pitch DOF, ϑ, starting from the center in positive direction and back
- *Bending left:* Motion along the roll DOF, φ, starting from the center in negative direction and back
- *Bending right:* Motion along the roll DOF, φ, starting from the center in positive direction and back

When displaying head displacement d against time t, the shape of the dominating DOF of each of these gestures can be approximated well by a Gaussian function that is given by

$$d = d_{max} \cdot e^{-\left(\frac{t-t_c}{w}\right)^2} \tag{1}$$

Maximum head displacement is expressed by the amplitude d_{max}, t_c is the centroid, and w is related to the peak width. Peak width can be influenced by the time the user needs to perform the gesture. The parameters are gesture- and user-dependent. Amplitude and time for gesture execution are important parameters to describe the gesture while the centroid is of minor importance for time series data. The goodness of fit can be expressed by means of the R^2-value. Correct gesture execution leads to high R^2-values. Overall, this leads to five important features to describe each gesture:

- Amplitude d_{max} of the dominating DOF
- Peak width w of the dominating DOF
- R^2-value of the fit
- Relative coupling Δ_1 of the first non-dominating DOF
- Relative coupling Δ_2 of the second non-dominating DOF

The relative coupling is defined as the range of a non-dominating DOF normalized by the amplitude d_{max} of the dominating DOF. Common values of these features can be obtained empirically as described in section III-B.

C. Windowing

We assume that activity is present if the magnitude of linear acceleration of the sensor is greater than a previously

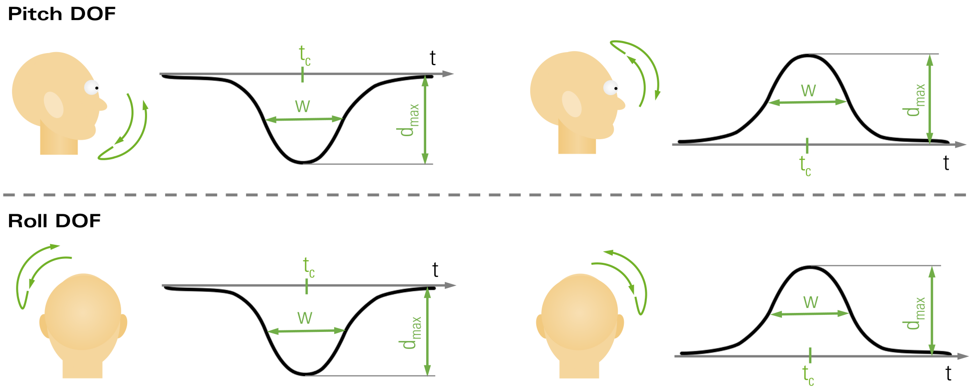


Fig. 2. The four gestures to be classified, and their shapes of the dominating DOF when plotting head displacement against time.

defined threshold $\varepsilon_a = 0.5\,\mathrm{mm\,s^{-2}}$. For every sample with index i a decision whether there is activity ($a_b(t=i)=1$) or not ($a_b(t=i)=0$) is made. In order to compensate for unwanted spikes, a lowpass filter has to be applied to this binary signal. Simple moving average (SMA), linear weighted moving average (LWMA) and quadratic weighted moving average (QWMA) have been considered due to their low latency in real-time applications. The spike removal is best using SMA, but it also leads to the highest latency. QWMA has the lowest latency but performs worst at smoothing. However, QWMA was still found to be the best compromise because the information content was highest for the resulting windows. A QWMA of order n can be written as

$$\begin{aligned} a_{QWMA}(t) &= \frac{1}{\beta}\sum_{i=1}^{n} i^2 \cdot a_b\left(t-n+i\right) \qquad (2)\\ \text{with } \beta &= \sum_{j=1}^{n} j^2 \end{aligned}$$

For the experiments, the order was chosen to be $n = 20$. In conclusion, windowing starts when the smoothed activity signal $a_{QWMA}(t)$ exceeds a certain threshold $\varepsilon_w = 0.4$ and ends when it falls below. All mentioned parameter values have been obtained empirically. Fig. 3(a) illustrates the real-time windowing procedure.

D. Analytically Derived Classification Algorithm

For every window it is checked whether one direction of movement is dominating at the beginning. If so, the direction is identified. A direction is considered dominating if its amplitude of linear acceleration a_j normalized by the softmax function exceeds a threshold $\varepsilon_s = 90\%$. The softmax function for DOF $j \in \{\varphi, \vartheta, \psi\}$ is given by

$$p_j(t) = \frac{e^{\lambda \cdot a_j(t)}}{\sum_k e^{\lambda \cdot a_k(t)}} \qquad \text{with } k = \varphi, \vartheta, \psi \qquad (3)$$

Where $\lambda = 25$ is the scaling factor. Based on the result, a preselection is made. This can be done before the entire window is recorded. At this stage, the four gestures which are used as control commands are already clearly separated. This is due to the fact that the direction of the initial linear acceleration is unique for each gesture.

TABLE I. USER INDEPENDENT CLASSIFICATION THRESHOLDS

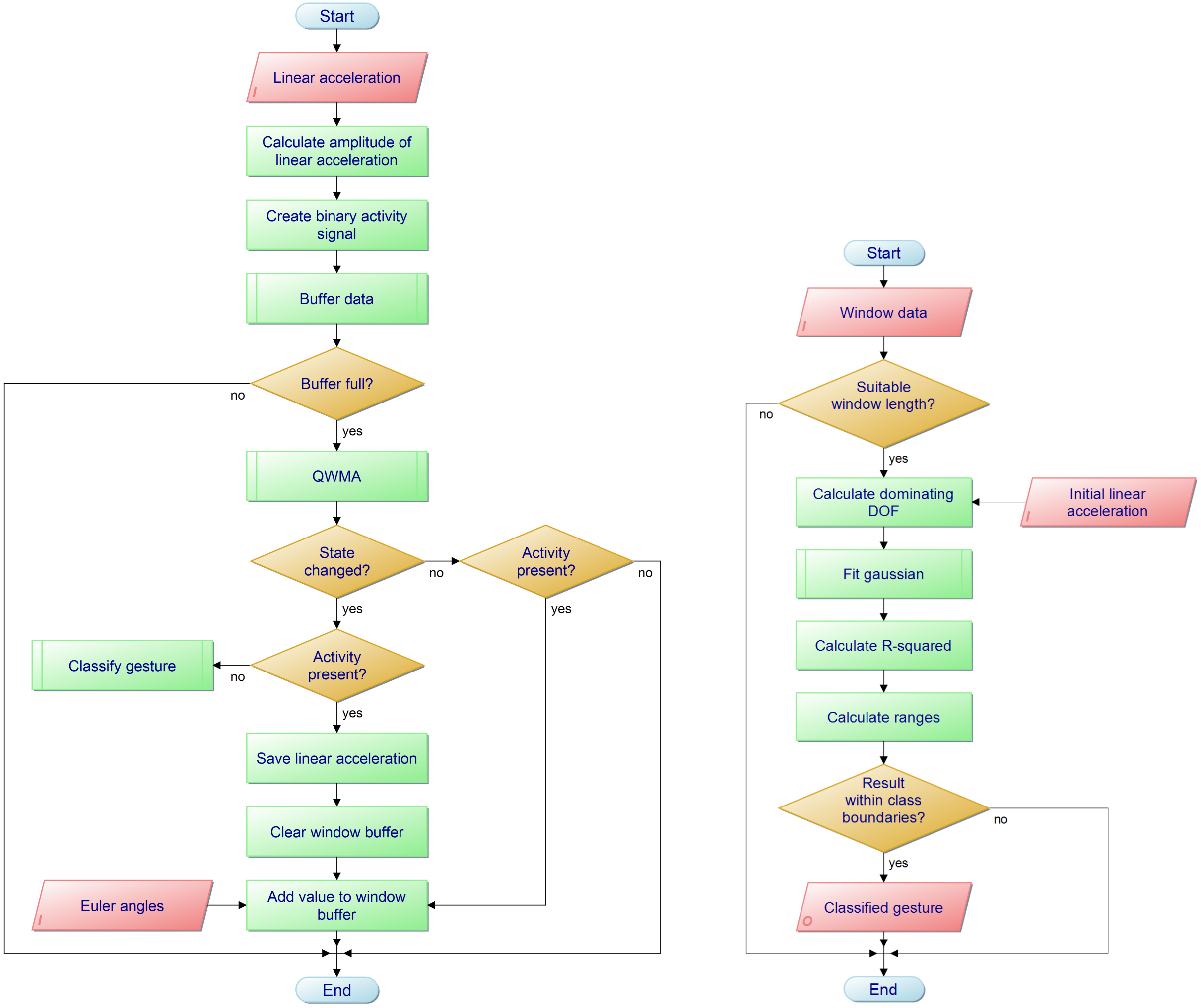
Description	Parameter	Value
Minimum amplitude	$\varepsilon_{d_{max}}$	25°
Maximum peak width	ε_p	0.28 s
Minimum R-squared value	ε_R	85 %
Max. relative coupling of 1st non-dominating DOF	ε_{Δ_1}	80 %
Max. relative coupling of 2nd non-dominating DOF	ε_{Δ_2}	80 %

However, the major challenge is to separate these gestures from unintended motion as well as from motion for control of the robot arm. For this reason, in the next step it is checked whether the window data fulfills the definition of the eligible gesture or not[2]. Due to the preselection, the dominating DOF as well as the direction of motion are known. That means, only relevant features need to be computed. As a result, preselection reduces the feature space so that no additional dimensionality reduction, known from common pattern recognition, is required.

In order to determine amplitude and peak width, a Gaussian function (Eq. 1) is fitted to the head angle data of the dominating DOF. This is done using the Levenberg-Marquardt algorithm[3]. After a successful fit, the R^2-value is calculated. In addition, the relative coupling of the non-dominating DOFs is computed. If the amplitude d_{max} is greater than a certain threshold $\varepsilon_{d_{max}}$, the peak width w below ε_p, the R^2-value greater than ε_R and the relative coupling of the non-dominating DOFs $\Delta_{1,2}$ below the thresholds $\varepsilon_{\Delta_{1,2}}$, the gesture is classified (Fig. 3(b)). That means, simple thresholds are used to separate the gestures from every other motion that might occur. For the presented experiments, the thresholds for the classification (Table I) have been obtained empirically as described in section III-B.

[2] For computational efficiency this only needs to be done if the window length is in the range of the gesture length.

[3] The starting point for the optimization is chosen to be identical with the expectation values, which may be obtained empirically. Assuming a gesture is present, the starting point is already close to the optimum, which speeds up computation and minimizes the risk of running into a local minimum.



(a) Activity-based windowing using smoothed accelerometer data.

(b) Feature calculation and classification after a window has been recorded. The feature space is reduced by calculating only features which are relevant for the expected gesture.

Fig. 3. Flow chart of the real-time implementation of the proposed analytical gesture recognition algorithm.

E. Integration of Gesture-Based Switching into the Robot Control Structure

As described in [1] during direct control of the robot, a sigmoidal transfer function is used between head displacement and robot velocity. That means, small head movements below a certain threshold do not result in physical robot motion. This zone is called dead zone and can be used for consistent integration of switching commands into the robot control structure.

The key idea is that a head gesture used for switching must clearly differ from robot control signals within the dead zone. In general, slow and smooth motion is used for robot control. As a consequence, the linear sensor acceleration always stays below a certain threshold during robot control. If the magnitude of linear sensor acceleration exceeds this threshold within the dead zone, head gesture recognition is initiated (Fig. 4). From then on, no robot control is possible until the magnitude of sensor acceleration falls below the acceleration threshold and the user's head returns to the dead zone.

If linear sensor acceleration is above the previously defined threshold while head displacement is beyond the dead zone, the head motion is assumed to be unintended and neither converted into robot motion nor into a switching command.

III. Methods

The following measurements were carried out using an FSM-9 by Hillcrest [11], which was mounted on a hairband. The sampling rate was set to f_s =125 Hz.

A. Subjects

10 able-bodied subjects were recruited from the University of Bremen to take part in the experiment. Three of them were females and the remaining seven males. Their ages ranged from 27 to 48. None of the subjects suffered from known neck motion limitations.

Able-bodied subjects were used because the availability of tetraplegics is low due to the low prevalence of tetraplegia.

TABLE II. MOVEMENTS WHICH ARE CANDIDATES FOR MISCLASSIFICATION (DISTURBANCES)

No.	Disturbance	Directions	Trials
1	Turn head slowly in one of the directions and back	$(+\varphi,0,0)$, $(-\varphi,0,0)$, $(0,+\vartheta,0)$, $(0,-\vartheta,0)$	4
2	Turn head quickly in one of the directions and slowly back	$(+\varphi,0,0)$, $(-\varphi,0,0)$, $(0,+\vartheta,0)$, $(0,-\vartheta,0)$	4
3	Turn head slowly in one of the directions and quickly back	$(+\varphi,0,0)$, $(-\varphi,0,0)$, $(0,+\vartheta,0)$, $(0,-\vartheta,0)$	4
4	Turn head quickly in one of the directions, then in opposite direction and back to resting position	$(+\varphi,0,0)$, $(-\varphi,0,0)$, $(0,+\vartheta,0)$, $(0,-\vartheta,0)$	4
5	Turn head quickly in one of the diagonal directions and back	$(+\varphi,+\vartheta,+\psi)$, $(-\varphi,+\vartheta,-\psi)$, $(-\varphi,-\vartheta,-\psi)$, $(+\varphi,-\vartheta,+\psi)$	4
6	Turn head quickly in one of the directions and back	$(0,0,+\psi)$, $(0,0,-\psi)$	2
7	Do not move head	-	3

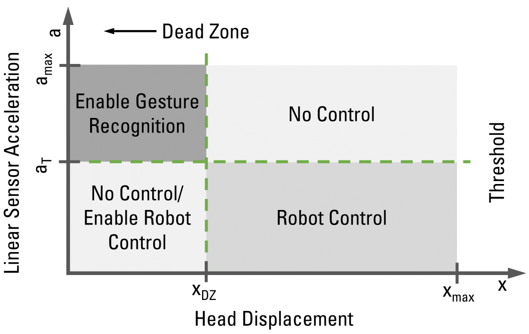

a
Dead Zone
a_max
a_T
Linear Sensor Acceleration
Enable Gesture Recognition
No Control
No Control/ Enable Robot Control
Robot Control
Threshold
x_DZ
x_max
x
Head Displacement

Fig. 4. Gesture recognition mode is entered when the threshold a_T is exceeded within the dead zone. Gesture recognition stops when linear acceleration falls below a_T. Robot control mode is enabled whenever the user's head is inside the dead zone while linear acceleration is below a_T. If he then slowly leaves the dead zone, robot control is carried out.

Furthermore, in this stage of development able-bodied subjects were considered sufficient to provide a proof-of-concept of the proposed algorithm.

B. Model Parameter Adjustment

In order to adjust the model parameters, the subjects repeated each of the four gestures (Fig. 2) for 5 times, resulting in 20 trials. They were instructed to perform the gestures ”quick” without any further description. The resulting data can be regarded as training data as used for supervised learning algorithms. However, signal-oriented models using supervised algorithms usually have to be re-trained occasionally due to signal variation. In contrast, the proposed model is designed in a way that re-training is not necessary. Within this work, we investigate whether the parameters have to be adjusted once per user or once in total.

As a consequence, the parameters were set both user specifically and user independently with data from all the subjects. The thresholds were chosen in a way that all training data with the exception of outliers were included. Outliers were detected by personal inspection of the data. Personal inspection was chosen because the human brain is able to classify data very reliably even though the database is small. Moreover, automation is not required when re-training is not needed. For both user specific and user independent thresholds the mean values of d_{max} and w of each gesture have been used as starting points for curve fitting.

The user independent thresholds are shown in Table I. Choosing appropriate thresholds for relative coupling turned out to be difficult because sometimes coupling was high although from optical inspection of the subjects during the experiments the subjects moved their heads in only one direction while sensor and neck yaw axes seemed to coincide. However, further investigation on coupled motion is ongoing.

C. Testing Data

For testing, the subjects repeated each of the four gestures (Fig. 2) for another 5 times. They were instructed to move the same way like during parameter adjustment. They did not know which parameter values were used for the model. Furthermore, the subjects were instructed to perform 25 predefined non-gestures (disturbances), as described in Table II. These disturbances are typical candidates for misclassification because they are very similar to the gestures and therefore quite hard to separate from them. In total, the test set contained five repetitions for each of the four gestures and in addition 25 disturbances.

IV. RESULTS AND DISCUSSION

User specific thresholds led to an average classification rate of $93.56\% \pm 4.96\%$. At this point, one should acknowledge that no confusion between gestures was encountered. The misclassifications resulted either from gestures which were not detected at all (False Negatives) or from disturbances which were too similar to the definition of a certain gesture (False Positives). This is true for disturbances 1-3. The total misclassification rates of disturbances 1 and 3 were both 5%. Disturbance 2 was misclassified in 42.5% of all cases. If this disturbance turns out to be relevant for real-time control, skewness may be added to the feature space in order to avoid misclassification. Disturbances 4-7 were not misclassified at all. Gestures 1-3 were not detected although they were present in 4% of all cases. Gesture 4 was not detected in 8% of all cases. In general, gestures were not detected whenever they were not performed as defined. As a result, this kind of error could be reduced by expanding the gesture definitions but for the price of a higher misclassification rate. However, the chosen thresholds are considered a reasonable tradeoff. In this context, one has to mention that this tradeoff mainly arises due to the fact that the user just did not always perform the gestures as instructed.

The average classification rate using user independent thresholds was $87.56\% \pm 8.90\%$. Fig. 5 contrasts the results using user specific and user independent thresholds for all the subjects. User independent thresholds led to slightly lower classification rates than user specific ones, but it is likely that

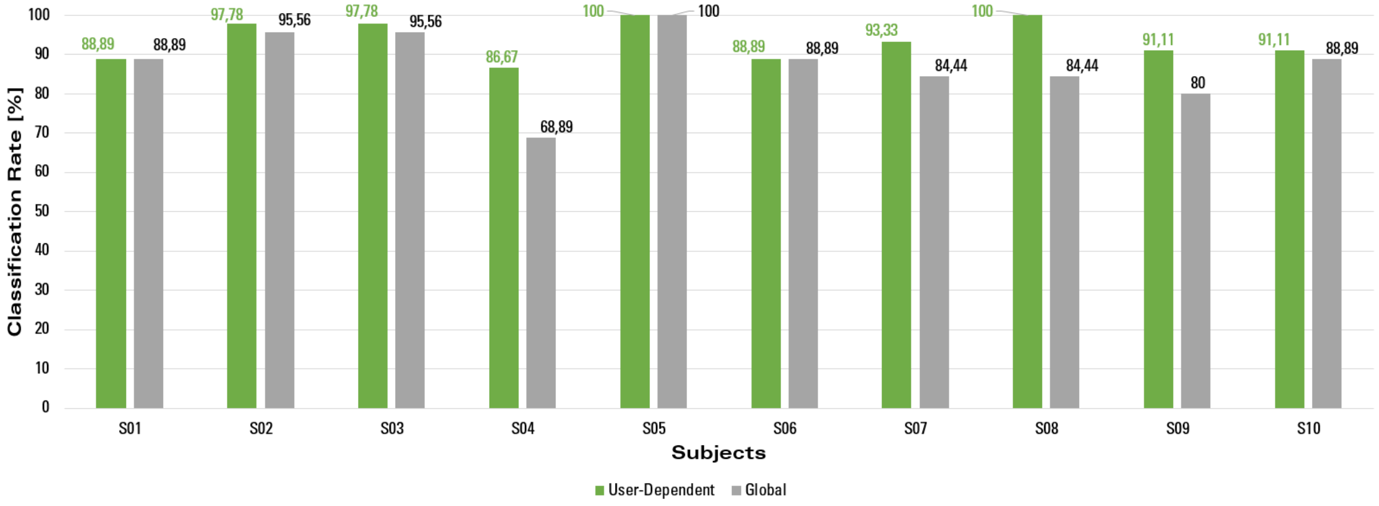


Fig. 5. Classification rates for all the subjects using user-dependent and global thresholds. Global thresholds were obtained with data from all the subjects.

users are able to adapt their movements to the gesture definitions if they do not suffer from neck movement limitations.

V. Conclusion

With the proposed analytical approach, head gestures are identified in a robust way with high accuracy. Therefore, the analytical approach is recommended to be used for detecting switching commands. Misclassifications of different gestures are not present. Such a good class separability can hardly be achieved when using purely signal-oriented ARCs with a comparable amount of data. This hypothesis has been verified during preliminary tests in which standard ARCs using for example k-Nearest Neighbor or Support Vector Machines have been evaluated.

False Positives were classified when a disturbance was too similar to the definition of one of the gestures. As a consequence, misclassification rate can be reduced by narrowing the gesture definitions. For this purpose, users need to learn to perform the gestures with higher reproducibility. As the used five features and the corresponding classification thresholds are interpretable from a physical point of view, they can be used for feedback in order to advise users how to adjust their movements for optimal classification. The possibility of giving meaningful feedback is a major advantage of the analytical approach. Using signal-oriented pattern recognition, the relationship between input signals and classification result is much harder to interpret. As a result, the user can hardly adjust his behavior in a way that classification performance is improved. This is not only true for user specific but also for user independent ARCs.

In contrast, with the presented analytical approach, user-learning will probably lead to a lower inter-subject variability. It is therefore likely that using the same thresholds for all users will not decrease performance significantly. However, future research is directed towards implementing the proposed classification algorithm into the control structure in order to investigate its performance in a real-life scenario.

Acknowledgement

The authors would like to thank Hillcrest Labs for their outstanding support and all the volunteers who took part in the experiments.

This work was funded by the Ministry for Innovation, Science and Research of North Rhine-Westphalia.

References

[1] N. Rudigkeit, M. Gebhard, and A. Gräser, “Towards a user-friendly AHRS-based human-machine interface for a semi-autonomous robot,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2014), Workshop on Assistive Robotics for Individuals with Disabilities: HRI Issues and Beyond*, Chicago, IL, 14 Sep. 2014.

[2] J. K. Aggarwal and M. S. Ryoo, “Human activity analysis: A review,” *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.

[3] A. Bulling, R. Blanke, and B. Schiele, “A tutorial on human activity recognition using body-worn intertial sensors,” *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[4] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, “Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey,” in *Architecture of computing systems (ARCS), 2010 23rd international conference on*. VDE, 2010, pp. 1–10.

[5] K. Altun, B. Barshan, and O. Tunçel, “Comparative study on classifying human activities with miniature inertial and magnetic sensors,” *Pattern Recognition*, vol. 43, no. 10, pp. 3605–3620, 2010.

[6] A. Zinnen, C. Wojek, and B. Schiele, “Multi activity recognition based on bodymodel-derived primitives,” in *Location and Context Awareness*. Springer, 2009, pp. 1–18.

[7] C. Amma, M. Georgi, and T. Schultz, “Airwriting: a wearable handwriting recognition system,” *Personal and ubiquitous computing*, vol. 18, no. 1, pp. 191–203, 2014.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[9] A. Zinnen, U. Blanke, and B. Schiele, “An analysis of sensor-oriented vs. model-based activity recognition,” in *Wearable Computers, 2009. ISWC’09. International Symposium on*. IEEE, 2009, pp. 93–100.

[10] L. Zhang, “Investigation of coupling patterns of the cervical spine,” Master’s thesis, University of Dundee, 2014.

[11] Hillcrest Laboratories, Inc. (2013, Sep.) FSM-9 Data Sheet. [Online]. Available: http://hillcrestlabs.com/