Inertial Motion Sensing Glove for Sign Language Gesture Acquisition and Recognition

Jakub Gałka, *Member, IEEE*, Mariusz Mąsior, Mateusz Zaborski, and Katarzyna Barczewska

***Abstract*— The most popular systems for automatic sign language recognition are based on vision. They are user-friendly, but very sensitive to changes in regard to recording conditions. This paper presents a description of the construction of a more robust system—an accelerometer glove—as well as its application in the recognition of sign language gestures. The basic data regarding inertial motion sensors and the design of the gesture acquisition system as well as project proposals are presented. The evaluation of the solution presents the results of the gesture recognition attempt by using a selected set of sign language gestures with a described method based on Hidden Markov Model (HMM) and parallel HMM approaches. The proposed usage of parallel HMM for sensor-fusion modeling reduced the equal error rate by more than 60%, while preserving 99.75% recognition accuracy.**

***Index Terms*— Inertial motion sensors, gesture analysis, sign language recognition, sensor glove.**

I. Introduction

HAND movement recognition, in its different approaches, has been a topic of research since the early 90s [1], [2]. Regardless of the passage of time, the topic is still relevant [3]–[7], most likely due to the tons of data provided by human limb movement (measured by different devices, such as IoT, CCTV, smart home electronics, etc.).

Researchers are trying to use the human hand as a precise controller of electronic devices. There are domains where this method of movement acquisition is in high demand [8]. The first example concerns medicine. For young interns, the possibility of surgery simulation, including hand movements, would be a valuable experience [9]. The second example, which is directly connected to the topic of this work, is sign language recognition. Deaf people, in order to maintain their independence, must be able to communicate with other people and interact with consumer devices. An efficient gesture recognition system, when correctly used, could improve their quality of life.

The most popular sign language recognition systems use contactless RGB cameras and image processing, as user movement is then not limited by any gear or additional equipment. Vision-based approaches allow for up to 95% correct recognition of sign language gestures [1], [10]. In the context of evaluation scenarios and testing conditions, the accuracy is reasonably high, although it is not enough in cases where a highly reliable and robust system is needed. Inertial and orientation sensors such as accelerometers, magnetometers, or gyroscopes are highly efficient. These devices are not influenced by environmental conditions such as illumination or the background, which are usually problematic in vision systems. Those sensors also allow for relatively easy acquisition of parameters which are hard to obtain in vision systems, such as hand shape or forward/backward movement (related to the image depth axis).

Inertial-based systems also have drawbacks. The sensors are mounted on the entire upper limb, which often introduces limitation in hand movement. Additionally, when using a wired solution, the user's freedom of movement may be limited. However, even with such drawbacks, a device based on inertial sensors could be employed in the first stages of a system project, e.g. in data acquisition support, gesture training, or as a validator of vision data. In this case, the use of a hybrid system should be considered, wherein the inertial sensor measurements are treated as support for the simultaneously acquired vision data.

II. Inertial Motion Sensors

Researchers have developed a multitude of different solutions based on diverse sensors. Some of them are commercialized, but there is no widespread and integrated solution used in gesture recognition yet. Many solutions employ only flex sensors, which are used mainly for hand movement or hand posture acquisition. One of such solutions for hand posture acquisition is presented in [11], where the sensing glove was equipped with fiber Bragg gratings sensors, which allowed for the measurement of finger bending. Another solution used an accelerometer sensor wristband to capture arm movements [12], [13]. In these works, the detected arm movements were used as an additional data stream along with the body joint positions extracted from the depth-image for fusion-based gesture recognition. This approach, however, used only one sensor and did not allow for hand and finger posture analysis required in more complicated sign language gesture modeling.

A sensor-based solution was employed in the presented project. The idea of the authors is to create a sensor glove, which will be used as an additional synchronous input in a hybrid system along with an RGB camera for sign language

Manuscript received April 11, 2016; revised June 6, 2016; accepted June 7, 2016. Date of publication June 22, 2016; date of current version July 18, 2016. This work was supported by the Polish National Centre for Research and Development through the Applied Research Program entitled Virtual Sign Language Translator under Grant PBS2/B3/21/2013. The associate editor coordinating the review of this paper and approving it for publication was Dr. Roozbeh Jafari. *(Corresponding author: Jakub Gałka.)*

J. Gałka, M. Mąsior, and M. Zaborski are with the Department of Electronics, AGH University of Science and Technology, 30-059 Kraków, al. Mickiewicza 30, Poland (e-mail: jgalka@agh.edu.pl; masior@agh.edu.pl; matzab@student.agh.edu.pl).

K. Barczewska is with the Department of Automatics and Biomedical Engineering, AGH University of Science and Technology, 30-059 Kraków, al. Mickiewicza 30, Poland (e-mail: kbarczew@agh.edu.pl).

Digital Object Identifier 10.1109/JSEN.2016.2583542

hand-gesture recognition. The glove tracks the movement of the entire upper limb due to an additional accelerometer placed on the arm, as well as the movement of each of the fingers, allowing for precise hand-gesture modeling and recognition.

As part of the design of the Accelerometer Glove, the designers want to be able to acquire an exact model of limb movement. There are several criteria the device must satisfy. The first one was sufficient number of sensors. Each sensor requires its own communication line. A significant amount of sensors requires a significant amount of communication lines, which creates more connections on printed circuit boards (PCBs). The second criterion concerns the surface of the PCBs. It should be as small as possible to avoid limb movement limitations.

Sign language introduces additional ergonomic requirements. A sign language user needs total freedom of movement in each direction for every joint of their upper limbs. Sign language gestures are highly dynamic and complex, with the signer moving their arms and fingers at the same time [14]. This simultaneous, complex movement is especially difficult to obtain when using solely vision systems [2].

From the user's point of view, an inertial system set up on the upper limb is less comfortable than using a contactless camera-based visual solution. However, in order to obtain information regarding precise limb movement, only the sensors placed on the wrist and fingers can be used. These sensors acquire information regarding the general arm movement and hand shape dynamics. The hand shape data are considered crucial information in the case of the more complicated gestures involving multiple rotations or joint bends, which are elusive for vision-based systems.

III. The Architecture of the System

In order to have a mechanically accurate model of an upper limb, a very complex model should be considered: seven degrees of freedom should be assumed for three joints in the upper limb (glenohumeral, elbow, and wrist) [15], and 23 degrees of freedom distal to the wrist [16], which results in a total of 30 degrees of freedom. To copy such an exact model, the use of at least 30 sensors should be considered. Not all information provided by such a big set of sensors would be needed in the recognition process. There are anatomical points whose behavior is more distinctive than others, so the number of sensors could be significantly limited. The glove made by the authors covers the most important degrees of freedom, as the arm, wrist, and fingers are all monitored. Particular parts of the upper limb are connected, which allows e.g. to estimate the position of the elbow, or the finger's proximal interphalangeal joint. The number of sensors used allows for sufficient hand-posture and gesture modeling, even for complicated sign-language gestures.

A. Hardware Description

The Accelerometer Glove is a device designed for sign language users. The device has seven active sensors, with five located on the fingers (one sensor on each finger), one on the wrist, and one on the arm (Fig. 1). Each of them is a 3-axis acceleration sensor. The device has a modular structure: the designed PCBs have zero insertion force (ZIF) sockets. The PCBs are connected by using ZIF connectors. Each sensor is on a separate board. All of the boards are electrically connected to the educational kit, which contains a microcontroller. The microcontroller manages measurement acquisition and streaming. The measurement is then sent through the universal serial bus (USB) to the PC and received via a graphical user interface (GUI). The entire acquisition system configuration is shown in Fig. 2.

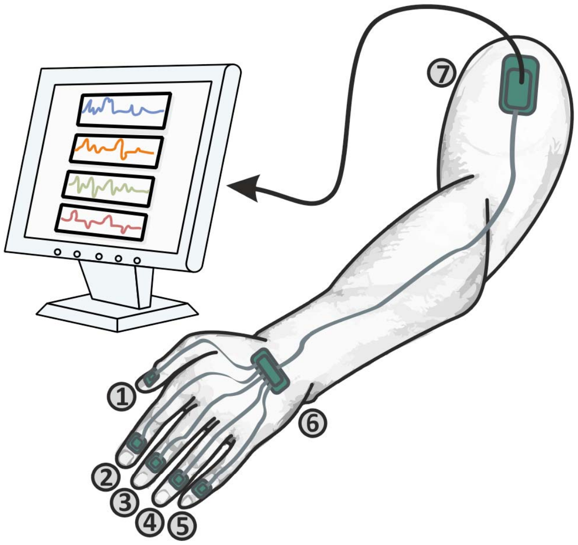


Fig. 1. Motion sensor placement.

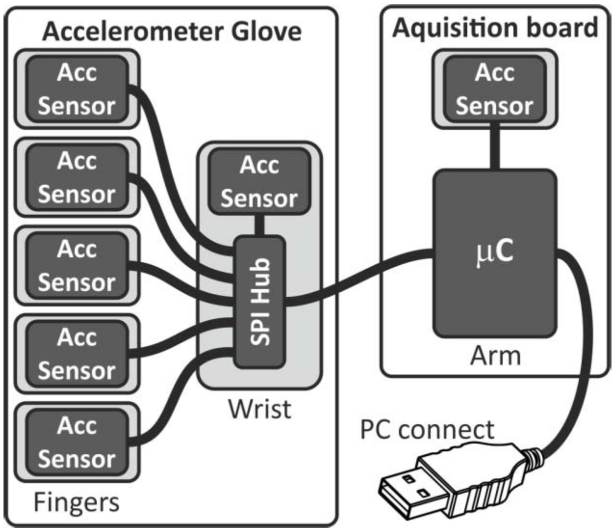


Fig. 2. Acquisition system component configuration.

Modularity allows for a fast relocation of sensors in case if a sensor is broken, or if the user wants to test another type of sensor. It is a rarely occurring solution in glove-type devices: in most gloves, the positions of the sensors are fixed.

The small sizes of the designed boards and proper attachment to the user's body cause the glove to not limit the freedom of movement. This is why sensor glove usage is not exhausting for the user during longer sessions.

B. Measurement Resolution

The manufacturer guarantees 10-bit resolution for the acceleration sensors, which translates to 1024 recognizable states. The transmission line length caused by the length of the human arm results in a high-frequency noise in the system. The highest values of noise reveal that the actual resolution is 7-bit (128 recognizable states). Another experiment, which tested the Signal to Noise Ratio (SNR), proved that the value of the SNR is 40 dB for most sensors (noise level was estimated as an average signal, measured with stationary glove in different positions). Both parameters indicate that the value of noise is 1/100 the value of the signal. If such reasoning is followed, then: 1024/100 ~10 – uncertain states in the 10-bit resolution range. 3 lowest bits: $2^3 = 8$ states. It can be therefore assumed that in most situations, the 3 lowest bits contain noise values. This confirms the resolution being 7-bit.

C. Signal Acquisition and Processing

All of the sensors employed in the Accelerometer Glove are 3-axis accelerometers. Each sensor is connected to a microcontroller by using the Serial Peripheral Interface (SPI) Bus. Data is acquired from the sensors synchronously. Following data collection, the entire set is sent to the PC through USB. The PC recognizes the device as a Serial Port (due to the Virtual COM Protocol implemented in the microcontroller). Then, the dataflow is intercepted by the GUI for acquisition and further processing.

The sensors are connected to a single SPI Bus. The measurements are collected with a frequency of 400 Hz. Each sensor is queried regarding the data in proper order, and, following a whole cycle, the measurements are sent to the PC. Due to time uncertainty, in a worst case scenario the data may be delayed by 2.5 ms. Such a delay is fully acceptable, because natural upper limb movement, and even rapid movement, is slow enough to be recorded by acceleration sensors.

The saved data is later processed by the PC. After calibration, the data is filtered through a low-pass Hamming-window-based running average digital filter. The length of the filter is 250 ms. This stage of pre-processing ensures the removal of the higher noise frequencies from the signal. The feature vector consists of 3D acceleration measured by each sensor. Prior to gesture modeling, a standardization procedure is applied to all of the features, separately for each gesture. The signals for the gesture "good", measured on the forefinger, are presented in Fig. 3.

IV. Sign Language Gesture Recognition

The authors' model isolates sign language gestures by using Parallel Hidden Markov Models (PaHMM), used at first in Automatic Speech Recognition (ASR) systems and described by [17] and [18] but also successfully adopted in Automatic Sign Language Recognition (ASLR) systems, [19]–[21].

Usually, PaHMM is used for the modeling of sign language gestures in accordance with sign language linguistics, taking into account the parallelism of elements of articulation indicated e.g. by Stokoe [14]. Each PaHMM channel corresponds to a group of features which describes different articulatory elements and is modeled as an independent HMM (Fig. 4).

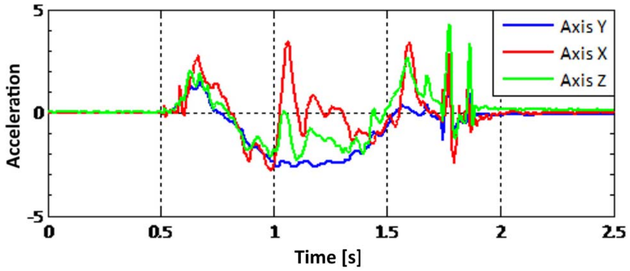


Fig. 3. Acceleration signals for gesture "good", measured on forefinger.

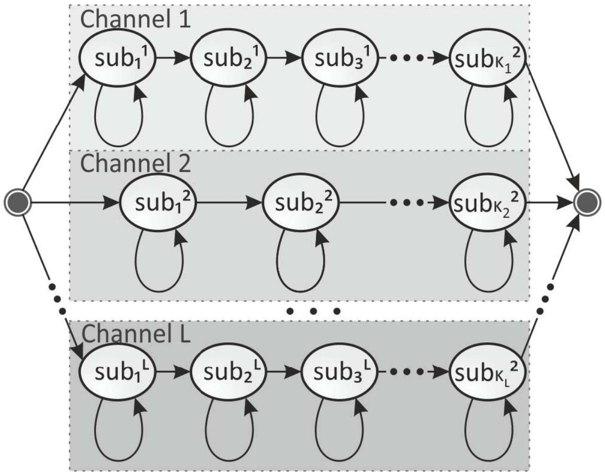


Fig. 4. Scheme of Parallel Hidden Markov Model with independent HMMs for each channel.

In the approach presented in this article, PaHMM channels correspond to multiple sensors attached to the signer's hand. The gesture in each channel is modeled as a sequence of subunits. After taking into account the results of the conducted experiments, a joint-feature HMM has also been attached to PaHMM in a separate channel.

Unlike in [12], where data fusion was performed at feature level and employed full joint-feature modeling only, we adopted another approach where the fusion of different sensor signals is performed at score level. A full joint-feature model is also included as an additional stream. This solution increases the robustness of the system significantly when compared to similar feature-level fusion (joint feature model) only. A comparison between both approaches is discussed later, in the evaluation section of the paper.

A. Recognition Architecture

Each gesture is modeled as a sequence of subunits, which are smaller elements, similar in speech analysis to phonemes. An isolated gesture can be transcribed as

$$gesture : sub_1 sub_2 \ldots sub_K . \qquad (1)$$

The simultaneous character of sign language causes that each independent articulatory element can be described separately in a parallel model. Independent parallel channels can correspond to independent parallel events which happen during signing, as well as to different measurement devices used in the experiment. Thus, gestures can be also

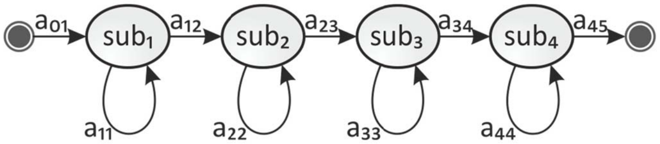


Fig. 5. Four-subunit left-to-right HMM gesture model.

transcribed as

$$gesture : \begin{cases} channel1 : sub_1^1 sub_2^1 \ldots sub_{K_1}^1 \\ channel2 : sub_1^2 sub_2^2 \ldots sub_{K_2}^2 \\ \ldots \\ channelL : sub_1^L sub_2^L \ldots sub_{K_L}^L \end{cases} \quad (2)$$

where L is the number of channels, and K_L the number of subunits in the L-th channel. In this article, the parallel channels correspond to different sensors attached to the signer's hand.

An isolated gesture is represented by a sequence of observations O, which consists of feature vectors o_t observed at each time frame t

$$O = o_1, o_2, \ldots, o_T \quad (3)$$

To recognize a gesture, one needs to find such a gesture model g^* for which

$$g^* = \arg\max_i \left(P\left(g_i|O\right)\right) \quad (4)$$

where $P\left(g_i|O\right)$ is unknown and can be calculated using the Bayes rule

$$P\left(g_i|O\right) = \frac{P\left(O|g_i\right) P\left(g_i\right)}{P\left(O\right)}. \quad (5)$$

$P\left(g_i\right)$ is the prior probability, assumed to be equal for different gestures, and $P\left(O|g_i\right)$ is a generative gesture model likelihood based on the sequence of observations, the probability of which, $P\left(O\right)$, can be calculated with

$$P\left(O\right) = \sum_i P\left(O|g_i\right) P\left(g_i\right). \quad (6)$$

Every subunit in a given channel is modeled as a single-state HMM with a Gaussian Mixture Model (GMM), denoted as $\lambda = \{\mu_i, \Sigma_i, \omega_i\}$, which describes the probability of the observation emission

$$p\left(o|\lambda\right) = \sum_{i=1}^{M} \omega_i p_i\left(o\right), \quad (7)$$

where o is a D-dimensional observed feature vector, ω_i are mixture weights, and M is the number of mixtures. Mean vector μ_i and diagonal covariance matrix Σ_i are parameters of unimodal probability densities $p_i\left(o\right)$.

For a single PaHMM channel, the feature vector is $D = 3$ dimensional, containing acceleration measured in each dimension by a single sensor. An HMM joint model contains signals from all sensors, which means that the feature vector is $D = 21$ dimensional (7 sensors × 3D acceleration).

To model an entire gesture, the models of subunits are sequentially connected into a composite left-to-right HMM (Fig. 5).

B. Gesture Model Training

A gesture model is trained separately in each parallel channel. Initially, the mixture parameters $\{\mu_i, \Sigma_i, \omega_i\}$ are assumed to be global values calculated for the whole training set and are equal in all composite HMM states (*flat start*). At first, GMMs have single components, and their number is incremented by one in each training step by using mixture splitting [22]. The parameters of the model are re-estimated in further training steps by the use of the Baum-Welch algorithm. In each step, the subunit borders are realigned, taking into account the best match of the new model to the observations. The subunits are not synchronized in time between channels, as the synchronization is done by performing a fusion of channel responses at the whole-gesture level.

C. Recognition

The recognition for a single l channel is performed by a token passing algorithm and an analysis of the N-best list which contains log-likelihood values (scores) obtained by each gesture model $g_{l,i}$. For single channels, the test sign is recognized as the one for which the log-likelihood value is the highest

$$g^* = \arg\max_i \left(\log\left(P\left(g_{l,i}|O\right)\right)\right). \quad (8)$$

To include information from different channels, a fusion of their responses is performed. To compare different channels, the scores must be scaled to a similar range by using score normalization. It is performed separately in each channel l, and within each tested sign i

$$score_{l,i} = \frac{\log\left(P\left(g_{l,i}|O\right)\right) - \mu_l}{\sigma_l} \quad (9)$$

where the mean μ_l and variance σ_l^2 values are estimated, taking into account all of the sign model scores from the N-best list, except for the highest one.

Fusion is performed as the weighted sum of normalized channel responses. The sign is recognized as the one for which the weighted sum of normalized scores has the highest value

$$g^* = \arg\max_i \left(\sum_{l=1}^{L} w_l score_{l,i}\right). \quad (10)$$

Weights w_l for different channels are proportional to recognition accuracy Acc_l, independently Acc$_\mathrm{l}$ obtained by a single channel

$$w_l = \frac{Acc_l}{\sum\limits_{r=1}^{L} Acc_r}. \quad (11)$$

V. Recognition Evaluation

A. Gesture Database

The quality of the entire sign language gesture recognition system has been verified on a set of specifically collected gestures, recorded with the designed Accelerometer Glove.

The purpose of creating a database of recordings was to verify recognition efficiency and the possibilities of the interoperability of inertial gesture recognition with a vision system (based on RGB cameras and infra-red depth sensors).

In the case of a recognition system adapted to the purpose of a dialogue system, there is no need for a large dictionary. The cardinality of a dictionary for video gesture recognition systems is often less than 100 [24], [25].

The dictionary gestures were matched to a dialog system from the use case of a deaf person booking a visit to a doctor's office. This decision has conditioned the type of gestures and determined their semantic content. The database contains isolated gestures describing days of the week, months, basic numerals, and names of medical specialties (pediatrician, cardiologist, dermatologist, etc.). Finally, 40 of such gestures were selected, which is an acceptable number for initial research.

It should also be mentioned that the gestures were chosen by taking into account the possibility of proper efficiency verification. The creation of a dictionary ensures that the gesture database covers and involves the entire space of all possible centers of articulation for sign language.

The database contains recorded gestures which either differ in the entire range of movements (shape, direction, and speed), or differ only in a small part of the total motion (e.g. the final movement of the hand, the number of taps or exposed fingers).

This approach to the design of the gesture database allows for reliable and consistent verification of the operation of the system and for the determination of its full applicability for the case of sign language recognition as well as of a supportive data stream for the development of vision-only gesture modeling.

Finally, the created recording database contains 10 repetitions of each of the 40 gestures registered for the 5 signers. This results in a total of 2000 recordings used for the validation of the solution.

The signers were in a sitting position in order to minimize the movement of the entire body, just as in vision-based setups. Each isolated gesture begins and ends in the same position (both hands rested on knees). Recordings were made for each person using a single Accelerometer Glove worn on the dominant hand (Fig. 6).

The gesture recording was divided into several recording sessions (taking place on different days). Gestures were shown in different order and with a maximum of 3 repetitions of the same gesture in a row during the course of a single recording session. This approach was designed to minimize the effect of the signers familiarizing themselves with the gesture and therefore signing it in a similar way.

The recording procedures were designed to proceed as smoothly as possible. Automatic computer software was created for this purpose (Fig. 7). It allows for simultaneous acquisition of the video representations of gestures (recorded from the RGB cameras and depth sensor).

B. Evaluation Procedures

The entire evaluation was performed in compliance with all required practices for the validation of algorithms in pattern recognition problems.

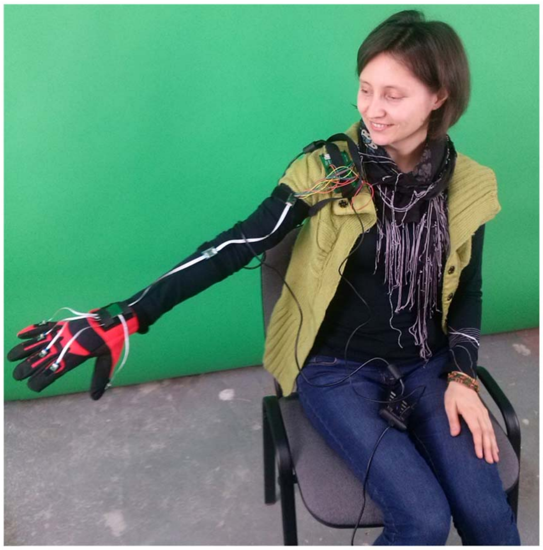
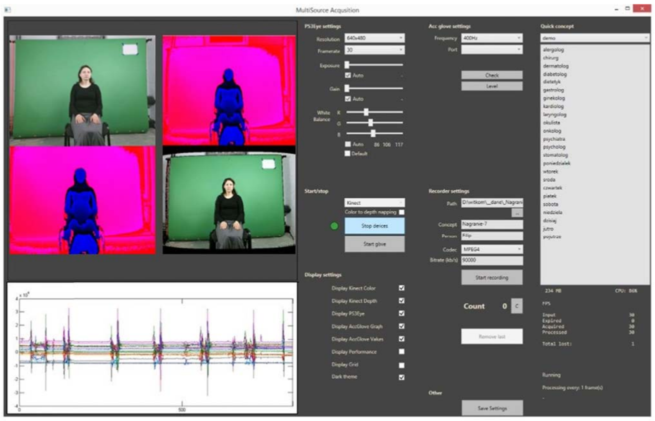


Fig. 6. Recording of example sign for evaluation corpus.



MultiSource Acquisition

PS3Eye settings
Resolution 640x480
Framerate 30
Exposure ☑ Auto
Gain ☑ Auto
White Balance R G B □ Auto 86 106 117
□ Default

Acc glove settings
Frequency 400Hz
Port

Quick concept
demo
alergolog
chirurg
dermatolog
diabetolog
dietetyk
gastrolog
ginekolog
kardiolog
laryngolog
okulista
onkolog
psychiatra
psycholog
stomatolog
poniedzialek
wtorek
sroda
czwartek
piatek
sobota
niedziela
dzisiaj
jutro
pojutrze

Start/stop
Kinect
Color to depth napping □

Recorder settings
Path D:\witkom_dane_Nagran
Concept Nagranie-7
Person Filip
Codec MPEG4
Bitrate (kb/s) 90000

Display settings
Display Kinect Color ☑
Display Kinect Depth ☑
Display PS3Eye ☑
Display AccGlove Graph ☑
Display AccGlove Values ☑
Display Performance □
Display Grid □
Dark theme ☑

Count 0

234 MB CPU: 86%
FPS
Input 30
Expired 0
Acquired 30
Processed 30
Total lost: 1

Running
Processing every: 1 frame(s)

Fig. 7. User interface of the multisource acquisition application. Sensor glove data visible in the bottom-left corner of the screen.

TABLE I
CLASSIFICATION RESULTS (IN %) FOR THE ACCELEROMETER DATABASE

Recognition	Acc	EER	F_1	Precision	Recall
HMM	99,75	1,00	98,56	98,61	98,50
PaHMM	99,75	0,50	99,76	99,77	99,75

As the scheme of 5-fold cross-validation was adopted, the gesture records were divided into training and test sets in the ratio of 80% – 20% in each validation. This division was made ensuring that the training and test sets consisted of gestures made during different recording sessions and with different people.

C. Evaluation Results

An experiment was conducted in accordance with the described evaluation procedure, training, and recognition scenarios. The recognition performance is presented in Table I. It contains recognition accuracy (Acc), equal error rate (EER), F_1 score, precision, and recall. The results also present the

TABLE II

CLASSIFICATION RESULTS (IN %) FOR THE VIDEO DATABASE

Recognition	Acc	EER	F_1	Precision	Recall
HMM	94,68	3,55	93,09	93,27	92,90
PaHMM	92,26	2,76	92,72	93,20	92,26

TABLE III

CLASSIFICATION RESULTS (IN %) OBTAINED FOR SEPARATE SENSORS COMPARED TO PaHMM AND JOINT-FEATURE HMM APPROACHES (SENSOR NUMBERS AS IN FIG. 1)

Recognition	Acc	EER	F_1	Precision	Recall
Sensor 1	90,50	3,72	85,99	88,91	83,25
Sensor 2	94,25	3,75	82,56	86,77	78,75
Sensor 3	**96,50**	3,75	89,89	92,14	87,75
Sensor 4	94,25	3,83	87,56	91,73	83,75
Sensor 5	76,25	4,73	75,17	76,11	74,25
Sensor 6	95,25	**2,50**	**93,78**	**95,09**	**92,50**
Sensor 7	77,25	10,96	71,19	73,52	69,00
HMM	99,75	1,27	98,56	98,61	98,50
PaHMM	**99,75**	**0,50**	**99,76**	**99,77**	**99,75**

recognition evaluation for the joint-feature HMM as a reference for the effectiveness of the PaHMM.

Recognition accuracy, precision, and recall are very high, while the EER is very small, which demonstrates the effectiveness of an accelerometer-based system for the recognition of isolated sign language gestures.

Using the PaHMM approach leads to an even lower value of EER in comparison to the joint-feature HMM, which achieves the same recognition accuracy. Low EER is particularly important for end-user applications with low false acceptance and false rejection rates. For comparison, the results obtained for a similar database, containing less gestures (31), but created for video-based recognition methods, are presented in Table II.

As can be observed, all of the parameters of the vision-based method are worse: lower accuracy, precision, and recall, and higher EER, even though less gestures were used in the experiment.

Because of the very high efficiency of recognition performed with fractures from all accelerometer sensors, the possibility of recognition of individual sensors in relation to the overall processing structure is worth taking into consideration. Table III presents the recognition performance achieved by separate sensors in comparison to the PaHMM and the joint-feature HMM.

The results obtained for the inertial features of separate sensors are worse than the results for the PaHMM and HMM approaches, which is also illustrated in the Detection Error Trade-off and precision-recall plots in Fig. 8 and Fig. 9 respectively. However, it can be observed that the best single sensors in terms of accuracy (Sensor 3 – middle finger) and in terms of EER (Sensor 6 – wrist) achieve significantly better results than by using the vision-based method.

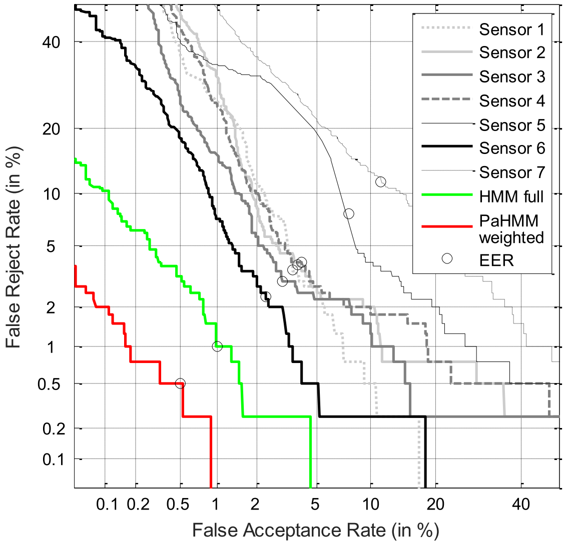


Fig. 8. DET plot for accelerometer features for joint-feature HMM, PaHMM, and recognition for separate sensors (Sensor numbers as in Fig. 1).

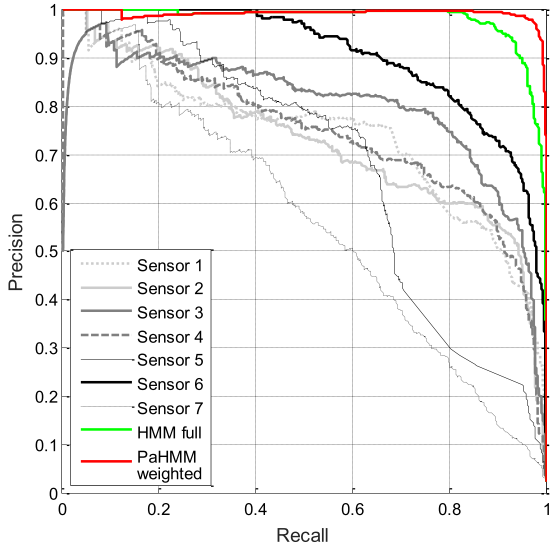


Fig. 9. Precision-recall plot for accelerometer joint-feature HMM, PaHMM, and recognition for separate sensors (Sensor numbers as in Fig. 1).

VI. CONCLUSION

The evaluation results of the described acquisition system and sign language gesture recognition, using accelerometer sensors, clearly shows that such an approach can result in an extremely high efficiency of recognition. The efficiency is much higher than in systems based solely on video sensors.

The very high stability and resistance to different conditions and variances of recording gestures, as well as the differences caused by recording different people (EER of 0.5%), leads to the conclusion that using inertial motion sensors may result in very high recognition confidence and robustness to data variability.

It would obviously be quite difficult and inconvenient to use this approach (the Accelerometer Glove) with dialog systems for deaf people. The vision system, as a contactless approach, is more convenient, and does not require any special sensors or devices for acquisition, except for an RGB camera.

The presented system can be used to improve the effectiveness of vision systems – e.g. at the stage of gesture model training. Temporal models (such as the HMM) can be assumed to be better suited to the variability of movement due to the high efficiency of recognition in the case of inertial movement parameters. For this reason, such a model (or even the time division of the gesture into segments) could be used as input information for model training for a system based on RGB cameras or depth sensors. This issue needs further investigation.

Some of the results (Table III, Fig. 8 and Fig. 9) allow for optimism in regards to the ability of using a single inertial sensor for gesture recognition (autonomously or in cooperation with another system). It could be quite efficient and ergonomic to use smartphones or smartwatches in the future.

References

[1] Y. Wu and T. S. Huang, "Vision-based gesture recognition: A review," in *Gesture-Based Communication in Human-Computer Interaction* (Lecture Notes in Computer Science), vol. 1739. Berlin, Germany: Springer, 1999, pp. 103–115.

[2] D. J. Sturman and D. Zeltzer, "A survey of glove-based input," *IEEE Comput. Graph. Appl.*, vol. 14, no. 1, pp. 30–39, Jan. 1994.

[3] H. Teleb and G. Chang, "Data glove integration with 3D virtual environments," in *Proc. ICSAI*, 2012, pp. 107–112.

[4] H. Zhou, H. Hu, N. D. Harris, and J. Hammerton, "Applications of wearable inertial sensors in estimation of upper limb movements," *Biomed. Signal Process. Control*, vol. 1, no. 1, pp. 22–32, 2006.

[5] Z. Lu, X. Chen, Q. Li, X. Zhang, and P. Zhou, "A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 2, pp. 293–299, Apr. 2014.

[6] S. Zhou *et al.*, "2D human gesture tracking and recognition by the fusion of MEMS inertial and vision sensors," *IEEE Sensors J.*, vol. 14, no. 4, pp. 1160–1170, Apr. 2014.

[7] Y.-C. Kan and C.-K. Chen, "A wearable inertial sensor node for body motion analysis," *IEEE Sensors J.*, vol. 12, no. 3, pp. 651–657, Mar. 2012.

[8] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE Sensors J.*, vol. 15, no. 3, pp. 1321–1330, Mar. 2015.

[9] R. C. King, L. Atallah, B. P. L. Lo, and G.-Z. Yang, "Development of a wireless sensor glove for surgical skills assessment," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 5, pp. 673–679, Sep. 2009.

[10] S. C. W. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 873–891, Jun. 2005.

[11] A. F. da Silva, A. F. Gonçalves, P. M. Mendes, and J. H. Correia, "FBG sensing glove for monitoring hand posture," *IEEE Sensors J.*, vol. 11, no. 10, pp. 2442–2448, Oct. 2011.

[12] K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of inertial and depth sensor data for robust hand gesture recognition," *IEEE Sensors J.*, vol. 14, no. 6, pp. 1898–1903, Jun. 2014.

[13] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors J.*, vol. 16, no. 3, pp. 773–781, Feb. 2016.

[14] W. C. Stokoe, Jr., "Sign language structure: An outline of the visual communication systems of the American deaf," *J. Deaf Stud. Deaf Edu.*, vol. 10, no. 1, pp. 3–37, 1960.

[15] E. J. Dijkstra, "Upper limb project, modeling of the upper limb," Dept. Eng. Technol., Univ. Twente, Enschede, The Netherlands, Tech. Rep. s0142395, Dec. 2010.

[16] D. J. Sturman, "Whole-hand input," Ph.D. dissertation, Media Arts Sciences Section, School Archit. Planning, Massachusetts Inst. Technol., Cambridge, MA, USA, 1992.

[17] H. Bourlard and S. Dupont, "A mew ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. IEEE ICSLP*, Philadelphia, PA, USA, Oct. 1996, pp. 426–429.

[18] H. Bourlard and S. Dupont, "Subband-based speech recognition," in *Proc. IEEE ICASSP*, Munich, Germany, Apr. 1997, pp. 1251–1254.

[19] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of American sign language," *Comput. Vis. Image Understand.*, vol. 81, no. 3, pp. 358–384, Mar. 2001.

[20] U. von Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss, "Recent developments in visual sign language recognition," *Universal Access Inf. Soc.*, vol. 6, no. 4, pp. 323–362, Feb. 2008.

[21] S. Theodorakis, V. Pitsikalis, and P. Maragos, "Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition," *Image Vis. Comput.*, vol. 32, no. 8, pp. 533–549, Aug. 2014.

[22] S. Young *et al.*, *The HTK Book (for HTK Version 3.4)*. Cambridge, U.K.: Eng. Dept. Cambridge Univ., 2006.

[23] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London, U.K.: Prentice-Hall, 1982.

[24] S. Bilal, R. Akmeliawati, A. A. Shafie, and M. J. E. Salami, "Hidden Markov model for human to computer interaction: A study on human hand gesture recognition," *Artif. Intell. Rev.*, vol. 40, no. 4, pp. 495–516, Dec. 2013.

[25] S. G. M. Almeida, F. G. Guimarães, and J. A. Ramírez, "Feature extraction in Brazilian sign language recognition based on phonological structure and using RGB-D sensors," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7259–7271, 2014.

[26] M. W. Kadous, "Temporal classification: Extending the classification paradigm to multivariate time series," Ph.D. dissertation, School Comput. Sci. Eng., Univ. New South Wales, Kensington, NSW, Australia, 2002.

Jakub Gałka (M'14) received the M.Sc. and Ph.D. degrees in telecommunications and electronic engineering from the AGH University of Science and Technology, Kraków, Poland, in 2003 and 2008, respectively. He has been with the Department of Electronics, AGH University of Science and Technology, where he is currently a Researcher and a Lecturer. In terms of his work, he was involved in several Polish and European research projects related to speech and audio processing. His research focus lies in speech and language processing and recognition, speaker recognition, multimedia signal processing, and data analysis. He is working on the development of commercially available ASR and speaker verification systems.

Mariusz Mąsior received the M.Sc. and Engineering degrees in telecommunications and electronic engineering from the AGH University of Science and Technology, Kraków, Poland, in 2010. He has been an Assistant Professor, a Lecturer, and a member of the Signal Processing Group, Department of Electronics, AGH University of Science and Technology. He specializes in signal processing, speech technology, embedded systems, and systems engineering.

Mateusz Zaborski received the Engineering degree from the Department of Electronics, AGH University of Science and Technology, Kraków, Poland, where he is continuing his education. His scientific interests concentrate in the areas of electronic design and sensors data acquisition.

Katarzyna Barczewska received the M.Sc. degree in biomedical engineering from the AGH University of Science and Technology, Kraków, Poland, in 2011. She is currently pursuing the Ph.D. degree with the Department of Automatics and Biomedical Engineering, AGH University of Science and Technology. Her research interests include statistical learning, machine learning, and gesture recognition.