Tomasz DZIUBICH, Julian SZYMAŃSKI

Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland

Automatic Classification of Polish Sign Language Words

Abstract. *In the article we present the approach to automatic recognition of hand gestures using eGlove device. We present the research results of the system for detection and classification of static and dynamic words of Polish language. The results indicate the usage of eGlove allows to gain good recognition quality that additionally can be improved using additional data sources such as RGB cameras.*

Streszczenie. *W artykule przedstawiono podejście do automatycznego rozpoznawania gestów migowych w oparciu o dedykowane do tego zadania urządzenie pod nazwą eGlove. Przeprowadzono analizę podejść do klasyfikacji gestów statycznych i dynamicznych. Uzyskane rezultaty wskazują, że opracowane urządzenie może zostać wykorzystane do analizy gestów języka mówionego, jednakże dla gestów dynamicznych ograniczeniem jest rozmiar słownika. (**Automatyczna klasyfikacja znaków Polskiego Języka Miganego**).*

Keywords: gesture recognition, wearable computing, classification, human – machine interaction.
Słowa kluczowe: rozpoznawanie gestów, elektronika ubieralna, klasyfikacja, interakcja człowiek – maszyna.

doi:10.12915/pe.2014.03.43

Introduction

In the domain of Human - Computer communication it is selected six typical methods of interaction: command language, natural language, menu selection, form filling, direct manipulation, and anthropomorphic interfaces [11]. All of these methods require the medium for transmission of information between human and machine [4]. The most popular is a graphic display that is typically related to the input-output devices: mouse and keyboard but microphone and cameras are also in use. As the domain of ubiquitous computing [18] grows there also increases the number of additional communication devices that can be used for building alternative ways for interaction with the machines. They are usually named Natural User Interfaces (NUI) [15]. They are based on monitoring human activities and extracting some information that is passed to the machine where it is interpreted.

A lot of electronics can now be accomplished to human activities, such as: smart watches (e.g. iWatch), bracelets (e.e Jawbone Up), eyewear (e.g. Star XL Vuzix) that are named together wearable computing [14]. Accessories incorporating advanced electronic technologies into clothing allow to build applications for monitoring and real time feedback in everyday life. Also other areas of communication channels are used: electroencephalography is used for brain signals analysis and build interfaces that allow to control electronic devices with thoughts [20]. Eye tracking devices [5] are gaining increasing popularity and they are widely added to modern smartphones and eBook readers. Some subset of NUI is Tangible User Interface (TUI) in which a person interacts with digital information through the physical environment/things [3]. This leads to spatial user interfaces (SUI) in which there are spatial input and 3D output, with an emphasis on the issues around interaction between humans and computer systems [19].

In our work in the domain of natural computing we built a interface based on hand glow accomplished with accelerometers and magnetometers that allows to monitor position of a hand. There is wide range of applications were the device can be employed: computer games, interaction between human and intelligent space, monitoring the changes of hand placement for medical analysis, recognition of hand gestures etc.

The most profitable area of applications are games. The sales on game consoles market such as: Xbox360, PlayStation, Wii, Nintendo are now mainly increased by accomplishing them with additional hardware that augments interaction between gamers and consoles. The most popular devices are MS Kinect, PlayStation Move, Wii Balance Board make games fare more attractive than classical games, thus instantly new devices are introduced. The accomplishing the consoles with eGlove seems to be nice another way for making them more attractive.

In this work we focus on gestures recognition that allows to analyze wide spectrum of it's possible applications. Research related to controlling machines with gestures is subject of extensive studies for three decades [16]. Gestures are very natural way of expressing messages for humans, thus introduction this separate medium into human-machine interactions considerably support lexical communication (based on spoken and written words). There is wide range of examples of gestures that cannot be easily replaced with single words e.g. pointing some objects in the space, as well as strengthening verbal communication e.g.: whisking by hand for waving farewell [2].

Usage of gestures for communication with machines differs from conventional input/output devices in that it operates with metaphors from the real human world not virtual one. In such systems, burden of adaptation to two-way communication is shifted to machine. System is easier to learn and comprehension and through its architecture does not limit human expressivity. The interfaces realize demand of transparency thus is very comfortable for usage. In the literature it is selected two main advantages of such a communication [1]:

- increase of work efficiency trough introduction of multimodal analysis of the human – machine interaction,
- improved ergonomics – without necessity of adaptation to the device, a person is not liable to the injuries related with long being in forced positions.

The gesture interface implementation uses a variety of ideas and input devices. The most popular of these is the analysis of the image obtained with a video camera, and analysis of the signals obtained from the set of sensors placed on human body.

Usage of cameras has several advantages. This kind of interface is completely uncoupled with the user. Second, the cameras are already ubiquitous, allowing for rapid implementation of any new solution. In addition, it is possible to use the knowledge and methods derived from the area of pattern recognition, which includes many other issues not directly related to the recognition of gestures. The use of RGB video has also some limitations. One of the major drawbacks is that there must be placed in a source of light. In order to observe the user in three-dimensional space it is required to use more than one camera and there is a need to consolidate and synchronize images what usually in not easy. Also some additional issues related to the precision of image analysis should be taken into consideration.

Set of sensors is one of the best alternative to usage of RGB cameras. In the form in which they are now found, we cannot call them transparent to the user because they are usually placed on a human body. However, progressive miniaturization will allow in the near future to place sensors into e.g. jewelry or directly on the nails or skin, so the problem will cease to be of importance [8]. The second drawback is the need to the charge of the battery, which in many applications can be a big hurdle. However, this type of interface allows one to analyze in any environmental conditions in three-dimensional space what is not easily achieved using cameras. Also usage of dedicated sensors significantly increases precision of position detection of monitored elements.

While this paper builds on the model presented in [7], the contribution of this paper is as follows:

- usage of single HMM or DTW instead of the hierarchical classifier,
- their integration with the Polish Hand Alphabet as a learning set as in [10, 13],
- evaluation of eGlove device usefulness in static and dynamic gesture recognition.

In this work, as an input interface we used a second type of device. In next chapter we describe our device called eGlove next the test environment and test results are presented respectively. The work ends with evaluation of the obtained results and conclusions.

eGLOVE DEVICE

The device for acquisition of the data from the user was made as an electronic glove and we called it eGlove. The device shown in left hand graph in Fig. 1 has been created at the Gdansk University of Technology. Its main element is a printed circuit board placed on top of the hand. On the board we placed a microcontroller ATmega 128 and Bluetooth (FLC-BTM403) for communication. The device employs 3-axis acceleration sensors, one of which is located directly on the plate, while the others are located at the ends of all fingers. On the board we install the three-axis magnetic field sensor. The exact location of the sensors is shown in Fig. 1 in right hand graph. All sensors are attached to a specially designed cotton glove, by which elements of the device are fixedly held near the hand. The acceleration value for each axis accelerometer is presented in the form of 8-bit signed number representing the acceleration of the range ±2g. The value of each of the three axis magnetometer is a 16-bit unsigned integer.

The microcontroller performs acquisition of the data from sensors with a maximal frequency of 100 samples per second. The eGlove software allows one to remotely set its two parameters:

The first is the number of packets with samples sent per second via Bluetooth; valid values are: 10, 20, 25, 50 and 100 samples per second.

The second parameter is the number of past samples to be taken into account by averaging filter. Valid values are 2, 4, 8, 16, 32, 64 and 128 samples. For example the parameter settings of up successively with 25 and 4 will be obtained every 40 ms data packet with the values averaged last 4 samples that were collected from the sensors at intervals of 10 ms.

Our accuracy evaluation is based on ATmega 128L chip, due to its popularity and ease of use. The board has a possibility of attaching various sensors. The main limitation is using I2C or SPI interface. In this work, we use three-axis accelerometer from VTI devices, CMA3000-D1. The accelerometer has a range of -8g to 8g and noise below 3.5mg when operating at 400Hz . But this mode is energy-consuming. In our work we use range 2g and in addition motion detection mode. The second type of sensor is MMC312xMQ, a tri-axis magnetic sensor with I2C interface, with range from -2g to 2g. and low power consumption. On the board are placed Li-On battery and charger-management controller (MCM73781). The whole device consumes about 120 mW (8MHz, 3V and 50 measurements per second). The board can send the acceleration data through Bluetooth 2.0 to a PC in real time. We implement eGlove software for Windows PC using Visual C# and Android 3.2.

Fig.1. eGlove device (left) and placement of the sensors (right)

eGlove gives out recognition result without perceptible delay in our experiments based on PCs. We measured the speed of eGlove implemented in C on multiple platforms. On PC with i5 650 3.2 GHz, it takes less than 1 ms for a template library of twelve gestures. On Asus TF101 (tablet) with Android 3.2 and NVIDIA® Tegra™ 2 1.0 GHz dual-core processor, it takes about 3 ms for the same vocabulary. Such latencies are too short to be perceptible to human users.

Data analysis flow

Data processing in our system that is dedicated to recognize gestures (described in next section) starts with obtaining a raw data from the board. They are normalized and filtered, unnecessary data are removed. The quantization is made in the degree that minimize introduced at this level distortions. Performed quantization is made in such a degree that it does not negatively impact the quality of further processing (in case described in this paper - gesture recognition). As a result we obtain reduced and filtered data. Thus demand for computing resources in the later stages is much lower as well as filtered data positively affect the results of classification.

The proposed approach extends the method used e.g.: in uWave system [12], that operates directly on a raw data from the accelerometer. At the stage of pre-processing we use averaging filter, in order to smooth out minor swings. They usually are noise or minor, unintentional movements that we eliminate to improve the computational representation of the eGlove position. In addition, the data were transformed from the direct acceleration values given in the form of a floating point numbers, into 32 discrete levels using the non-linear function.

The extraction characteristic of the data during its processing creates parameter vector (feature vector) that describes eGlow current position that is more compact method than others, e.g. one introduced by Vieriu [17] that consists of approximate contour of a hand given as an image created from a series of segments that are straight lines. As a result of our approach we obtain set of data that precisely describe required aspects of the device and in that form they can be used directly as an input data to the classification algorithm. Final step enables the feature vector normalization and standardization.

Classification is the final stage of the processing. The input is a feature vector of a gesture, output - identifier that describes the class to which the gesture has been assigned. To enable the classification of the predefined

gestures we built a set of standardized gestures. Depending on the classification method, they may be directly compared with gesture representation collected from eGlove (DTW, Dynamic Time Wrapping) or the classification is performed after acquisition of all signals and then compared with internal representation of sign (other methods). As the result information about detection of particular gesture is provided. Alternatively, it may have additional attributes such as certainty of the classification rate or the rate of movement.

Gesture dictionaries

The eGlove can be used as an alternative interface for human-machine communication. The most important issue for its wide usage is evaluation of spectrum of gestures that can be analyzed and thus properly interpreted by the machine. For that purpose we decide to perform the experiments that allow to recognize Polish Hand Gestures described in [6].

In our system we use the dictionary containing both the finger alphabet and basic set of fingerprints poses for the sign language. Polish Hand Alphabet consists of a subset of hand poses. It is very similar to International Hand Alphabet (Polish fingerprint 'A' equals International 'Am'). The missing letter (e.g. ą, ć) is obtained by adding the element of movement to some of them. It contains 48 sings. Classifier for all poses could be easily extended to detect only simple movements, so as to distinguish the complete alphabet. In addition, the effective detection of all poses may provide a great foundation for future recognition systems complete set of dynamic gestures. Thus we decide to use the dictionary containing a set of all 48 poses.

The dynamic signs are more complex. Due to their large number that is equal to the number of the words in a language we use only their limited subset. To evaluate the ability of recognition dynamic signs we select 25 test gestures. The decision about which gesture is suitable for learning set was made based on video materials made available by the Internet TV ONSI.tv for deaf-mute people within PJM online dictionary [9]. At the very beginning we select these gestures which used only in the right hand. Then we select the final set of 25 of them. This was performed by a set of gestures were both similar to and strongly different ones are put, according to the authors of judgment. Such a dataset allows to examine gestures classifiers both in terms of easier and more difficult cases.

Evaluation datasets and test procedure

Each of the gestures from our dictionaries need to described by a sufficient number of recordings, hereinafter referred to as samples. Some of them were used to train classifier (80%), the rest of the validation (20%). It was decided that each gesture has created a static 25 samples. For a dictionary consisting of a total of 48 gestures (fingerspelling) were created so a total of 1200 records. In the case of dynamic gestures created 50 samples per gesture, giving a total of 1250 records. Differentiation of the number of records between static and dynamic gestures has several issues. The first is the fact that the latter contain much more information, and thus, individual samples can be much more distributed. As a result, dynamic gesture recognition algorithms require more samples of training and verification. The second reason is the large number of samples generated by a single static gesture recording.

In order to maintain consistent terminology, in this work we use the word sample to describe the entire recording both static gestures and dynamic. Each sample is recorded as a sequence of frames (a single measurement). While in the case of dynamic gestures recognized the whole sequence, whereas the static gesture is necessary to determine how to deal with multiple frames. We consider the choice of one representative frame and usage each of recording frames as a separate sample. Selection of representative frames is an additional issue that should be taken into consideration and may provide some noise to the analyzed data. E.g. in case as during the recording a hand is put in wrong position it is a risk that it can be selected, but this frame will not reflect the characteristics of the palm. Therefore it was decided that each frame is treated as a separate sample. In the consequence we obtain relatively large samples sizes for each of static gesture.

In the case of classification of sample records, as a result will be a class that gain the majority of frames from the video. As during the time of signs expression they may change even if performed by same person, it could also influence the quality of the classification [12]. To take this into consideration it was decided to dissipate during the recording sessions gestures. It was decided that a total of five sessions will be conducted in two-day intervals. This means that when one of them will be recorded after 5 samples for static gestures and 10 for dynamic.

Polish gestures of Sign Language are difficult to be correctly processed. If they are performed by the unfamiliar with the sign language person it do not produce reproducibility, gained while a people who naturally use these language. To avoid distortion of the results, in order to produce recordings of dynamic gestures we use the assistance of a person related with the deaf community. A static gestures we considered as simple enough to perform by ourselves.

For a static gestures we use following procedure: put the hand in eGlove in the position describing the gesture, start recording (using the other hand), wait about 3 seconds, stop recording using the other hand.

In case of dynamic gesture recording procedure was as follows: hand placement in a starting position, start recording using the other hand, making a gesture, a return to the starting position, stop recording using the other hand.

For recording, static gestures eGlove was configured to transmit 25 frames per second, averaging 8 frames. Frames should be assumed to be similar to each other, through the use of averaging can thus alleviate any interference. In the case of dynamic gestures will be essential to accurately capture the details of the movement. It was therefore decided that these parameters have values of 100 and 4. Each of recordings is stored in the database where it is described with identifier, gesture name, session information, and number of the recording within the session.

We have conducted six tests. Tests 1 - 5 have been used for creation a set of learning samples of each gesture with one session and classify all other of this and other sessions. In each of these tests, samples were selected using following procedure: Test no. 1, broke down into days of recording, allowed to verify the effectiveness of the change detection over time. All tests in this area have been used for selection of samples that determine the impact on the quality of classification during the learning phase. In the 6th set samples for creating the classifier were selected after several recording every gesture of every day. They were re-classified for all other recordings. This test allows to determine the effectiveness of a scenario in which the classifier was trained using full set of samples from different days. Theoretically, the effectiveness of recognition should be the best.

All tests described above are common for all evaluated classification methods as they are used during their learning and testing phase. In case of some classification algorithms we need to tune some parameters. For this specific cases we use additional tuning sets.

The results and their discussion

In our experiments we perform evaluation for tests 1-5 of static and dynamic gestures classification. The first conclusion we draw from the achieved results is the ANN (Artificial Network Neural) classifier achieved worse results than SVM (Support Vector Machine). In perspective of the fact that the data used in this experiment samples came from one day sampling, and evaluated with the all-session set, it can be concluded that the SVM classifier coped better with the generalization. This means that it better reflects the information contained in a limited training set to varying degrees in different samples. Nevertheless, the results can still be considered as good for both SVM and SSN. Performance achieved in the worst case are respectively 87% and 82%. These values are lower than in test 6, which indicates that more and more varied, as recorded on different days, samples allows to achieve better results.

Noticeable is the big difference in performance, depending on which sample came learning session. Moreover, it is exactly reflected in the graphs for both classifiers. For ANN dispersion is 10, and 7% for SVM. Selection of training samples is therefore crucial. Selecting the only one recording session creates a risk that they will not properly capture the characteristics of gestures, because some days of recordings provide some noise in the data, what can be caused e.g. by the mood of person that performs the signs.

For dynamic gestures classifier, results for these limited learning is much weaker than in the test 6. In the best case DTW classifier efficiency reached 48%, and 53% HMM (Hidden Markov Models) classifier. The worst results were lower by 20 and 9%. DTW method turns out to be so much more sensitive to the selection of learning samples.

Summary and future work

The results of the experiments indicated that both the artificial network neural and support vector machines can be used for classification of static gestures. Neural networks, however, require time-consuming tuning process, the unlike the other methods. The second conclusion is that the angular coordinates are not suitable for components of the feature vector for the SVM method and the hidden Markov models possible to achieve satisfactory results in the classification of dynamic gestures.

Achieving very high quality, however, would required the development of method over its standard form. DTW algorithm is not suitable for classifying complex movements. Its efficiency is low and the duration of action very long. Some classification errors can be eliminated by use of the modified or more advanced hardware interface that provide additional information.

The further work is planned to integrate classification method based on data from electronic sensors with the visual based method. In addition, work is underway the use of Kinect camera in order to obtain images of depth and use to further increase the effectiveness of recognition of static and dynamic hand gestures. Efficiency achieved by these systems is very different and varies between 80.9% and 99.2%.

This work has been supported by the National Center for Research and Development (NCBiR) under research Grant No. SP/I/1/77065/1 SYNAT: "Establishment of the universal, open, hosting and communication, repository platform for network resources of knowledge to be used by science, education and open knowledge society"

REFERENCES

[1] Beckhaus S., Kruijff E., Unconventional human computer interfaces*, ACM SIGGRAPH 2004 Course Notes,* (2004)

[2] Bhuiyan M., Picking R., Gesture-controlled user interfaces, what have we done and what's next?, *Proceedings of the Fifth Collaborative Research Symposium on Security, E-Learning, Internet and Networking* (SEIN 2009), Darmstadt, Germany (2009), 26-27

[3] Brave S., Ishii H., Dahley A., Tangible interfaces for remote collaboration and communication, *Proceedings of ACM conference on Computer supported cooperative work, CSCW '98*, New York, USA, (1998), 169-178

[4] Dix A., Finlay J.E, Abowd G.D., Beale R., Human-Computer Interaction (3rd Edition). *Prentice-Hall, Inc., Upper Saddle River, NJ, USA,* (2003)

[5] Duchowski A.T., Eye tracking methodology: Theory and practice, *Springer*, (2007)

[6] Hendzel J. K., Słownik Polskiego Języka Miganego*, Wydawnictwo OFFER,* (1995)

[7] Hernandez-Rebollar J.L., Lindeman R.W, Kyriakopoulos N., A multi-class pattern recognition system for practical finger spelling translation, *Proceedings. Of Fourth IEEE International Conference on Multimodal Interfaces,* (2002), 185-190

[8] Hollar J., Perng J.K., Pister K., Wireless static hand gesture recognition with accelerometers - the acceleration sensing glove, *Berkeley Sensor & Actuator Center, University of California, Berkeley* (2001)

[9] http://www.onsi.tv/slownik.htm. *Słownik Polskiego Języka Migowego. (on line*, 10 may 2013)

[10] Kapuściński T., Rozpoznawanie polskiego języka miganego, *In Phd Thesis. Zielona Góra : Uniwersytet Zielonogórski*, (2006)

[11] Krumm J., Ubiquitous Computing Fundamentals. *Chapman & Hall CRC Press*, 1st edition, (2009)

[12] Liu J., Zhong L., Wickramasuriya J., Vasudevan V., uWave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5 (2009), n.6, 657-675

[13] Marnik J., The polish finger alphabet hand postures recognition using elastic graph matching. *In Computer Recognition Systems 2, Advances in Soft Computing Volume Springer,* 45 (2007), 454-461

[14] Roggen D., Magnenat S., Waibel M., Troster G., Wearable computing, *IEEE Robotics & Automation Magazine*, 18 (2011), n. 2, 83-95

[15] Steinberg G., Natural user interfaces, *University of Auckland Seminar Reports*, (2012)

[16] Tutenel T., Smelik R., Lopes R., de Kraker K., Bidarra R., Generating consistent buildings: a semantic approach for integrating procedural techniques, *IEEE Transactions on Computational Intelligence and AI in Games,* 3 (2011), n. 3, 274-288

[17] Vieriu R-L., Goras R., Goras L., On HMM static hand gesture recognition*, 10th International Symposium ISSCS 2011 - In Signals, Circuits and Systems (2011),* 1-4

[18] Weiser M., Ubiquitous computing, *Computer*, 26 (1993), n. 10, 71-72

[19] Wingrave Ch.A., Williamson B., Varcholik P.D., Rose J., Miller A., Charbonneau E., Bott J., LaViola J., The wiimote and beyond: Spatially convenient devices for 3d user interfaces, *IEEE Comput. Graph. Appl.*, 30 (2010), n. 2, 71-85

[20] Zander T.O., Kothe Ch., Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general, *Journal of Neural Engineering*, 8 (20011), n. 2, 025005

Authors:
dr inż. Tomasz Dziubich, dr inż. Julian Szymański, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, 11/12 Gabriela Narutowicza street, 80-233 Gdańsk, Poland, E-mail: tomasz.dziubich@eti.pg.gda.pl, julian.szymanski@eti.pg.gda.pl.