STUDIA INFORMATICA 2016
Volume 37 Number 1 (123)

Jakub GAŁKA, Przemysław WĘGRZYNOWICZ, Mariusz MĄSIOR
AGH University of Science and Technology

ACQUISITION OF MULTIMODAL DATA CORPUS FOR AUTOMATIC SIGN LANGUAGE PROCESSING[1]

Summary. This paper presents the creation of a Polish Sign Language corpus suitable for recognition research and automatic translation of sign language. The recording approach used and the captured data modalities are presented, as well as the description of the acquisition system implementation. The evaluation of the collected corpus is presented and compared to other available resources.

Keywords: sign language recognition, data analysis, data collection, image processing

REJESTRACJA MULTIMODALNEGO KORPUSU DANYCH DLA AUTOMATYCZNEGO PRZETWARZANIA JĘZYKA MIGOWEGO

Streszczenie. Artykuł prezentuje utworzenie korpusu nagrań gestów Polskiego Języka Migowego na potrzeby badań możliwości rozpoznawania i automatycznego tłumaczenia języka migowego. Zaprezentowano podejście i metodykę tworzenia bazy nagrań oraz opisano implementację systemu akwizycji. Przedstawiono także ewaluację zebranych danych pod kątem rozpoznawania gestów języka migowego oraz porównano z innymi, dostępnymi zasobami.

Słowa kluczowe: rozpoznawanie języka migowego, analiza danych, gromadzenie danych, przetwarzanie obrazów

1. Introduction

In recent years, progressive work has been carried out on automatic sign language recognition systems – ASLR. According to the World Federation of the Deaf, there are approx-

[1]This work was supported by the Polish National Centre for Research and Development – Applied Research Program under Grant PBS2/B3/21/2013 titled: "Virtual sign language translator".

imately 70 million deaf people who use sign language as a way of communication [1]. In everyday life, they have to face and overcome the problem of being misunderstood. Due to communication problems, many deaf people have difficulties in access to information, education, and employment.

Sign language is counted among visual-spatial languages. Utterances are constructed using body movements and facial expressions. It is independent from the national spoken language and has its own, different grammar rules. This is the main reason for the difficulties in mutual automatic translation.

The article presents multimodal corpus data which was prepared as a part of the "Virtual Sign Language Translator" (WITKOM) research project. The creation of a practical solution for the communication between deaf and hearing people was undertaken by Polish researchers from the AGH University of Science and Technology and the VoicePIN.com company. The system was developed to recognize gestures of Polish Sign Language – PJM.

The preparation of an automatic sign language recognition system is divided into several stages. The first step in sign language sentence recognition is data acquisition. It is believed that every sign is a combination of articulatory components [2]. Due to this fact, as well as the simultaneous character of utterances in sign language, it was decided to use multimodal data acquisition. Gestures were recorded with the use of a single structural light sensor (Microsoft Kinect), RGB cameras, and a data glove with a series of accelerometers. Recording assembly and database management are complicated due to the individual variability of the same gestures among different signers.

High recognition accuracy can be achieved only by gathering a substantial training dataset. It should consist of multiple repetitions of every sign for a large number of signers. The whole process of data acquisition should be performed in a specific environment which allows to create the best possible repetition of recording conditions in a whole data corpus. Moreover, the entire recording set should be described with a proper linguistic annotation.

The next stage is the parameterization process of the assembled data. Image processing algorithms produce features which fully describe the manual character of the gesture. The extracted features are utilized for building the models of each gesture. In the presented project, every sign is modeled separately using Parallel Hidden Markov Models (PaHMMs). The same approach was previously taken by different research groups [3-5]. The classification method using PaHMMs was described and validated in Automatic Speech Recognition systems first, and then adapted to sign language recognition. Modeling in parallel HMM channels allows to group the features with reference to different articulatory elements. Another approach involves modeling the parameterized signal of every recording device in each chan-

nel of the PaHMM. It is assumed that a hybrid system allows to fuse information recorded by different measurement devices.

2. Databases of gestures for automatic sign language recognition

There is no widely available database of Polish Sign Language gestures. Most of the resources found were published as sign dictionaries. This caused the conditions of video recording to not be standardized during sessions. Moreover, there are no repetitions of each sign. However, the task of constructing such a database was undertaken for other sign languages. The existing recording sets were carefully analyzed and, finally, compared by the recognition system as a form of an independent validation benchmark.

One of the biggest and best-described sign language gesture databases is the AUSLAN database [6]. It is available online at the website of the Center for Machine Learning and Intelligent Systems [7]. This data set consists of gestures for Australian Sign Language. Signs were performed by native signers. They were captured using instrumented gloves and high-quality position trackers. The feature vector contains 22 parameters, including the position and orientation of both hands. In addition, every finger bend was measured. The sample rate of the complete system is described as 100 frames per second. The database consists of 27 repetitions of 95 Auslan signs and, in total, has 2565 recordings. The average length of each sign was approximated and is equal to 57 frames (0.57 s).

Some available databases are published as training and validation sets in gesture recognition challenges. They are formed as collections of different gestures, but only some of them belong to sign language. One of the examples is a large video dataset used in the Multi-modal Gesture Recognition Challenge 2013 [8]. The database consists of 13 858 gestures from a lexicon of 20 Italian gesture categories. They were recorded with the use of a Kinect camera. The device provides a multimodal data stream which includes the following: RGB image, depth image, skeletal model, user mask, and audio. The gestures in the dataset were performed by 27 users and were recorded in sequences. The corpus demands user-independent learning methods to acquire promising recognition results.

The majority of datasets are recorded using the single data acquisition method. Feature extraction and the recognition process are often limited to one mode. Multi-modal corpus data makes it possible to construct methods of parallel recognition and fuse classification results. The presented approach definitely extends the capabilities of building an efficient, automatic Polish Sign Language recognition system.

3. Acquisition system architecture

The implementation of the multichannel acquisition system was one of the main prerequisites established at the beginning of the WITKOM project. Performing separate recording sessions for each sensor would be expensive and time-consuming. It was assumed that simultaneous acquisition from multiple devices would be a better solution.

The selected approach causes several difficulties. The use of different sensors requires appropriate time synchronization. It should be considered that parallel data streams have to be initially processed and then saved to a storage device.

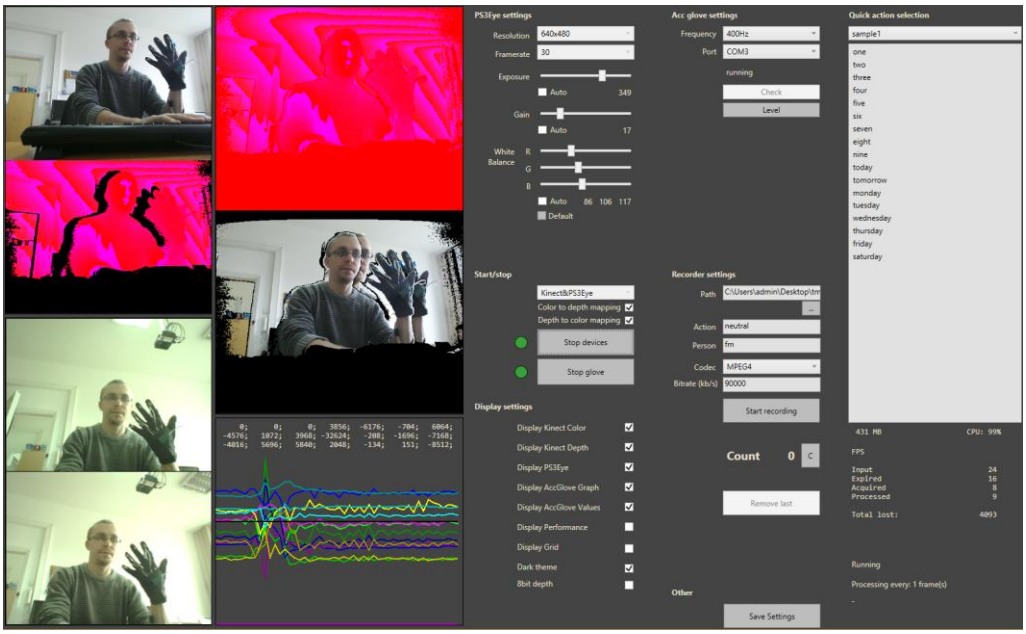

Fig. 1. User interface of the multimodal acquisition system
Rys. 1. Interfejs użytkownika multimodalnego systemu akwizycji

The acquisition system was developed by using the .NET Framework and will be publicly available for academic and non-commercial purposes. The system can simultaneously support four devices: a Microsoft Kinect 2 sensor, two Sony PS3 Eye cameras, and a custom-made Accelerometer Glove with an ARM microcontroller. The system can be expanded with any type of sensor, as long as it can be connected to a PC and drivers are available for programmatic access to the device. A dedicated mechanism was implemented to synchronize the stream from all of the cameras. The data stream form the Accelerometer Glove is handled independently, based on its timestamps. Accurate synchronization of the video signal of stereo cameras is essential for the depth reconstruction methods.

3.1. The recording studio

The acquisition studio was built for the purpose of the WITKOM project. It is a separate, small room with two-point diffusion lighting. Shadow-free illumination and the existence of a green screen make the recording conditions more stable and allows for chroma-key compositing. Background uniformity is a crucial aspect for the post processing of the collected video. All cameras were placed at the same height with a spacing of 15 cm.

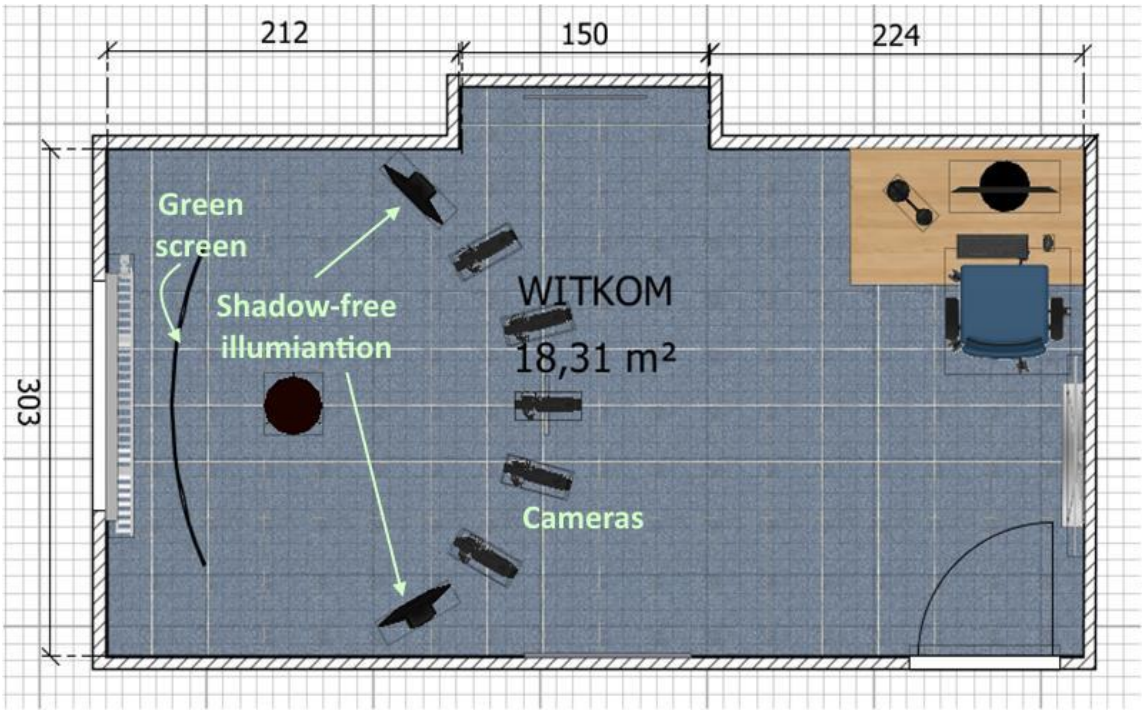

Fig. 2. Recording studio arrangement plan
Rys. 2. Plan aranżacji studia nagraniowego

3.2. HD video signal acquisition

Initially, the acquisition of the video signal was performed with relatively high image resolution. For the purpose of the recording, 8 GoPro Hero 3+ Black Edition cameras were used. The cameras were connected to a capture card with HDMI connectors and were controlled via WiFi. The presented composition allowed to simultaneously record video at 1920x1080 resolution (Full HD) from several devices. It was assumed that multiple high quality video sources form different directions are sufficient to depict the whole movement in time. However, the applied approach had some hardware limitations and a decision was made to reduce the resolution at the acquisition stage, especially due to the fact that the designed recognition system required a lower data rate.

3.3. The acquisition of the video signal at webcam quality

A recorded Full HD image of hand movement is believed to be redundant. The performed experiments confirmed that efficient vision processing of such data requires image downsampling. Furthermore, acquisition requirements were limited to simplify the resultant business implementation of the complete system. Nowadays, electronic devices are equipped with a simple webcam. Because of this, video signal capturing was performed at VGA resolution, and the previously-used GoPro devices were replaced by PlayStation 3 Eye digital cameras.

3.4. Image acquisition from the depth sensor

Motion controllers are often used in the field of computer vision. They were developed for consoles to allow hand-free control of electronic devices. Tracking the movement of objects in three dimensions is made possible by the use of a structured-light 3D scanner. The simple depth sensor is composed of an infrared projector and a monochrome CMOS detector. The obtained high-quality depth map and RGB video image are separate streams of data which can be modeled independently and then fused.

Fig. 3. Depth image and color mapped to depth projection captured by the Kinect sensor
Rys. 3. Obraz głębi i koloru zmapowany na projekcję głębi, zarejestrowany przez Kinect

At the beginning of the project, acquisition was performed with the use of an Asus Xtion PRO LIVE sensor. The depth image size is equal to 640x480 (VGA) with a 30 fps frame rate. The resolution of the video image captured by the RGB camera is 1280x1024 (SXGA). The next depth sensor which was used in the acquisition process is the Microsoft Kinect 2. It is a popular device in the field of motion recognition. Relatively simple access to the device output is possible by using the Kinect for Windows SDK 2.0. The captured color image is at Full HD resolution with a 30 fps frame rate. The Kinect device provides a high-quality depth map with a size of 512x424 at 30 Hz frequency. In addition, it has a separate data stream which contains the tracking positions of human body joints. Because of this, it is possible to

describe body motion in three dimensions, including the movements of palms and thumbs. The stability of this feature is important for further fingerspelling recognition.

In the end, the Microsoft Kinect 2 device was used for depth acquisition. Three data streams were collected: color bitmap, depth image, and joint positions. The resolution of the color image was decreased to VGA size to reduce the computational complexity.

More accurate skeleton motion capture systems exist, which also could be employed in data acquisition process, however usage of hi-end motion capture system would be difficult in any imaginable consumer-centric applications of the technology, and would produce a significant cost increase. Therefore, only consumer-compliant skeleton motion capture solutions were considered in the project.

3.5. Motion data acquisition by Accelerometer Glove

The Accelerometer Glove is a device designed for sign language users. The device has seven active sensors, five of which are on the fingers (one sensor on each finger), one on the wrist, and the last one on the arm (Fig. 4). Each of them is a 3-axis acceleration sensor.
The construction of the motion sensors and the acquisition system guarantees 10-bit resolution of signal with a 400 Hz sampling frequency (with SNR at a level of 40 dB).

Fig. 4. Accelerometer Glove and inertial sensors placement
Rys. 4. Rękawica akcelerometryczna oraz rozmieszczenie czujników inercyjnych

4. The WITKOM sign language gesture corpora

The described database of selected gestures of Polish Sign Language is one of the major achievements of the WITKOM project. It should be pointed out that 20 native signers were involved in the recording sessions. Due to the fact that they are deaf, the attendance

of a translator was required. During recording sessions, several breaks were scheduled to improve the attention of the signers. Efficient acquisition over a large number of sessions was a practical and logistical challenge.

The total volume of the collected recordings is equal to 1.6 TB. The approximate duration of all recordings is more than 20 hours. 24 372 movement samples were captured during all sessions. The dictionary consists of 390 different signs. Half of them were performed by more than 10 signers, which is very important for signer-independent system development and evaluation. The selected vocabulary has potential for further application – it contains signs of medical specialties, days of the week, months, numerals, and the fingerspelling alphabet (which is enough for building a simple automatic dialogue system). In each session, a single sign was repeated multiple times.

Table 1

Statistics and parameters of the WITKOM sign language gesture corpora

WITKOM sign corpora		**Acquisition devices**
Number of unique signs (gestures)	390	Microsoft Kinect 2: - color image VGA (640x480) - depth image (512x424) - color mapped to depth (512x424) - single body data (25 joints) 2 PlayStation 3 Eye cameras: - color image VGA (640x480)
Number of native signers	20	
Number of samples	24 372	
Number of data files	97 488	
Total duration	20 h	
Number of gesture classes performed by more than 10 users	158	
Number of unique signs with more than 10 repetitions per user	114	

The data streams from the multimodal acquisition system were saved to 4 files for every performed gesture. First of all, the collected data contains the video in color from the Kinect RGB camera and the vertically combined video captured by the PS3 Eye cameras. The third file contains the depth image of the Kinect sensor and a color map applied to the depth image. The last one is a numerical array containing joint orientation. The entire database consists of almost 100 thousand files.

The structure of the collected data is hierarchical. The first level is divided into independent user sessions. Within a session, there were collected directories of different gestures (unique dictionary entries). The lowest level contains sets of repetitions of each sign.

The presented database allows to perform diversified validation scenarios. A large number of gesture repetitions makes it possible to model the variability of movement. Due to the diversity of the signers, it is proposed to carry out user independent gesture recognition. It is

much more demanding than the standard user-dependent approach which is typical for the majority of other sign language databases.

The semantic annotation of every sign was performed during the recording sessions. There are no standardized versions of signs. There is an individual variability in the trajectories of movements. The invited signers performed natural signs. The aim of the annotation process was to describe the different variants of gestures with the same meaning. Native signers were presented only with the meaning of the gestures using images or inscriptions, in such a way as to minimize the impact on the performed gesture.

The purpose of the WITKOM project is to make the database publicly available. Intensive work has been carried out in the field of parameterization of collected video recordings. In addition to the multimodal database, we plan to publish the sets of features describing every recorded motion. These visual features are computed using object detection and the tracking approach. Also developed was a method of handshape feature estimation. The most promising is the approach of calculating the features of a motion field, which is outlined based on the optical flow algorithm [9]. This advanced method will be described in further publications. The collected features will enrich the database and may become useful for the machine learning research community. The WITKOM corpora will be available to research community since October 2016. Please contact authors for details.

5. Sign language gesture recognition evaluation

Several validation scenarios were conducted to exhibit the recognition potential of multimodal acquisition systems. The purpose of presented scenarios is to assess the sign language isolated gesture classification performance and how such classification system can benefit from different data modalities and features, especially when data fusion was applied.

To assess model accuracy and classification confidence, typical performance measures were calculated. The following measures were chosen: classification accuracy, error equal rate (EER), F1-measure, precision, and recall. A subset of the WITKOM sign corpus was selected and vision processing was performed.

Different methods of image feature extraction (Motion Field – MtnFld, Object Detection – ObjDet and Hand Shape – HndShp) were used [10–12]. Each data stream was modeled by PaHMM. The impact of the multimodal acquisition approach on recognition results was tested by the score-level fusion of PaHMMs. Classification scores obtained using models trained on different data streams were fused using the Artificial Neural Network (ANN) which performed late score fusion. The ANN contained one hidden layer and was trained to maximize the overall recognition accuracy given the development subset of the data training.

The recognition results of the data subset from the WITKOM corpus were summarized in Table 2. The test scenario verifies the performance of recognition among 94 gestures performed by a single user. In the presented tests, a 5-fold cross-validation procedure was applied.

The selected subset is similar to AUSLAN database in terms of the number of gestures. It was evaluated using 5-fold cross-validation on its reference database. As before, PaHMMs were used as the classification system. The results are comparable and more satisfactory than reported in [10]. The classification method based on Derivate Dynamic Time Warping achieved 87.7% accuracy [10].

Additionally, the cross-validation of a small set of gestures recorded with the use of a custom-made accelerometer data glove was conducted. The data set contains 40 unique gestures, with 10 repetitions each. Table 3 presents the recognition results for both data glove systems.

Table 2

Recognition results of the multimodal data subset of the WITKOM corpus

Data features used in recognition								**Performance measures (%)**				
Camera RGB Kinect			**Depth sensor Kinect**	**PS3 Eye Top camera**		**PS3 Eye Down camera**						
Mtn Fld	**Obj Det**	**Hnd Shp**	**KinBod**	**Mtn Fld**	**Obj Det**	**Mtn Fld**	**Obj Det**	**Accuracy**	**EER**	**F1**	**Precision**	**Recall**
+								80,13	4,55	80,67	81,19	80,15
	+							85,97	2,71	86,70	87,45	85,97
		+						94,16	2,23	94,43	94,71	94,15
			+					93,30	1,77	93,55	93,81	93,30
+	+	+	+					**98,30**	**0,64**	**98,37**	**98,44**	**98,30**
				+				81,40	3,72	81,91	82,41	81,41
					+			81,19	3,51	81,83	82,50	81,18
						+		80,02	4,04	80,54	81,06	80,02
							+	81,19	3,78	82,12	83,07	81,19
				+	+	+	+	**88,42**	**2,02**	**88,54**	**88,68**	**88,40**
+	+	+	+	+	+	+	+	**98,30**	**0,65**	**98,37**	**98,45**	**98,30**

Table 3

Recognition results for systems using instrumented gloves

System	**Accuracy**	**EER**	**F1**	**Precision**	**Recall**
Data glove - AUSLAN	98,26	1,42	98,24	98,28	98,21
Data glove - WITKOM	99,75	0,01	99,76	99,77	99,75

The use of multimodal video acquisition methods allows to achieve rewarding results for automatic sign language recognition systems. The highest increase of accuracy can be observed for the fusion of the RGB video image and depth projection. The performance of the vision system equals the results achieved by using data gloves only.

6. Conclusions

The purpose of the WITKOM project is the development of a Polish Sign Language corpus. The vocabulary and the number of signers will constantly grow. A comprehensive database of gestures will be the first step to build a complete ASLR system. Strong emphasis will be put on the improvement of user-independent results.

In sign language recognition research, every single sign should be considered a spatial phenomenon spread in time. The processing of human communication requires huge collections of language samples. It is essential to prepare appropriate training and development sets. These collections should include the entire diversity and specificity of the language. Combining different acquisition methods seems to be the way of describing the entire spatial variability of the phenomenon.

An important objective for future work is the implementation of a framework for automatic data collection and processing. Cloud computing should be used to share the service and take advantage of collective data production. This approach will enable rapid development of the database and the processing methods.

It is believed that only a comprehensive approach ensures success in the field of sign language processing. The high effectiveness of ASLR systems depends strongly on the data preparation stage.

BIBLIOGRAPHY

1. World Federation of the Deaf: Sign Language. Available at: http://wfdeaf.org/human-rights/crpd/sign-language (accessed 26 January 2016).
2. Stokoe W.C.: Sign Language Structure: An Outline of the Visual Communication System of the American Deaf. J. Deaf Stud. Deaf Educ.. Vol. 10(1), 1960, p. 3÷37.
3. Vogler C., Metaxas D.: A Framework for Recognizing the Simultaneous Aspects of American Sign Language. Computer Vision and Image Understanding, Vol. 81, No. 3, 2001, p. 358÷384.
4. Theodorakis S., Pitsikalis V., Maragos, P.: Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. Image and Vision Computing, Vol. 32, No. 8, 2014, p. 533÷549.
5. Von Agris U., Zieren J., Canzler U., Bauer B., Kraiss K.-F.: Recent developments in visual sign language recognition. Universal Access in the Information Society, Vol. 6, No. 4, 2008, p. 323÷362.

6. Kadous M.W.: Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series. PhD dissertation, School of Computer Science and Engineering, Univ. of New South Wales, Australia 2002.
7. Lichman M.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science, Irvine, CA 2013.
8. Escalera S. et al.: Multi-modal Gesture Recognition Challenge 2013: Dataset and Results. ICMI, 2013.
9. Horn B.K.B., Schunck B.G.: Determining optical flow. Artificial Intelligence, Vol. 17, 1981, p. 185÷203.
10. Barczewska K.: The Automatic Recognition of Isolated Sign Language Signs Based on Gesture Components and DTW Algorithm. Challenges of Modern Technology, Vol. 5, No. 3, 2014, p. 1÷8.
11. Tsai D.M., Chiu W.Y., Lee M.H.: Optical flow-motion history image for action recognition. Signal, Image and Video Processing, Vol. 9, No. 8, 2014, p. 1897÷1906.
12. Yeo H.S., Lee B.G., Lim H.: Hand Tracking and Gesture Recognition System for Human-Computer Interaction Using Low-Cost Hardware. Journal of Multimedia Tools and Applications, Vol. 74, No. 8, 2013, p. 2687÷2715.

Omówienie

Artykuł prezentuje multimodalny korpus nagrań gestów języka migowego, przygotowany jako część realizacji projektu „Wirtualnego Tłumacza Języka Migowego" (WITKOM).

Zaprezentowano całościowe podejście do konstrukcji zbiorów danych umożliwiających rozwój, ewaluację i walidację rozwiązań w dziedzinie automatycznego rozpoznawania i tłumaczenia języka migowego.

W trakcie pracy nad systemami automatycznego rozpoznawania zjawisk, takich jak mowa lub język migowy, odpowiednie podejście do konstrukcji danych treningowych i rozwojowych jest kluczowe we właściwej realizacji przedsięwzięcia. Artykuł prezentuje procedurę i metodykę tworzenia bazy nagrań gestów Polskiego Języka Migowego, zaprezentowano także opis systemu akwizycji i stworzonych na jego potrzebę narzędzi.

Przedstawiono wyniki ewaluacji systemu rozpoznawania na zebranych danych, w celu weryfikacji możliwości pracy algorytmów automatycznego rozpoznawania gestów, działających dla danych multimodalnych.

Addresses

Jakub GAŁKA: AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, Poland, jgalka@agh.edu.pl.

Przemysław WĘGRZYNOWICZ: AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, Poland.

Mariusz MĄSIOR: AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, Poland, masior@agh.edu.pl.