Feature vector or time-series – comparison of gestures representations in automatic gesture recognition systems

Katarzyna Barczewska[1], Wioletta Wójtowicz[2], Tomasz Moszkowski[1]

1 AGH University of Science and Technology,
Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering,
Department of Automatics and Biomedical Engineering,
al. Mickiewicza 30, 30-059 Krakow, Poland
e-mail: kbarczew@agh.edu.pl, tmoszkow@agh.edu.pl
2 Cracow University of Technology,
al. Jana Pawła II 37, 31-864 Cracow, Poland
e-mail: wioletta.wojtowicz@mech.pk.edu.pl

In this paper, we performed recognition of isolated sign language gestures - obtained from Australian Sign Language Database (AUSLAN) – using statistics to reduce dimensionality and neural networks to recognize patterns. We designated a set of 70 signal features to represent each gesture as a feature vector instead of a time series, used principal component analysis (PCA) and independent component analysis (ICA) to reduce dimensionality and indicate the features most relevant for gesture detection. To classify the vectors a feedforward neural network was used. The resulting accuracy of detection ranged between 61 to 87%.

Key words: principal component analysis (PCA), independent component analysis (ICA), Neural Networks, sign language, automatic recognition

Introduction

The recent advances in sensor technologies led to new solutions of human-computer interfaces like touch screens, sensors, game controllers and motion-tracking cameras and data gloves. They can not only mediate in interacting with computer games but also help disabled or deaf people in communication with the healthy by gathering data about the performance of gestures. Automatic gesture recognition can use this data for systems augmenting learning of sign languages as well as translating gestures into words of spoken or written language. Many research groups strive to develop a system that would automatically recognize and translate sign language into text or speech [1,2,3,4,5]. The approaches of automatic gesture recognition vary depending on the used data sets (e.g. simple gestures measured with accelerometer [6], fingerspelled gestures [3], isolated signs, gesture sequences [1,7,8,9], facial expressions [10, 11]), data set characteristics (e.g. time series [4,6,12], feature vectors [2,13], subunits (cherems) [7,9,12]) human-machine interfaces (e.g. cameras [2,3,7,8,9,13], game controllers [5,14], data gloves [1,12]), and classifiers (e.g. Hidden Markov Model [3,7,8,13], neural networks [2], statistical models and dynamic time warping [4,12]).

Unfortunately, in order to develop a gesture recognition algorithm that would operate in real-time, the approaches should consider reducing the computing burden by reducing the number of dimensions of represented gestures e.g. by reduction of the number of distinctive features.

The goal of this study is to compare results of two different approaches to automatic recognition of gestures. First, described in [15], where gestures are represented as time-series and second, where gestures are represented by feature vectors, that include just characteristics of these time-series. The first approach uses classification based on measuring similarity of the time series – each gesture is compared to the set of exemplars in terms of distance, which is calculated using dynamic time warping methods. The second method, which we propose in this article, uses a feed-forward neural network to recognize sign-language gestures.

We extracted gesture features from the Australian Sign Language data set (Section Materials), reduced dimensionality and analyzed the distinction of features using principal and independent component analysis (PCA and ICA) (Section Reduction), trained different setups of neural networks (Section Classification), analyzed the results of classification using performance measures and compared with results obtained in [15] (Section Results).

Materials

The data set and features

To make the results comparable, we used the same database that was used in [15]. Australian Sign Language signs data

set (AUSLAN) consists of 95 gestures represented by signals collected from one native Australian signer using two data gloves [12,16] which contained bending sensors placed on each finger and position trackers. In each measurement session (one session weekly for a total of nine weeks), the volunteer repeated each of 95 gestures three times (2565 measurements in total). 22-variable time series – of 58-frame average length sampled at 100 Hz – represented each gesture sample which consisted of the following parameters:

- position in three-dimensional space (x, y, z) expressed in meters, with the origin of the coordinate system below the chin;
- angular orientation in Euler angles (yaw, pitch and roll) expressed in degrees;
- finger extension expressed by a value ranging from 0 to 1, where 0 means straight and 1means totally bent.

Figure 1 shows an example of signals characteristic to the sign different. We used 94 gestures from the AUSLAN Database – one sign had an insufficient number of repetitions – and divided the database into two sets: training (663 sessions – 18 samples of each gesture) and testing (333 sessions – 9 samples of each gesture).

Feature vector

We characterized raw and filtered data from both hands using 70 features and grouping them for each gesture – thus, we did not need to analyze the full 22 – dimensional signal time-series. We built the gesture vectors using the minima and maxima of position (x, y, z) and angular orientation (roll, pitch, yaw) – 24 variables, 12 per palm; the number of slope sign changes – 24 variables, 12 per one palm; the number of zero-crossings – 12 variables, 6 per one palm; and the number of bending events of fingers (with a threshold of 0.8) – 10 variables, 5 per one palm. We constructed training (X_{train} – 1692 samples of 70 features) and testing (X_{test} – 846 samples of 70 features) matrices using the resulting feature vectors. In [15] each gesture was represented as a matrix containing the set time-series: 30×22 (the number of frames per one gesture, different for different gestures × number of measured components: x, y, z, roll, pitch, yaw, bending of 5 fingers; for each of 2 hands).

Dimensionality reduction

We applied principal and independent component analysis (PCA and ICA) to the training (X_{train}) and testing (X_{test}) data matrices to project their rows – containing feature vectors – from a 70-dimensional space to several lower-dimensional spaces. For clarity, we will refer to both the training and testing data as X.

The PCA component axes were found by $R_k = X * P_k$, where P_k is the transformation matrix (which contains eigenvectors of X covariance matrix), k is the number of defined principal components (and also corresponds to the number of columns in P_k). We used $k = 32, 24$ and 17 corresponding to *90, 80* and *70*% of the explained variance in the gesture data, respectively (Fig. 2). Afterwards , the

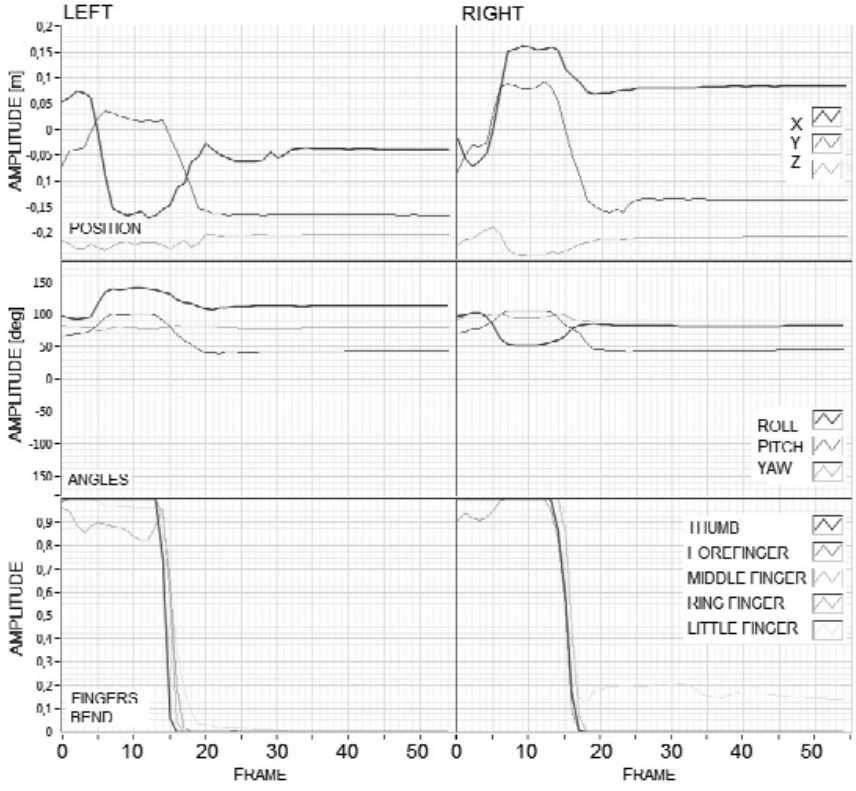

Figure 1. Signals for right and left hand while performing the sign different

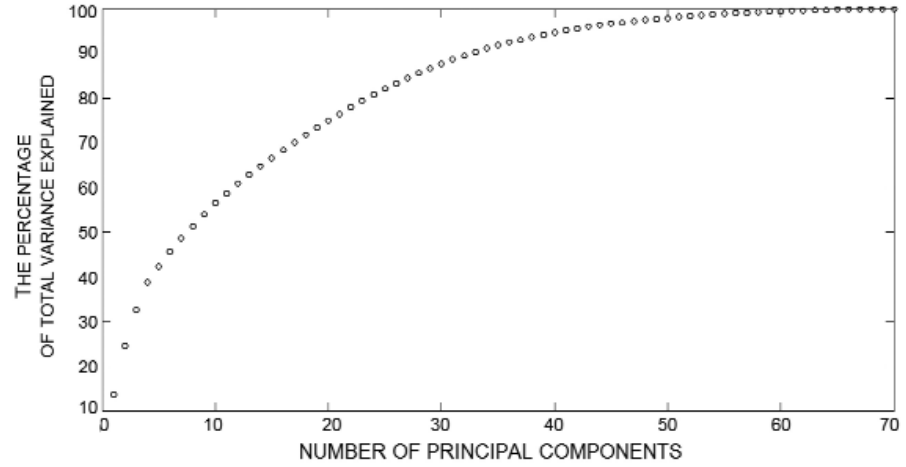

Figure 2. Number of principal components corresponding to the percentage of the total explained variance of the Xtrain matrix

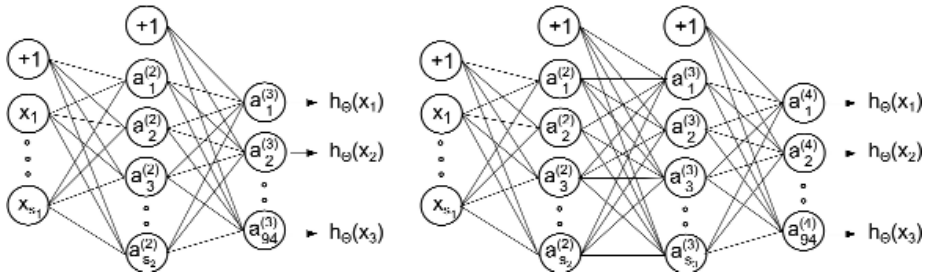

Figure 3. The structure of feedforward neural networks. sj is the number of units in the j-th layer; in the network on the right $s_2=s_3$

Table 1. The recognition accuracy depending on the number of features used (percent of the variance variability explained by reduced vectors)

Reduction Method	Variance	Features	Accuracy [%]	
			NNl	NN2
none	100%	70	55.08	-
PCA	90%	32	84.90	86.96
	80%	24	73.52	77.03
	70%	17	61.35	66.47
PCA-ICA	90%	32	86.13	87.23
	80%	24	73.72	77.62
	70%	17	62.33	65.41

Table 2. The recognition accuracy of gestures represented as time-series [15]. In the case indicated by the asterisk, the representation of gesture contained also hand velocities for each direction. Dynamic time warping (DTW) and derivative dynamic time warping (DDTW) were used to measure similarity to the set of exemplars

Reduction Method	Variance	Features	Accuracy [%]	
			DTW	DDTW
none	100% 100%	30x22 30x28	63.5 85.6	69.5 87.7

X was reconstructed as $X^{rec} = R_k * P^T_k$, which is a minimum square error approximation.

Then we performed ICA using FASTica [17, 18] algorithm that yielded a matrix W such that $W * P^T_k = U$ – where U is the matrix of statistically independent sources - from eigenvectors in P_k. We could reconstruct the matrix of transformation $P^{recT}_k = W^{-1} * U$ and the approximation of the data $X^{rec} = R_k * W^{-1} * U$ both containing statistically independent sources. Independent components represented the gestures in the rows of the matrix $B = R_k * W^{-1}$.

Classification

We used the feature vectors to teach (in a supervised learning process) the feed-forward neural networks using the backpropagation algorithm according to [19]. The networks contained one or two hidden layers of neurons with monopolar sigmoid activation function (see Fig. 3). In the case of the neural network with two hidden layers, we assumed that both of them should have an equal number of neurons. The number of inputs corresponded to the number of gesture features in the input vectors obtained after dimension reduction; the number of outputs corresponded to 94 different gestures from the AUSLAN database.

The training of the neural network was performed in three steps: 1) constant value of hidden units and variable values of the regularization parameter (responsible for the tradeoff between bias and variance), 2) constant value of regularization parameter and a variable number of hidden units, and 3) training for optimal parameters. Minimization of the cost function - which depended on the number of hidden layer units and the regularization parameter – yielded optimal parameters for the neural networks.

Optimal parameters obtained during training were used to classify the gestures from the test set. The output of the neural network is a vector of 94 elements with the values ranging from 0 to 1. Each output unit corresponds to one of 94 gestures from the AUSLAN database - numbered from 1 to 94. The algorithm considered the output unit of the highest classification response as the winner and, as a result , returned its index. Recognition accuracy was the sum of all correctly recognized gestures divided by the sum of all gestures from the test set . Experiments were conducted in the Mathworks Matlab R2012a environment.

Results

The accuracy of tested methods - summarized in Table 1 – spanned between 553 and 873: the lowest observed in the two-hidden-layer neural network fed with 32-dimensional representation of gestures projected using ICA, the highest observed in the one-hidden-layer neural network fed with the full feature vector with no feature projection. In most cases, the use of ICA resulted in more accurate classification of gestures in contrast to PCA, although the number of dimensions remained the same. Reducing the number of features from 70 to 32 improved recognition substantially, but reducing it further yielded lower accuracy. In all cases, the two-hidden-layer neural network classified the inputs more accurately. For comparison, in Table 2 we presented recognition results obtained in [15].

Conclusions and Discussion

In the study we have elaborated the construction of vector representation of sign-language gestures and their classification using neural networks, which can be used in a system of automatic gesture recognition. The presented approach was compared with a previously proposed method ([15]) using a multidimensional time series representation of gestures from the same data set. Despite using different representation of the data and different classifiers, we obtained comparable maximal recognition accuracies – 87.7% and 87.23% for multivariate time series and feature vectors representations, respectively. These results showed that significant reduction of data dimension using vectors containing only physical properties (features) of signals instead of full time series representation can greatly reduce the computation demand while maintaining similar recognition accuracy.

Simultaneously, it should be noted that the selection of signal features used during construction of gesture vectors was performed arbitrarily. As described in previous sections, the feature vectors contained rather the physical properties of particular signals disregarding the majority of information regarding the signals that describe the tested sign language vocabulary. As a result, it is not surprising that initially the 70-dimensional representation of gestures provide very poor recognition accuracy (around 55%), which indicates that many of the vector features have introduced irrelevant information into the training data set. Therefore, the main focus of our experiment was moved to the examination if lower dimensional vectors that can be obtained using PCA and ICA projection techniques, can give recognition results which are similar to the results obtained in [15]. Considering three alternative lower-dimension spaces, we found that reducing the amount of variance in the data set firstly led to increased recognition accuracy, which means that the disregarded variance contained much of the information interfering with classification. However, further reduction of explained variance resulted in degradation of accuracy - this variance apparently contained information vital for classification. Interestingly, the projections themselves also influenced the accuracy: uncorrelated projections (ICA) seem to further reduce the amount of irrelevant information for classification, while correlated projections (PCA) seem to introduce – perhaps through duplication - more error. Thus, as mentioned in previous paragraph, dimensionality reduction techniques give reduced representation of gestures, that contains such amount of information, that allows obtaining the same recognition accuracy as method which needs as an input whole signals.

Our study shows that the accurate of classification of sign language gestures can be improved by finding an appropriate subspace of arbitrarily selected gesture features. To further improve the obtained results, we propose revising or extending the feature vector. Furthermore, analyzing higher numbers of feature projections, testing different neural network architectures (e.g. counter-propagation networks, learning vector quantization) and different classifiers (e.g. Support Vector Machine) could be also considered . Regarding the training method, an N-fold cross-validation might also further improve the estimation of the recognition accuracy

Acknowledgements

The work was supported by AGH University of Science and Technology project number 11.11.120.612.

References

[1] Liang R.H., Ouhyoung M., Wolthusen S.D., *Real-time Continuous Gesture Recognition System for Sign Language*, Proc. 1998 Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 558-567, April 1998.

[2] Gweth Y.L., Plahl C., Ney H., *Enhanced Continuous Sign Language Recognition using PCA and Neural Network Fea tures,* Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. Providence, pp. 55-60, June 2012.
[3] Liwicki S., Everingham M., *Automatic Recognition of Fingerspelled Words in British Sign Language*. Proc. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 50-57, June 2009.
[4] Lichtenauer J.F., Hendriks E.A., Reinders M.J.T., *Sign Language Recognition by Combining Statistical DTW and Independent Classification,* IEEE Transactions on Pattern Analysis and Machine Inteligence, 30(11) , pp. 2040-2046, 2008.
[5] KinecTranslator, [web page] http://www.kinectranslator.com/pl/technologia/. [Accessed on 31 Jul.2013.].
[6] Barczewska K., Drozd A., Folwarczny Ł., *Rozpoznawanie gestów z wykorzystaniem czujników inercyjnych o 9 stopniach swobody*, Pomiary, Automatyka, Kontrola, 59(3), pp. 235-238, 2013.
[7] Theodorakis S., Pitsikalis V., Rodomagoulakis I., Maragos P., *Recognition with raw canonical phonetic movement and handshape subunits on videos of contiunuous sign language*, Proc. 2012 IEEE International Conference on Image Processing (ICIP), pp. 1413-1416, 2012.
[8] Kapusciński T., *Rozpoznawanie polskiego języka migowego w systemie wizyjnym*, PhD dissertation. Uniwersytet Zielonogórski, Wydział Elektrotechniki, Informatyki i Telekominukacji , Zielona Góra 2006.
[9] Oszust M., *Zastosowanie grupowania szeregów czasowych do rozpoznawania wypowiedzi w języku migowym na podstawie sekwencji wizyjnych*, PhD dissertation. AGH WIET, Kraków 2013.
[10] Nguyen T.D., Ranganath S., *Facial expressions in American sign language: Tracking and recognition*, Pattern Recognition, 45(5), pp. 1877-1891, 2012.
[11] von Agris U., Knorr M., Kraiss K.F., *The significance of facial features for automatic sign language recognition*, IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1-6, Sept. 2008.
[12] Kadous M.W., *Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series*, Doctoral Dissertation, School of Computer Science and Engineering, University of New South Wales, 2002 .
[13] Trmal J., Hruz M., Zelinka J., Campr P., Muller L., *Feature Space Transforms for Czech Sign-Language Recognition,* Proceedings of the Interspeech 2008, Bris- bain, Australia, pp. 2036-2039, 2008.
[14] Chai X., Li G., Lin Y., Xu Z., Tang Y., Chen X., *Sign Language Recognition and Translation with Kinect,* [web page] http://vipl.ict.ac.cn/sites/default/files/papers [Accessed on 30 Nov . 2013].
[15] Barczewska K, *Automatic Recognition of Isolated Sign Language Signs Based on Gesture Components and DTW Algorithm,* Challenges of Modern Technology 5(3), 2014.
[16] AUSLAN data set, [web page] http://archive.ics.uci.edu/ml/machine-learning-databases/auslan2-mld/auslan.data.html [Accessed on 31 Sep. 2013.].
[17] Hyvarinen A., Fast *and robust fixed-point algorithms for independent component analysis*, IEEE Trans. Neural Networks, 10(3), pp. 626-634, 2011.
[18] FastICA algorithm, [web page] http://research.ics.aalto.fi/ica/fastica/ [Accessed on 31 Aug. 2013].
[19] Ng A., *Machine Learning* materials from on-line course [web page] https://www.coursera.org/ [Accessed on 30 Jun. 2013].
[20] Baek K., Draper B.A., Beveridge J.R., She K., *PCA vs ICA: A comparison on the FERET data set*, Proc. of the 4th International Conference on Computer Vision, ICCV '02, 2002.
[21] Hyvarinen A., Oja E., *Independent component analysis: Algorithms and Applications,* Neural Networks, 13(4-5), pp. 411-430, 2013.
[22] Hyvarinen A., *Independent component analysis: recent advances,* Phil. Trans. R. Soc„ 371(1984), 2013.

Katarzyna Barczewska, MSc Eng – is a PhD candidate and an academic assistant at AGH University of Science and Technology in the Department of Automatics and Biomedical Engineering at AGH, Krakow, Poland. She received her MSc degree in biomedical engineering at AGH University in year 2011. Her research interests include statistical and machine learning, gesture recognition and human-computer interaction.

Tomasz Moszkowski, MSc Eng – is a PhD candidate at AGH University of Science and Technology, The Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, Poland. He obtained his MSc degree in biomedical engineering in year 2012. His research focuses on electrical stimulation of neural structures, recording of biopotentials and modeling of electrical phenomena within human tissues.

Wioletta Wójtowicz, MSc – is an academic assistant at Institute of Applied Computer Science, Cracow University of Technology and a PhD candidate at AGH University of Science and Technology, The Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, Krakow, Poland. She received her MSc in Mathematics from AGH University of Science and Technology in 2011. Her research interests include digital watermarking, image analysis, biometrics and machine learning methods.