Recognition of Human Gestures Represented by Depth Camera Motion Sequences

Adam Świtoński[1,2], Bartosz Piórkowski[2], Henryk Josiński[2], Konrad Wojciechowski[1], and Aldona Drabik[1]

[1] Polish-Japanese Academy of Information Technology, Aleja Legionów 2, Bytom, Poland,
{aswitonski, kwojciechowski}@pjatk.edu.pl
[2] Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland,
{adam.switonski, henryk.josinski}@polsl.pl,

Abstract. The method of gesture recognition useful in construction of hand controlled user interfaces is proposed in the paper. To validate the method, a database containing seven selected gestures, performed by six subjects is collected. In acquisition Microsoft Kinect device was used. In the first stage of introduced method, linear dimensionality reduction in respect to depth images of human silhouettes is carried out. Furthermore, the proper recognition is based on Dynamic Time Warping technique. To assess general features of human gestures across subjects, different strategies of dividing captured database into training and testing sets are taken into consideration. The obtained classification results are satisfactory. The proposed method allows to recognized gestures with almost 100% precision in case of training set containing data of classified subjects and with 75% accuracy otherwise.

Keywords: motion capture, motion analysis, times series classification, artificial intelligence, dimensionality reduction, dynamic time warping, depth imaging

1 Introduction

Human-computer interaction is gaining more importance in construction of computer systems and their user interfaces. Quite new possibilities in this field are brought by still developing motion acquisition techniques, which allow to recognize gestures made by a human. Best precision of measurements of human motions is obtained by optical marker based motion capture systems. However such an acquisition has serious limitations related to requirements of mocap laboratory, markers which have to be attached on a body and calibration stage. In contrary there is a motion capture based on depth imaging. It gives worse precision of the measurements, but only shortcoming of the acquisition process corresponds to the limited distance between a human and capturing device.

Gestures can be defined as expressive, meaningful body motions involving physical movements of the fingers, hands, arms, head, face or other body parts

[6]. The problem of human gesture recognition is broadly studied in the world literature for years. Below only the selected approaches which utilize Kinect depth imaging are described.

In [3] histogram based features of 14x14 grid placed on the determined foreground are computed. What is more simple differences between subsequent frames called motion profiles are calculated. The proper classification is carried out by multiclass Support Vector Machine. The proposed method is strongly simplified, because it does not consider the time domain and recognition corresponds only to single frames. This leads to worse accuracy, since there are some poses which are very similar in different gestures. To discriminate them analysis of whole sequence of poses is necessary. The similar approach is introduced in [8]. The static gestures are recognized by Finger-Earth Mover's Distance and template matching.

In [7] Neural Network, Support Vector Machine, Decision Trees and Naive Bayes classifiers are trained on the basis of parameters of skeleton model determined by Kinect software rather than raw depth images. Once again only single poses are recognized. The obtained classification accuracy is very high. in the best case it has even 100% of correctly classified instances, because very simple set of only three static poses - stand, sit down and lie down - is taken into consideration.

Whole motion sequences are analyzed for instance in [13], which deals with the problem of hand gesture recognition. In a preliminary stage palm is detected on every motion frame and its tangent angle at different time instants is calculated. Tangent angle relates to hand movement across images axes. The final classification is carried by Hidden Markow Models (HMM) - in the training phase single model for every class is determined and in the recognition stage the model with greatest probability is searched. Another approach based on HMM, but this time in respect to features extracted for four joints angles of the skeleton model: left elbow yaw and roll, left shoulder yaw and pitch - is presented in [5]. Prior to HMM modeling data is segmented by K-Means clustering into thirty groups and HMM in further processing takes into consideration only cluster centroids. Very naive method in which a multidimensional, skeleton model based motion sequences are transformed into single feature vector is described in [2]. To recognize gestures supervised classification is carried out. However such a method has the strong limitations related to the specified number of frames of the recordings and it does not considers possible local shifts between subsequent gesture phases.

There are also examples of Dynamic Time Warping (DTW) applications to Kinect data. In [1] DTW is used to align motion sequences to a median length, reference one. Further Gaussian Mixture Models (GMM) are utilized to construct feature vectors of subsequent time instants. In the final stage DTW once again is carried out with soft-distance based on the probabilities of a point belonging to G components of GMM. In [4] an approach of recognition of hand written digits is proposed. It consists of following steps: hand detection of depth motion sequences, feature extraction and DTW classification.

In the paper a strict machine learning based method for gesture recognition which analyzes whole motion sequences is introduced. In the preliminary stage dimensionality of extracted monochromatic silhouettes representing depth data is reduced. In the classification stage dynamic normalization of the time domains of compared motion sequences is carried out by Dynamic Time Warping (DTW) transform.

2 Gesture Database

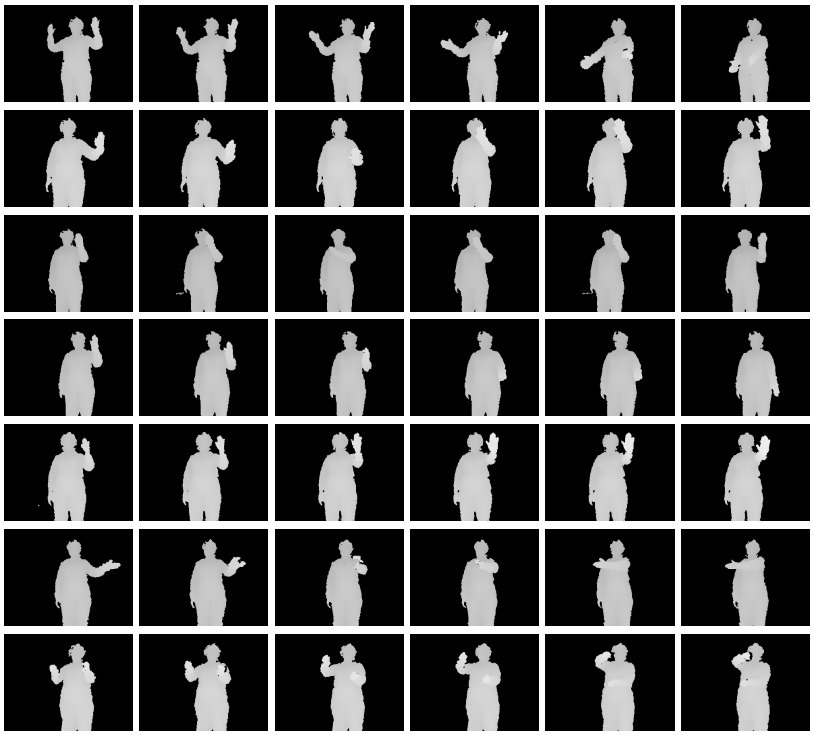


Fig. 1: Types of gestures

For the preliminary validation of proposed method database containing data of six subjects and seven selected types of gestures is collected. In the acquisition a Kinect camera was used and raw data consist of sequences of low resolution (160x120) depth images. The chosen types of gestures visualized in subsequent rows of Fig. 1 are related to possible control actions used in construction of user interfaces. They are as follows: arm zoom, hand spinning, hand wave, sliding down, sliding forward, sliding inward and wheel steering. Hand spinning and waving are performed by left upper limb. In total there are 196 different instances of recordings wherein complete gesture is performed only once.

Fig. 2: Bounding box detection

To remove the influence of a location of a human on the recognition performed, the bounding box is determined as presented in Fig. 2. It is based on geometric center of silhouette depth image in which background points are labeled by zero value.

The captured database is available online at http://as.pjwstk.edu.pl/aciids2015.

3 Classification Method

As described above, there are two main stages of the proposed classification method - dimensionality reduction of silhouettes appearing in subsequent time instants and proper classification of reduced motion sequences. On the basis of our previous experiences on pose classification [11] and diagnosis of gait abnormalities [12], classical linear Principal Component Analysis (PCA) technique was chosen for a dimensionality reduction. Because of satisfactory results obtained in human gait identification challenge problem by classifier which compares dynamically scaled motion sequences [10], in proper recognition Dynamic Time Warping transform is applied.

3.1 Principal Component Analysis

The PCA method determines linear independent combinations Y of the input attributes X with the greatest variances, which are called principal components. New base of an input space is established. It turns out, that the base created by the eigenvectors v^T ordered according to corresponding eigenvalues of the covariance attributes matrix satisfies the demands [14].

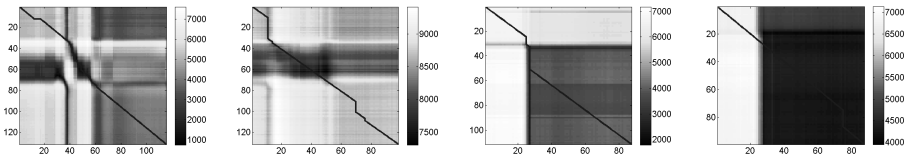
$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = V^T \cdot X = \begin{bmatrix} v_1^T \\ v_2^T \\ \cdots \\ v_n^T \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} \tag{1}$$

The variance of resultant attributes is denoted by corresponding eigenvalues λ_i, thus in dimensionality reduction only specified number of the first principal components is considered. What is more to evaluate the transformation, the variance cover vC(k) of the first k PCA attributes in respect to dimensionality n of the input space can be calculated. It assesses how the input space is explained by PCA features and it is expressed by the ratio constructed on the basis of proper eigenvalues:

$$vC(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{2}$$

3.2 Dynamic Time Warping

Dynamic Time Warping firstly applied in spoken word recognition [9] scales time domains of analyzed pair of motion sequences. It tries to match most similar subsequent time instants. The transformation is represented by a warping path which indicates corresponding times instants of both sequences. The path is assessed by its cost - aggregated total distance between matched silhouettes. As the result a warping path with minimal cost is determined and the cost can be stated as a dissimilarity between whole compared motions. During DTW com-



(a) HS vs HS, SU (b) HS vs HS, DU (c) AZ vs SF, SU (d) AZ vs SF, DU

Fig. 3: Example DTW similarity matrices and DTW paths (HS-Hand spining, AZ-arm zoom, SL-sliding forward, SU-single user, DF-different users)

putation, similarity and cost matrices are calculated. The first one contains data of distances between every pair of silhouettes of compared sequences which is approximated by Euclidean metric applied to reduced PCA space. The specified cell c(n,m) of cost matrix corresponds to warping path with minimal cost for time series reduced to the first n and m time instants. What is more DTW is monotonic transformation, which means moving backward in time domain is not allowed. So, to determine specified value c(n,m) the minimum of three previous possible neighbors c(n-1,m), c(n,m-1) and c(n-1, m-1) is found and it is increased by a proper value of the similarity matrix. Thus, it is sufficient to carry on calculation in a proper order for subsequent rows or columns and the last cell of cost matrix corresponds to wanted cost of a complete warping path.

In case when motion dissimilarity is approximated, nearest neighbor classification scheme [14] can be applied - an object is being assigned to the class most common among its k nearest neighbors.

Example DTW paths determined with corresponding similarity matrices for 32 dimensional PCA space and input data reduced by 80x110 bounding box are visualized in Fig. 3. In case when the same gestures are compared (Fig. 3a and 3b) the path is matched in the narrow, dark region of more similar silhouettes, which is located in the left-upper quarter of similarity matrices. The right-lower

quarter is related to standing, final pose because duration of the recordings is longer than gestures activity. That is a reason why it quite uniformly distributed. There is no such an observation for different gestures in Fig. 3c and 3d. Data of

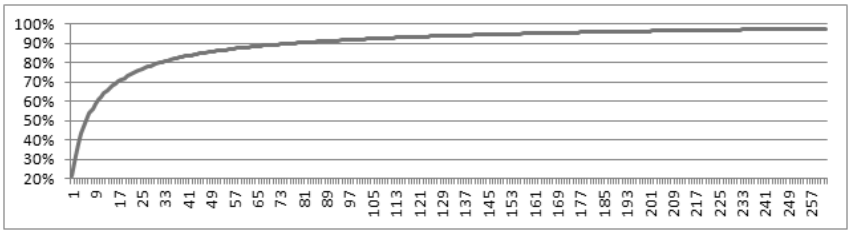


Fig. 4: Aggregated, total variance cover of the first n principal components.

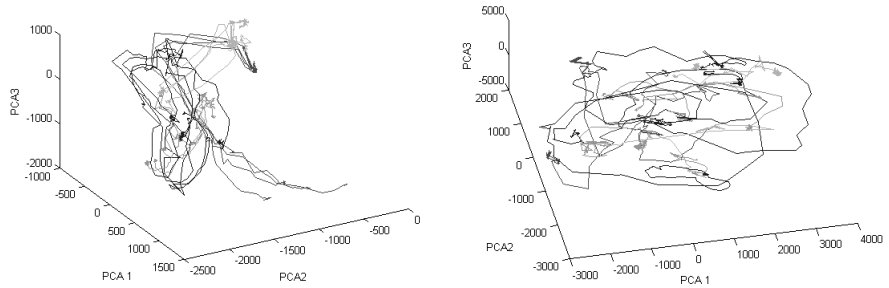
separate users (Fig. 3a and 3b) by default are discriminated by visible greater scale, especially if the same gestures are analyzed as for instance in Fig. 3a.

4 Results and Conclusions

In Fig. 4 the aggregated, total variance cover of the first n principal components reflected by the horizontal axis is presented. The single attribute is sufficient to store over than 20% percent of variance, to preserve 50% only six components are required, 60% - 10, 70% - 19, 80% - 38 and 90% - 114. The results seem to be very satisfactory, though further components are noticeably less informative and even 512 of them is insufficient to represent 99% of variance. It is very naive to expect that only few attributes are able to preserve all details of movements. However a main assessment of the dimensionality reduction is related to an obtained accuracy of classification based on reduced silhouette spaces.

In Fig 5 3D trajectories of the first three principal components for four chosen gestures labeled by different colors are shown. In case of data of a single user trajectories are clearly separated in visualized reduced space. It is more difficult to notice such a discriminative features for trajectories of different users, however it seems still to be possible to partially recognize gestures. Surely it is difficult to state final conclusions on the basis of such a visualization, because charts contain only small part of collected data, thus they are strongly simplified.

The proper classification results are presented in Tab 1 and Tab. 2. The validation experiments are iterated across different dimensionalities of reduced spaces, two different sizes of applied bounding box in the preprocessing stage and also for complete silhouette images not reduced by a bounding box detection. What is more, there are two strategies of dividing collected dataset into the training and testing sets. In the first one further called S1 the classical leave one out method is utilized, which means that training set also contains data of classified user. It is a less convenient approach in respect to practical deployments, however we think it to be still acceptable. To directly reproduce such an implementation, a prior to appropriate system running a preliminary calibration stage in which user has to perform complete set of exemplary gestures



(a) Data of a single user (b) Data of different users

Fig. 5: PCA 3D trajectories for arms zoom, hand wave, hand spinning and sliding forward gestures.

is required. In the second strategy S2 there are six iterations for every user of collected database. In a single iteration step, all data of specified user are moved to a testing set and remaining data are considered to be a training part. It means gestures are recognized on the basis of exemplary instances of other users, captured for instance in production stage of the system, thus no any extra calibration is required.

Table 1: Classification results for subsequent dimensionalities and in respect to different applied sizes of bounding box and training sets

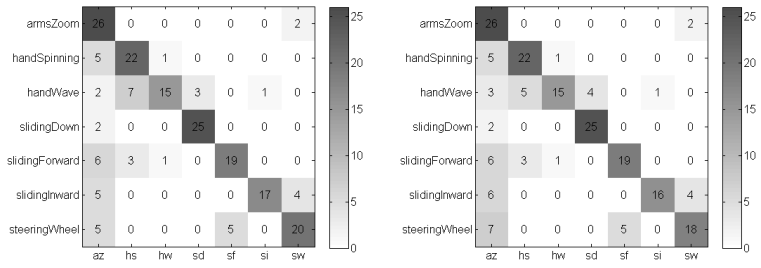
Dimensionality	No Bounding Box		Boudning Box120x120		Boudning Box80x110	
	S1	S2	S1	S2	S1	S2
1	55,61	15,31	74,49	16,33	76,02	16,33
2	65,31	14,80	87,76	34,18	88,78	34,18
4	78,57	20,92	96,94	59,69	96,94	59,69
8	85,20	28,06	97,96	71,94	97,45	69,90
16	86,22	32,14	99,49	70,92	99,49	72,45
32	89,29	37,76	99,49	71,94	99,49	73,47
64	89,29	38,27	99,49	69,39	99,49	71,43
128	89,80	37,24	99,49	69,90	99,49	70,92
256	89,80	38,27	99,49	68,88	99,49	71,94
512	89,80	38,27	99,49	69,39	99,49	71,94

For the strategy S1 very high 99.49% accuracy of recognition, expressed by percentage of correctly classified instances of a testing set, is obtained. It means only a single missclassified gesture of 196. In such a case dimensionality reduction is also very efficient. Only the first principal component allows to classify with over than 75% precision and best results are achieved for 16 dimensional spaces. Thus, general conclusion can be stated - it is possible to precisely recognize gestures on the basis of reduced sequences of human silhouettes by machine

learning in case when training data contains exemplary instances of gestures of a classified user. As it is expected it is much more difficult otherwise. It is caused by variations in movements across users for the interpreted individually gestures and different anthropometric features and postures of a human body which makes the recognition to be more challenging. Much more representative training set with greater number of users would probably minimize the influence of the first reason and classification based on the preprocessed model based skeleton data instead of raw silhouette images - the second one. Though obtained accuracy 73.48% for the strategy S2 in respect to considered seven classes is surely promising. Another observation is related to bounding box detection,

Table 2: Classification results for subsequent users

Subject	No Bounding Box		Boudning Box120x120		Boudning Box 80x110	
	DIM 16	DIM 32	DIM 16	DIM 32	DIM 16	DIM 32
1	30,30	24,24	81,82	78,79	84,85	78,79
2	17,65	26,47	29,41	32,35	32,35	32,35
3	35,48	41,94	96,77	93,55	100,00	100, 00
4	35,48	45,16	83,87	87,10	80,65	87,10
5	8,57	17,14	57,14	57,14	57,14	60,00
6	68,75	75,00	81,25	87,50	84,38	87,50



(a) Bounding Box 80x110 (b) Bounding Box 120x120

Fig. 6: Confusion matrices for 32 dimensional PCA spaces and S2 strategy.

which noticeably improves the performance of the classification. It is so, because bounding box removes from an input data their dependency on strict human location and it initially reduces the dimensionality of the input space, which makes further processing to be more efficient. What is a bit surprising smaller bounding box of the size 80x110 gives bit better results, though in some cases of very tall users or side stretched hands of T-pose silhouettes are cropped too much. Once again it can be explained in the same way - the drawback is balanced by lower dimensionality of input space.

As shown in Table 2, there is a user with id 2, which has vary poor classification accuracy of about 30%. In contrary the gestures of user 3 obtain 100% precision. If all the instances of users 2 and 3 are removed from the testing set, the average results are very alike.

To investigate similarities between gestures and their discriminative features confusion matrices, sensitivity and specificity [14] of subsequent classes are determined and presented in Fig. 6 and in Table 3 respectively. Arm zoom has a lot of false positives which are distributed quit uniformly across other gestures, but it has only two false negatives. It means there are many gestures instances which are badly recognized as arm zoom in contrary to only two missclassified arm zoom instances. It is analogous with hand wave which in some cases is badly recognized as to be hand spinning, but hand spinning is never recognized as to be hand wave. Acceptable level of sensitivity rate is obtained only for arm zoom and sliding down, but the first one has noticeably worse specificity.

Table 3: Sensitivity and specificity of gestures classes for 32 dimensional PCA spaces

	Boudning Box 80x110		Bounding Box 120x120	
	Sensitivity	Specificity	Sensitivity	Specificity
arm zoom	92,86	85,12	92,86	82,74
hand spinning	78,57	94,05	78,57	95,24
hand wave	53,57	98,81	53,57	98,81
sliding down	92,59	98,22	92,59	97,63
sliding forward	65,52	97,01	65,52	97,01
sliding inward	65,38	99,41	61,54	99,41
steering wheel	66,67	96,39	60,00	96,39

Summarizing, the obtained results are quit promising. In case of strategy S1, which requires precalibration stage of every new user, the proposed method is very precise. For much more convenient in respect to practical implementation and usage strategy S2, the further improvements have to be proposed before deployments. As described above it is expected that model based data would obtain better accuracy of recognition. Collected database is multimodal, it also contains parameters of assumed skeleton model determined by Kinect software. Thus, stated hypothesis will be verified in the next step. The linear PCA transformation seems to be sufficient to efficiently reduce dimensionality of the silhouette spaces, however it is still possible that some other nonlinear techniques would outperform it. What is more the proposed method is going to be verified on the basis of chosen challenge gesture database with greater number of instances and classes, as for instance Chalearn (http://gesture.chalearn.org).

Acknowledgments. The work was supported by The Polish National Science Centre and The Polish National Centre of Research and Development on the

basis of decision number DEC-2011/01/B/ST6/06988 and project UOD-DEM-1-183/001.

References

1. Bautista, M.n., Hernndez-Vela, A., Ponce-Lpez, V., Perez-Sala, X., Bar, X., Pujol, O., Angulo, C., Escalera, S.: Probability-based dynamic time warping for gesture recognition on rgb-d data. In: Jiang, X., Bellon, O.R.P., Goldgof, D.B., Oishi, T. (eds.) WDIA. Lecture Notes in Computer Science, vol. 7854, pp. 126–135. Springer (2012)
2. Bhattacharya, S., Czejdo, B., Perez, N.: Gesture classification with machine learning using kinect sensor data. In: Third International Conference on Emerging Applications of Information Technology. pp. 348–351. Kolkata (2012)
3. Biwas, K.K., Basu, S.K.: Gesture recognition using microsoft kinect. In: 5th International Conference on Automation, Robotics and Applications. pp. 100–103. Wellington (2011)
4. Doliotis, P., Stefan, A., McMurrough, C., Eckhard, D., Athitsos, V.: Comparing gesture recognition accuracy using color and depth information. In: Betke, M., Maglogiannis, I., Pantziou, G.E. (eds.) PETRA. p. 20. ACM (2011)
5. Gu, Y., Do, H., Ou, Y., Sheng, W.: Human gesture recognition through a kinect sensor. In: IEEE International Conference on Robotics and Biometrics. pp. 1379–1385. Guangzhou (2012)
6. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics - part C: Applications and Reviews 37, 311–324 (2007)
7. Patsadu, O., Nukoolkit, C., Watanapa, B.: Human gesture recognition using kinect camera. In: Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE). pp. 28–32. Bangkok (2012)
8. Ren, Z., Meng, J., Yuan, J., Zhang, Z.: Robust hand gesture recognition with kinect sensor. In: 19th ACM international conference on Multimedia. pp. 759–760. New York (2011)
9. Sakoe, H., Chuba, S.: Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 8, 43–49 (1978)
10. Switonski, A., Joinski, H., Zghidi, H., Wojciechowski, K.: Selection of pose configuration parameters of motion capture data based on dynamic time warping. In: 12th International Conference of Numerical Analysis and Applied Mathematics. Rodos (2014)
11. Switonski, A., Joisski, H., Jedrasiak, K., Polanski, A., Wojciechowski, K.: Classification of poses and movement phases. Lecture Notes in Computer Science 6374, 193–200 (2010)
12. Switonski, A., Joisski, H., Jedrasiak, K., Polanski, A., Wojciechowski, K.: Diagnosis of the motion pathologies based on a reduced kinematical data of a gait. Electrical Review 57, 173–176 (2011)
13. Wang, Y., Yang, C., Wu, X., Xu, S., Li, H.: Kinect based dynamic hand gesture recognition algorithm research. In: 4th International Conference on Intelligent Human-Machine Systems and Cybernetics. pp. 274–279. Nanchang (2012)
14. Witten, I., Frank, E., Hall, M.: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Boston, Massachusetts (2011)