Hand Gesture Recognition using a Convolutional Neural Network

Eirini Mathe[1], Alexandros Mitsou[3], Evaggelos Spyrou[1,2,3] and Phivos Mylonas[4]
[1]*Institute of Informatics and Telecommunications National Center for Scientific Research – "Demokritos,"* Athens, Greece
[2]*Department of Computer Engineering, Technological Education Institute of Sterea Ellada,* Lamia, Greece
[3]*Department of Computer Science, University of Thessaly,* Lamia, Greece
[4]*Department of Informatics*, Corfu, Greece
email: emathe@iit.demokritos.gr, amitsou95@gmail.com, espyrou@iit.demokritos.gr, fmylonas@ionio.gr

***Abstract*—In this paper we present an approach towards hand gesture recognition that uses a Convolutional Neural Network (CNN), which is trained on Discrete Fourier Transform (DFT) images that result from raw sensor readings. More specifically, we use the Kinect RGB and depth camera and we capture the 3D positions of a set of skeletal joints. From each joint we create a signal for each 3D coordinate and we concatenate those signals to create an image, the DFT of which is used to describe the gesture. We evaluate our approach using a dataset of hand gestures involving either one or both hands simultaneously and compare the proposed approach to another that uses hand-crafted features.**

I. Introduction

Poses and gestures are one the basic means of communication between humans while they may also play a crucial role in human-computer interaction, as they are able to transfer some kind of meaning. The research area of pose and gesture recognition aims to recognizing such expressions, which typically involve some posture and/or motion of the hands, arms, head, or even skeletal joints of the whole body. In certain cases, meaning may differ, based on a facial expression. Several application areas may benefit from the recognition of a human's pose or the gestures she/he performs, such as sign language recognition, gaming, medical applications involving the assessment of a human's condition and even navigation in virtual reality environments. There exist various approaches and techniques which involve some kind of sensor, either "worn" by the subject (e.g., accelerometers, gyroscopes etc.), or monitor the subject's motion (e.g., cameras). In the latter case, the subject may also wear special "markers" which are used to assist in the identification of several body parts and/or skeletal joints. However, during the last few years several approaches rely solely on a typical RGB camera, enhanced by depth information. One such example is the well-known Kinect sensor. Therefore, the user only needs to stand in front of the camera without wearing any kind of sensor. Several parts of her/his body are detected and tracked in the 3D space. Typically, features are extracted and are used for training models to recognize poses and/or gestures.

In this paper, we present a gesture recognition approach that focuses on hand gestures. We propose a novel deep learning architecture that uses a Convolutional Neural Network (CNN). More specifically, we use the Kinect sensor and its Software Development Kit (SDK) in order to detect and track the subject's skeletal joints in the 3D space. We then select a subset of these joints, i.e., all that are involved at any of the gestures of our data set. Then, we create an artificial image based on these 3D coordinates. We apply the Discrete Fourier Transform on these images and use the resulting ones to train the CNN. We compare our approach with previous work [19], where a set of hand-crafted statistical features on joint trajectories had been used. Finally, we demonstrate that it is possible to efficiently recognize hand gestures without the need of a feature extraction step. Evaluation takes place using a new dataset of 10 hand gestures.

The rest of this paper is organized as follows: Section II presents related research works in the area of gesture recognition using skeletal data, focusing on those that are based on deep learning. Section III presents the Kinect sensor and its SDK which was used for the extraction of raw data, the methodology for converting those data to the input image and also the proposed CNN that was used for hand gesture recognition. The dataset and the experimental results are presented in Section IV. Finally, in Section V, we draw our conclusions and discuss plans for future work.

II. Related Work

The problem of hand gesture recognition has attracted many research efforts during the last decade. In this section our goal is to present works that aim to recognize simple hand gestures, i.e., as the ones we also aim to recognize in the context of this work. We shall present both approaches that use traditional machine learning techniques and also approaches that are based on deep learning.

Feature-based works typically rely on traditional machine learning approaches such as artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs) or K-nearest neighbor classifiers (KNN). In [3] SVMs and DTs are trained on 3D skeletal joint coordinates, while in [10] distances in reference to the *Spine Center* are used as features with a KNN classifier. Within the approach of [13], cascades of ANNs are used to firstly classify the gesture side (left/right) and then recognize the gesture type. In the work of [14] SVMs are to recognize distinctive key poses, while Decision Forests are then used to recognize gestures as sequences of such poses. Other approaches [4], [21], [20], [7] exploit the fact that a gesture may be considered as a temporal sequence that may vary in terms of speed and use the Dynamic Time Warping Algorithm (DTW) [2], while in

some cases, a machine learning algorithm complements the classification scheme. Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) have also been frequently used [1], [5], [6], [25], [28] since they are able to solve temporal pattern recognition problems. Their input typically consists of 3D positions of skeletal joints or body parts. In previous work [19], we proposed a novel approach for feature extraction using measurements on skeletal joints and evaluated it using several machine learning algorithms.

In [18], the authors propose an approach for the recognition of hand gestures from the American Sign Language using CNNs and auto encoders. 3D CNNs are used in [15] to detect drivers' hand gestures and in [9] for continuous gesture recognition. Moreover, in [22] a CNN is used to for hand gesture recognition, which would then serve as an estimation of hand pose and orientation. Several CNN architectures to predict 3D joint locations of a hand given a depth map are investigated in [17]. A synthetically trained CNN is presented in [26]. A method based on multi-scale and multimodal deep learning for is presented in [16]. A two-stage approach that combines feature learning from RGB and depth with CNNs and PCA to get the final features is presented in [12]. Deep dynamic NNs are used in [27] for both gesture segmentation and recognition. A Deep Belief Network and a CNN are used in [24] for learning features that are insensitive to movement, scaling, and rotation from hand posture images.

III. Hand Gesture Recognition

In the context of this work, we may define a *gesture* as the transfer of a subset of joints from point A to point B, using a vaguely predefined trajectory. Note that a gesture may also be defined as a continuous stream of poses, between A and B, i.e., an ordered of temporary motion suspensions where body joints assume a distinct configuration in space. Also, we limit our recognition in hand gestures, i.e., gestures that are based on motion of hands, yet involving palms, wrists, elbows and shoulders. In this section we present the proposed approach for classification of hand gestures. In brief, from a given set of video frames we obtain the 3D coordinates of a set of skeletal joints. These are concatenated to form a *signal* image which is then transformed to an *activity* image using the Discrete Fourier Transform. The proposed Convolutional Neural Network is trained using these activity images. A graphic representation of the proposed methodology is illustrated in Fig. 1.

A. The Microsoft Kinect SDK

The Microsoft Kinect sensor [29] is used to capture the raw skeletal joint motion data. Kinect is both an RGB and depth camera. Moreover, using its SDK, one may extract the 3D position (i.e., x, y and z coordinates of a human's skeletal joints in real time. More specifically, a structured graph of joints is continuously streamed. Graph nodes correspond to the most representing body parts (e.g., skeletal joints of arms, legs, head, etc.), while graph edges follow the structure of joints within the human body. A parent–child relationship is implied from top to bottom, i.e., *Head* is the parent of *Shoulder Center*, which is the parent of *Shoulder Left* and *Shoulder Right* etc. In Fig. 2 we illustrate the 20 human skeleton joints, that are extracted with the Kinect SDK.

B. Data processing

From the extracted set of skeletal joints, we select a subset of them that is typically involved in hand gestures. More specifically, from the aforementioned 20 skeletal joints we use the following 8: *Hand Left*, *Wrist Left*, *Elbow Left*, *Shoulder Left*, *Hand Right*, *Wrist Right*, *Elbow Right* and *Shoulder Right*.

Inspired by the work of Jiang and Yin [8], we first create an activity image, which based on the signals that are produced for each coordinate of each skeletal joint. This way, 24 raw signals are concatenated to produce a signal image. Then, the 2D Discrete Fourier Transform (DCT) is applied to the signal image. We then create an "activity" image by discarding phase information, i.e., keeping only the magnitude. Example signal images and activity images for each hand gesture are illustrated in Fig. 3. Note the visual difference of these images which may not be significant, yet allows the CNN to learn the differences between two classes.

Our work focuses only on the classification of a gesture into a set of predefined classes. Therefore, we should clarify that it does not perform any temporal segmentation to detect the beginning and the ending of a possible gesture; instead we consider this problem as solved. We use pre-segmented videos and only aim to recognize gestures per segment; each may contain at most one gesture. Since gestures typically vary in terms of duration even when performed by the same user, we perform a linear interpolation step to ensure the same length on all gesture sequences in order to concatenate them to form the signal image. This way, all signal images have size equal to 24×60.

C. CNN architecture

Deep learning approaches have been recently playing a key role in the field of machine learning in general. The are of computer vision is among those that have benefited the most. Several deep architectures have been proposed during the last few years. Nonetheless, the Convolutional Neural Networks (CNNs) [11] still remain the dominant deep architecture in computer vision. A CNN resembles a traditional neural network (NN), yet it differs since its goal is to learn a set of convolutional filters. However, training takes place as with every other NN; a forward propagation of data and a backward propagation of error do take place to update weights.

The key component of a CNN are the *convolutional* layers. They are formed by grouping neurons in a rectangular grid. The training process aims to learn the parameters of the convolution. *Pooling* layers are usually placed after a single or a set of serial or parallel convolutional layers and take small rectangular blocks from the convolutional layer and subsample them to produce a single output from each block. Finally, *dense* layers (also known as "fully-connected" layers) perform classification using the features that have been extracted by the convultional layers and then have been subsampled by the pooling layers. Every node of a dense layer is connected to all nodes of its previous layer.

Moreover, our approach makes use of the *dropout* regularization technique [23]. Following this technique, at each training stage several nodes are "dropped out" of the net in

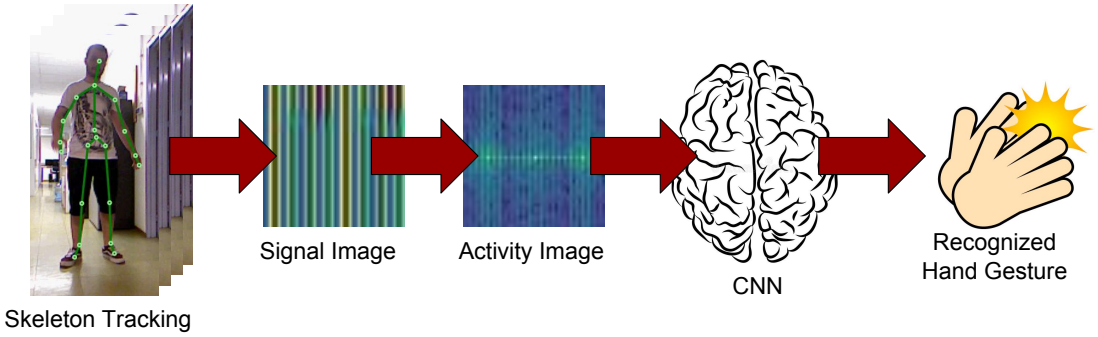

Fig. 1. Steps of the proposed method. Skeleton joints are tracked during a gesture. Then, the signal image is constructed. Upon application of the DFT, the activity image is constructed and given as input to the CNN. Finally, the CNN classifies the activity image to a class, i.e., *clapping*.

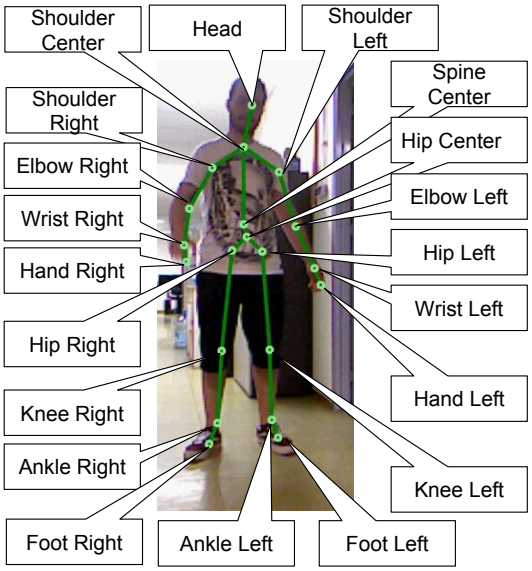

Fig. 2. Extracted human skeleton 3D joints using the Kinect SDK.

order to reduce or even eliminate overfitting by preventing complex co-adaptations on training data.

The architecture of our proposed CNN is presented in detail in Fig. 4. The first convolutional layer filters the 24×60 input activity image with 16 kernels of size 3×3. A second convolutional layer filters the 24×60 resulting image with 16 kernels of size 3×3. Then the first pooling layer uses "max-pooling" to perform 3×3 subsampling. The third convolutional layer filters the 11×29 resulting image with 32 kernels of size 3×3. Then, the fourth convolutional layer filters the 11×29 resulting image with 128 kernels of size 3×3. A second pooling layer uses "max-pooling" to perform 2×2 subsampling. Then, a flatten layer transforms the output image of size 5×14 of the second pooling to a vector, which is then used as input to a dense layer using dropout. Finally, a second dense layer using dropout produces the output of the network.

IV. Experiments

A. Data Set

In order to evaluate the proposed approach, we have constructed a real-life dataset. More specifically, we have involved 10 users (5 male and 5 female, aging between 22 and 30 years old) in the process. In order to obtain spontaneous gestures, none of these users was neither involved in the research conducted in the context of this work, nor had been previously involved in similar tasks. Moreover, we only provided them with an intuitive description of the way gestures should be performed, allowing "noise" in the collection process. This way, we feel that the collected gestures were as "natural" as possible. The set of gestures consists of the following: *Swipe-left*, *Swipe-right*, *Swipe-up-left*, *Swipe-up-right*, *Swipe-down-left*, *Swipe-down-right*, *Swipe-diagonal-left*, *Clapping*, *Two-hands-raise* and *Two-hands-circle*. In Fig. 5 we illustrate a sequence of RGB frames with overlaid the skeleton for several gestures; the remaining may be easily conceived, since they are symmetric to some the aforementioned.

Note that we have asked all users to perform each gesture at least 10 times and then we manually cleaned the data set by removing the wrongly performed gestures and keeping only 5 samples per gesture, per user. This resulted to a total of 500 gestures. All samples were manually segmented using a video editor; sample gestures begin approx. 100msec upon the beginning of the corresponding video, while end approx. 100msec prior to its ending. This way we were able to isolate gestures from other movements. For each gesture, we also recorded the relevant skeletal data and extracted the 3D coordinates and the corresponding sets of features of [19], for all the aforementioned joints involved in these gestures, i.e., Shoulders, Elbows, Wrists and Hands. 8 users were used for training, while the 2 remaining ones were used for testing.

B. Implementation Details

For the implementation of the CNN we used Keras[1] running on top of Tensorflow[2]. The aforementioned data processing step was implemented in Python 3.6 using NumPy[3] and SciPy[4].

[1] https://keras.io/
[2] https://www.tensorflow.org/
[3] http://www.numpy.org/
[4] https://www.scipy.org/

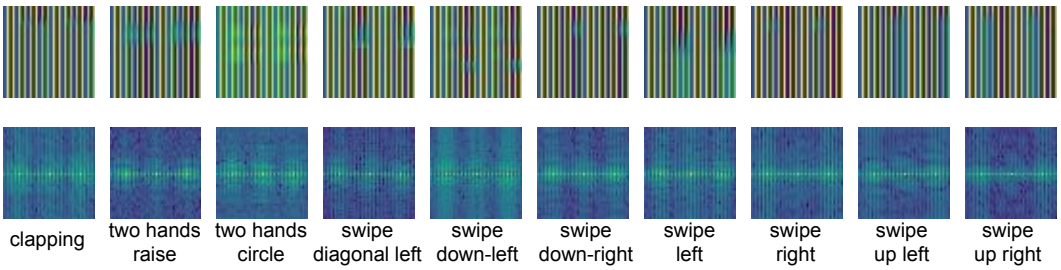

Fig. 3. Sample images from all 10 gestures used throughout our experiments: upper row: signal images; lower row: activity images.

TABLE I. EXPERIMENTAL RESULTS.

	proposed	[19] - ERT	[19] - SVM
accuracy	**0.90**	0.82	0.73

C. Experimental Results

Using the aforementioned dataset, we have evaluated the proposed CNN architecture. We compared it to our previous work[19] that was based on a set of hand-crafted features. More specifically, we used an SVM with RBF (i.e., non-linear) kernel and also an ERT. ERTs showed the best performance among a set of well-known classifiers. However, the proposed CNN showed superior performance to both these approaches. Results are summarized in Table I

V. CONCLUSION

In this paper we presented initial results of a novel approach for recognition of hand gestures, using a convolutional neural network. Our approach was based on the creation of "activity" images. Each activity image resulted upon application of the discrete Fourier Transform to a given signal image. This single image was constructed upon concatenation of raw signals that correspond to the location of skeletal joints in the 3D space.

We compared the proposed approach to previous work that used a set of hand-crafted features and two machine learning algorithms, namely extremely randomized trees (ERTs) and support vector machines (SVMs) with non-linear kernels. To this goal we used a custom dataset that consisted of 10 gestures involving either one or both hands. Although the performance of the ERT was satisfactory, as expected, still it was inferior to the one of the proposed CNN architecture. However, we should note that the produced models from the ERT and the SVM were far less complex than the one of the CNN, which means that less time would be needed for prediction, something crucial for certain applications.

Future work will focus on the following: a) investigation on methods for creating the signal image, possibly with the use of other types of sensor measurements such as wearable accelerometers, gyroscopes etc.; b) investigation on methods for transforming the signal image to the activity image, such as wavelets, discrete cosine transformation etc.; c) evaluation of the proposed approach on larger public datasets; and d) application into a real-like assistive living environment, for real-time detection of human actions.

ACKNOWLEDGMENT

We acknowledge support of this work by the project SYNTELESIS "Innovative Technologies and Applications based on the Internet of Things (IoT) and the Cloud Computing" (MIS 5002521) which is implemented under the "Action for the Strategic Development on the Research and Technological Sector", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

REFERENCES

[1] Anuj, A., Mallick, T., Das, P.P. and Majumdar, A.K., 2015, December. Robust control of applications by hand-gestures. In Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2015 Fifth National Conference on (pp. 1-4). IEEE.

[2] Albrecht, T. and Muller, M., 2009. Dynamic time warping (dtw). Information Retreival for Music and Motion, pp.70-83.

[3] Bhattacharya, S., Czejdo, B. and Perez, N., 2012, November. Gesture classification with machine learning using kinect sensor data. In Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on (pp. 348-351). IEEE.

[4] Celebi, S., Aydin, A.S., Temiz, T.T. and Arici, T., 2013, February. Gesture recognition using skeleton data with weighted dynamic time warping. In VISAPP (1) (pp. 620-625).

[5] Gonzalez-Sanchez, T. and Puig, D., 2011. Real-time body gesture recognition using depth camera. Electronics Letters, 47(12), pp.697-698.

[6] Gu, Y., Do, H., Ou, Y. and Sheng, W., 2012, December. Human gesture recognition through a kinect sensor. In Robotics and Biomimetics (ROBIO), 2012 IEEE International Conference on (pp. 1379-1384). IEEE.

[7] Ibanez, R., Soria, ., Teyseyre, A. and Campo, M., 2014. Easy gesture recognition for Kinect. Advances in Engineering Software, 76, pp.171-180.

[8] Jiang, W. and Yin, Z., 2015, October. Human activity recognition using wearable sensors by deep convolutional neural networks. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 1307-1310). ACM.

[9] Camgoz, N.C., Hadfield, S., Koller, O. and Bowden, R., 2016, December. Using convolutional 3d neural networks for user-independent continuous gesture recognition. In Pattern Recognition (ICPR), 2016 23rd International Conference on (pp. 49-54). IEEE.

[10] Lai, K., Konrad, J. and Ishwar, P., 2012, April. A gesture-driven computer interface using Kinect. In Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on (pp. 185-188). IEEE.

[11] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

[12] Li, S.Z., Yu, B., Wu, W., Su, S.Z. and Ji, R.R., 2015. Feature learning based on SAEPCA network for human gesture recognition in RGBD images. Neurocomputing, 151, pp.565-573.

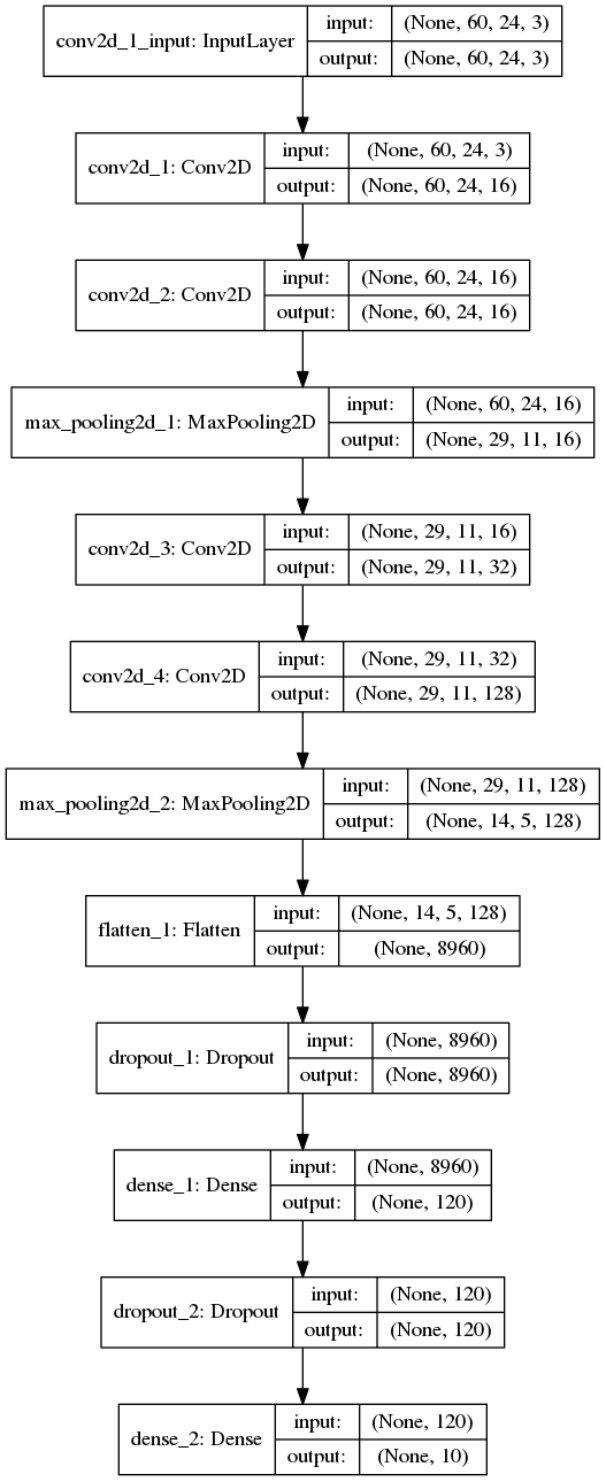

Fig. 4. The proposed CNN architecture.

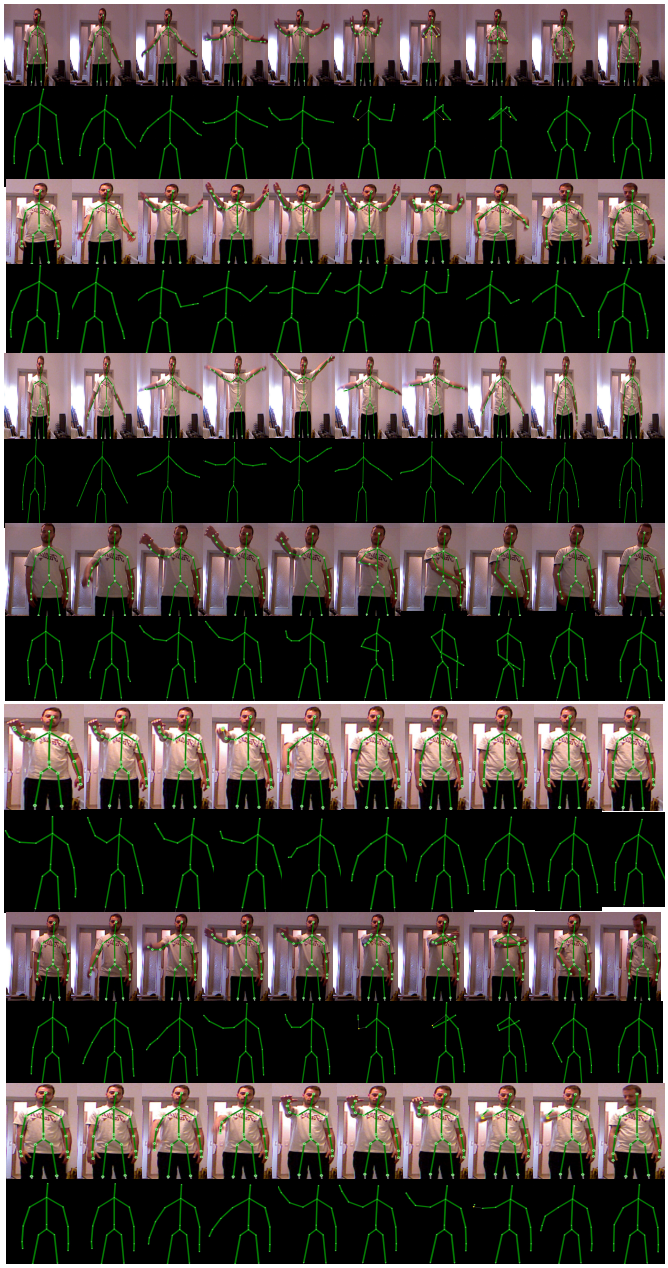

Fig. 5. A subset of the available gestures within the data set used. From top to bottom: *clapping*, *two hands raise*, *two hands circle*, *swipe diagonal left*, *swipe down left*, *swipe left*, *swipe up left*.

[13] Mangera, R., Senekal, F. and Nicolls, F., 2014. Cascading neural networks for upper-body gesture recognition. Avestia Publishing.

[14] Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A.W. and Campos, M.F., 2012, August. Real-time gesture recognition from depth data through key poses learning and decision forests. In Graphics, patterns and images (SIBGRAPI), 2012 25th SIBGRAPI conference on (pp. 268-275). IEEE.

[15] Molchanov, P., Gupta, S., Kim, K. and Kautz, J., 2015. Hand gesture recognition with 3D convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 1-7).

[16] Neverova, N., Wolf, C., Taylor, G.W. and Nebout, F., 2014, September. Multi-scale deep learning for gesture detection and localization. In Workshop at the European conference on computer vision (pp. 474-490). Springer, Cham.

[17] Oberweger, M., Wohlhart, P. and Lepetit, V., 2015. Hands deep in deep learning for hand pose estimation. arXiv preprint arXiv:1502.06807.

[18] Oyedotun, O.K. and Khashman, A., 2017. Deep learning in vision-based static hand gesture recognition. Neural Computing and Applications, 28(12), pp.3941-3951.

[19] Paraskevopoulos, G., Spyrou. E. and Sgouropoulos, D., 2016. A Real-Time Approach for Gesture Recognition using the Kinect Sensor. Proceedings of the 9th Hellenic Conference on Artificial Intelligence

(SETN).

[20] Rib, A., Warcho, D. and Oszust, M., 2016. An Approach to Gesture Recognition with Skeletal Data Using Dynamic Time Warping and Nearest Neighbour Classifier. International Journal of Intelligent Systems and Applications, 8(6), p.1.

[21] Reyes, M., Dominguez, G. and Escalera, S., 2011, November. Featureweighting in dynamic timewarping for gesture recognition in depth data. In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on (pp. 1182-1188). IEEE.

[22] Sanchez-Riera, J., Hsiao, Y.S., Lim, T., Hua, K.L. and Cheng, W.H., 2014, July. A robust tracking algorithm for 3d hand gesture with rapid hand motion through deep learning. In Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on (pp. 1-6). IEEE.

[23] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), pp.1929-1958.

[24] Tang, A., Lu, K., Wang, Y., Huang, J. and Li, H., 2015. A real-time hand posture recognition system using deep neural networks. ACM Transactions on Intelligent Systems and Technology (TIST), 6(2), p.21.

[25] Tran, C. and Trivedi, M.M., 2012. 3-D posture and gesture recognition for interactivity in smart spaces. IEEE Transactions on Industrial Informatics, 8(1), pp.178-187.

[26] Tsai, C.J., Tsai, Y.W., Hsu, S.L. and Wu, Y.C., 2017, May. Synthetic Training of Deep CNN for 3D Hand Gesture Identification. In Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), 2017 International Conference on (pp. 165-170). IEEE.

[27] Wu, D., Pigou, L., Kindermans, P.J., Le, N.D.H., Shao, L., Dambre, J. and Odobez, J.M., 2016. Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE transactions on pattern analysis and machine intelligence, 38(8), pp.1583-1597.

[28] Yin, Y. and Davis, R., 2014, July. Real-time continuous gesture recognition for natural human-computer interaction. In Visual Languages and Human-Centric Computing (VL/HCC), 2014 IEEE Symposium on (pp. 113-120). IEEE.

[29] Zhang, Z., 2012. Microsoft kinect sensor and its effect. IEEE multimedia, 19(2), pp.4-10.