The 23rd IEEE International Symposium on Robot and
Human Interactive Communication
August 25-29, 2014. Edinburgh, Scotland, UK,

HandSOM - Neural Clustering of Hand Motion for Gesture Recognition in Real Time

German I. Parisi, Doreen Jirak and Stefan Wermter

Abstract— **Gesture recognition is an important task in Human-Robot Interaction (HRI) and the research effort towards robust and high-performance recognition algorithms is increasing. In this work, we present a neural network approach for learning an arbitrary number of labeled training gestures to be recognized in real time. The representation of gestures is hand-independent and gestures with both hands are also considered. We use depth information to extract salient motion features and encode gestures as sequences of motion patterns. Preprocessed sequences are then clustered by a hierarchical learning architecture based on self-organizing maps. We present experimental results on two different data sets: command-like gestures for HRI scenarios and communicative gestures that include cultural peculiarities, often excluded in gesture recognition research. For better recognition rates, noisy observations introduced by tracking errors are detected and removed from the training sets. Obtained results motivate further investigation of efficient neural network methodologies for gesture-based communication.**

I. INTRODUCTION

Human beings perform gestures rather unconsciously in everyday life, for instance, when we explain the shape of an object or the way to a station. Thus, gestures provide a useful visual complement to support language. Gestural understanding is a significant part in our communication and is often referred to as co-speech [1]. Gestures are also the essential visual channel for hearing-impaired and deaf people who must rely on sign language.

The underlying neural processes which are involved in gesture understanding are still under investigation [2]. However, from a bio-psychological perspective, humans pay attention to a particular object for tracking [3]. The retina uses motion information to follow the region of interest, i.e. the perceiver ultimately processes a trajectory [4]. This is an effective way to separate the foreground from the background, thus attractive to be integrated in gesture recognition systems. From different motion trajectories we can infer the gestural meaning by classification.

In the last decades, researchers explored multiple approaches and used different devices to provide interfaces, ranging from mouse-based tools to data gloves, hand- and arm markers, and multi-camera setups. However, cables and extra devices are quite cumbersome, and cost factors and calibration issues additionally hinder a natural gesture interface. Vision-based approaches provide the most intuitive interfaces for the recognition of gestures without the use of invasive devices. On the other hand, they are characterized by high computational demand and may therefore not perform in real time. The emergence of new and cost-effective sensor technologies such as Microsoft Kinect[1] and ASUS Xtion[2] simplifies access to color and depth information, allowing better performance in terms of computational complexity for estimating the position of objects in real-world coordinates.

We propose a learning framework for the recognition of hand gestures with the following major interests:

1) An intuitive and robust interface for HRI using a depth sensor and no other additional equipment;
2) The representation of different types of gestures, i.e. simple commands and cultural co-speech signs, taking into account not only hand positions but also, e.g., head distance and arm angle;
3) The automatic segmentation and recognition of a set of training gestures with low latency, providing real-time characteristics.

For our approach, we extract hand motion from depth map videos and encode gestures as hand-independent motion sequences. We only extract information of the most salient moving hand. In case that both hands are used, the type of interaction between the hands is considered. To collect a training set of gestures, the system is presented a number of video clips from which the gestures are automatically segmented. Motion sequences are then clustered by a two-stage learning architecture based on self-organizing maps (SOM) trained with labeled samples for classification purposes. Firstly, noisy observations introduced by tracking errors are detected and removed from the set of motion vectors. Secondly, labeled sequences are processed through the hierarchy of SOM networks in terms of motion trajectories. An overall overview of the learning framework is depicted in Fig. 1. We run experiments on two data sets with command-like gestures for HRI scenarios and Italian communicative gestures, each set with 10 different gesture classes.

II. RELATED WORK

Using depth information to develop a gesture-based digit recognition system was proposed in [9]. Experiments included a subject performing a gesture and at least one other non-performing person in the background. The authors used the RGB image to detect skin-color for hand- and face recognition using histograms and motion information. Depth

German I. Parisi, Doreen Jirak and Stefan Wermter are with the Department of Informatics, University of Hamburg, Vogt-Koelln-Strasse 30, D-22527 Hamburg, Germany {parisi,jirak,wermter}@informatik.uni-hamburg.de

[1]Kinect for Windows. http://www.microsoft.com/en-us/kinectforwindows

[2]ASUS Xtion Pro Live. http://www.asus.com/Multimedia/Motion_Sensor/Xtion_PRO_LIVE/

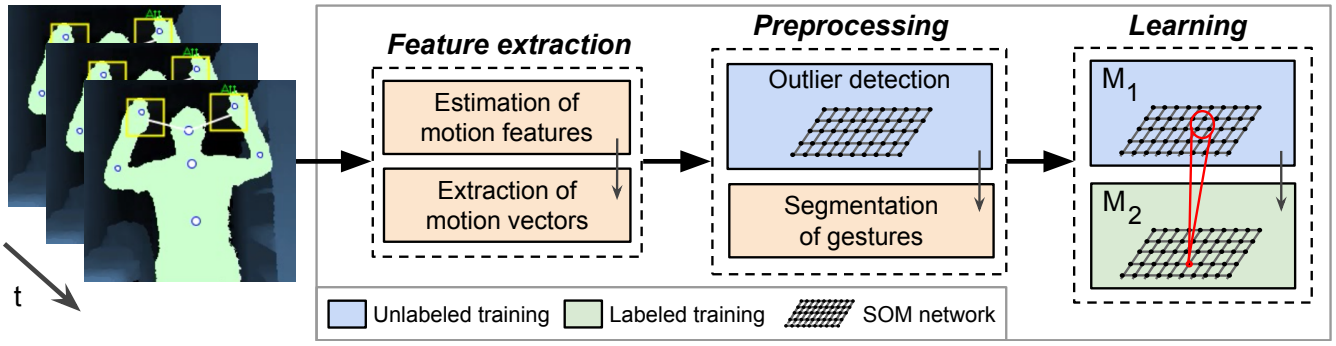

Fig. 1. Overview of the HandSOM framework: feature extraction from depth information, preprocessing for the unsupervised detection of outliers and the segmentation of gestures, and SOM-based hierarchical learning of gesture sequences.

information was used to estimate the hand position. For the training, gesture data derived from subjects wearing a green glove needed to generate expressive trajectories. While testing the system this condition was excluded. The actual classification was carried out using Dynamic Time Warping (DTW), which due to the nature of this method accounts for inter-user variability in the length of individual gesture performance. The authors showed that their detector using depth is comparable to a classification with the accurate trajectories derived for training purposes. However, no automatic temporal segmentation is provided, thus the start- and end frames needed to be manually labeled. In addition, the gesture set was quite limited, so that no real HRI-scenario could be established.

Using SOMs and modeling gestures as time-series was combined in a Self-Organizing Markov Model (SOMM) approach developed in [5]. Tracked hand coordinates and hand direction, computed as the optical flow, were used as input for the SOM network. The Best-Matching-Unit nodes of the SOM grid served then as state trajectories in the subsequent implemented Hidden Markov Model (HMM). The system was tested using 30 artificial gestures, as the goal was to account for intra- and inter subject variability in gesture performance. In [6] the authors proposed a system for hand gesture recognition which makes use of attention-based properties of the human visual system. The authors used saliency-maps computing potential hand candidates, which were further integrated to provide the central-foci features coding for the motion. and the classification with a SVM. The gesture vocabulary consisted of 7 signs, e.g. rectangle, performed frontally to the camera and with static background. Start- and end frame were again annotated manually.

A recent gesture interface for HRI comprising Growing Neural Gas (GNG) and reinforcement learning was introduced in [7]. Using 3D information from the Kinect sensor an assistive robot should learn user specific motion and commands. Three gestures were provided as initial steps towards the development of the underlying scenario and implementation. The motion descriptors were estimated using the concept of dynamic instants (DI), which entail acceleration information. The descriptors were then fed into a GNG clustering procedure. Finally, the user provides feedback to their system enabling the robot to perform the according action. As the approach at an early stage, claimed by the author as proof-of-concept, there is no thorough evaluation of their architecture in terms of HRI-related scenarios.

III. MOTION REPRESENTATION

We extract motion properties that describe the dynamics of gestures in terms of sequences of spatiotemporal features. For this purpose, we consider a set of motion descriptors for a given set of tracked body joints, i.e. hands and head. The descriptors are encoded with a saliency-based approach and subsequently processed as motion sequences.

We estimate the position of body joints from a 3D model of the human skeleton. Body joints are represented as a point sequence of real-world coordinates $C = (x, y, z)$. Head and hand joints are tracked as the 3D position of their estimated centers of mass. We obtain the joints R_i and L_i for the right and left hand respectively and use them to calculate a set of hand motion descriptors at time i. We compute the pixels' difference $D_i = (d^x, d^y, d^z)$ between two consecutive frames for J_i and J_{i-1}, and then estimate the intensity of motion with respect to the sensor as

$$V_i = \{\frac{d^x}{v}, \frac{d^y}{v}, \frac{d^z}{v}\} \tag{1}$$

where $v = \sqrt{(d^x)^2 + (d^y)^2 + (d^z)^2}$. To estimate the position of a hand J_i with respect to the head on the image plane, we firstly compute the polar angle between J_i and the head joint H_i as

$$\varphi_i = arctan(\frac{J^y - H^y}{J^x - H^x}). \tag{2}$$

We finally compute the hand-head distance as

$$h_i = \sqrt{(H_i^x - J_i^x)^2 + (H_i^y - J_i^y)^2}. \tag{3}$$

This approach allows to describe spatially articulated gestures, in which not only motion but also the positions of joints with respect to the body are relevant. This additional joint provides therefore a reference for a more informative hand position with respect to the upper body, important for semantic gestures in human communication (e.g. Italian gestures, where also the head is used as a reference point.).

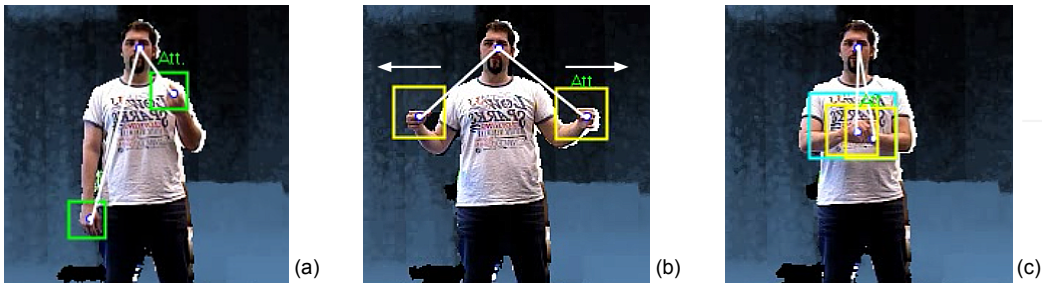

Fig. 2. Saliency-based gesture encoding and hand interactions: (a) Only the information from the most salient hand is encoded (Att.), (b) symmetric interaction when moving hands in opposite directions and (c) joint (physical and symmetric) interaction on "washing hands" gesture.

A. Saliency-based Encoding

A number of gesture recognition models encode motion information from both hands without taking into account which hand is performing a gesture [8]. This common approach may often add avoidable complexity to the model, e.g. leading to motion representations of unnecessary higher dimensionality. For our approach, we encode only the information of the most salient hand in terms of apparent motion (Fig. 2.a), which is more likely to attract the attention of the observer. For gestures in which both hands are jointly used we consider three types of interaction: physical, symmetric, and joint interaction (Fig. 2.b-c). Physical interaction occurs when the hands touch or overlap in the visual space as perceived by an aware observer. On the other side, two hands moving asynchronously or in opposite directions can be linked by the semantic meaning of the specific gesture being performed. Symmetric interactions aim to model the cognitive process in which two hands, even if not perceptually overlapped, are tracked as a whole since they contribute to a single gesture (Fig. 2.b). Joint interaction occurs when both physical and symmetric properties are detected (Fig. 2.c).

For a tracked target at time i, we obtain the following motion vector

$$M_i = (s_i, V_i, \varphi_i, h_i, \lambda_i) \quad , \tag{4}$$

where s_i is the type of hand interaction (0=none, 1=physical, 2=symmetric, 3=joint), λ_i is the annotated gesture label, and V_i, φ_i, h_i are defined by Eq. 1, 2, and 3 respectively.

The purpose of our saliency-based encoding is threefold. Firstly, we reduce the amount of information required to represent a gesture. Secondly, we are able to capture hand-independent gestures by considering the most active hand. Lastly, we obtain a length-invariant representation of gestures convenient for neural network based clustering.

B. Preprocessing of Gesture Sequences

To collect a data set of gestures, the system is presented a set of video streams from which the gestures are extracted in terms of motion vectors. Each video sequence contains a specific gesture being performed an arbitrary number of times. After annotating gesture labels, we apply two preprocessing steps: 1) detection and removal of outliers from extracted motion vectors, and 2) segmentation of gestures from the set of denoised motion vectors to create the training set.

The first step aims to address tracking errors that may lead to outliers in the data. An outlier is seen as an observation that does not follow the pattern suggested by dominating data clouds [11]. We consider inconsistent changes in hand joint velocity, e.g. isolated peak values, to be caused by tracking errors rather than actual tracked motion. Therefore, we use a SOM-based approach to detect and remove outliers from the training and the test set (see Section IV).

The second step consists of the automatic segmentation of gestures from the denoised motion vectors. The idea is to keep only the motion vectors that belong to the execution of a gesture, while removing the others. Our assumption is that gestures are performed in a given area of interest, where significant hand motion occurs during the training sessions. For this purpose, we define a dynamic area of interest expressed in terms of distance from the head with the empiric threshold value

$$\alpha = max(H) - \sigma(H) \quad , \tag{5}$$

where H is the set of hand-head distance values for a the encoded hand joint, $max(H)$ is the maximum value of H (maximum distance from the head), and $\sigma(H)$ is the standard deviation. Thus, our training set includes only hand motions (Eq. 4) for which the condition $h_i < \alpha$ is satisfied. This process leaves out motion vectors for which, e.g., hands are hanging at the sides of the body.

IV. LEARNING FRAMEWORK

The SOM algorithm [12] has shown to be a compelling approach for clustering motion expressed in terms of multi-dimensional flow vectors [10], [11]. The traditional SOM is unsupervised and allows to obtain a low-dimensional discretized representation from high-dimensional input spaces. It consists of a layer with competitive neurons connected to adjacent units by a neighborhood relation. The network learns by iteratively reading each training vector and organizes the units so that they describe the domain space of input observations. Each unit j is associated with a d-dimensional model vector $m_j = [m_{j,1}, m_{j,2}, ..., m_{j,d}]$. For each input vector $x_i = (x_1, ..., x_n)$ presented to the network, the best matching unit (BMU) b for x_i is selected by the smallest

Euclidean distance as

$$b(x_i) = \arg\min_j \|x_i - m_j\| \quad . \tag{6}$$

For an input vector x_i, the quantization error q_i is defined as the distance of x_i from $b(x_i)$. We consider two-dimensional networks with units arranged on a hexagonal lattice in the Euclidean space. Each competitive network is trained with a batch variant of the SOM algorithm. This iterative algorithm presents the whole data set to the network before any adjustments are made. The updating is done by replacing the model vector m_j with a weighted average over the samples:

$$m_j(t+1) = \frac{\sum_{i=1}^{n} h_{j,b(i)}(t) x_i}{\sum_{i=1}^{n} h_{j,b(i)}(t)} \quad , \tag{7}$$

where b is the best matching unit (Eq. 6), n is the number of sample vectors, and $h_{j,b(i)}$ is a Gaussian neighborhood function:

$$h_{b,i}(x) = exp\left(\frac{-\|r_b - r_i\|^2}{2\sigma^2(t)}\right) \quad , \tag{8}$$

where r_b is the location of b on the map grid and $\sigma(t)$ is the neighborhood radius at time t. Since the SOM algorithm uses the Euclidean distance to measure distances between vectors, variables with different range of values must be equally important. To avoid range-biased clustering during the training phase, we perform a standard score transformation to normalize the training vectors.

A. Outlier Detection

The presence of outliers in the training set may negatively affect the SOM-based clustering by decreasing the sparsity of the projected feature map. Therefore, the first SOM network in our framework (Fig. 1) aims to remove outlier values from extracted motion vectors (Eq. 4) with an unsupervised scheme. This network is trained with values for hand velocity V_i (Eq. 1) only, while the rest of the attributes are not taken into account. The goal is to approximate the distribution of the observations with a trained SOM. Outliers will tend to be mapped into segregated units in the feature map. After the training of the network has been completed, the training set is processed again for detecting outliers (see [13] for a complete description of the algorithm). If an outlier is detected, it is removed from the training set. Using the trained SOM network as reference, also outliers in the test set are detected and removed.

B. Hierarchical SOM Learning

We propose a hierarchical SOM-based approach to cluster gesture sequences. We first train the network M_1 with motion vectors (Eq. 4) from the denoised training set. After this training phase, chains of labeled best matching units (Eq. 6) for ordered training sequences produce time varying trajectories on the network map. We empirically define a BMU trajectory for a training vector x_i as

$$\tau_i = (b(x_{i-2}), b(x_{i-1}), b(x_i), \lambda(x_i)) \quad , \tag{9}$$

where $\lambda(x_i)$ is the label of x_i. We denote the set of all trajectories for the training set X as $T(X)$. The second network M_2 is trained with a supervised variant of the SOM algorithm, where the inputs for the network are the labeled training trajectories from $T(X)$. This final step produces a mapping with labeled gesture segments from consecutive standalone samples. We compute the set of labeled trajectory prototypes as

$$P = \{\langle p_k, \lambda(p_k)\rangle : k \in [1..w]\} \quad , \tag{10}$$

where w is the number of training trajectories.

C. Gesture Classification

At recognition time, new extracted samples are processed separately. For the last three denoised observations, a new test trajectory τ_{i+1} is obtained from M_1 and then fed to M_2. We then compute P_{j+1} from M_2 and return the label $\lambda(p_{j+1})$ associated to the unit $b(p_{j+1})$. We consider the last 3 test sequence labels and calculate the statistical mode as output label for the classified gesture as:

$$Mo(\lambda(p_{j+1}), \lambda(p_{j+2}), \lambda(p_{j+3})) \quad . \tag{11}$$

A new output label of a classified gesture will therefore be returned every 9 observations, which corresponds to approximately less than 1 second of captured motion. As shown by our experiments, this approach significantly increases classification accuracy.

V. EXPERIMENTAL RESULTS

Depth images were acquired with an ASUS Xtion sensor installed on a fixed platform 1,40 meters above the ground. The depth map resolution was 640x480 pixels and the depth operation range was from 0.8 to 3 meters. The video sequences were sampled at a constant frame rate of 30 Hz. The segmentation and tracking of the user, and the estimation of body joints were addressed with the publicly available OpenNI/NITE framework[3]. To reduce sensor noise, we computed the median value of the last 3 measurements resulting in a total of 10 frames per second.

For our experiments, we collected two data sets of labeled gestures. The first data set consisted of a 10 command-like gestures for HRI scenarios (Fig. 3) and the second was composed of 10 common Italian communicative gestures that introduce cultural peculiarities. Every gesture class was performed 30 times by 4 different actors for a total of 1200 gestures for each data set. Single-hand gestures were performed hand-independently and the distance from the sensor varied from 1 to 3 meters. For a better understanding of the dynamics of the considered gestures, a number of experiments are reported on video[4].

The system was trained and tested on the two data sets separately. For the training sessions, we used 900 training samples and 300 testing samples. During the recognition, each gesture class was performed 30 times in a random order

[3] http://www.openni.org/software

[4] http://www.informatik.uni-hamburg.de/WTM/videos/videos.shtml

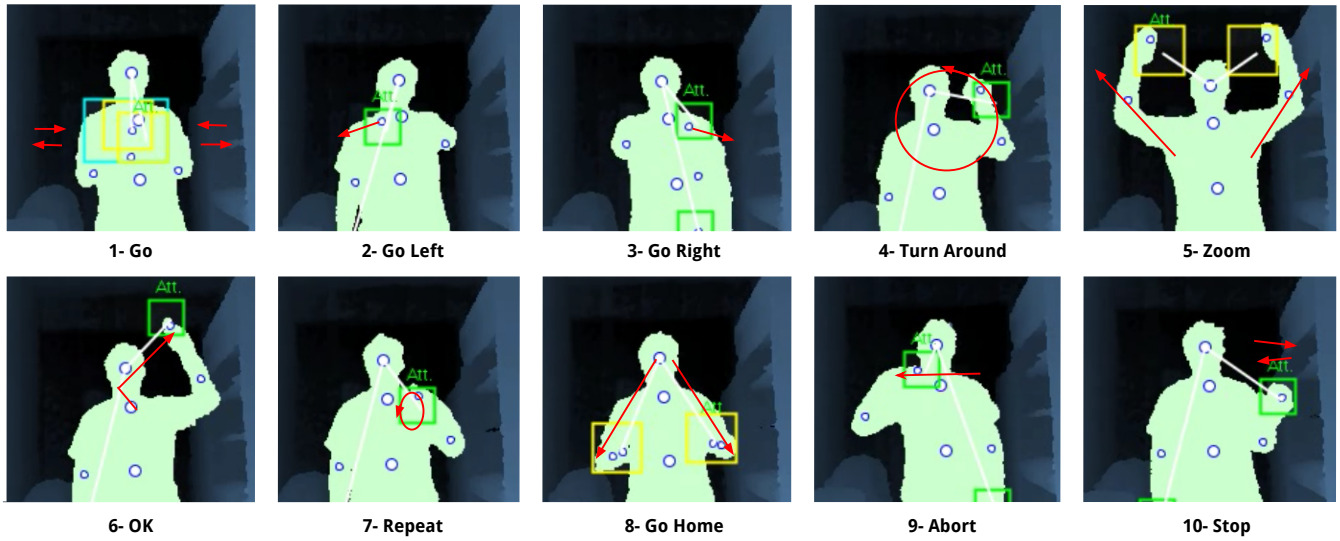

Fig. 3. Set of 10 gestures for hand-independent HRI scenario: hand interaction (colored squares) and hand motion trajectories (red arrows).

and at different sensor distances. Annotated ground truth data was used to decide the successful recognition of the gesture, i.e. when the detection took place between the first and last frame of its performance. The SOM structure and the training parameters are as described in [13]. To evaluate the impact of tracking errors and outliers, we performed experiments with denoised and raw training data.

Our results show that the system successfully recognized gesture classes with high recognition rates. For our data sets with HRI and Italian gestures we obtained averaged recognition rates by 80% and 82% respectively. As shown in Fig. 4 and 5, the removal of outliers from the training data increased recognition by 14.7% and 13.85% respectively. Misclassification mostly occurred among gestures that shared similar motion properties or that were performed at significant different speeds than during the training phase.

From a SOM perspective, gestures with very similar representations will be mapped into close regions of the subspace and will thus tend to be more easily misclassified. For the evaluation we did not consider tracking conditions in which at least one of the body joints couldn't be estimated by the tracking framework. In this case, the 3D position of the joint would be missing, leading to the incorrect estimation of the motion descriptors and therefore compromising the training phase. To address this issue, the information of the missing body joint could, for instance, be estimated as the interpolation of the joint trajectory during the preprocessing of the gesture sequence. On the other hand, our recognition algorithm showed to be robust to noisily estimated joints caused by tracking errors.

VI. DISCUSSION

We presented a neural network approach using motion sequences for learning and recognizing gesture. The choice of our gesture data set was motivated on the one hand for the application in HRI scenarios, where verbal input like "Turn around" is replaced by a gestural command. On the other hand we also addressed gestures in the context of communication (iconic gestures). For the gestures serving as co-speech, we chose different Italian gestures, which necessarily make also use of head information. With that, we could show the flexibility of our gesture recognition system concerning different information input and gesture types. In addition, we introduced physical and symmetric hand gestures to differentiate usage of one or more hands to perform a gesture. We used motion information and an attention-driven scheme for tracking spatiotemporal properties of gesture sequences. Two preprocessing steps were implemented to reduce sensor noise in the data and extract gesture sequences from video streams. Our system does not need to compute skin-color regions, which makes our approach more robust to different skin types and prevents additional processing steps. One motivation against the use of biologically inspired approaches is in fact the time-consuming training procedures. Our system provides real time recognition while concurrently taking into account temporal information without demanding high training times. This aspect is significant, as in a HRI scenario time delays hinder fluent communication. For example, a robot should react on a command the way humans do, i.e. turning its head to an object being pointed by a person.

Generally, the set of gestures used for recognition experiments is rather constrained. For real scenarios in HRI, detecting only ciphers as in [9] is insufficient, since their expressiveness is quite limited. Rather random or unnatural gestures [5] or the pure differentiation between two gestures cannot be adequately employed. Our effort in recognizing gestures aims to control a humanoid robot with visual commands or enable it to understand gestures as part of the language. Although interesting scenarios for gestures have been shown in, e.g., [6], we claim that research in the field of gesture recognition should provide more sensible operations to enable the integration of humanoid robots and other

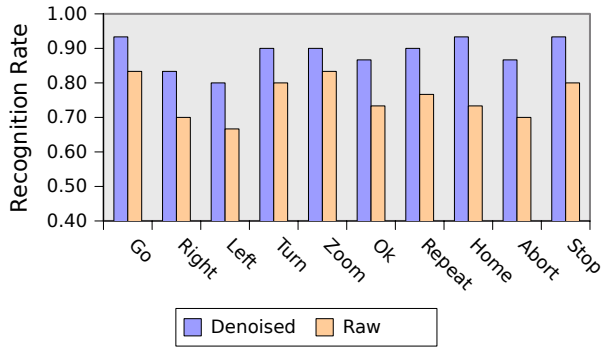

Fig. 4. Evaluation on the recognition of 10 HRI gesture classes with 300 testing samples for raw and denoised training data.

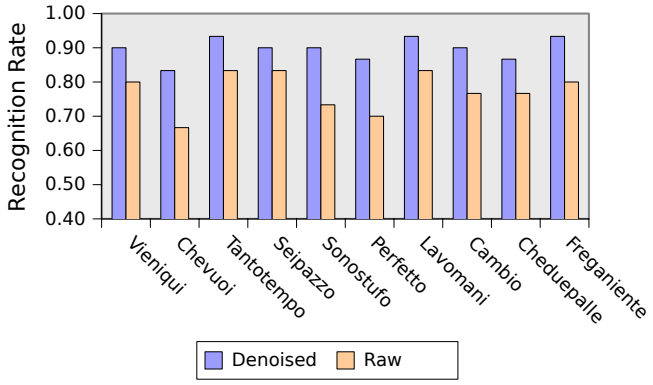

Fig. 5. Evaluation on the recognition of 10 Italian gesture classes with 300 testing samples for raw and denoised training data.

supporting technologies in our everyday lives. Therefore, we aim to test stable systems on novel robotic platforms.

At the current stage, our implementation requires the performer to stand frontal to the sensor. However this is not necessarily a limiting factor, since in gesture-based communication persons generally face each other.

VII. CONCLUSION AND FUTURE WORK

Our work contributes to the research area of gesture recognition in several aspects. We showed how the combination of depth information for motion extraction and the use of neural network architectures can be a prominent tool for recognition of gesture sequences with real time characteristics. In this context, the integration of biological mechanisms to computer vision approaches has shown to be a powerful approach to increase robustness and performance of a number of HRI-oriented applications. Our system can recognize a set of learned command-like gestures and Italian co-speech gestures, which were of major interest in our work. At the current stage, the system provides a natural and intuitive interface without the need of additional equipment, e.g. data gloves. The recognition of gestures is hand-independent and gestures with the use of both hands were also considered. Our framework can easily be extended to work on multi-user scenarios as in [13].

For future experiments we will extend our gesture vocabulary both in terms of command-like gestures and co-speech. Our goal is to investigate the bidirectional influence of gestures for complementing language and vice versa, i.e. how auxiliary phrases such as "Grasp the bottle" are helpful in learning and understanding gestures. As a further extension of our recognition framework, we could also consider situations in which more persons are present in the scene.

From the neural processing point of view, brain activity in the frontal and the temporal lobe has been found for both gestures and languages [14]. These findings imply that both modalities are processed in the same areas, thus they may have an influence on each other. Future neural network approaches could consider these crucial aspects in terms of development of multi-modal neural architectures to increase the robustness and reliability in human-robot communication.

REFERENCES

[1] A. S. Dick, S. Goldin-Meadow, A. Solodkin, and S. L. Small. Gesture in the developing brain, Developmental Science, 15(2):165-180, 2012.
[2] M. F. Villarreal, E. A. Fridman, and R. C. Leiguarda. The effects of the visual context in the recognition of symbolic gestures. PLoS One, 7(2), 2012.
[3] G. A. Alvarez and B. J. Scholl. How does attention select and track spatially- extended objects? new effects of attentional concentration and amplification. Journal of Experimental Psychology, 132:461-476, 2005.
[4] B. P. Olveczky, S. A. Baccus, and M. Meister. Segregation of object and background motion in the retina, Nature, 419:401-408, 2003.
[5] G. Caridakis, K. Karpouzis, A. Drosopoulos, and S. Kollias. Somm: Self organizing markov map for gesture recognition. Pattern Recognition Letters, 31(1):52-59, 2010.
[6] M. Ajallooeian, A. Borji, B. N. Araabi, M. N. Ahmadabadi, and H. Moradi. Fast hand gesture recognition based on saliency maps: An application to interactive robotic marionette playing. In Proc. IEEE International Symposium on Robot and Human Interactive Communication, pp. 841-847, 2009.
[7] P.M. Yanik, J. Manganelli, J. Merino, A.L. Threatt, J.O. Brooks, K.E. Green, and I.D. Walker, I.D. A gesture learning interface for simulated robot path shaping with a human teacher. IEEE Transactions on Human-Machine Systems, 44(1):41-54, 2014.
[8] J. Suarez and R. Murphy. *Hand gesture recognition with depth images: A review*. In Proc. IEEE International Symposium on Robot and Human Interactive Communication, pp. 411-417, France, 2012.
[9] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, and V. Athitsos. Comparing gesture recognition accuracy using color and depth information. In Proc. International Conference on Pervasive Technologies Related to Assistive Environments, pp. 201-207, USA, 2011.
[10] W. Hu, D. Xie, and T. Tan. A hierarchical self-organizing approach for learning the patterns of motion trajectories. In Proc. IEEE Transactions on Neural Networks, 15(1):135-144, 2004.
[11] A. K. Nag, A. Mitra, and S. Mitra. Multiple outlier detection in multivariate data using self-organizing maps title. Computational Statistics, 2(2):245-264, 2005.
[12] T. Kohonen. Self-organizing map. Springer-Verlag, 2nd ed., 1995.
[13] G. I. Parisi and S. Wermter. Hierarchical SOM-based detection of novel behavior for 3D human tracking. In Proc. IEEE International Joint Conference on Neural Networks, pp. 1380-1387, USA, 2013.
[14] J Xu, P. J. Gannon, K. Emmorey, J. F. Smith, and A. R. Braun. Symbolic gestures and spoken language are processed by a common neural system. In Proc. of the National Academy of Sciences (PNAS), 106(4):20664-20669, 2009.