Real-Time Single Camera Hand Gesture Recognition System for Remote Deaf-Blind Communication

Giuseppe Airò Farulla[1](✉), Ludovico Orlando Russo[1], Chiara Pintor[1], Daniele Pianu[2], Giorgio Micotti[3], Alice Rita Salgarella[4], Domenico Camboni[4], Marco Controzzi[4], Christian Cipriani[4], Calogero Maria Oddo[4], Stefano Rosa[1], and Marco Indaco[1]

[1] Politecnico di Torino, Turin, Italy
{giuseppe.airofarulla,ludovico.russo,stefano.rosa, marco.indaco}@polito.it,chiara.pintor@studenti.polito.it
[2] Institute of Electronics, Computer and Telecommunication Engineering, National Research Council of Italy, Padova, Italy
daniele.pianu@ieiit.cnr.it
[3] Politecnico di Milano, Milano, Italy
giorgio.micotti@mail.polimi.it
[4] The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy
{a.salgarella,d.camboni,m.controzzi,ch.cipriani,oddoc}@sssup.it

Abstract. This paper presents a fast approach for marker-less Full-DOF hand tracking, leveraging only depth information from a single depth camera. This system can be useful in many applications, ranging from tele-presence to remote control of robotic actuators or interaction with 3D virtual environment. We applied the proposed technology to enable remote transmission of signs from Tactile Sing Languages (i.e., Sign Languages with Tactile feedbacks), allowing non-invasive remote communication not only among deaf-blind users, but also with deaf, blind and hearing with proficiency in Sign Languages. We show that our approach paves the way to a fluid and natural remote communication for deaf-blind people, up to now impossible. This system is a first prototype for the PARLOMA project, which aims at designing a remote communication system for deaf-blind people.

Keywords: Real-time Markerless Hand Tracking · Hand Gesture Recognition · Tactile Sign-Language Communication · Haptic Interface

1 Introduction

The problem of tracking human hands joints, recognizing a wide set of signs, from single marker-less visual observations is of both theoretical interest and practical importance. In the last years promising results in terms of performances and robustness have been achieved due also to rapid advances in modern sensing technologies.

L.T. De Paolis and A. Mongelli (Eds.): AVR 2014, LNCS 8853, pp. 35–52, 2014.
DOI: 10.1007/978-3-319-13969-2_3

Many approaches have been presented in the hand gesture recognition area [16,20,27]; they differ in the used algorithm, in the type of camera, in the theoretical justifications, etc. Due to recent lowering in prices, new data sources have become available for mass consumers, such as time-of-flight [7] and structured light cameras [5], which ease the task of hand gesture recognition. Indeed, a robust solution has yet to be found, as existing approaches very often require an intensive tuning phase, the usage of coloured or sensitized gloves, or a working framework which embed more than one imaging sensor. Currently, traditional vision-based hand gesture recognition methods do not achieve satisfactory performances for real-life applications [15]. On the other hand, the development of RGB-D cameras, able to generate depth images with few noise also in very low illumination conditions, has recently accelerated the process of investigating for innovative solutions. In addition, the advent of modern programming frameworks for GPUs enable real-time processing even for complex approaches (i.e., that do not rely on too simplistic assumptions), that otherwise would be much slower if executed in CPUs.

Hand gestures are a natural part of human interaction, both with machines and other humans. As they represent a simple and intuitive way of conveying information and commands (e.g., zoom in or out, drag...), hand gesture recognition is of great importance for Human Machine Interaction (HMI) as well. Human interaction is widely based on hand gestures, above all when subjects with severe disabilities are involved and speech is absent, as in Sign Language (SL) based interaction. In both these fields (HMI and SL based interaction) it is necessary to provide support for real-time unaided gestures recognition, as markers or gloves are cumbersome and represent a hindrance to natural interaction. It is preferable to develop a system which relies on a single camera and does not require any calibration or tuning phase. Extensive initialization would represent a barrier for users which are not comfortable with technology, in particular when severe disabilities such as deaf-blindness are targeted.

Indeed, deaf-blind people can use neither vocal mean nor standard SLs, in the latter case because they are not able to perceive the meaning expressed by the signer. For this reason their communication is based on a different mechanism: the receiver's hands are placed on the ones of the signer in order to follow the signs made. Since the communication is still based on SL, but with tactile feedback, this variant is called tactile SL (tSL). Therefore, while it is possible for two normal speakers or two deaf signers to communicate in presence or remotely (either through phone calls or video-calling systems), as of now, there is no way for two deaf-blind persons to communicate with each other if they are not in the same place, given the basic need to touch each other's hands. Moreover, one-to-many communication is not possible, and the same signs must be repeated in front of each *listener* if the same message should reach many different persons [19].

In this context, the PARLOMA project[1] aims at designing a system able to capture messages produced in SL and reproduce them remotely in tSL, in order

[1] http://www.parloma.com

to overcome the spatial limitation posed by tSL communication. Indeed, the project poses the bases for the experimentation of a "telephone for deaf-blind people".

In this paper, we present a sophisticated approach to make the remote communication between deaf-blind people feasible. The proposed solution is based on a reliable marker-less hand gestures recognition method, which is targeted to recognize static signs from Italian SL (LIS) and is able to work up to 30 fps, that is the maximum operating frequency of the Kinect sensor[2].

To show its effectiveness, in addition to a quantitative and qualitative analysis, we also present an experimental apparatus in which signs from a subset of LIS hand spelling alphabet are recognized and sent remotely over the Internet, so that a compatible robotic actuator can reproduce them and any *listener* with proficiency in LIS can understand the meaning of what is signed. This work is a first step toward complete remote deaf-blind communication, in which more complex and also dynamic signs will be recognized.

This paper is organized as follows: Section 2 lists already existing related works, Section 3 discusses theoretical background and practical implementation of our solution, Section 4 presents results from our experiments and summarizes the pipeline of the remote communication system we developed and finally Section 5 presents some conclusion.

2 Related Works

This paper relies on hand tracking, anthropomorphic robots and data transmission over the web. In this Section, state-of-the-art approaches on these topics are briefly discussed.

2.1 Hand Tracking

Object tracking techniques can be divided into two main classes: *invasive* approaches, based on tools physically linked to the object (e.g., sensitized gloves [18] or markers [28]), and *non-invasive* approaches. The former are usually fast and computationally light, but also very expensive and cumbersome. The latter require more computational resources, but are based on low-cost technologies and do not require a physical link to the object to be tracked.

Non-invasive approaches proposed in literature (as in [22]) can be classified according to the kind of information in input they need (2D or 3D) and output they provide. Obviously, as real world life is embedded in a 3D universe, best performances are obtained when 3D features characterization is performed. Moreover, relying on 3D input information makes a visual system more robust and accurate [13].

Thanks to technology evolution though, it is nowadays possible to obtain reliable features extraction from confused backgrounds by trying to isolate and

[2] http://msdn.microsoft.com/en-us/library/jj131033.aspx

segment the object to be recognized (e.g., a human hand). In fact, 3D information may be obtained by depth maps that are automatically calculated by acquisition system using cheap RGB-D cameras.

Non-invasive approaches are classified into *partial tracking* and *full tracking*: tracking is defined as partial when it requires only information on a subset of input DoF [22] (e.g., only thumb and index in case of partial hand tracking), while it is said to be full when all the DoF are taken into account for computation. Of course, full tracking approaches are the best in terms of accuracy and robustness, but they also require a lot of computational resources [10]. Full tracking solutions can be divided into *model-based* and *appearance-based* approaches.

Model-based approaches [12] are based on a 3D model representing the object to be tracked. These approaches seek for the 3D model parameters that minimize the discrepancy between the appearance and 3D structure of hypothesized instances of the model and actual observations. The problem is usually solved using probabilistic optimization approaches [25] or evolutionary algorithms [20], leading to accurate results, but requiring a lot of computational resources.

Appearance-based techniques are based on special points (features) extracted from input data. An algorithm tries to directly map the extracted features to the hand configuration (i.e., the pose). Appearance-based algorithms are often implemented using machine learning [24] or database-retrieval [4] techniques. Learning how to map features and model configurations is the most computational intensive phase. Since this task is performed off-line, appearance-based techniques easily achieve real-time performances. The accuracy of these algorithms is strongly related to the quality of the training set or database, particularly to its variety and capacity to cover the set of poses.

2.2 Anthropomorphic Haptic Interfaces

Haptic devices elicit human perception through the sense of touch. Haptics therefore extends the communication channels for human-machine interaction, in addition to the typical senses, such as vision and hearing. Haptics includes wearable devices, such as gloves, and robotic devices, such as arms and hands. With respect to robotic hands, despite the significant progress in the last decades in electronic integrated circuits and in applied computer science, current systems still lack in dexterity, robustness and efficiency, as well as in matching cost constraints [9]. Examples of dexterous robotic hand for humanoid robotics are the Awiwi Hand [14] and the Shadow hand [26].

In prosthetics, electro-actuated hands have been commercially available since the early 70s: the major manufacturer is Ottobock (Austria), followed by other few companies. Recently two commercial prosthetic hands with greater Degree of Freedom (DoF) have been introduced to the market: the i-limb (developed by Touch Bionics in 2007) and BeBionic (developed by RSL Steeper in 2010) prostheses. Both hands are capable of different grasping patterns thanks to five individually-powered digits, but their functionality is limited by the passive movement of the thumb abduction/adduction joint. Most of current commercial prosthetic hands are very simple grippers with few active DoF and poor cosmetic appearance, however major research progresses are being achieved [21].

2.3 Transmission

Remote control of robotic actuators is nowadays a well studied task, investigated above all in surgery [3]. For what concerns the transmission of human movements to the robotic hand in this project, we evaluated different scenarios trying to maintain the infrastructure as simple as possible, as this will naturally lead to a simple, fast to develop and robust pattern of communication.

3 The Developed Solution

The proposed system is designed in order to accomplish to three different tasks: (1) sign acquisition and recognition (front-end), (2) sign conversion and transmission, (3) sign synthesis (back-end); that are performed by three different sub-blocks, i.e., the *input module*, the *transmission module* and the *robotic hand module*.

The input module is connected to a depth camera (the acquisition device) and is able to identify signs made by the human hand in front of the device. The transmission module is in charge of encoding the information generated by this first block, sending them through the web, and decoding them in a way that is suitable for the last block. Finally, the robotic hand module is composed by the robotic haptic interface and by a controller that uses the information from the first module to control robotic hand in a proper way.

3.1 The Input Module

The proposed implementation of the input module follows the work proposed in [16], where authors propose a full-DoF appearance-based hand tracking approach that uses a random forest (RF) classifier [23]. RF is a classification and regression technique that has become popular recently due to its efficiency and simplicity [16].

In the proposed system, a low-cost depth-camera (see Fig. 1), is used as only input to the hand segmentation phase, that is the task of isolating hand from the background (RGB information is discarded). Once foreground pixels have been recognized and separated from background, the hand pose can be reconstructed, resorting to two main blocks, that are the *hand labelling block* and the *joints position estimating block*. Hand labelling is an appearance-based method that aims at recognizing single sub-parts of the hand in order to isolate the joints, while the joints positions estimation block aims at approximating the joints 3D position starting from the noisy labelling and depth measurements. As done in [23], in our approach the RF classifier is employed to label pixels of the depth-image according to the region of the hand they should belong to, and than clusters each region in order to find the position of the centre of that

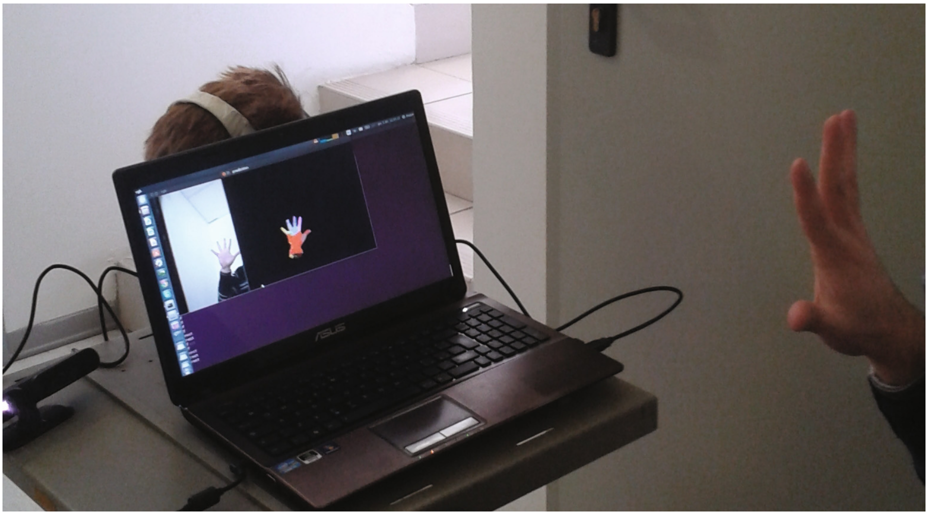


Fig. 1. The hand tracking input system

region. Regions are chosen in order to be centred over the joints of the hand, so that, at the end of the clustering process, the algorithm outputs the 3D position of each joint of the hand.

The developed code can recognize 22 different sub-parts of the hand, which are palm, wrist and 4 joints for each of the fingers. Each part is centred around a specific joint. Parts are tagged with different encoding and the tags are visually represented by different colours.

The hand is first segmented by thresholding depth values. The segmented hand is isolated from the background and tracked resorting to OpenNi tracker [1]. Finally, a point cloud for further processing is obtained, taking into considerations all the points within the sphere centred in the centre of the tracking and with a conservative radius τ.

To label the hand, an approach based on machine learning algorithms has been developed. Basically, at the very beginning, a RF classifier [6] is trained on thousands of different hands performing different signs, also turned or oriented differently. The classifier reads and examines such signs, and calculates the same feature for all of them; then, it keeps the more discriminative features. Such features can be later used to distinguish, with a certain confidence, the different hand sub-parts, and especially pixels that belong to different labels. Finally, the joints position is approximated applying the mean shift clustering [8] algorithm on the hand sub-parts. This approach provides promising results: first experiments with real-world depth map image show that it can properly label most parts of the hand in real time without requiring excessive computational resources.

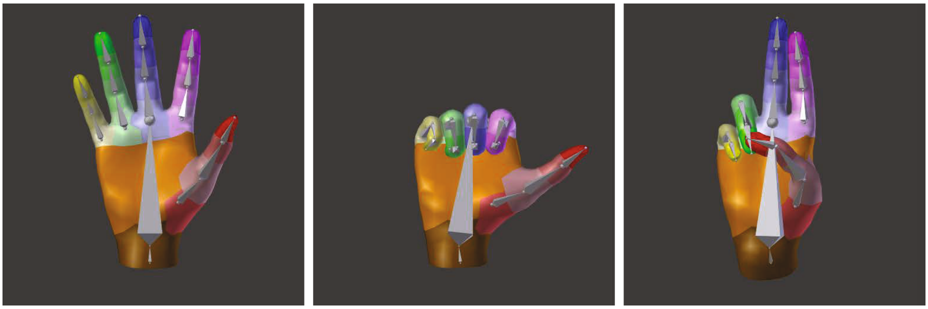


Fig. 2. 3D model in different poses used to generate the synthetic training set

In our approach we perform a per-pixel classification, where each pixel $\mathbf{x}$ of the hand is described using the following feature

$$\mathcal{F}(\mathbf{x}) = \left\{ F_{\mathbf{u},\mathbf{v}}(I, \mathbf{x}) = \sum_{\mathbf{j} \in (\mathbf{u},\mathbf{v})} I\left(\mathbf{x} + \frac{\mathbf{j}}{I\left(\mathbf{x}\right)}\right), ||\mathbf{u}|| < R, ||\mathbf{v}|| < R \right\}, \qquad (1)$$

where I is the depth-image so that $I(\cdot)$ represent the depth value of the image at a given point, while $\mathbf{u}, \mathbf{v}$ are two offset limited to a finite R length.

We use this feature because, in combination with RF, it has proved to very quickly succeed in discriminating hand parts, as shown in [16]. Hand poses can be estimated by labelled segmented hands resorting on mean shift [8]. Also, we resort on the mean shift local mode finding algorithm (as in [24]) to reduce the risk of outliers, that might have a large effect on the computation of the centroids for the pixel locations belonging to a hand part. In such a way, we obtain a more reliable and coherent estimation of the joints set $\mathcal{S}$.

Note that (1) is not invariant to rotations, while in the other hand it is invariant to distance and 3D translations (thanks to the normalization factor $I\left(\mathbf{x}\right)$). So, it is necessary to build a wide training set composed of the same sign framed from different point of view; for this reason, we have also investigated ways to effectively and automatically build comprehensive large train sets.

To train the algorithm, a training set with labelled samples is necessary. Since manually building a dataset is a tedious, time-consuming and error-prone process, a system able to create a synthetic training set was developed. Such system is based on the 3D model of a human hand shown in Fig. 2. Some examples of the outcomes of the synthetic training tool are shown in Fig. 3.

Main parameters describing the RF we trained were chosen as the ones providing best results after several tests and are summarized in Table 1. Each tree we use is trained with 2'000 random pixels from each training image. Offset vectors u and v from (1) are sampled uniformly between -30 and 30 pixels.

Finally, using a look-up table, the module converts the recognized hand pose in a list of 19 joints positions, that represents the angular positions that each

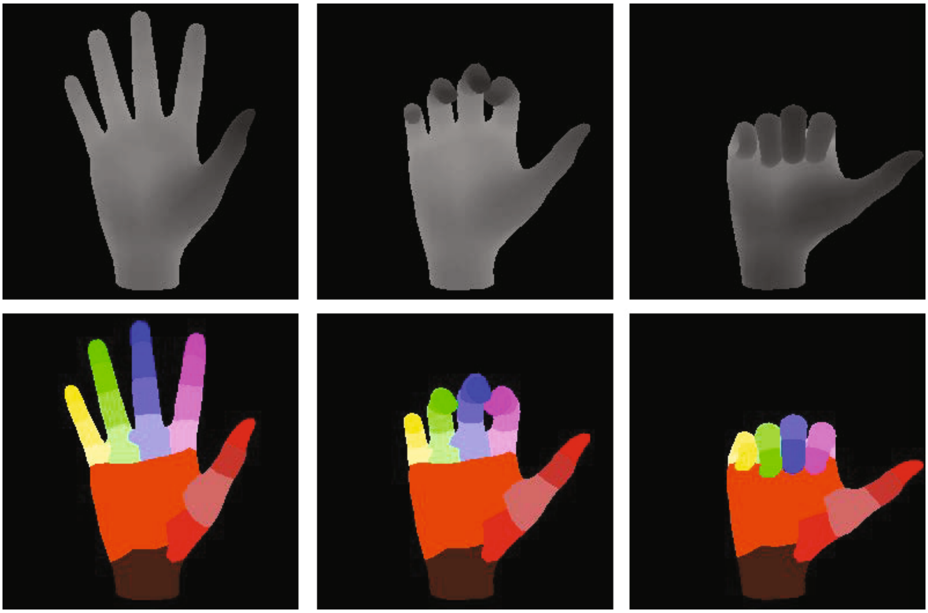


Fig. 3. Outcomes from the synthetic training tool: depth images and related labeling in 3 different poses

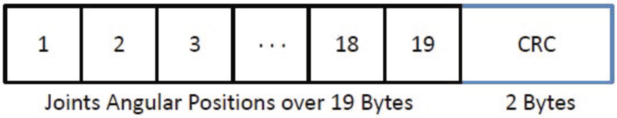


Fig. 4. Structure of the packet with the joints positions

joint of the hand have to reach in order to perform the sign. Global hand rotation (3 DOF) is at the moment discarded as the robotic hand used cannot rotate over the palm base.

3.2 The Transmission Module

Remote communication is implemented using a client - server socket architecture. The client is in charge of coding the sign coming from the input module, creating a proper packet for remotely sending needed data; this packet is also encrypted for secure communication. On the other side, the server is in charge of receiving and decrypting packets, and to decode commands in a suitable way for the hand control module.

The client gets the 19 joints positions, coming from the input module (as shown in Section 3.1). Each position ranges from 0 to 180 degrees, so it is encoded

Table 1. Optimal values we propose to train the RF classifier

Parameter	Value
U,V Offsets	30 pixel
Features extracted per image	2'000
Threshold	10
Sample pixels per image	2'000
Tree depth	18
Number of trees	3

with an unsigned Byte, where the value “180 degrees” is mapped with the maximum number representable (255) and linear scaling is used. Then a 16-bit CRC check is applied to detect potential errors in the transmitted packet. The generated CRC signature is joined to the packet (as in Fig. 4), and then encrypted using a robust cipher algorithm, AES [11]. The output of the cypher operation is an unintelligible string over 108 Bytes.

A TCP socket is opened to build a communication bridge between the client itself and the server, to send the encrypted list of positions.

The server is in charge of receiving and decrypting packets in order to retrieve the position of each of the robotic fingers. Finally, the received information are shared with the hand controller. Then, the robotic hand module is triggered when new data are available.

Failures in the network, such as unilateral unattended errors or crashes within the communication, are well managed by the code, in which ad-hoc exceptions handlers are implemented.

3.3 The Robotic Hand Module

Particular attention must be paid on what concerns the robotic hand, as this is one of the most important points in the solution communicative chain. The hand must be solid, very precise and user-friendly, combining well packaged mechanical components with a cosmetic glove able to mimic as much as possible the characteristics of human skin.

To develop a first prototype of the solution, a programmable anthropomorphic human-sized hand, EH1 Milano series, has been used. This is a versatile device for multiple research scenarios, produced by Prensilia s.r.l..

Such hand comes with 6 DoF, and the five compliant fingers are independently driven by electrical motors, by means of tendon transmission. The thumb abduction/adduction actuator is placed within the palm, whereas the fingers bending/extension motors are hosted into what could be thought as the forearm, a mechanical platform that represents a support for the hand itself and contains all the electronics needed to control the six motors and to communicate with a PC [2].

Communication with the hand is performed using the serial standard; in order to connect the hand to a standard laptop, an USB to serial converter has been used. The serial commands that have to be sent to the hand-side serial port, in order to make the hand move and reproduce signs, have been encoded in a demo program by means of a vocabulary that associates the list of six motor positions with the sign to be reproduced.

Each time the controller is triggered by the transmission module, it synthesizes the joints position in the commands needed by the actuators of the robotic device. Finally, it sends the commands to the robotic hand.

3.4 Implementation of the System

For the tests that are described in the next Section, a Raspberry Pi acts as server and hand controller, while a Laptop PC is used as client-side to compute sign acquisition.

Raspberry Pi[3] is a credit card size computer with low-cost hardware running Linux. The choice is motivated by the fact that, operations such as package decrypting and hand controlling do not require a powerful device.

A Notebook PC equipped with Intel Core i5-2450M @2.50GHz and 8 GB of RAM, running Ubuntu 13.10 OS is used to run all the algorithms regarding the input module, since image processing (coming from a PrimeSense[4] depth camera) requires much more computational power. On this machine, the whole acquisition system processes 30fps, thus achieving real-time performances. All the code is written in C/C++ and Python 2.7, using Open Source Software.

A video showing the proposed solution can be seen at the PARLOMA YouTube channel page[5].

4 Experimental Results

This Section summarizes results of some of the experiments we have performed to test the effectiveness of the proposed approach. These experiments aimed at: (1) evaluating the ability in recognizing signs in LIS (input module); (2) tuning the remote control of the robotic hand (robotic hand module); (3) assessing the effectiveness in transmitting the information over the whole pipeline; (4) getting feedbacks in order to fix potential errors and problems. In particular, the input module has been more intensively tested, as the ability of recognizing reliably and quickly SL signs is of crucial relevance for the whole system.

4.1 Input Module Validation

For what concerns the classification, we report both the average per-class accuracy and the hand gesture recognition accuracy. The first metric highlights how many times each pixel is labelled correctly by the classification layer. Results,

[3] http://www.raspberrypi.org

[4] http://www.primesense.com

[5] http://www.youtube.com/watch?v=6MGJb_GqauU

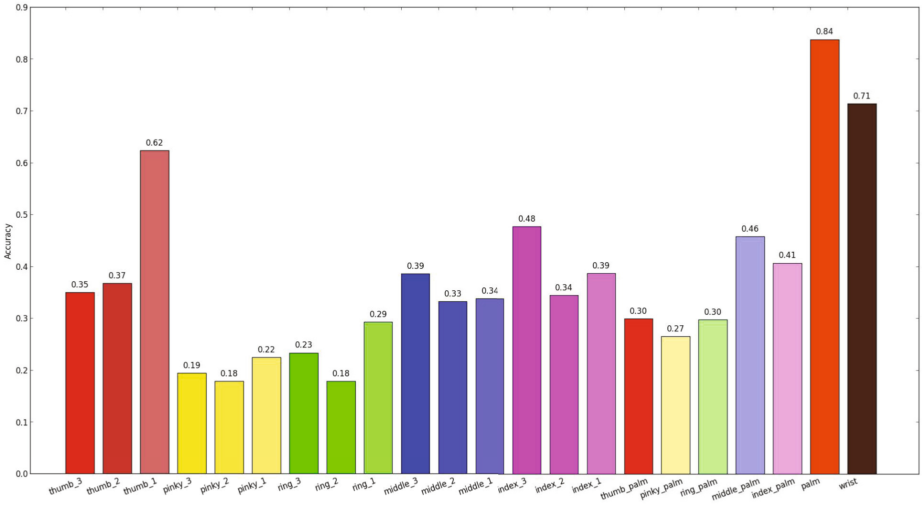
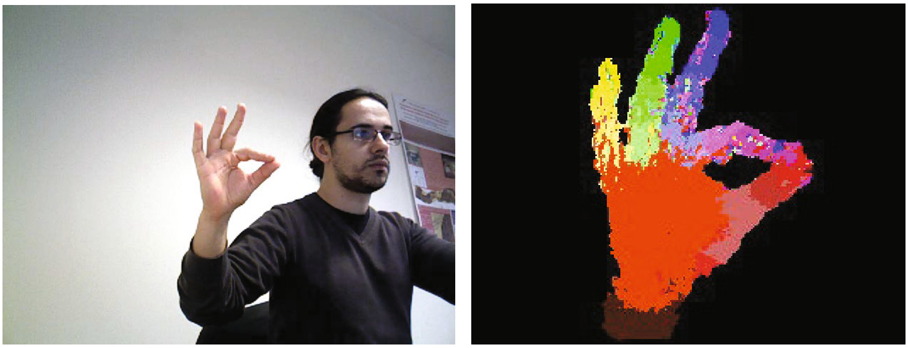


Fig. 5. Average per-pixel classification accuracy for each hand part. In the x axis, for each finger, palm subscripts identify the metacarpophalangeal joints (MCP), while indexes 1, 2 and 3 identify respectively proximal interphalangeal joints (PIP), distal interphalangeal joints (DIP) and fingertips. The y axis represent, in percentage, how many time the hand part is correctly labelled.

presented in Fig. 5, show that our system is usually able to discriminate among fingers and reach peaks of accuracy in discriminating palm and wrist. Little fingers, as ring and pinky, are obviously more difficult to track, especially for the self-occlusions that are experimented in many poses, and so the accuracy in their labelling is lower. Data presented in Fig. 5 represent the average accuracy of our system with respect to a ground truth set composed by 42 depthmaps, manually labelled. Hand labelling example is given in Fig. 6.

Average accuracy obtained by of our system in per-pixel classification is slightly worse than the one achieved in [16], but this is just due to the fact that we used a much smaller training set, composed of less than 15'000 images, while in [16] 200'000 images are used (and authors could not use more for memory constraints).

However, the experiments confirm that the average accuracy reached by our approach is sufficient to effectively track the hand and discriminate among hand gestures, even if similar. To this extent, Fig. 7 shows a graph summarizing the hand gesture recognition accuracy. Such data are computed using one against everything else cross-correlation validation, a process in which data from one subject is used for testing and all the others are used for training. This is the same metric used in [17] and allowed us a comparison between our approach and the results obtained by Kuznetsova et al. on real data. Error rate that authors report for multi-layered RF relying on decision trees with depth fixed to 20 do



(a) RGB image **(b)** Labeled Hand image

Fig. 6. Hand labelling example. Depthmap corresponding to the RGB image (a) is processed by the input block: in the labelled image (b), the background is removed and each identified different sub-part of the hand is coloured with the corresponding colour from the model.

not go below 49%. Such results are outperformed by our approach, since we achieve an average error rate of 46% in the same operating conditions.

As shown in Fig. 7, our approach is able to accurately recognize a sign most of the times. Even if accuracy is practically never over 90% in our experiment, we notice that a precision of nearly 35% is always guaranteed and it is sufficient to accurately recognize signs. For instance, Fig. 8 shows two example of a hand labelled by our approach. As it is shown in Fig. 8d, the T letter is easily discriminated even if average classification accuracy is slightly more than 40%.

4.2 Experimental Method

To test the whole pipeline of the system, comprehensive experiments were performed. In these experiments, subjects, not expert in TSL (blindfolded and with ears covered with headphones), are required to recognize the signs performed by the robotic hand, while a proficient in SL performs the signs in front of the input device in another room of the same building.

Each test subject is visually trained for five minutes on the subset of chosen signs with the proficient in LIS (the robotic hand is not used in this phase). After this first phase, the subject is blindfolded and his/her ears are covered; then, the subject is introduced in the room where there is the robotic hand. Note that the subject does not see the robot hand when training.

The message is sent to the robotic hand through a net (local network in the experiments). The results collected so far show that most of the times signs are correctly sent over the network and successively recognized in few seconds touching the robotic hand, even by non expert people.

In these experiments, we use only a subset of the LIS alphabet, consisting of characters S, U, V, W, F, because these are the signs recognized with more

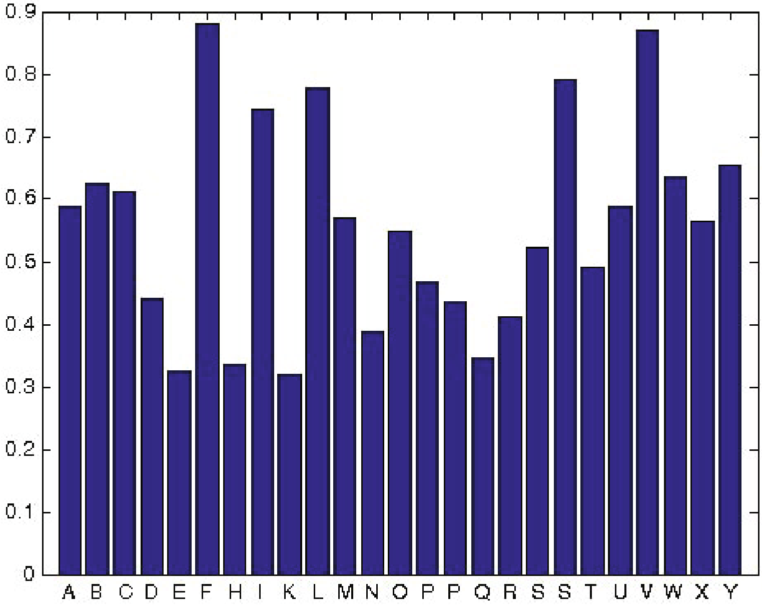


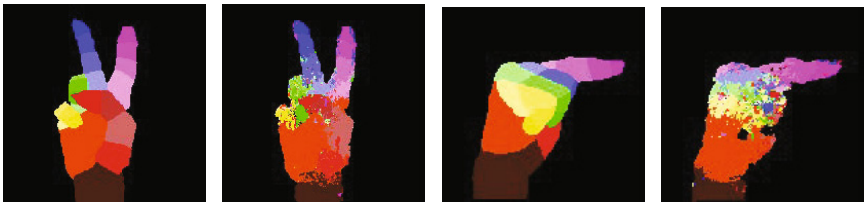
Fig. 7. Average percentage classification accuracy for different hand gestures, the sign is on the x-axes; please note that signs P and S are repeated two times because LIS admits two ways of performing those signs

accuracy and are also reproducible by the robotic hand, which is a first prototype with 6 DoF and cannot reproduce the whole possible static signs from LIS.

In particular, the experiments show that the system is able to work without errors for hours, and pose the basis for a more intensive session with deaf-blind subjects.

A test application collects data about the experiments. An example of recorded data is provided in Table 2. The test application randomly produce a List of 40 Signs (`LoS`) and the signer is asked to perform these signs, one by one, in front of the acquisition device. Moreover, it records the list of Signs Recognized by the Input module (`SRI`).

Recognized signs are transmitted to the controller of the hand by the transmission module. The robotic hand performs the sign and holds it for 5 seconds, and then it comes back to a rest position (open hand) and waits for the next sign. The subject has to recognize the sign by touching the robotic hand (using one or both hands), and then pronouncing the sign he/she understands. An experimenter records the Signs Performed by the Hand (`SPH`) and the relative Signs Recognized by the Subject (`SRS`). Each experiment lasts about 20 minutes



(a) Ground truth **(b)** Labelling **(c)** Ground truth **(d)** Labelling

Fig. 8. Classification and ground-truth for two poses of the LIS alphabet. (a,b) V letter; (c,d) T letter. The classification accuracy is, respectively, 81% and 46%.

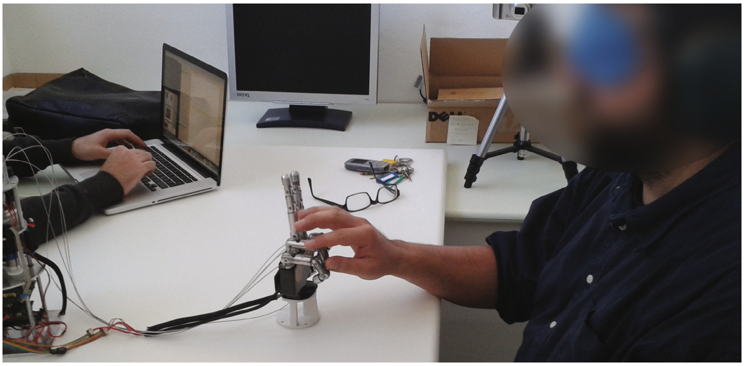


Fig. 9. Subject (blindfolded and with headphones) interacting with the robotic hand during the test session (V letter is performed in the picture)

per subject. Fig. 9 shows a subject during a test, while Fig. 10 illustrates the pipeline of the described experimental apparatus.

The experiments consist in repeating the procedure previously described for 10 subjects, for total amount of 400 signs produced. At the end of the experiments, four list of signs (`LoS`, `SRI`, `SPH` and `SRS`) are available. Hence, the recorded lists have been compared among each other and the results summarized in Table 3.

`LoS` VS `SRI` refers to the percentage of signs correctly recognized by the input module. This comparison to evaluate the effectiveness of recognition module. Here errors are due to finger occlusions, that sometimes deceive the recognition algorithm, but mainly to mistakes of signer. An example of erratic recognition is shown in column 3 of Table 2.

`SRI` VS `SPH` refers to the percentage of signs correctly sent to the hand. This comparison to evaluate the effectiveness of transmission module and the robotic hand module. No errors happened in this stage during the experiments.

Table 2. Example of recorded data during real experiments. Here are reported the List of Sign (LoS) to be produced, the list of Signs Recognized by the Input module (SRI), the list of Signs Performed by the Hand (SPH) and the list of Signs Recognized by the Subject (SRS).

#	1	2	3	4	5	6	7	8	9	10	...
LoS	F	W	V	S	V	F	S	S	V	S	...
SRI	F	W	W	S	V	F	S	S	V	S	...
SPH	F	W	W	S	V	F	S	S	V	S	...
SRS	W	W	W	S	V	F	S	S	V	S	...

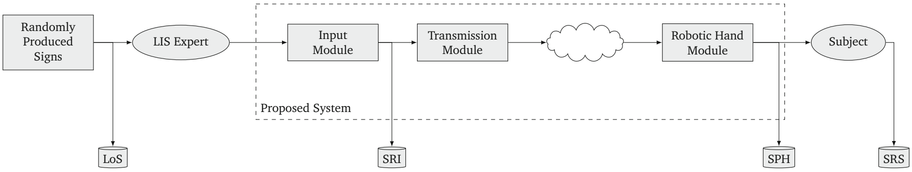


Fig. 10. Pipeline of the experimental apparatus

SPH VS SRS refers to the percentage of signs correctly recognized by the subjects. More of the errors in this phase can be ascribed to the subjects' lack of experience in Tactile LIS. In particular, the sign W is often confused with the sign F, since in both cases three fingers are opened.

Finally, LoS VS SRS measures the efficiency of the whole experimental apparatus. Here the percentage of success synthesizes the other percentages.

As shown in Table 3, overall success rate is 91%. Such result proves the general robustness of our system. In addition, we are confident that success rate will be higher in communication with real deaf-blind persons, that would surely make less mistakes in both performing and recognizing signs and letters from their SL alphabet.

Table 3. Reliability performances of the proposed system

LoS VS SRI	95%
SRI VS SPH	100%
SPH VS SRS	96%
LoS VS SRS	91%

5 Conclusions

This paper presents a system to allow non-invasive remote control of a robotic hand by using low-cost acquisition devices. The system is able to recognize human hand poses and can send them over the Internet, in order to control a robotic hand, that is able to reproduce poses in real time. This system does not require any tuning phase. Despite further optimizations which are still required, our approach shows great accuracy in discriminating even similar poses and achieves real-time performances.

Such system can be useful in many different fields, as for example human-machine interaction, and may easily and intuitively allows interaction with 3D virtual environments.

The paper presents also an early set of experiments demonstrating the efficiency of the system. The preliminary collected results demonstrate that more than 90% of times signs are correctly sent over the network and successively recognized by the test subjects touching the robotic hand. Note that errors in sign recognition by the subjects are not a validation penalty, since they are not Tactile LIS experts. The system will be evaluated through future experiments, when deaf-blind persons will be involved as well. Nevertheless, performed experiments were very useful to preliminarily assess the feeling of the subjects in touching the haptic interface while performing the sign recognition task.

References

1. Openni. http://www.openni.org/
2. Prensilia s.r.l., datasheet eh1 milano series (2010). http://www.prensilia.com/index.php?q=en/node/41
3. Abbou, C.C., Hoznek, A., Salomon, L., Olsson, L.E., Lobontiu, A., Saint, F., Cicco, A., Antiphon, P., Chopin, D.: Laparoscopic radical prostatectomy with a remote controlled robot. The Journal of Urology **165**(6), 1964–1966 (2001)
4. Athitsos, V., Sclaroff, S.: Estimating 3d hand pose from a cluttered image. In: Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, p. II-432. IEEE (2003)
5. Bray, M., Koller-Meier, E., Van Gool, L.: Smart particle filtering for 3d hand tracking. In: Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 675–680. IEEE (2004)
6. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
7. Breuer, P., Eckes, C., Müller, S.: Hand gesture recognition with a novel ir time-of-flight range camera–a pilot study. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 247–260. Springer, Heidelberg (2007)
8. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(5), 603–619 (2002)

9. Controzzi, M., Cipriani, C., Carrozza, M.C.: Design of artificial hands: A review. The Human Hand as an Inspiration for Robot Hand Development. STAR, vol. 95, pp. 219–246. Springer, Heidelberg (2014)
10. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. Computer Vision and Image Understanding **108**(1), 52–73 (2007)
11. Frankel, S., Glenn, R., Kelly, S.: The aes-cbc cipher algorithm and its use with ipsec. RFC3602 (2003)
12. Gavrila, D.M., Davis, L.S.: 3-d model-based tracking of humans in action: A multi-view approach. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1996, pp. 73–80. IEEE (1996)
13. Goncalves, L., Di Bernardo, E., Ursella, E., Perona, P.: Monocular tracking of the human arm in 3d. In: Proceedings of Fifth International Conference on Computer Vision, pp. 764–770. IEEE (1995)
14. Grebenstein, M.: The awiwi hand: An artificial hand for the DLR hand arm system. In: Grebenstein, M. (ed.) Approaching Human Performance. STAR, vol. 98, pp. 67–136. Springer, Heidelberg (2014)
15. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with microsoft kinect sensor: A review (2013)
16. Keskin, C., Kıraç, F., Kara, Y.E., Akarun, L.: Real time hand pose estimation using depth sensors. In: Consumer Depth Cameras for Computer Vision, pp. 119–137. Springer (2013)
17. Kuznetsova, A., Leal-Taixe, L., Rosenhahn, B.: Real-time sign language recognition using a consumer depth camera. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 83–90 (2013)
18. Lorussi, F., Scilingo, E.P., Tesconi, M., Tognetti, A., De Rossi, D.: Strain sensing fabric for hand posture and gesture monitoring. IEEE Transactions on Information Technology in Biomedicine **9**(3), 372–381 (2005)
19. Mesch, J.: Signed conversations of deafblind people
20. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient model-based 3d tracking of hand articulations using kinect. In: BMVC, pp. 1–11 (2011)
21. Raspopovic, S., Capogrosso, M., Petrini, F.M., Bonizzato, M., Rigosa, J., Di Pino, G., Carpaneto, J., Controzzi, M., Boretius, T., Fernandez, E., Granata, G., Oddo, C.M., Citi, L., Ciancio, A.L., Cipriani, C., Carrozza, M.C., Jensen, W., Guglielmelli, E., Stieglitz, T., Rossini, P.M., Micera, S.: Restoring natural sensory feedback in real-time bidirectional hand prostheses. Science Translational Medicine **6**(222), 222ra19 (2014)
22. Rehg, J.M., Kanade, T.: Digiteyes: Vision-based hand tracking for human-computer interaction. In: Proceedings of the 1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects, pp. 16–22. IEEE (1994)
23. Rodriguez-Galiano, V., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.: An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing **67**, 93–104 (2012)
24. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM **56**(1), 116–124 (2013)

25. Stenger, B., Thayananthan, A., Torr, P.H., Cipolla, R.: Model-based hand tracking using a hierarchical bayesian filter. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(9), 1372–1384 (2006)
26. Walkler, R.: Developments in dextrous hands for advanced robotic applications. In: Proc. the Sixth Biannual World Automation Congress, Seville, Spain. pp. 123–128 (2004)
27. Wang, R., Paris, S., Popović, J.: 6d hands: markerless hand-tracking for computer aided design. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 549–558. ACM (2011)
28. Wang, R.Y., Popović, J.: Real-time hand-tracking with a color glove. ACM Transactions on Graphics (TOG) **28**, 63 (2009)