2012 IEEE International Conference on
Multisensor Fusion and Integration for Intelligent Systems (MFI)
September 13-15, 2012. Hamburg, Germany

Hand gesture recognition using color and depth images enhanced with hand angular pose data *

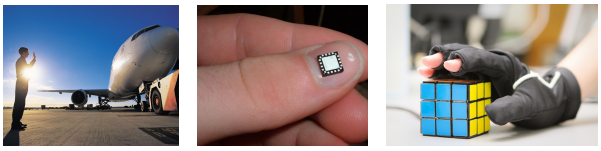
Pedro Trindade[1], Jorge Lobo[2] and João P. Barreto[2]

Abstract— **In this paper we propose a hand gesture recognition system that relies on color and depth images, and on a small pose sensor on the human palm. Monocular and stereo vision systems have been used for human pose and gesture recognition, but with limited scope due to limitations on texture, illumination, etc. New RGB-Depth sensors, that reply on projected light such as the Microsoft Kinect, have overcome many of those limitations. However, the point clouds for hand gestures are still in many cases noisy and partially occluded, and hand gesture recognition is not trivial. Hand gesture recognition is much more complex than full body motion, since we can have the hands in any orientation and can not assume a standing body on a ground plane. In this work we propose to add a tiny pose sensor to the human palm, with a minute accelerometer and magnetometer that combined provide 3D angular pose, to reduce the search space and have a robust and computationally light recognition method. Starting with the full depth image point cloud, segmentation can be performed by taking into account the relative depth and hand orientation, as well as skin color. Identification is then performed by matching 3D voxel occupancy against a gesture template database. Preliminary results are presented for the recognition of Portuguese Sign Language alphabet, showing the validity of the approach.**

I. INTRODUCTION

Monocular and stereo vision systems have been used for human pose and gesture recognition [1], [2], but with limited scope due to limitations on texture, illumination, etc. With the development of RGB-Depth (or simply RGB-D) sensors, such as the Kinect [7], it became possible to get the 3D point cloud of the observed scene. This means some of the common monocular and stereo vision limitations are partially resolved due to the nature of the depth sensor. Also Inertial Measurement Units (IMU) sensors have become less intrusive and available whichs allows them to be used in a more precise and robust hand gesture recognition system. An example of these sensors is seen in Fig. 1b where a small MEMS accelerometer is no bigger than a human fingernail. Gloves have also been used for inertial sensing (Fig. 1c) although they may clutter the natural human hand movement.

We propose a robust use of a RGB-D sensor point cloud aided with IMU sensors on the back of the hand for gesture recognition. We capture the point cloud using the RGB-D camera and filter it to get the relevant blob. We use the IMU information to apply a rotation and thus reduce the size of the search space. Our application focuses on the recognition of the Portuguese Sign Language alphabet.



(a) Airport ramp (b) Accelerometer (c) Acceleglove

Fig. 1: Communication by gestures, and minute accelerometers to capture hand gestures.

II. RELATED WORK

Vision systems have been previously used for human body parts tracking and recognition. Good algorithms have been developed for face tracking [3] but recent trends in technology have made recognition of body motion and gestures a popular research topic.

In recent years, sign language has been a popular topic in gesture recognition. Many works have emerged, like the American Sign Language recognition [4], the Portuguese Sign Language [5] and also the Indian Sign Language [6]. Vision-based systems have been extensively researched and have been recently complemented with RGB-D sensors like the Kinect [7]. Many research works have already made use of these popular sensors such as Virgile Hogman [8] that presented the building of a 3D map out of RGB-D sensors. These sensors have also been used for hand pose recovery and estimation [9], [10], [11]. Also recently Shotton et al. [12] and Girshick et al. [13] have been researching human pose recognition and activity with results used in Microsoft entertaining platforms.

When filtering information related to the detection of a human body part, an important feature is skin color. Vezhnevets et al. [14] presents a survey on the techniques for identifying the human skin range detection.

Although the smart sensor concept envisions a minute self contained sensing system, current technology uses accessories such as gloves and other worn based option. A survey on these worn system is presented by Dipetro et al. [15].

Our previous work already made use of inertial information. We used acceleration sensing to extract hand angular pose for gesture recognition [5] and grasp type identification [16]. Although acceleration sensing can provide angular pose information in 2 degrees of freedom [17], it cannot provide the third degree of freedom due to the inobservance of rotation around gravity's axis. To overcome this limitation

* The research leading to these results has been partially supported by the HANDLE project, which has received funding from the European Community's 7th Framework Programme under grant agreement ICT 231640.

[1]Pedro Trindade is with ISR - Institute of Systems and Robotics, University of Coimbra, Portugal pedrotrindade@isr.uc.pt

[2]Jorge Lobo and João P. Barreto are with ISR - Institute of Systems and Robotics and the Department of Electrical and Computer Engineering, University of Coimbra, Portugal {jlobo,jpbar}@isr.uc.pt

several techniques were used based on our previous works [18][19] where calibration between cameras and inertial sensing was performed using quaternions of rotation, base in Horns closed-form solution for absolute orientation [20].

III. Methodological Approach

The approach followed by this work is depicted in Fig. 2 and is detailed in the next sections. After getting the point cloud from RGB-D sensor we apply a filtering process to get the hand blob only. At the end of filtering by applying k-means we compute the aproximate center of the hand cluster which we expect to be close to the actual hand center. Using the information from the IMU sensing we normalize the hand angular pose for proper comparison of the observed gesture against the library of gestures. This comparison was tested using ICP and by binary voxel comparison.

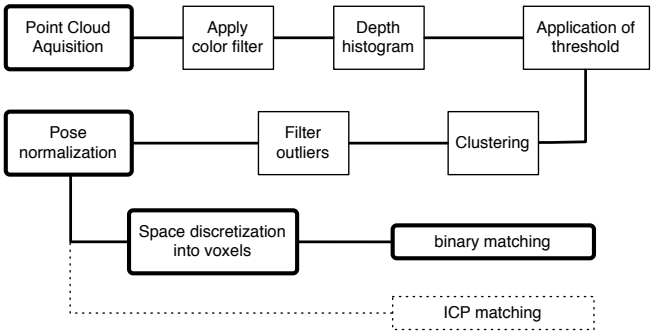


Fig. 2: Diagram with the workflow of the approach followed. Key stages are highlighted in the diagram.

The optimal solution is found by minimizing the error in those comparisons. This optimal solution therefore identifies the correct gesture performed.

A. Skin Color Filtering

Analysis of color information from the RGB image allows to narrow in the pixels representing the hand skin. A skin color detection techniques survey work [14] provided the information for filtering out the non relevant color information of the observed point cloud O, thus reducing the search space S and the observing point cloud that becomes O_r. Applying the RGB color restriction in Eq. (1) the search space S should become very largely reduced where typically the scene is not skin color colored except for the human body parts. Each point $\mathbf{p_i} \in O$ is classified as skin if:

$$\left\{ \begin{array}{c} R > 95 \\ G > 40 \\ B > 20 \\ \max\{R,G,B\} - \min\{R,G,B\} > 15 \\ |R - G| > 15 \\ R > G \\ R > B \end{array} \right\}, \quad (1)$$

where (R,G,B) are the components of the color space C for each $\mathbf{p_i}$ belonging to observation space O.

B. Depth filtering

Depth information can be used to distinguish between the different body parts not filtered by the previous color skin detection process. When the RGB-D sensor has front planar view of the person we assume that the hand performing the gesture is closer to the sensor than the other body parts. This can be seen in Fig. 3 where the gesture is seen closer to the RGB-D sensor. By applying a threshold filter to a histogram of depth information it is possible to extract the hand blob that performs the gesture.

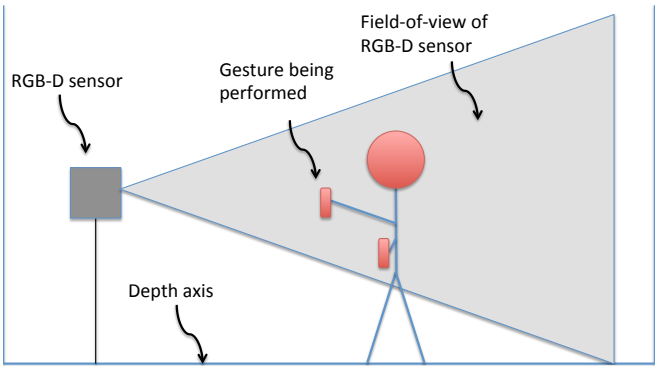


Fig. 3: Scene diagram for depth position understanding between gesture and RGB-D sensor.

Let $\mathbf{p_i} \in O_r$, s_j a planar section at depth $j \in \mathbb{R}$, Δ a neighbor size parameter for creating a RGB-D front planar box B_j with depth size of 2Δ centered at j, and λ the threshold value for the count of points inside B_j. The new observing space S_h will be given by Eq. (2).

$$S_h = \{\mathbf{p_i} \in B_k\}, \quad (2)$$

where k represents the depth of the first skin blob, defined by Eq. (3):

$$k = \min j : H_j > \lambda, \quad (3)$$

with H_j being the histogram count at box B_j as defined in Eq. (4):

$$H_j = \sum_{i,j} \mathbf{p_{i,j}}, \quad \mathbf{p_{i,j}} \in B_j \quad (4)$$

C. Clustering and outlier removal

In order to identify the center of the hand point cloud we cluster the whole point cloud S_h. Using k-means we are able to identify the approximate center of the hand point cloud since it is the area where most points are centered. Let C be the set of cluster centers $\mathbf{c} \in \mathbb{R}^n : |C| = k$. The aim of this algorithm is to minimize the objective function in Eq. (5):

$$\min \sum_{\mathbf{x} \in X} \|f(C, \mathbf{x}) - \mathbf{x}\|^2, \quad (5)$$

where X is a set of examples $\mathbf{x} \in \mathbb{R}^n$. $f(C, \mathbf{x})$ procures the nearest cluster center $\mathbf{c} \in C$ to $\mathbf{x}$ using L^2-norm, i.e. the Euclidean distance.

D. Angular pose normalization

The IMU provides valuable information for determining angular pose. It has been previously demonstrated that acceleration information can be used for finding angular pose information [5][17]. This information comes with the limitation of not knowing the rotation around the gravity axis. In a cartesian frame of reference let α be the rotation angle around x-axis and β the rotation around y-axis in the sensor coordinate system. Let $R_{\beta,\alpha}$ be the rotation matrix around those two axis. $R_{\beta,\alpha}$ is therefore defined in Eq. (6)

$$R_{\beta,\alpha} = \begin{bmatrix} \cos\beta & \sin\beta \cdot \sin\alpha & \sin\beta \cdot \cos\alpha \\ 0 & \cos\alpha & -\sin\alpha \\ -\sin\beta & \cos\beta \cdot \sin\alpha & \cos\beta \cdot \cos\alpha \end{bmatrix} \quad (6)$$

Let $\mathbf{g} = [0, 0, -a]^T$ be the gravity vector, with a being the gravity's acceleration. If we define $\mathbf{m} = [m_x, m_y, m_z]^T$ as the measured acceleration by the sensor we get the relation in Eq. (7).

$$\mathbf{m} = R^T_{\beta,\alpha} \cdot \mathbf{g} \Leftrightarrow \begin{bmatrix} m_x \\ m_y \\ m_z \end{bmatrix} = \begin{bmatrix} a \cdot \sin\beta \\ -a \cdot \cos(\beta) \cdot a \cdot \sin\alpha \\ -a \cdot \cos(\beta) \cdot a \cdot \cos\alpha \end{bmatrix} \quad (7)$$

From Eq. (7) roll and pitch angles (α and β respectively) can be easily extracted. From the magnetometer we can directly extract the yaw value for the rotation, thus creating the RPY (Roll-Pitch-Yaw) angles for normalizing the observing gesture rotation using gravity as a reference.

E. Matching algorithms

1) ICP matching: The ICP - Iterative Closest Point is an algorithm originally presented by Zhang [21]. It revises the necessary rigid transformation in order to minimize the distance between two different point clouds. As detailed in Eq. (8) it aims to minimize the distance r between a reference point cloud Q and an iteratively rotated R and translated T point cloud P.

$$r = \min \sum_i \|(R \cdot P + T) - Q\| \quad (8)$$

In our work the Q point cloud refers to a gesture in a library of gestures. P refers to one of the possible rigid transformation of an observing point cloud corresponding to a performed trial gesture. Since we are able to previously normalize the hand rotation, as detailed in section III-D, we can simplify the initial ICP procedure to the algorithm shown in Fig. 4.

2) Voxel quantization and binary matching: Voxel quantization allows to organize space information in a more useful manner for comparison techniques. By defining a voxel width w space quantization can be performed by a simple equation, as in Eq. 9:

$$\mathbf{idx} = f(\frac{\mathbf{p}_i}{w}), \quad (9)$$

where $\mathbf{idx}$ are the indexes of the voxel matrix V_m containing the voxel ocupacy information of the corresponding observed point p_i and f the round to the nearest integer number function.

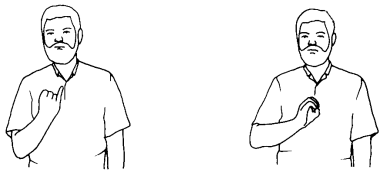
Direct binary comparison between an observed gesture and one from a library of gestures allows to rapidly evaluate the correct match against between all the gestures from the library. The number of the maching voxels will be the metric for the gesture recognition.

```
begin
    while (not termination-condition) do
        begin
            r ⟵ Compute closest points from between
            P and Q;
            evaluate r;
            alter T;
            next iteration;
        end
    end
end
```

Fig. 4: Simplified version of ICP algorithm assuming hand rotation normalization

F. Gesture recognition

In this paper we propose a gesture recognition system. More specifically the recognition of the Portuguese Sign Language alphabet. The letters of this alphabet are characterized by being represented as a static hand pose. Two examples are presented in Fig. 5.



(a) Letter 'i' (b) Letter 'o'

Fig. 5: Two gestures from the Portuguese Sign Language

IV. Experimental Results

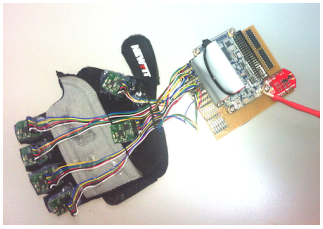
The experimental setup consisted of a RGB-D sensor [7] and a magnetic tracker sensor [22]. In previous work we used the Anthrotronix Acceleglove to extract acceleration information. Currently we are developing a sensing system consisted in accelerometers, magnetometers in order to have a sensing glove. This is shown in Fig. 6c and 6d. The information provided by this system can be converted into angular pose information (see section III-D). However for this work it was not available yet, and the experimental results were obtained from a magnetic tracker, that provides the angular pose data.

Depicted in Fig. 6, Kinect is the RGB-D sensor used and Polhemus sensor is the motion tracker sensor that provides

(a) Polhemus Liberty Tracking System

(b) Kinect and other RGB-D sensors

(c) Currently development of our IMU sensor



(d) Future integration of IMU sensors in a glove

Fig. 6: Experimental setup for acquiring data and prototype sensors currently under development.

angular pose data. In this work we do not use the IMU for INS (Inertial Navigation System). Since our use is in determining angular pose and not position the magnetic and gravity vertical references are static and not subject to drift issues. Thus our use of a Polhemus sensor, that provides angular pose data simulates completely an IMU sensor, since currently we cannot recur to our under development IMU sensor. [4]

A. Skin color and depth filtering process

In Fig. 7 it is possible to see the result of skin color filtering. From the whole point cloud it was possible to remove the majority of points without compromising the hand point cloud information.

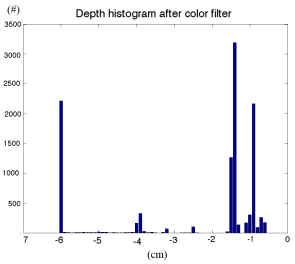
(a) Original point cloud

(b) Skin color filter applied

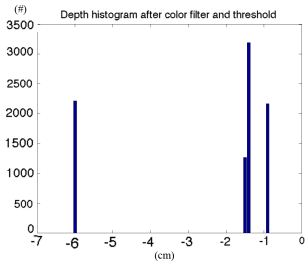
Fig. 7: Original and Color filtered images for a trial gesture of letter "Y".

After skin color filtering an histogram threshold was applied. The histograms presented in Fig. 8 represent the point count along the depth axis, before and after applying a threshold. This threshold retains only the bigger blobs of information. From that we choose the closest one to the RGB-D sensor because it represents the first human body part detected: the hand that is performing the gesture.

[4]To complement the viewing in this section a webpage has been created. It shows all the images and histograms in this section and also the 3D files for better 3D visualization. http://mrl.isr.uc.pt/people/pedrotrindade/mfi2012



(a) Original histogram

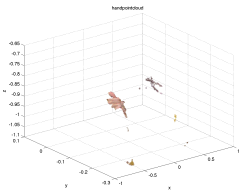


(b) Threshold filtered histogram

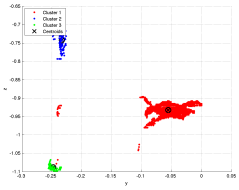
Fig. 8: Original and filtered histogram along depth axis of letter "Y" gesture trial.

B. Clustering process and outliers removal

By applying k-means we are able to estimate the center of the hand blob. Because there is still noise near the hand blob, k-means clustering provides the center of the cluster with greater density which is then considered as the hand blob. As seen in Fig. 9 the most relevant centroid is the one we are interested. The coordinates of the centroid are also the center of a surrounding box to filter outliers from the point cloud.



(a) Hand blob with noise point cloud



(b) k-means cluster areas identification

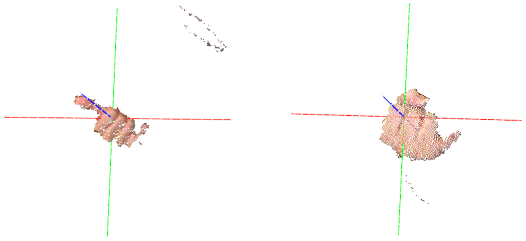
Fig. 9: Original and Cluster identification for trial point cloud of gesture corresponding to letter "Y".

C. Angular pose normalization

By utilizing IMU sensors such as accelerometers and magnetometers we are able to get the hand angular pose. This information is an important feature as we are therefore able to normalize all hand angular pose for comparison. This results in dimensionality reduction thus simplifying and accelerating the search process. The result of the normalization of two different gestures with respect to the gravity and magnetic field references can be seen in Fig. 10.

D. ICP analysis results

After the process of angular pose normalization ICP is used for matching. While ICP has been used with success in some applications it typically needs to be tuned for a good performance. This is particularly true in the case of incomplete point clouds. In our case we tried to take advantage of our inertial sensing. This feature simplifies the process of rotation between point clouds in ICP. Yet, by looking at Fig. 11 we see the non-matching of the two trials for the same gesture (both representing letter "Y"). Although it is the same gesture in space there is too much incomplete



(a) Gesture ”Y” angular pose normalization (b) Gesture ”I” angular pose normalization

Fig. 10: Normalization of angular pose of gestures corresponding to letters ”Y” and ”I”.

information in the point clouds of the two trials. This results in the ICP not being able to find the rigid transformation to properly align the two point clouds.

Fig. 11: Results of using ICP for comparison of two trial gestures to the same letter ”Y”.

E. Voxel quantization and binary analysis results for gesture recognition

An alternative method to the ICP for matching is binary voxel matching after quantizing the point cloud. Voxel quantization reduces the search space for matching the two point clouds. By using a step window of 0.3cm we were able to get a positive match, something we were not able to get when using ICP. The quantization using voxels simplified the search in R^3.This result can be seen Fig. 12.

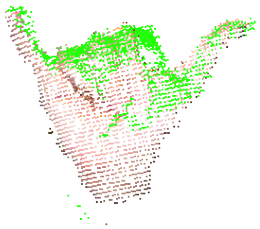


Fig. 12: Results of using binary comparison of two trial gestures to the same letter ”Y”.

Using voxel quantization and binary comparison we tested the ”Y” trial gesture against two Library gestures: the ”Y” and ”A” gesture. In Fig. 13 the two Library gestures are displayed.

(a) Gesture ”Y” from the Library of Gestures (b) Gesture ”A” from the Library of Gestures

Fig. 13: Two Library gestures: ”Y” and ”A”.

These two gestures were compared with a trial gesture ”Y”. The result of the search for matching resulted in the overlay displayed in Fig. 14.

(a) Overlay of Library ”Y” with trial ”Y” (b) Overlay of Library ”A” with trial ”Y”

Fig. 14: Overlay of two gestures from the Library of gestures ”Y” and ”A” against a trial gesture ”Y”.

In relation to the ”Y”-”Y” matching (Fig. 14a), 152 voxels of perfect match were found. As expected, between gestures ”Y”-”A” only 76 voxels of perfect matching were found. Although any point cloud will partially match another point cloud, the more voxels we get for a gesture, greater number of gestures will be possible to match.

Fig. 15 provides a more complete analysis of the binary comparison between more trial gestures and more gestures from the Library. It is clear from the voxel count how the best match has a major voxel count. In Fig. 16 the relation of the voxel count with the total number of matchable voxels for each comparison is presented. This information provides a quality indicator for the voxel matching result.

V. Conclusions and Future Work

We have presented here the preliminary results of enhancing RGB-Depth information with inertial sensing. The major problem of performing gesture recognition has been to deal with the matching in 6 dimensions' (position and angular pose) space. With our approach we are able to reduce dimensionality by using inertial sensing to get accurate angular pose. Using this information we were able to normalize the angular pose of every gesture performed. For the matching of gestures in 3D space we applied ICP and also quantize the space in order the perform binary comparison between voxels. This showed potential results to be further developed.

While ICP proves not to be effective, other variants of this

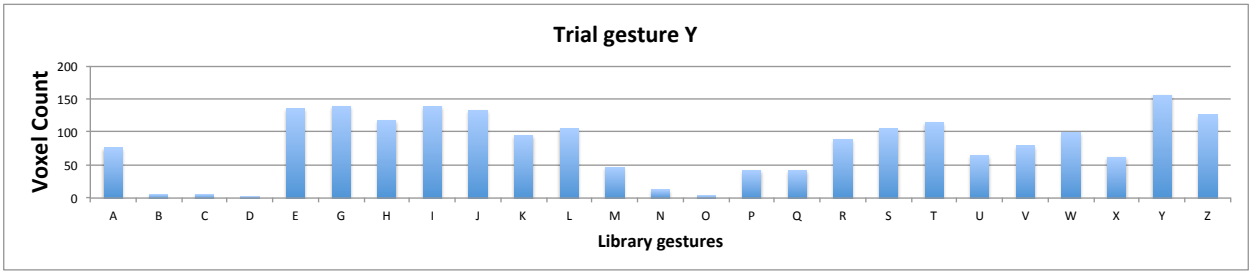


Fig. 15: Number of voxel match between trial gesture and the library of gestures.

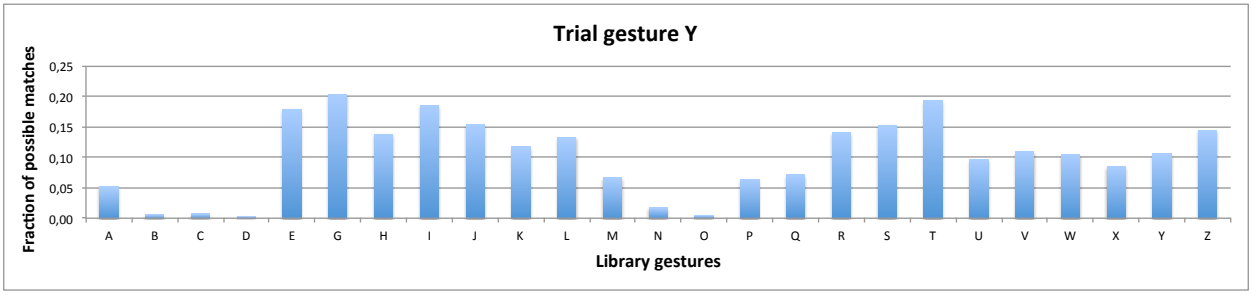


Fig. 16: Fraction of possible matched voxels between trial gesture and the library of gestures.

algorithm [23] may optimize the search space leading to better results.

Scale factor should be pursued in future work and it can be seen as an expansion of search space from R^3 to R^4.

Using more inertial sensor could lead to a segmented search space allowing a more precise matching of voxels for each segment. Also signature recognition for the normalized point cloud could provide an interesting solution for searching in a higher dimensionality search space.

REFERENCES

[1] Robert Y. Wang. Real-time hand-tracking as a user input device. 2008.

[2] Robert Y. Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM Trans. Graph.*, 28:63:1–63:8, July 2009.

[3] Gary R Bradski. Computer vision face tracking for use in a perceptual user interface. *Interface*, 2(2):1221, 1998.

[4] F. Ullah. American sign language recognition system for hearing impaired people using cartesian genetic programming. In *Proc. 5th Int Automation, Robotics and Applications (ICARA) Conf*, pages 96–99, 2011.

[5] P. Trindade and J. Lobo. Distributed accelerometers for gesture recognition and visualization. In *DoCEIS'11 - Doctoral Conference on Computing, Electrical and Industrial Systems*, pages 215–223, Lisbon, Portugal, February 2011.

[6] A. S. Ghotkar, R. Khatal, S. Khupase, S. Asati, and M. Hadap. Hand gesture recognition for indian sign language. In *Proc. Int Computer Communication and Informatics (ICCCI) Conf*, pages 1–4, 2012.

[7] Microsoft Kinect. http://www.xbox.com/kinect, 2012.

[8] Virgile Hgman. Building a 3d map from rgb-d sensors. Master's thesis, KTH Royal Institute of Technology, 2011.

[9] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Markerless and Efficient 26-DOF Hand Pose Recovery. In *Asian Conference on Computer Vision*, Queenstown, New Zealand, 2010.

[10] Nikolaos Kyriazis Iason Oikonomidis and Antonis Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *Proc. BMVC*, pages 101.1–101.11, 2011. http://dx.doi.org/10.5244/C.25.101.

[11] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Tracking the Articulated Motion of Two Strongly Interacting Hands. In *Computer Vision and Pattern Recognition*, Providence, Rhode Island, USA, 2012.

[12] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. June 2011.

[13] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D Mapping: Using depth cameras for dense 3D modeling of indoor environments. In *Proceedings of the 12th International Symposium on Experimental Robotics*, 2010.

[14] Vladimir Vezhnevets, Vassili Sazonov, and Alla Andreeva. A survey on pixel-based skin color detection techniques. In *IN PROC. GRAPHICON-2003*, pages 85–92, 2003.

[15] L. Dipietro, A.M. Sabatini, and P. Dario. A survey of glove-based systems and their applications. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(4):461 –482, july 2008.

[16] J. Lobo, P. Trindade, and J. Dias. Observing hand grasp type and contact points using hand distributed accelerometers and instrumented objects. In *IEEE/ICRA 2011: Workshop on Autonomous Grasping*, 2011.

[17] T. Hoshi, S. Ozaki, and H. Shinoda. Three-dimensional shape capture sheet using distributed triaxial accelerometers. In *Networked Sensing Systems, 2007. INSS '07. Fourth International Conference on*, pages 207 –212, june 2007.

[18] J. Lobo and J. Dias. Relative pose calibration between visual and inertial sensors. *The International Journal of Robotics Research (IJRR) Special Issue from the 2nd Workshop on Integration of Vision and Inertial Sensors.*, 26:561–577, 2007.

[19] J. Lobo and J. Dias. Vision and inertial sensor cooperation using gravity as a vertical reference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1597–1608, December 2003.

[20] B.K.P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, 4(4):629–462, April 1987.

[21] Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *Int. J. Comput. Vision*, 13(2):119–152, October 1994.

[22] Polhemus Liberty Electromagnetic Motion Tracking System. http://www.polhemus.com, 2012.

[23] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the ICP algorithm. In *Third International Conference on 3D Digital Imaging and Modeling (3DIM)*, June 2001.