# EVALUATION OF REAL-TIME HAND MOTION TRACKING USING A RANGE CAMERA AND THE MEAN-SHIFT ALGORITHM

Hervé Lahamy * and Derek Lichti

Department of Geomatics Engineering, University of Calgary
2500 University Dr NW, Calgary, Alberta, T2N1N4 Canada
(hdalaham, ddlichti)@ucalgary.ca

**Commission V, WG V/3**

**KEY WORDS:** Hand motion tracking, Range camera, Mean-shift algorithm, Accuracy assessment

**ABSTRACT:**

Several sensors have been tested for improving the interaction between humans and machines including traditional web cameras, special gloves, haptic devices, cameras providing stereo pairs of images and range cameras. Meanwhile, several methods are described in the literature for tracking hand motion: the Kalman filter, the mean-shift algorithm and the condensation algorithm. In this research, the combination of a range camera and the simple version of the mean-shift algorithm has been evaluated for its capability for hand motion tracking. The evaluation was assessed in terms of position accuracy of the tracking trajectory in x, y and z directions in the camera space and the time difference between image acquisition and image display. Three parameters have been analyzed regarding their influence on the tracking process: the speed of the hand movement, the distance between the camera and the hand and finally the integration time of the camera. Prior to the evaluation, the required warm-up time of the camera has been measured. This study has demonstrated the suitability of the range camera used in combination with the mean-shift algorithm for real-time hand motion tracking but for very high speed hand movement in the traverse plane with respect to the camera, the tracking accuracy is low and requires improvement.

## 1. INTRODUCTION

Different sensors have been used to improve the interaction between man and machine. While Breuer et al., (2007) use an infra-red range camera, Luzanin and Plancak (2009) considered specific gloves. A webcam was used by Sung et al. (2008). Elmezain et al. (2009) used 3D depth map computed using left and right images captured with a camera that provides a stereo pair of images. In this research, the range camera is selected for its characteristics described in Section 2.

Among the different algorithms existing in literature for tracking moving objects over time, three are described in the following sub sections.

The Kalman filter is the most commonly used technique. While tracking independent clusters, Heisele and Ritter (1999) assume the motion along the X, Y and Z directions to be decoupled and therefore predicted by separate Kalman filters. The motion of the clusters is assumed to have nearly constant velocity. To account for slight changes in the velocity, the continuous-time acceleration is modelled as white noise. The parameters of the filter are the process noise and the measurement noise. Nguyen et al. (2005) use the Kalman filter to predict the hand location in one image frame based on its location detected in the previous frame. The Kalman filter is used to track the hand region centroid in order to accelerate hand segmentation and choose the correct skin region. Using a model of constant velocity motion, the filter provides and estimates the hand location, which guides the image search for the hand.

Another technique used for tracking a hand segment within acquired images is the condensation algorithm. Isard and Blake (1996) argue that trackers based on Kalman filters are of limited use because they are based on Gaussian densities which are unimodal. They suggest the condensation algorithm which is highly robust in tracking agile motion in the presence of dense background clutter. The condensation algorithm (conditional density propagation) allows quite general representations of probability. One of the most interesting facets of the algorithm is that it does not compute every pixel of the image. Rather, pixels to process are chosen at random, and only a subset of these pixels ends up being processed.

The mean-shift method is a powerful and versatile, non parametric and iterative algorithm that has been used for tracking hand motion. For each data point, the mean-shift algorithm associates it with the nearby peak of the dataset's probability density function. The mean-shift defines a window around it and computes the mean of the data point. Then it shifts the center of the window to the mean and repeats the algorithm till it converges. After each iteration, the window shifts to a denser region of the dataset. At the high level, the mean-shift algorithm can be summarized as follows: fix a window around each data point, compute the mean of data within the window and shift the window to the mean and repeat till convergence. The classic mean-shift algorithm is time intensive. Many improvements have been made to the mean shift algorithm to make it converge faster. This method has been used by Elmezain et al. (2009) in association with the Kalman filter.

The final goal of this research is to design a virtual environment application where range cameras are used for real-time and automatic hand gesture recognition. In this paper, the objective is to evaluate the suitability of the range camera

* Corresponding author.

associated with a fast and efficient hand motion tracking algorithm. A simple version of the mean-shift method has been considered. Section 2 describes the sensor used. Section 3 estimates the camera warm-up time. Section 4 focuses on the segmentation process meaning the extraction of the region of interest. Section 5 highlights the principle of the proposed tracking method. While in Section 6 the accuracy of hand gesture tracking is evaluated under different speeds, distances and integration times, Section 7 estimates the time difference between the image acquisition and the hand segment display while varying the same parameters. Conclusions and future work are provided in Section 8.

## 2. RANGE CAMERA

The sensor considered in this research is the SR4000 range camera (Figure 1). In contrast to stereo cameras where 3D information is obtained from overlapping images, the SR4000 produces a 3D point cloud in every single frame acquired from a single sensor. This ability to provide 3D dense data motivates its choice.

The SR4000 constantly emits an amplitude-modulated infrared light source. Objects in the field of view of the camera at different distances are reached by different parts of the sinusoidal wave which is reflected back. Both range and amplitude images are simultaneously captured by the SR4000 by the means of an integrated sensor. The SR4000 has a low resolution of 176 × 144 pixels. Once the images are acquired, the range information is used to generate the coordinates for every pixel. The range camera produces images at a rate of up to 54 frames per second.

Figure 1. SR4000

## 3. CAMERA WARM-UP TIME

The objective of this section is to test whether a prior warm up of the camera is required in order to get a good accuracy while tracking a hand gesture. The experiment has been performed several times after the camera has been cooled down for more than eight hours. During 160 minutes, 27283 images of a wall were captured at a rate of 2.8 images per second with an integration time of 25ms. The camera was placed perpendicular to a wall at a distance around 1m. After fitting a plane to the obtained point cloud for every image acquired, it has been noticed that the orthogonal distance between the camera and the target has been reduced from 5mm after 40 minutes and remains stable for an hour and half (Figure 2). As a conclusion, for high precision hand motions, a warming up of the SR4000 for 40 minutes is recommended. For most of the considered hand motions where 5mm difference is not an issue, there is no need for a warm-up.

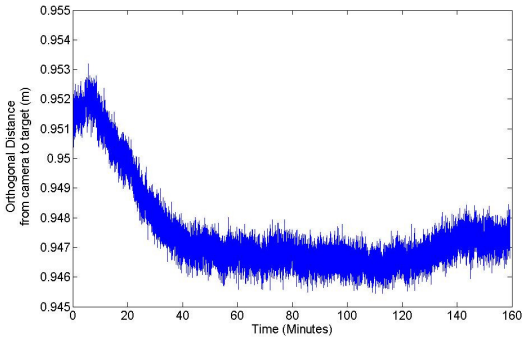



Figure 2: Determination of camera warm-up time

Similar results were obtained by Chiabrando et. al, (2009) where they conclude that a period of 40 minutes warming-up was necessary for the SR4000 to achieve a good measurement stability. Additional tests have been performed regarding the sensitivity of the range camera relative to the integration time, lighting conditions and surrounding objects (Lahamy and Lichti, 2010).

## 4. SEGMENTATION

Segmentation is the process of grouping points that belong to the same object into segments. The idea here is to extract from the point cloud, the set of points that describe the user's hand, the object to be tracked (Figure 3).

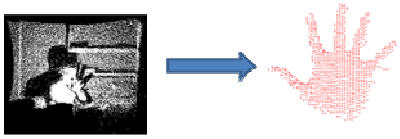


Figure 3: Objective of Segmentation

The most commonly used technique for hand segmentation is colour-based segmentation as demonstrated by Xiaoming and Ming (2001). The skin colour is a distinctive cue and is invariant to scale and rotation. Human hands have almost the same hue and saturation but vary in their brightness. Another method is based on image differencing between consecutive video frames (Zhang et al., 2008). Qiuyu et al. (2008) proposed a hand gesture detection and segmentation method for video sequences coming from a stationary camera with complex backgrounds. The hand segment is extracted based on a threshold grey value calculated from the image's intensity histogram. In Holte and Moeslund (2007), the hand motion is detected using double difference range images. To extract the hand information from a range image obtained from an active camera, Xia and Fujimura (2004) make use of a depth constraint to separate the foreground and the background of the image. Ghobadi et al. (2007) propose a robust segmentation technique based on fusion of range and intensity images. According to the authors, none of the intensity, range and amplitude data delivered by the camera can be used alone to make robust segmentation. In this paper, a multiple-step range based segmentation has been designed (Figure 4).


Figure 4: Multiple-step range based segmentation

This segmentation process has been applied outside of the tracking process in order to determine the initialization position; in other words, the position of the centroid of the hand from which the tracking process starts. The same process has also been used to determine the true positions of the hand centroid after saving the tracked images during the real-time experiments.

### 4.1 Range-based segmentation

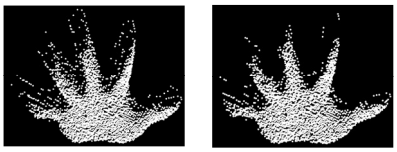
The underlying principle is that there shouldn't be any object between the camera and the hand. Thus the hand appears in the foreground of the image. The algorithm is designed based on the following two key points:

a) Find the first 3000 points closest to the camera using the range. This threshold was obtained from the analysis of the total number of points describing a hand gesture with respect to the distance from the camera to the hand;
b) Assuming that an average human's hand can fit within a 3D cube bounding box of 20cm side; a sub-selection is extracted to achieve this objective; the idea being to get rid of an eventual part of the user's arm;

An example of the result of this algorithm is presented in Figure 5 (Image before noise removal).

### 4.2 Noise Removal

The results obtained contain the appropriate information but appear noisy due to the presence of isolated points (Figure 5). The point density of the hand is much higher than the one of the isolated points. The point cloud obtained from the range-based segmentation is split into voxels (3D cells) of same size. Voxels that have a low point density (maximum of two points) are discarded from the segment.



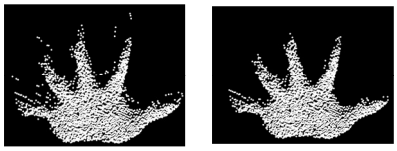
Before... After...

Figure 5: Example of noise removal using the point density

.

### 4.3 Connected Component Analysis

This step is an additional noise removal algorithm. Connected component labelling is used in computer vision to detect unconnected regions. It is an iterative process that groups neighbouring elements into classes based on a distance threshold. A point belongs to a specific class if and only if it is closer within the distance threshold to another point belonging to that same class. After the noise removal, the hand segment appears to be the biggest one in the dataset. An example of the results from and connected component analysis is provided in (Figure 6).



Before... After...

Figure 6: Remaining noise removal using the connected component analysis

## 5. PRINCIPLE OF TRACKING

To avoid a time-consuming segmentation on every acquired frame, tracking the hand gesture appears to be an appropriate method. A simple version of the mean-shift algorithm is implemented. The segmentation process described in Section 4 is used to determine the initial coordinates of the centroid of the hand segment. A 3D cube bounding box of 20cm side centred on this initialization point is then used to collect from the first frame the points expected to belong to the user's hand. Once selected, the centroid of the new set of points is determined. In order to identify the segment in the following frame, the newly determined centroid is considered as the center on the following hand segment and thus the center of the bounding box. Thus iteratively, hand segments are extracted and centroids computed. This method takes its advantage from the fact that the range camera provides the x, y z coordinates of every pixel and for every frame in the same camera frame. In addition, because of the high frequency of the frames, it is assumed that the centers of the hand segments in two consecutive frames are quite close to each other in such a way that the centroid of the hand segment in the first frame can be used as centroid of the segment in the second frame. Consequently, this method applied iteratively enables a real-time tracking that is evaluated in the following Sections.

## 6. EVALUATION OF TRACKING ACCURACY

To evaluate the accuracy of the tracking process, tracked positions of the hand centroid have been compared with the corresponding true positions. The tracked positions are obtained using the results of the proposed tracking methods as described in Section 5 while the true positions are independently computed offline using the segmentation methodology described in Section 4. The acquired images are saved and the robust segmentation process is applied on each of them. The tracked positions have been obtained during the real-time tracking while the true positions have been computed independently of the real-time process. Three parameters were considered in the evaluation: the speed of the hand movement, the distance between the camera and the hand and the integration time of the camera, the objective being to check whether the accuracy of the tracking is a function of those parameters.

### 6.1 Evaluation of the tracking process with respect the speed of the hand movement

In this first experiment, the hand was moved back and forth with respect to the camera in the range direction and was also kept static for a while. Figure 7 shows that the speed of the hand does not influence the accuracy of the tracking. From Table 1, it can be concluded that the overall tracking accuracy is better than 1cm with the hand moving at up to 40cm/s, which is accurate enough for the intended application.

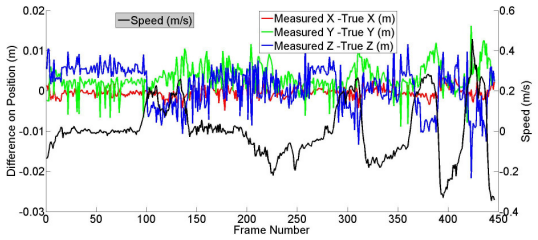



Figure 7: Evaluation of the tracking process while varying the speed of the hand movement in the range direction.

While moving the hand in the left and right directions with respect to the camera, that is in the traverse plane, and varying the speed of the movement (Figure 8), the independence of the accuracy with respect to the speed is observed only when the speed is lower than 10cm/s. For higher speeds, the difference between the true and the measured positions can reach 20cm in the Z direction (Figure 8 and Table 1) but the overall accuracy is 7cm. This lower performance can be explained by the fact that the speed of the hand becomes too high compared to the frame rate causing the number of selected points in the bounding box lower than it should. As a consequence the centroid computed is wrong and thus displaced from the true position of the hand. Examples of the tracked hand segments as the hand speed gets higher is provided in Figure 9.

The current method is thus limited when moving the hand very fast in the left and right directions with respect to the position of the camera. The tracking model implemented doesn't take into account the dynamics involved in the hand movement. In future work, the Kalman filter that can model this dynamics will be included in order to improve the tracking accuracy in case of high speed hand movement.

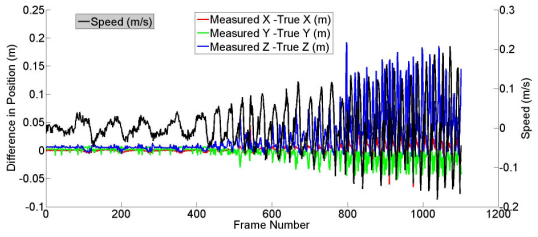



Figure 8: Evaluation of the tracking process while varying the speed of the hand movement in the transverse plane.

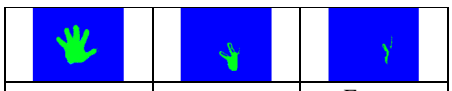



Figure 9: Examples of tracked hand segments as the speed gets higher

Table 1: Accuracy of hand motion tracking with variation in speed of hand movement

| Hand Motion | RMS(X) (mm) | RMS(Y) (mm) | RMS(Z) (mm) | RMS(XYZ) (mm) |
|---|---|---|---|---|
| Along the range direction | 2.9 | 4.8 | 8.3 | 10.0 |
| In the transverse plane (<10 cm/s) | 1.7 | 3.3 | 5.7 | 6.8 |
| In the transverse plane (>10 cm/s) | 13.4 | 22.5 | 72.7 | 77.2 |

### 6.2 Evaluation of the tracking process with respect to the distance Hand-Camera

Figure 10 and Table 2 show the results of the experiment where the hand has been tracked back and forth up to a distance greater than 2m at an average speed of 10cm/s. When the hand is at a distance higher than 50cm from the camera, the overall accuracy is better than 1cm. This accuracy decreases when the distance is lower than 50cm. Indeed, at such a distance, the intensity of light reaching the hand is too high causing some saturated pixels resulting in a wrong segmentation. To solve this issue, the integration time which was set at 27.2ms has to be adjusted constantly, no matter the distance between the camera and the hand. The adjusting method, described in (Lahamy and Lichti, 2010) removes the drawback noticed and makes the tracking independent on the distance.

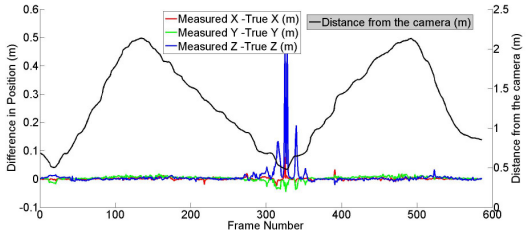



Figure 10: Evaluation of the tracking process while varying the distance Camera-Hand

Table 2: Accuracy of hand motion tracking with variation in the distance Hand-Camera

| Distance Camera-Hand | RMS(X) (mm) | RMS(Y) (mm) | RMS(Z) (mm) | RMS(XYZ) (mm) |
|---|---|---|---|---|
| > 50 cm and < 2.1m | 3.7 | 6.4 | 5.5 | 9.2 |
| <50 cm | 28.1 | 22.6 | 244.6 | 247.2 |

### 6.3 Evaluation of the tracking process with respect to the integration time of the Camera

The integration time is the period of time during which the pixels are allowed to collect light. For the SR4000, the integration times varies between 17.2ms and 119.2ms. Being aware that an appropriate integration time depends on the distance between the camera and the target, Figures 11 and 12 as well as Table 3 reveal that the accuracy of the tracking is independent of the integration time. Indeed, by tracking the hand over different integration times selected through the possible range available, it can be noticed that the overall accuracy in the worst case is around 1cm; which is good enough for the intended application.

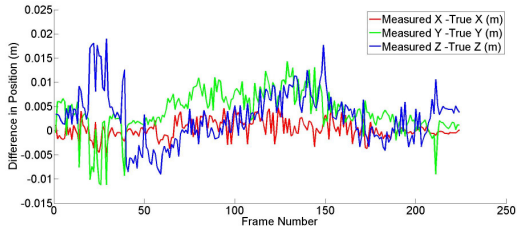



Figure 11: Evaluation of the tracking process with the integration time = 32ms

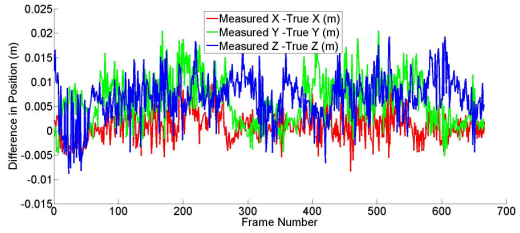



Figure 12: Evaluation of the tracking process with the integration time = 96ms

Table 3: Accuracy of hand motion tracking with variation in the integration time

| Integration Time (ms) | RMS(X) (mm) | RMS(Y) (mm) | RMS(Z) (mm) | RMS(XYZ) (mm) |
|---|---|---|---|---|
| 32 | 1.9 | 5.8 | 5.9 | 8.5 |
| 48 | 2.0 | 5.8 | 6.4 | 8.9 |
| 64.8 | 2.3 | 7.2 | 5.1 | 9.1 |
| 80 | 2.2 | 5.8 | 6.2 | 8.8 |
| 96 | 3.1 | 7.7 | 8.1 | 11.6 |

## 7. REAL-TIME CAPABILITY OF THE RANGE CAMERA

The evaluation of the hand motion tracking has also been assessed by measuring the time difference between the image acquisition and image display. This study has been made by varying the three parameters previously described: the speed of the hand movement, the distance between the camera and the hand and the integration time of the camera.

Figures 13, 14, 15 and 16 as well as Table 4 show that the time between the image acquisition and the hand segment display is independent of the speed of the hand movement, the integration time of the camera and the distance between the camera and the hand gesture. It varies from 9ms to 21ms which is a good rate for a real-time application where a maximum rate of 25ms is expected.

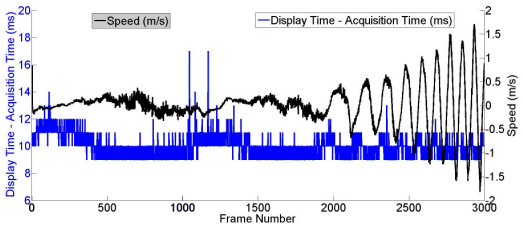



Figure 13: Evaluation of the difference of time between image acquisition and segment display while varying the speed of the hand movement.

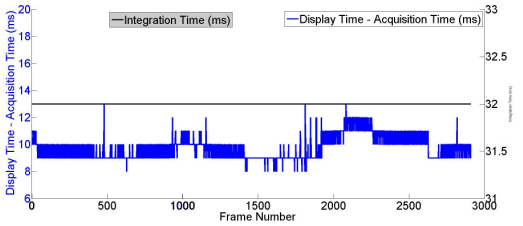



Figure 14: Evaluation of the difference of time between image acquisition and segment display with the integration time = 32ms

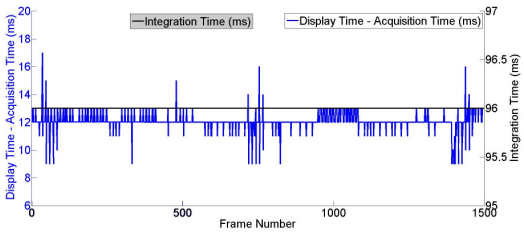



Figure 15: Evaluation of the difference of time between image acquisition and segment display with the integration time = 96ms

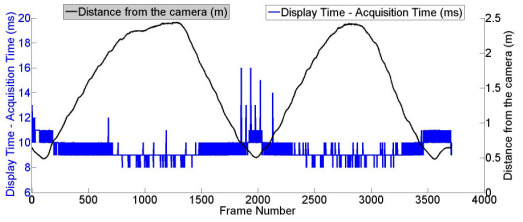



Figure 16: Evaluation of the difference of time between image acquisition and segment display while varying the distance Camera-Hand

Table 4: Difference of time between image acquisition and segment display

| | Speed (ms) | Distance (ms) | Integration Time (32 ms) | Integration Time (96 ms) |
|---|---|---|---|---|
| Minimum | 8 | 8 | 8 | 9 |
| Average | 9 | 9.4 | 10 | 12 |
| Maximum | 18 | 16 | 15 | 20 |

## 8. CONCLUSION AND FUTURE WORK

Tracking hand gestures using a range camera and the proposed tracking algorithm turns out to be quite accurate (around 1cm accuracy and 12ms in average to display the acquired image). This accuracy does not depend on the distance between the hand and the camera or on the integration time of the camera when properly set. But the accuracy decreases when the speed of the hand movement is higher than 10cm/s in the transverse plane with respect to the position of the camera. No prior warming up of the camera is required as the 5mm displacement that occurs after 40 min warming-up is not an issue when tracking hand gestures. Finally the time between image acquisition and hand segment display is small enough for real-time applications.

Future work includes a comparative analysis with the method mostly used which is the Kalman filter.

## ACKNOWLEDGEMENT

This work was supported by the Werner Graupe International Fellowship, the Computer Modelling Group LTD and the Natural Sciences and Engineering Research Council of Canada (NSERC).

## REFERENCES

Breuer, P.; Eckes, C. and Muller, S., Hand gesture recognition with a novel IR time-of-flight range camera - a pilot study. in the Proceedings of the Mirage 2007, Computer Vision/Computer Graphics Collaboration Techniques and Applications, Rocquencourt, France, 28-30 March, pp.247-260.

Chiabrando, F.; Chiabrando, R.; Piatti, D. and Rinaudo, F. Sensors for 3d imaging- metric evaluation and calibration of a CCD/CMOS time-of-flight camera, Sensors, 2009, 9(12), pp. 80–96.

Elmezain, M.; Al-Hamadi, A. and Michaelis, B.; Hand Trajectory-Based Gesture Spotting and Recognition Using HMM. 16th IEEE International Conference on Image Processing, 7-12 Nov. 2009, IEEE pp3577-80.

Ghobadi, S.; Loepprich, O.; Hartmann, K. and Loffeld, O. Hand segmentation using 2d/3d Images. In Proceedings of Conference Ivcnz 2007, Hamilton, New Zealand, 2007.

Heisele, B. and Ritter, W., Segmentation of range and intensity image sequences by clustering. In Proceedings of International Conference on Information Intelligence and Systems, 31 Oct.-3 Nov. 1999, IEEE Comput. Soc Pp223-5.

Holte M.B., and Moeslund T.B. Gesture Recognition using a range camera. Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark, 2007.

Isard M., Blake, A., Contour Tracking By Stochastic Propagation Of Conditional Density, Proceedings Of The 4th European Conference On Computer Vision-Volume I, P.343-356, April 15-18, 1996

Lahamy H. and Lichti D. Hand Gesture Recognition using Range Cameras, In Proceedings of the 2010 ISPRS Commission I Mid-term Symposium on Image data acquisition, sensors and platforms, Calgary Canada 15-18 June 2010.

Luzanin, O. and Plancak, M. Enhancing gesture dictionary of a commercial data glove using complex static gestures and an MLP ensemble. Strojniski Vestnik, 2009, 55(4), 230-6.

Nguyen D. B.; Enokia S. and Toshiaki E. Real-Time Hand Tracking and Gesture Recognition System. In Proceedings of the International Conference on Graphics, vision and image Processing (Gvip-05), Pages 362-368, December 2005.

Qiuyu Z.; Fan C. and Xinwen L. Hand Gesture Detection and Segmentation Based on Difference Background Image with Complex Background, Embedded Software and Systems, 2008. International Conference on, pp. 338 – 343, 29-31 July 2008.

Stenger, B.; Mendonca, P.R.S.; and Cipolla, R. Model-Based 3d Tracking of an articulated hand, Computer Vision and Pattern Recognition. In Proceedings of the 2001 IEEE Computer Society Conference, ppii-310-Ii-315 Vol.2.

Sung, K.K.; MI, Y.N. and PHILL, K.R. Color based hand and finger detection technology for user interaction. In Proceedings of the International Conference on Convergence and Hybrid Information Technology (ICHIT), 28-29 Aug. 2008, IEEE pp229-36.

Xia, L. and Fujimura, K. Hand Gesture Recognition Using Depth Data, In Proceedings of sixth IEEE International Conference on Automatic Face And Gesture Recognition, 17-19 May 2004, pp. 529 – 534.

Xiaoming Yin and Ming X. Hand Gesture Segmentation, Recognition and application. Computational Intelligence In Robotics and Automation. In Proceedings 2001 IEEE International Symposium on 2001, pp438-443.

Zhang, Q.; Chen, F. and Liu, X. Hand Gesture Detection and Segmentation based on Difference Background Image with Complex Background. International Conference on Embedded Software and Systems, 29-31 July 2008, IEEE pp338-43.