ROBUST REAL-TIME AND ROTATION-INVARIANT AMERICAN SIGN LANGUAGE ALPHABET RECOGNITION USING RANGE CAMERA

H. Lahamy * and D. Lichti

Department of Geomatics Engineering, The University of Calgary, 2500 University Dr NW - (hdalaham, ddlichti)@ucalgary.ca, Calgary AB Canada T2N 1N4

Commission V, WG V/3

KEY WORDS: Gesture recognition, Range camera, Hand motion tracking, 3D Signature, Rotation Invariance

ABSTRACT:

The automatic interpretation of human gestures can be used for a natural interaction with computers without the use of mechanical devices such as keyboards and mice. The recognition of hand postures have been studied for many years. However, most of the literature in this area has considered 2D images which cannot provide a full description of the hand gestures. In addition, a rotation-invariant identification remains an unsolved problem even with the use of 2D images. The objective of the current study is to design a rotation-invariant recognition process while using a 3D signature for classifying hand postures. An heuristic and voxel-based signature has been designed and implemented. The tracking of the hand motion is achieved with the Kalman filter. A unique training image per posture is used in the supervised classification. The designed recognition process and the tracking procedure have been successfully evaluated. This study has demonstrated the efficiency of the proposed rotation invariant 3D hand posture signature which leads to 98.24% recognition rate after testing 12723 samples of 12 gestures taken from the alphabet of the American Sign Language.

1. OBJECTIVE AND RELATED WORK

The objective of the current study is to design a system where the alphabet of the American Sign Language can be recognized in real-time. Contrary to existing methods, the current one allows hand posture recognition independently of the orientation of the user's hand. It makes use of a 3D signature, considers only one training image per posture and uses a significant number of testing images for its evaluation.

A lot of research has been conducted on this topic but most of it doesn't consider using the advantage that a 3D signature can provide. For example, in Yunli and Keechul (2007), after generating a point cloud of a hand posture from data captured with four web cameras, the authors use cylindrical virtual boundaries to randomly extract five slices of the point cloud. Each slice is processed by analyzing the point cloud distribution and the hand posture is recognized from this analysis. By doing so, though the hand postures are represented by a 3D point cloud, the full 3D topology is not considered in the recognition process. Other researchers, though using a 3D sensor, do not consider at all the third dimension in the features used to represent the hand postures. That is the case of Guan-Feng et al., (2001) where the authors use a 3D depth camera but only consider the 2D outline of the hand segment in their recognition process.

The design of a rotation invariant system has not been successfully achieved so far. Indeed many researchers consider the principal component analysis to evaluate the orientation of the 2D hand image but, as acknowledged by Uebersax et al. (2012), this method is not always accurate. Not only has the estimation of the rotation of a 2D hand segment not been successful so far but, furthermore the evaluation of the orientation of a 3D hand segment is not considered in most of existing approaches.

To test their hand motion classification using a multi-channel surface electromyography sensor, Xueyan et al. (2012) only consider five testing images per gesture. Contrary to most of the studies on this topic, a significant number of testing samples has been considered to validate the proposed algorithm. Indeed, testing 1000 images instead of 5 provides more evidence on the robustness of the methodology.

The proposed method considers only one single training image with the objective of showing the robustness of the method and also its appropriateness for a real-time application.

To track the hand motion during the real-time process the Kalman filter has been proposed with a detailed explanation on how the process noise and the measurement noise have been modelled.

In order to achieve these objectives, the sensor considered is the SR4000 range camera because of its ability to provide 3D images at video rates. For further details, please refer to (Lange, 2001) who provide an exhaustive explanation on the SR4000's principles.

This paper is structured as follows: Section 2 describes the set up of the experiment and section 3, the methodology for tracking the hand motion and its evaluation. In section 4, the recognition principle is depicted. The rotation invariance algorithm is highlighted in section 5. The experimental results

* Corresponding author.

and their analysis are shown in section 6 while a comparison with results from other papers is discussed in section 7. Conclusion and future work are provided in section 8.

2. EXPERIMENT SETUP

The experiment has been conducted in laboratory conditions. The camera was mounted on a tripod situated approximately at 1.5m from the user sitting on a chair and facing a desktop where the images acquired by the SR4000 camera as well as the results from the hand posture recognition are displayed in real-time. No particular background is required. Similarly, the user does not have to wear neither long sleeves nor specific gloves as required in some similar experiments as the segmentation is not based on colour. The integration time which is the length of time during which the pixels are allowed to collect light was set to an integer time of 15 (1.8ms) corresponding to a theoretical frame rate of 43 images per second.

3. HAND MOTION TRACKING

In order to recognize in real-time the hand postures, the hand blob has to be tracked within the acquired images. Considering that the hand is undergoing a linear movement with constant velocity and applying the Newton's law of motion, the linear discrete Kalman filter has been used for tracking the centroid of the hand. Using a 3D cube of 20cm side centred on the tracked point, all points falling within this cube are assumed to belong to the hand. The sate vector **(x)** comprises three states describing the position of the centroid of the hand segment in the camera frame and three other states corresponding to the velocity of the hand movement. The initial position coordinates are obtained from an initialization process where the user indicates approximately the starting position of the hand. An image of the hand is acquired and a segmentation process described in Lahamy and Lichti (2010) is applied to extract the hand segment from which the centroid is computed. The initial velocity is assumed to be null.

The change in the state vector is expressed as follows:

$$\mathbf{x_k} = \mathbf{\Phi_{k-1}} \mathbf{x_{k-1}} + \mathbf{w_{k-1}} \quad (1)$$

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{I_3} & \mathbf{\Delta t I_3} \\ \mathbf{O_3} & \mathbf{I_3} \end{pmatrix} \quad (2)$$

where $\boldsymbol{\Phi}$ is the transition matrix, $\mathbf{w}$ the process noise considered as white and $\mathbf{\Delta t}$ the time elapsed between the acquisition of the two consecutive frames k-1 and k.

To update the state vector, the measurement used **(z)** is made of the three position coordinates of the centroid of the hand segment.

$$\mathbf{z_k} = \mathbf{H_k x_k} + \mathbf{v_k} \quad (3)$$

$$\mathbf{H} = \begin{pmatrix} \mathbf{I_3} & \mathbf{O_3} \end{pmatrix} \quad (4)$$

where **H** is the observation matrix and **v** is the observation white noise that is assumed to be uncorrelated with **w**. The process noise and the measurements errors are respectively characterized by the covariance matrices **Q** and **R**. **Q** is computed by considering the accelerations $\mathbf{a_x}$, $\mathbf{a_y}$ and $\mathbf{a_z}$ of the hand and the equations (5) and (6). These accelerations are computed as the corresponding differences of velocity divided by the elapsed time between three consecutive images. Similarly, the velocities are computed as differences of corresponding positions divided by the appropriate time. Accelerations can only be computed if a minimum of three consecutive positions of the hand centroid have been recorded. **R** is obtained by considering that the observed coordinates have a precision of 1cm (Equation 7). This value has been obtained by imaging a static object (Spectralon target) hundred times within similar set up conditions as for the hand posture recognition. For several pixels, the standard deviations have been computed in X, Y and Z directions of the camera frame. The variations for the central pixel range from 2 mm in Y direction to 7 mm in Z direction.

$$\mathbf{Q} = \mathbf{G}\mathbf{G}^{\mathbf{T}}\boldsymbol{\sigma}_{\mathbf{a}}^{\mathbf{2}} \quad (5)$$

$$\mathbf{G}^{\mathbf{T}} = \begin{pmatrix} \frac{\Delta t^2}{2}\mathbf{a_x} & \frac{\Delta t^2}{2}\mathbf{a_y} & \frac{\Delta t^2}{2}\mathbf{a_z} & \mathbf{\Delta t} & \mathbf{\Delta t} & \mathbf{\Delta t} \end{pmatrix} \quad (6)$$

$$\mathbf{R} = \boldsymbol{\sigma}^{\mathbf{2}}\mathbf{x}\mathbf{I_3} \quad (7)$$

The Kalman filter state prediction and state covariance prediction are computed as follows:

$$\bar{\mathbf{x}}_{\mathbf{k}} = \mathbf{\Phi_{k-1}}\hat{\mathbf{x}}_{\mathbf{k-1}} \quad (8)$$

$$\bar{\mathbf{P}}_{\mathbf{k}} = \mathbf{\Phi_{k-1}}\hat{\mathbf{P}}_{\mathbf{k-1}}\mathbf{\Phi}_{\mathbf{k-1}}^{\mathbf{T}} + \mathbf{Q_{k-1}} \quad (9)$$

where $\hat{\mathbf{x}}_{\mathbf{k}}$ denotes the estimated state vector; $\bar{\mathbf{x}}_{\mathbf{k}}$ is the predicted state vector for the next epoch; $\hat{\mathbf{P}}_{\mathbf{k}}$ is the estimated state covariance matrix; $\bar{\mathbf{P}}_{\mathbf{k}}$ is the predicted state covariance matrix.

The Kalman filter update steps are as follows:

$$\mathbf{K_k} = \bar{\mathbf{P}}_{\mathbf{k}}\,\mathbf{H}_{\mathbf{k}}^{\mathbf{T}}(\mathbf{H_k}\,\mathbf{P_k}\mathbf{H}_{\mathbf{k}}^{\mathbf{T}} + \mathbf{R_k})^{-\mathbf{1}} \quad (10)$$

$$\mathbf{v_k} = \mathbf{z_k} - \mathbf{H_k}\bar{\mathbf{x}}_{\mathbf{k}} \quad (11)$$

$$\hat{\mathbf{x}}_{\mathbf{k}} = \bar{\mathbf{x}}_{\mathbf{k}} + \mathbf{K_k v_k} \quad (12)$$

$$\hat{\mathbf{P}}_{\mathbf{k}} = (\mathbf{I} - \mathbf{K_k H_k})\bar{\mathbf{P}}_{\mathbf{k}} \quad (13)$$

Where $\mathbf{K_k}$ is the Kalman gain, which defines the updating weight between the new measurements and the prediction from the system dynamic model.

The tracking process was assessed in terms of position accuracy of the tracking trajectory in X, Y and Z directions in the camera space. The coordinates of the true positions have been compared with those of the tracked positions, the true position is assumed to be the position of the hand segment centroid obtained from every segment.

In this assessment, the hand was moving back and forth in the X, Y and Z directions of the camera. The movement of the hand is assumed to be linear with a nearly constant velocity in all three directions. Any deviation from the assumptions of linearity and constant velocity are compensated by the process noise and the measurement noise. The elapsed time between

two consecutive acquired images including the image acquisition and the recognition processing time is 28ms in average, which corresponds to 35 images per second. This frame rate is good enough for a real-time visualization as 24FPS is the progressive format now widely adopted by film and video makers, Read and Meyer (2000).

From Figures 1, 2 and 3, a general matching between the tracked positions and the true positions is noticeable. However two phenomena can be observed: The mismatch observed at the beginning of the curves showing that the initial position of the hand is different from where it is expected to be; which is expected as the user cannot start exactly at the position defined during the initialization step. The other mismatches occur when the hand movement is changing direction. Though identifiable on the curves, these phenomena do not change the overall good RMSE of the tracking which are 3.5mm, 4.5mm and 7.0mm in X, Y and Z directions respectively. These accuracies demonstrate the appropriateness of the modelling of the process noise and the measurement noise.

In terms of overall accuracy, this algorithm performs slightly better than the mean-shift algorithm that was described and evaluated in Lahamy and Lichti (2011a). However, tracking hand gestures using a range camera and the mean-shift algorithm turns out to be quite accurate too (around 1cm accuracy) and in addition, 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.

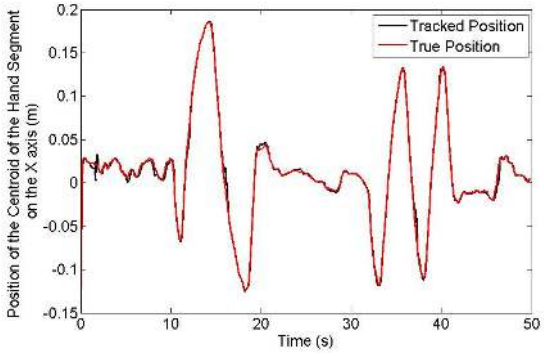

Figure 1. Evaluation of the tracking method: Comparison between tracked and true positions of the hand centroid on the X axis

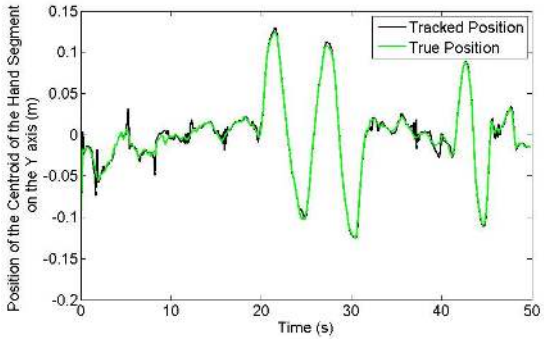

Figure 2. Evaluation of the tracking method: Comparison between tracked and true positions of the hand centroid On the Y axis

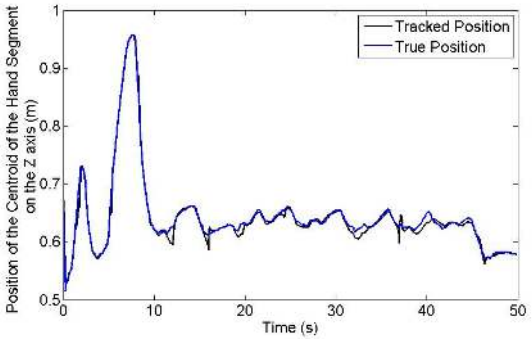

Figure 3. Evaluation of the tracking method: Comparison between tracked and true positions of the hand centroid On the Z axis

4. PRINCIPLE OF POSTURE RECOGNITION

Several recognition methods most commonly appearing in the literature have been described and evaluated in Lahamy and Lichti (2011b). The maximum overall recognition rate obtained is 97.29% when using the outline of the 2D hand segment and its corresponding distance transformation image as features and the chamfer distance as classifier. But unfortunately these features are not appropriate for real-time application. In this research, the real-time recognition of the postures has been achieved by using an heuristic and voxel-based algorithm. To derive a signature of a hand gesture from its point cloud, a 3D bounding box is generated and transformed into a regular grid. Only voxels containing at least one point from the hand segment's point cloud are considered in the hand gesture's signature. A column vector is thus generated in which every voxel is represented by a Boolean value, true when it contains at least one point and false when it is empty. This vector contains the full 3D topology of the hand gesture as voxels are stored following a predefined order. This signature is equivalent to the original gesture as it can be used to reconstitute the 3D structure of the original gesture. It considers the position and orientation of every single part of the hand segment. In addition, it has the advantage of containing less information and consequently it is easier to store and is more time-efficient for processing. It is a suitable parameter to measure the similarity between hand gestures. With this signature, hand gestures are compared by considering their 3D structure. Figure 4 shows an example of the generation of a hand signature. A 30X30X30 representation has been considered because after testing empirically several combinations, it appears to be the one that provides the better results.

To compare two hand gestures, one of the two point clouds is translated onto the second one by matching their centroids. The idea here is to define the same bounding box for both gestures to be compared. The comparison is achieved by evaluating the percentage of similarity between the two gestures in other words, the percentage of voxels containing at least one point in both datasets. This percentage is a suitable parameter to measure the similarity between hand gestures. Thus, hand gestures are compared by considering their 3D structure.

For each of the training images, the hand segment is computed and stored in addition to the corresponding class. For every image in the testing database, the gesture recognition is performed by comparing the current gesture to the training ones previously stored. The similarity measure between the candidate and all the templates in the training database are calculated and the highest score indicates the recognized posture. The selected gesture is the one from the training database that is closest in 3D topology to the current gesture. However, for this result to be considered acceptable, the likelihood should be higher than 50% which means that both gestures have 3D structures that are at least 50% similar to each other. In case this minimum requirement is not achieved for none of the training images, the result of the classification is "Gesture not recognized".

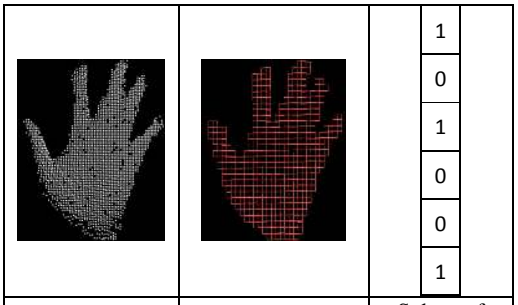

Figure 4. Generation of a hand segment signature

5. ROTATION INVARIANCE

The performance analysis of different methods for hand gesture recognition using range cameras, Lahamy and Lichti (2011b), show that none of the features considered are rotation invariant, which is not realistic for real-time applications where the user will not be allowed to rotate his hand while interacting with a computer.

In this research, the rotation-invariance of the gesture is achieved by measuring the orientation of the segmented point cloud and by removing that rotation before entering the recognition process. Once the orientation is removed, the segment is un-rotated and its signature can be computed. The recognition is thus performed with an un-rotated hand segment. The evaluation of the orientation is achieved in two steps. Using the principal component analysis, the primary axis of the hand segment is derived by considering the eigenvector corresponding to the largest eigenvalue. The angle between this principal axis and the Y axis of the camera frame is used to rotate the segment by bringing them into coincidence. At this step, the rotation is not completely removed as the angle around the Y axis between the direction the gesture is facing and the one it should be facing (the Z axis of the camera frame) is not yet evaluated. To measure the latter, the centroid of the hand segment is determined. All points within 3 cm of the centroid are selected and assumed to belong to the hand palm. Using least square regression, a plane is fitted within the obtained set of points. The perpendicular direction to this plane is supposed to be the direction the gesture is facing. An example is provided in Figure 5 where the original hand segment, the result of a random rotation applied to it, the result of the first rotation removal as well as the final result are shown.

One ambiguity appears in this methodology for evaluating the rotation of the hand segment: The orientations of the two axes determined. For example, after the first rotation removal, it is not always clear whether the hand segment is pointing upwards or downwards. The same problem has been notated for the second rotation removal. It has not been possible to rigorously solve this ambiguity and as a consequence for both rotation axes, the two orientations have to be considered every time; which results in four different possibilities for the un-rotated hand segment. An example of an incorrect, un-rotated hand segment that justifies this chosen solution is provided in Figure 6.

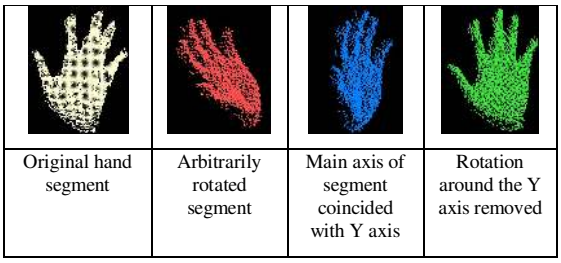

Figure 5. Example of rotation removal of a hand segment

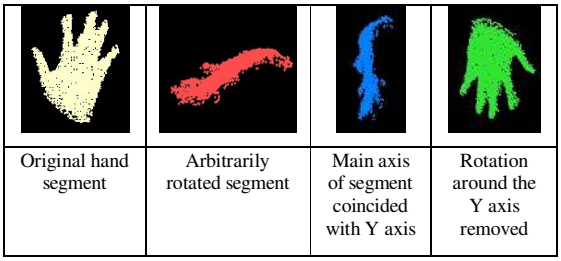

Figure 6. Example of rotation removal of a hand segment

To validate this methodology, the 33 postures that appear in the American Sign Language alphabet (Figure 7) have been considered. Only one template per posture has been captured with the range camera followed by the segmentation of the hand. After generating three random values representing the angles around the X, Y and Z axes, the latter are applied to the original template to obtain an arbitrarily rotated hand segment. One thousand rotated segments were generated randomly per posture. The recognition principle described earlier is then used to classify the 33 000 rotated images. The overall recognition rate of 97.88 % (Figure 8) clearly demonstrates that the methodology used to evaluate a hand posture's orientation is accurate enough to be used in a real-time application.

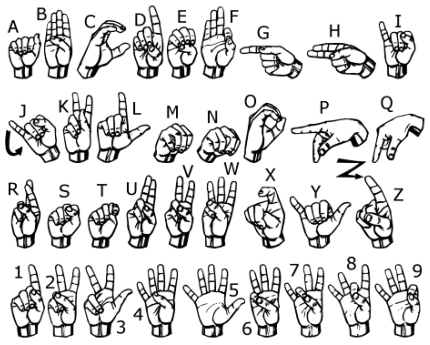

Figure 7. Considered gestures
(Source: http://www.lifeprint.com/asl101/fingerspelling/fingerspelling.htm accessed in April 2010)

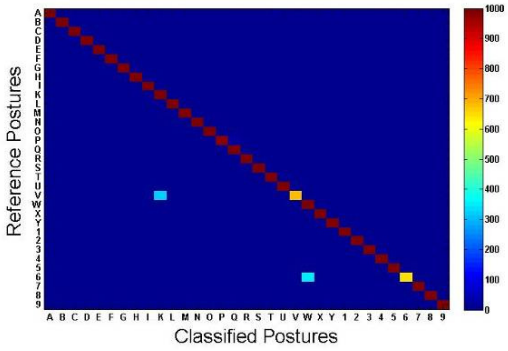

Figure 8. Confusion Matrix of the 33000

An analysis of the confusion matrix obtained from this classification (Figure 8) shows that 315 "V' images have been classified as "K" and 346 "6" images classified as 'W". The high similarity existing between those sets of postures as shown by Figure 7 and the fact that there might still be a very slight rotation in the un-rotated segment are the reasons of this misclassification.

6. EXPERIMENTAL RESULTS AND ANALYSIS

The signature herein described has been applied on 18 postures in Lahamy and Lichti (2012), 84% recognition rate was obtained after testing around 30000 samples which were not rotated. To evaluate this signature in combination with the algorithm for rotation removal, only 12 postures out of the 33 (Figure 7) have been considered. The main reason is to avoid postures that look alike and cause some misclassifications. For example, Figure 7 shows the resemblance between "A", "M", "N", "S" and "T" and "E". The same observation can be made with "D" and "1", "W" and "6" or with "G" and "H". In this case, only one training image has been used per posture while more than 1000 images were tested for each selected gesture. Some snapshots from the real-time application showing the original range image, the segmented rotated hand blob, the un-rotated segment as well as the recognized hand postures are presented in Figure 9. Figure 10 and Figure 11 present respectively, the confusion matrix and the recognition rates.

Because of the careful selection of the postures that avoids any resemblance, the overall recognition rate is 98.24%. Very few mismatches have been noted as shown by the Figure 10.

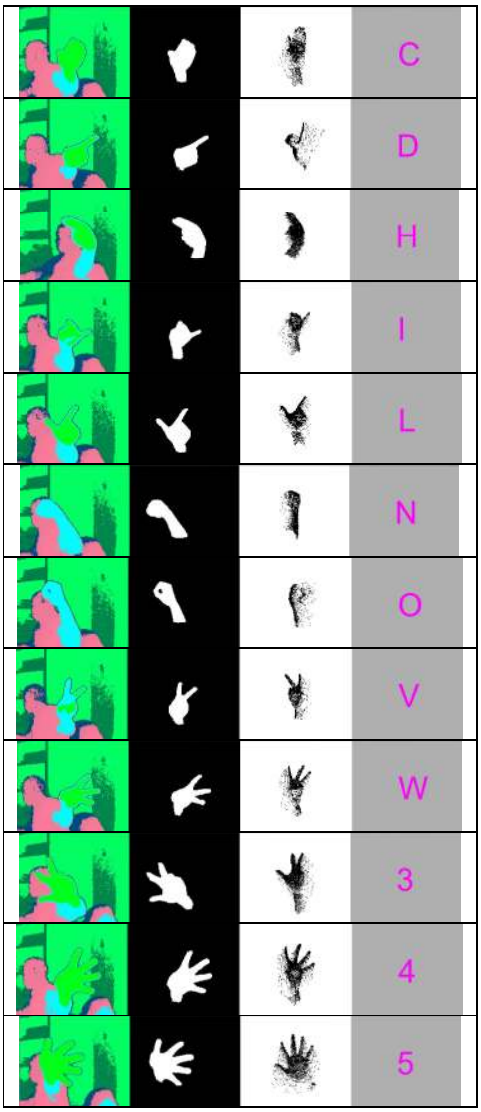

Fig. 9. Snapshots from real-time application showing the original range image, the segmented hand blob and the recognized hand postures

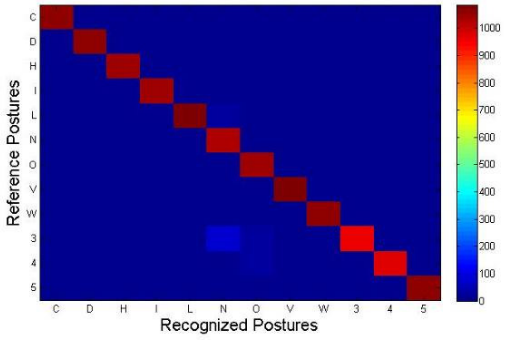

Figure 10. Confusion Matrix

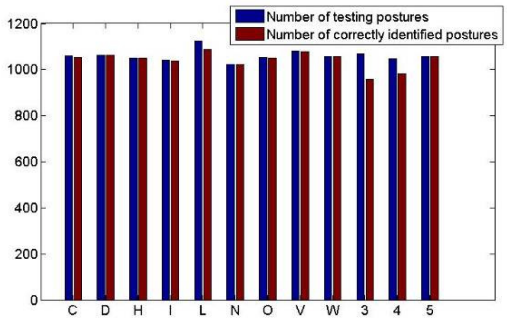

Figure 11. Recognition Rates

7. COMPARISON WITH EXISTING METHODS

Bourennane and Fossati, (2010) have performed a comparative study where different shape descriptors for hand posture recognition have been evaluated using different classification methods. Using a database made of 11 gestures and 1000 images per gesture taken from different users, the hand posture recognition has been performed with Hu-moments, Zernike moments and Fourier descriptors as features and Bayesian classifier, Support vector machine, k-Nearest Neighbors (k-NN) and Euclidian distance as classifiers. The best result achieved is 87.9% using the k-NN and Fourier descriptors. Another study, Bastos and Dias, (2008) show 88.8% overall recognition rate after testing 12 gestures with 20 images for each resulting in 240 tested images. In this study, the authors use the orientation histogram as feature and the gesture matching has been accomplished using Normalized Cross Correlation. Both papers claim a rotation-invariant process but there is no clear evidence on this objective.

Though the methodology described herein is tested on similar number of postures, it provides a higher recognition rate and has several advantages over these methods: the use of 3D images, a segmentation process independent on the skin of the colour, on the background of the image and on whether the user needs to wear long sleeves or not. In addition, only one training image is required compared to 500 in Bourennane and Fossati, (2010).

8. CONCLUSION AND FUTURE WORK

The current paper addresses the following question: How to recognize hand postures independently of the hand orientation while using a better representation of the hand compared to the mostly available ones in Literature? Though simplistic, the proposed signature associated with the rotation invariant algorithm has been successful in recognizing 12 postures taken from the American Sign Language alphabet. Indeed, 98.24% of the 12723 postures tested have been correctly recognized. This method uses a 3D representation of the hand and it has been proven the robustness of the rotation invariant algorithm. In addition, the objective was to design a real-time application and thus reduce as much as possible the recognition process time. To achieve the latter, only one training image has been considered in the supervised classification.

In future work, the focus will be made on the improvement of the signature as well as well as the recognition process in order to achieve all the 33 gestures appearing in the alphabet of the American Sign Language. Furthermore, dynamic gestures involving one or two hands and also multiple cameras will be addressed.

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

Bastos R. and Dias M.S., 2008. Skin color profile capture for scale and rotation invariant, hand gesture recognition 7th International Gesture Workshop, Lisbon. pp. 81-92.

Bourennane S. and Fossati C., 2010. Comparison of shape descriptors for hand posture recognition. Conference on Signal Image and Video Processing, Springer, London. pp. 1-11.

Guan-Feng H., Sun-Kyung K., Won-Chang S., Sung-Tae J., 2011. Real-time gesture recognition using 3D depth camera. IEEE 2nd International Conference on Software Engineering and Service Science, Beijing, pp. 187-190.

Lahamy H. and Lichti D., 2010. Real-Time Hand Gesture recognition using range camera. Canadian Geomatics Conference and the International Symposium of Photogrammetry and Remote Sensing (ISPRS) Commission (I) Calgary, Canada, June 14-18.

Lahamy H. and Lichti D., 2011a. Evaluation of hand gesture tracking using a range camera. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Calgary, Canada, Vol. XXXVIII-5/W12.

Lahamy H. and Lichti D., 2011b. Performance analysis of different methods for hand gesture recognition using range cameras. SPIE 8085, 80850B Munich, Germany.

Lahamy H. and Lichti D., 2012. Heuristic and voxel-based signature for hand posture recognition using a range camera the International Conference on Image Analysis and Recognition, ICIAR, Aveiro, Portugal.

Lange, R. and Seitz, P., 2001. Solid-state time-of-flight range camera, IEEE Journal of Quantum Electronics, 37(3), pp. 390–397.

Read P. and Meyer M., 2000. Restoration of motion picture film, Conservation and Museology. Gamma Group, Butterworth-Heinemann. pp. 24–26.

Uebersax, D., Gall, J., Van den Bergh, M., Van Gool, L., 2012. Real-time Sign Language Letter and Word Recognition from Depth Data. IEEE Workshop on Human Computer Interaction, Barcelona, pp. 383- 90.

Xueyan, T., Yunhui, L., Congyi, L. and Dong, S. 2012, Hand Motion Classification Using a Multi-Channel Surface Electromyography Sensor. Sensors, 12(2), pp. 1130-1147.

Yunli L. and Keechul J., 2007. 3D Posture Representation Using Meshless Parameterization with Cylindrical Virtual Boundary. PSIVT, Springer, Berlin. pp. 449-461.