# KERNEL BASED HAND GESTURE RECOGNITION USING KINECT SENSOR

Daniela Ramírez-Giraldo*, Santiago Molina-Giraldo*, Andrés M. Álvarez-Meza*
Genaro Daza-Santacoloma* †, and Germán Castellanos-Domínguez*
*Signal Processing and Recognition Group, Universidad Nacional de Colombia sede Manizales, Manizales, Colombia
†Instituto de Epilepsia y Parkinson del Eje Cafetero - Neurocentro, Pereira, Colombia
e-mail: daramirezgi, smolinag, amalvarezme, gdazas, cgcastellanosd {@unal.edu.co}

***Abstract*—Category 4. A machine learning based methodology is proposed to recognize a predefined set of hand gestures using depth images. For such purpose, a RGBD sensor (Microsoft kinect) is employed to track the hand position. Thus, a preprocessing stage is presented to subtract the region of interest from depth images. Moreover, a learning algorithm based on kernel methods is used to discover the relationships among samples, properly describing the studied gestures. Proposed methodology aims to obtain a representation space which allow us to identify the dynamic of hand movements. Attained results show how our approach presents a suitable performance for detecting different hand gestures. As future work, we are interested in recognize more complex human activities, in order to support the development of human-computer interface systems.**

***Keywords*— Depth sensor, human motion, kernel methods.**

## I. INTRODUCTION

Interacting with machines and environments is a task of interest in computer vision systems. In fact, being able to detect human activities using computer vision techniques allow us to suitable built human-computer interfaces, which can be useful fields like medicine, sport training, entertainment, controlling process, robotics design, among others [1], [2], [3]. Nonetheless, even when some of the current computer vision systems have provided the ability to realize an interactive human body tracking, the challenge is to develop a low-cost system, reliable in unstructured home settings, and also straightforward to use.

The most common and ancient method of human communication have been gestures. In recent years, the gestures have also employed for interacting with machines or computer-assisted systems, instead of the traditional use of devices such as keyboard, mouse, joystick, etc. The human gesture interaction has several benefits such as free movements, no wired device limitations, free hands to use other important tools. In order to track human full-body pose in real-time, camera-based motion capture systems can be used that typically require a person to wear cumbersome markers or suits [4]. There exist several limitations in the past approaches. Garg [5] uses 3D images in his method to recognize the hand gesture, but this process was complicated and inefficient. The focus should be on efficiency with the accuracy as processing time is a very critical factor in real time applications. Yang [6] analysis the hand contour to select fingertip candidates, then finds peaks in their spatial distribution and checks local variance to locate fingertips. This method was not invariant to the orientation of the hand. Then, the human gestures recognition (particularly hand gestures) is still a challenging task due to the complexity (degrees of freedom) and unpredictability of human movements.

978-1-4673-2761-9/12/ $31.00 

Recent advances have developed depth cameras that allow acquiring dense, three-dimensional scans of a scene in real-time, without the need for multi-camera systems. Such depth images are almost independent of lighting conditions and variations in visual appearance, e.g. due to clothing. In every image pixel, these cameras provide a measurement of the distance from the camera sensor to the closest object surface [4].

In this paper, we propose a machine learning based methodology to recognize a predefined set of hand gestures. For such purposes, we use a RGBD sensor (Microsoft Kinect) as the input sensor, and we present a learning algorithm based on kernel methods to discover the relationships among samples to infer the studied gestures. The goal of the proposed methodology is to obtain a representation space which allow us to identify properly the dynamic of the hand movements, which are captured by the Kinect. Attained results show how our approach presents an acceptable performance for detecting different hand gestures.

The remainder of this paper is organized as follows. In section II, proposed methodology for estimating hand position from depth images, and the kernel based framework used for recognizing hand gestures are described. In section III, the experimental conditions and the obtained results are shown. Finally, in sections IV and V, the discussion and conclusion are presented.

## II. RECOGNIZING HAND GESTURES

### A. Data Acquisition and Preprocessing

Kinect sensor has been widely used in computer vision tasks, due to the several advantages offered by the depth camera included in it [7]. The main advantages of depth sensors over traditional intensity ones are: enhance data representability by introducing a new characteristic (depth information), straightforward 3D reconstruction, capability of work in low light level scenes, and simplify the task of background subtraction. In order to take advantage of the kinect properties, a

preprocessing procedure is proposed to highlight the region of interest (hand) from depth images. In this regard, four different regions are extracted.

The former (gray region) is a dead zone configured by the user, in which the depth points are not taken into account. In the next region (yellow), the kinect sensor searches for the nearest depth point, in our case the hand. Note that, the yellow region does not contain depth data points. Then, in the green region, which is called the region of interest, the kinect establishes a working range, where it is expected to find the hand of the subject. Hence, as result an image containing only the data points that are presented in the green region are obtained. Finally, the last region (red) is also considered as a dead zone, where any object is captured by the sensor. Note that the length of the gray and the green regions can be fixed by the user. The above mentioned depth regions are summarized as in Fig. 1.

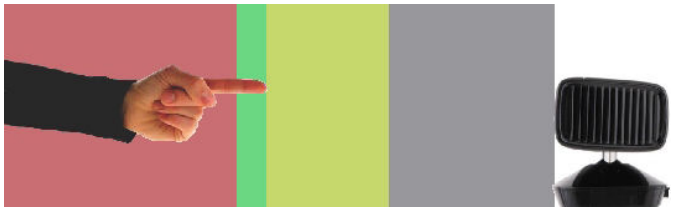


Fig. 1. Kinect sensor depth regions.

Given a depth image matrix $\mathbf{D} \in \Re^{h \times w \times 3}$, all the pixels that belong to the green region are fixed to the $\mathbf{g_r} \in \Re^{1\times 3}$ depth value. Therefore, the binary matrix $\mathbf{B} \in \Re^{h\times w}$ can be computed as in (1)

$$B_{ij} = \begin{cases} 1 & \|\mathbf{d}_{ij} - \mathbf{g}_r\| = 0 \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where $\mathbf{d}_{ij} \in \Re^{1\times 3}$ is the depth intensity vector of the $(i, j)$ pixel, with $i = 1, \ldots, h$ and $j = 1, \ldots, w$. Then, in order to identify the temporal dynamics of the hand movement, the centroid $(i_c, j_c)$ of the detected object is estimated as

$$i_c = \frac{1}{N_{g_r}} \sum_{i=1}^{h}\sum_{j=1}^{w} iB_{ij}, \quad j_c = \frac{1}{N_{g_r}} \sum_{i=1}^{h}\sum_{j=1}^{w} jB_{ij}; \tag{2}$$

being $N_{g_r}$ the number of elements in $\mathbf{B}$ that are equal to one. Regarding, let $n > 0$ the number of analyzed frames in a hand gesture, thus, the trajectory matrix $\mathbf{V} \in \Re^{n\times 2}$ is calculated with row vectors $\mathbf{v}_t = [i_c^t, j_c^t]$, being $(i_c^t, j_c^t)$ the centroid of the detected object in frame $t$, and with $t = 1, \ldots, n$.

Furthermore, a conventional lineal interpolation method is used to properly compare hand gesture trajectories with different sizes. Then, the matrix $\mathbf{S} \in \Re^{T\times 2}$ is obtained from interpolating the columns of $\mathbf{V}$, being $T > 0$ the fixed time trajectory size. Finally, a dynamic range normalization is used over each column of $\mathbf{S}$ to achieve consistency for comparing different trajectories. Therefore, the matrix $\mathbf{X} \in \Re^{T\times 2}$ is estimated as

$$X_{l1} = \frac{2\left(S_{l1} - \bar{\mathbf{s}_1}\right)}{\max\left(\mathbf{s}_1\right) - \min\left(\mathbf{s}_1\right)}, \quad S_{l2} = \frac{2\left(S_{l1} - \bar{\mathbf{s}_2}\right)}{\max\left(\mathbf{s}_2\right) - \min\left(\mathbf{s}_2\right)} \tag{3}$$

with $l = 1, \ldots, T$, and being $\mathbf{s}_1$ and $\mathbf{s}_2$ the first and second column of $\mathbf{S}$, respectively. Fig. 2 shows the proposed acquisition and preprocessing framework for predicting hand gestures trajectories from depth images.

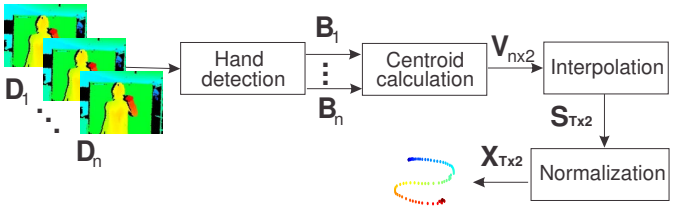



Fig. 2. Data acquisition and preprocessing scheme.

### B. Gesture recognition based on Kernel Representation

The use of kernel functions to infer relationships among samples have been widely used for machine learning procedures [8]. Here, we propose to use a kernel representation to unfold the hand gesture trajectories similarities. Recently, some machine learning approaches have shown that using multiple kernels to infer the data similarities instead of just one, can be useful to improve the data interpretability [9]. Given a pair of hand trajectory matrices $\mathbf{X}^p$ and $\mathbf{X}^p$, and assuming $Z$ kernel functions, the multiple kernel representations - MKR based methods aim to infer the combined kernel function $\kappa_\xi\left(\mathbf{X}^p, \mathbf{X}^q\right) = \sum_{z=1}^{Z} \xi_z \kappa_z\left(\mathbf{X}^p, \mathbf{X}^q\right)$, subject to $\xi_z \geq 0$, and $\sum_{i=1}^{Z} \xi_z = 1$ $(\forall \xi_z \in \Re)$. Thereby, the input data is analyzed from different information sources by means of a convex combination of basis kernels.

Using the above described MKR framework, we propose to combine two different kernels, $\kappa_a$ and $\kappa_o$, to estimate the abscissa and ordinate similarities among hand trajectories. Hence, a combined kernel function is computed as

$$\kappa\left(\mathbf{X}^p, \mathbf{X}^q\right) = \xi_a \kappa_a(\mathbf{x}_a^p, \mathbf{x}_a^q) + \xi_o \kappa_o(\mathbf{x}_o^p, \mathbf{x}_o^q), \tag{4}$$

where the vectors $\mathbf{x}_a^p$ and $\mathbf{x}_a^q$ correspond to first column of $\mathbf{X}^p$ and $\mathbf{X}^q$, respectively, and $\mathbf{x}_o^p$ and $\mathbf{x}_o^q$ contain the second ones. Moreover, $p, q = 1, \ldots, N$, being $N$ the number of given trajectories. Thus, the kernel matrix $\mathbf{K} \in \Re^{N\times N}$ can be estimated by (4).

## III. Experimental Set-up and Results

To test the performance of the proposed methodology to characterize time-series data, a hand gesture recognition database was recorded using the kinect sensor. We employ the kinect camera which gives a $640\times480$ image at 30 frames per second, using a depth resolution of $3[mm]$. The database contains 3 different hand gesture symbols performed by 2 subjects. The chosen symbols are the letters O,S and L, and each subject performs each symbol 10 times.

Data is extracted using the libfreenect software provided by OpenKinect [1], and the OpenCV C++ library is used for the image processing operations [2]. The data acquisition is made by using the region scheme explained in section II-A. The gray zone is set to approximately $1[m]$ (suggested distance by Microsoft). The length of the green region is small enough fixed for obtaining more accurate results, approximately $1[cm]$. The centroid of this region is determined by obtaining the mean of the row and column coordinates of the segmented data points by using equation 2. Moreover, to remove outlier data, we used a median filter over the abscissa and ordinate signals (each column of $\mathbf{V}$), with a fixed window of 12 samples. Each signal is scaled and interpolated with $T = 80$ (see section II-A). For each symbol recording, $n$ frames are taken according to each user symbol length. Fig. 3 shows a segmented image using the proposed acquisition and preprocessing framework, and Fig. 4 presents some preprocessed hand gesture trajectories.

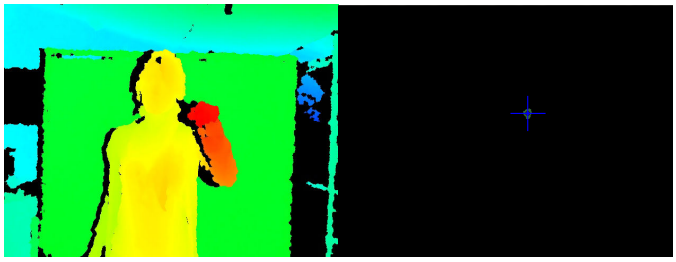


Fig. 3. Hand trajectory prediction example.

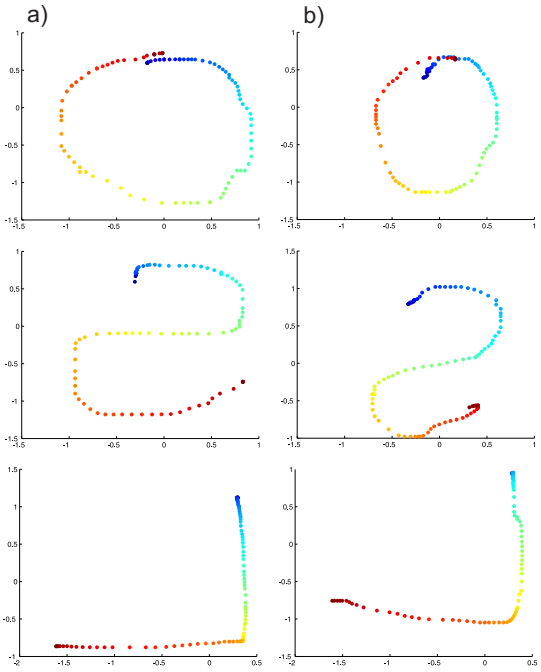



Fig. 4. Some preprocessed hand trajectories. a) Subject one. b) Subject two.

The MKR scheme explained in section II-B is used to represent, as well as possible, the obtained information. A gaussian kernel is used as basis to estimate the relationships among hand trajectories in (4). For concrete testing, the kernel band-width $\sigma$ is empirically fixed to 3. Besides, $\xi_a$ and $\xi_a$ are set to $0.5$ in (4). The resulting kernel matrix $\mathbf{K}$ of the studied dataset can be seen in Fig. 5

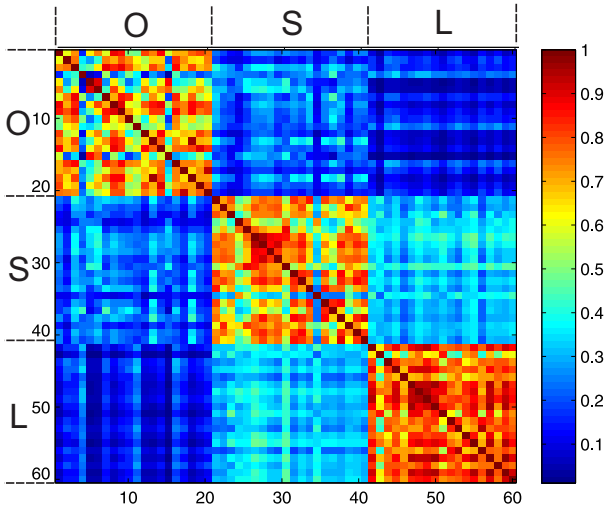



Fig. 5. Gaussian Kernel Matrix.

After that, a Kernel-Principal Component Analysis - KPCA is applied over $\mathbf{K}$ [8], obtaining a low-dimensional feature space $\mathbf{E} \in \Re^{60\times 3}$. Finally, a $k$-nearest neighbors classifier - knnc is trained over the low-dimensional space. It is important to note, that the system performance is tested using a 10-folds cross validation scheme. In Fig.6 a 3D representation of the studied data is presented.

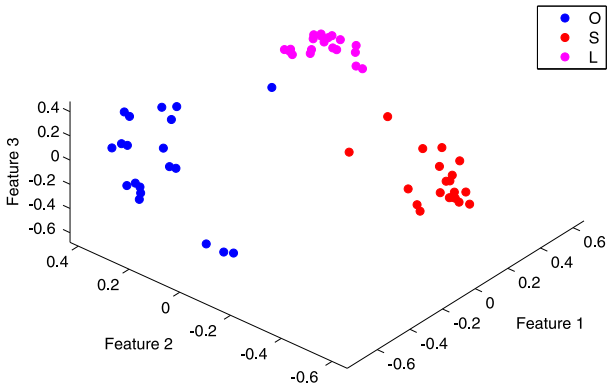



Fig. 6. Low-dimensional KPCA projection - knnc test accuracy = 100[%].

## IV. Discussion

According to the preprocessing results show in Fig. 4, it is possible to notice the capability of our approach for characterize the hand trajectory. Due to the region based methodology for inferring hand position, the estimate trajectory is smooth enough for further characterization stages.

On the other hand, the resulting similarity measure obtain by MKR using a gaussian kernel (Fig. 5) confirms that the similarity among signals from the same class is very high, with a mean similarity of $0.69$ (orange color). Again, the class that

[1] http://openkinect.org
[2] http://opencv.willowgarage.com/wiki/

exposes the highest intra-similarity corresponds to the symbol L with a mean similarity of 0.78. The classes more similar between them are the S and the L, exposing a mean similarity of 0.32.

The above given measures properties are corroborated by the estimated KPCA low-dimensional projection presented in Fig. 6. It can be notice how the MKR framework facilitates in a major way the classification process. The resulting feature space exhibits an appropriate separation among different classes. It is also noted that the L symbol (third class) shows the highest intra-similarity among all the signals.

## V. Conclusions

A machine learning based methodology for recognizing hand gestures using depth images captured by a kinect sensor was proposed. In this sense, a region based acquisition scheme using depth images was employed in order to obtain an accurate segmentation of the region of interest. Moreover, a MKR framework was proposed to combine into a single similarity matrix, the abscissa and ordinate features inferred from the centroid trajectories of hand gestures. Attained results showed that the proposed acquisition methodology obtains very accurate data points, properly identifying the dynamic of the gesture. Furthermore, the proposed MKR framework enhances the separability of the classes, facilitating further classification process. As future work, it should be interesting to include more hand gesture symbols, and also it will be useful to apply a similar MKR approach for skeleton tracking using depth images.

## Acknowledgments

This research was carried out under grants provided by a Msc. and a PhD. scholarships, and the project "ANÁLISIS DE MOVIMIENTO EN SISTEMAS DE VISIÓN POR COMPUTADOR UTILIZANDO APRENDIZAJE DE MÁQUINA", funded by Universidad Nacional de Colombia.

## References

[1] R. Urtasun and T. Darrell, "Sparse probabilistic regression for activity-independent human pose inference," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[2] T. Jaegglia, E. Koller-Meier, and L. Gool, "Learning generative models for multi-activity body pose estimation," *Int. J. Comput. Vis*, vol. 82, pp. 121–134, 2009.

[3] R. Kehl and L. Gool, "Markerless tracking of complex human motions from multiple views," *Comput. Vis. Image Underst.*, vol. 104, pp. 190–209, 2006.

[4] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab, "Human skeleton tracking from depth data using geodesic distances and optical flow," *Image and Vision Computing*, vol. 20, no. 1, pp. 217–226, 2012.

[5] P. Garg, N. Aggarwal, and S. Sofat, "Vision based hand gesture recognition," *World Academy of Science, Engineering and Technology*, vol. 49, pp. 972–977, 2009.

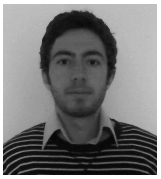
[6] D. Yang, L. Jin, J. Yin *et al.*, "An effective robust fingertip detection method for finger writing character recognition system," in *Proceedings of the Fourth International Conference On Machine Learning And Cybernetics*, 2005, pp. 4191–4196.

[7] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, vol. 2, 2011, p. 7.

[8] B. Scholkopg and A. J. Smola, *Learning with Kernels*. Cambridge, MA, USA: The MIT Press, 2002.

[9] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

**Daniela Ramírez-Giraldo** student of electronic engineering from the Universidad Nacional de Colombia sede Manizales. Her research interests are image and video processing using kinect sensor.



**Santiago Molina-Giraldo** received his undergraduate degree in electronic engineering from the Universidad Nacional de Colombia sede Manizales in 2012. Currently, he is pursuing a M.Sc at the same university. His research interests are nonlinear dimensionality reduction and kernel methods for motion analysis and signal processing.

**Andres Marino Alvarez-Mesa** received his undergraduate degree in electronic engineering with honors, and his M.Sc. engineering-industrial automation with honors from the Universidad Nacional de Colombia sede Manizales, in 2009 and 2012. Currently, he is pursuing a Ph.D at the same university. His research interests are nonlinear dimensionality reduction and kernel methods for signal processing.

**Genaro Daza-Santacoloma** received his undergraduate degree in electronic engineering in 2005, the M.Sc. degree in engineering-industrial automation with honors in 2007, and the Ph.D. degree in engineering-automatics with honors in 2010, from the Universidad Nacional de Colombia sede Manizales. Currently, he is an engineering researcher at the Instituto de Epilepsia y Parkinson del Eje Cafetero - Neurocentro, Pereira - Colombia. His research interests are feature extraction/selection and motion analysis for training pattern recognition systems.

**German Castellanos-Dominguez** received his undergraduate degree in radiotechnical systems and his Ph.D. in processing devices and systems from the Moscow Technical University of Communications and Informatics, in 1985 and 1990 respectively. Currently, he is a professor in the Department of Electrical, Electronic and Computer Engineering at the Universidad Nacional de Colombia at Manizales. He is Chairman of the GCPDS at the same university. His research interests include information and signal theory, digital signal processing and bioengineering.