A new Approach for Dynamic Gesture Recognition using Skeleton Trajectory Representation and Histograms of Cumulative Magnitudes

Edwin Escobedo; Guillermo Camara;
Department of Computer Science (DECOM)
Federal University of Ouro Preto
Ouro Preto,MG, Brazil
Email: edu.escobedo88,gcamarac@gmail.com

***Abstract*—In this paper, we present a new approach for dynamic hand gesture recognition that uses intensity, depth, and skeleton joint data captured by Kinect™ sensor. This method integrates global and local information of a dynamic gesture. First, we represent the skeleton *3D* trajectory in spherical coordinates. Then, we select the most relevant points in the hand trajectory with our proposed method for keyframe detection. After, we represent the joint movements by spatial, temporal and hand position changes information. Next, we use the direction cosines definition to describe the body positions by generating histograms of cumulative magnitudes from the depth data which were converted in a point-cloud. We evaluate our approach with different public gesture datasets and a sign language dataset created by us. Our results outperformed state-of-the-art methods and highlight the smooth and fast processing for feature extraction being able to be implemented in real time.**

***Keywords*-hand gesture recognition; spherical coordinate system; keyframes; global and local features; direction cosines; histogram of cumulative magnitudes.**

I. Introduction

Currently, hand gesture recognition is a challenging problem in computer vision and represents an active research. It is applied in sign language recognition systems, games, virtual reality, robotics, *etc* [1], [2]. Hand gesture communication involves hand and arm motion information; two approaches are commonly used to interpret them: sensor-based and vision-based [1]. Sensor-based methods use sensory gloves and motion trackers to detect hand shapes and body movement while vision-based methods use standard cameras to capture and classify hand shapes and body movements. Unfortunately, sensor-based methods require extensive calibration, they also restrict the natural hand movements and often are very expensive. Therefore, video-based methods are more used, but new problems arise: intensity images are vulnerable to illumination variations and cluttered backgrounds, hindering hand detection and tracking. However, with the recent appearance of cheap depth sensors, such as Microsoft Kinect™ [3] which provides intensity data, depth data, and skeleton joints positions, overcome these problems.

Nowadays, there are many studies focused on the analysis of hand gesture [4]. In [5], a new method was proposed with a particular emphasis on the trajectory analysis, the extraction of the relative movements of the elbow and wrist related to the hand. The Axis of Least Inertia concept is used by Geetha et al. [6], where 25 points are extracted as global features, and the Eigen distance from each fingertip to the center of the palm is proposed as local features. In [7], skeleton features are computed from the angles between two skeleton joints and the hand convex hull area represents the local features. In [8], the authors describe the joint trajectories by using spherical coordinates and describing the spatial and temporal information of the movements.

The extraction of hand shape features based on gradient value instead of standard *2D* shape features and arm movement features based on angles between each joint are proposed by Takimoto et al. [9]. A bag of visual and depth words is introduced in [10], together with a novel probability-based Dynamic Time Warping (PDTW), produce a Human Gesture Recognition pipeline. In [11] is proposed an online Sequential Extreme Machine Learning approach (OS-ELM) by using the features of upper body joints (head, hands, wrists, elbows, shoulders) and the projection of the angle position coordinates to the shoulder center and hip center, this information is used as input to K-means algorithm to generate hand features. In [12], a multimodal RGB-D data is proposed. Multiple hand features using both the body and hand masks (RGB and depth frames) are extracted together with skeletal features.

Lately, in [13] was proposed a novel method called Deep Dynamic Neural Networks (DDNN) for multimodal gesture recognition. They used deep neural nets to automatically extract the relevant information from the data and integrates two distinct feature learning methods, one for processing skeleton features and the other for RGB-D data. Besides, they used feature learning model with an HMM to incorporate temporal dependencies.

Contributions: In this paper, we propose a novel hand gesture recognition system combining local and global information obtained from depth, RGB and skeleton data captured by a Kinect™ device. In contrast to the previous works,

2377-5416/16 $31.00 © 2016 IEEE
DOI 10.1109/SIBGRAPI.2016.35

we exploit the spatial information of both arms, detecting variations of the dominant hand. Our approach also involves a new method for keyframe detection, based on the analysis of the skeleton *3D* trajectory, to identify more relevant points and reduce the processing time. The *3D* trajectory is converted to spherical coordinates by shifting the origin from the Kinect™ to the shoulder center, introducing new advantages as natural point normalization and user's translation invariance. Moreover, we use the direction cosines concept to generate Histograms of cumulative magnitudes from depth data which describe body positions for each keyframe. Our proposed method becomes independent of the repeated use of time series techniques as Hidden Markov Model (HMM) [14] or Dynamic Time Warping (DTW) [15].

Finally, we present a new Brazilian Sign Language (LIBRAS) dataset, which consists of 20 different signs used to test our proposed method. This challenging dataset is based on complex hand and trajectory configurations that make difficult the recognition process.

The remainder of this paper is organized as follows. In Section II, we describe and detail our proposed hand gesture recognition system. Experiments and Results are presented in Section III. The Conclusions and future works are presented in section IV.

II. Proposed Model

This section describes our approach for dynamic hand gesture recognition. First, we preprocess the gesture information recorded by a Kinect™ device (depth, intensity and skeleton data) to obtain the keyframes. Then, global features are computed to joint trajectories and local features from the body cloud-point. Finally, these features are used as input to our classifier. Fig. 1 shows the process of our proposed model.

A. Gesture Preprocessing

Based on [8], we use the skeleton data to generate spherical trajectories of upper joints corresponding to the head, elbow, wrist and hand of both arms (right and left).

The trajectories are converted into spherical coordinates, to avoid problems with user position changes, by assuming the shoulder center (SC) as the new coordinates origin. This coordinate conversion makes our method invariant to translation.

Definition 1 (Spherical Coordinates)**.** *The spherical coordinates of a point P in the gesture trajectory are defined by three components:*

- *The radius or radial distance r is the Euclidean distance from the origin (SC) to P, where x, y, z are the Cartesian coordinates of P.*

$$r = \sqrt{x^2 + y^2 + z^2} \tag{1}$$

- *The inclination (or polar angle) θ, is the angle between the zenith direction and the line segment SC-P.*

$$\theta = \arccos\left(\frac{z}{r}\right) \tag{2}$$

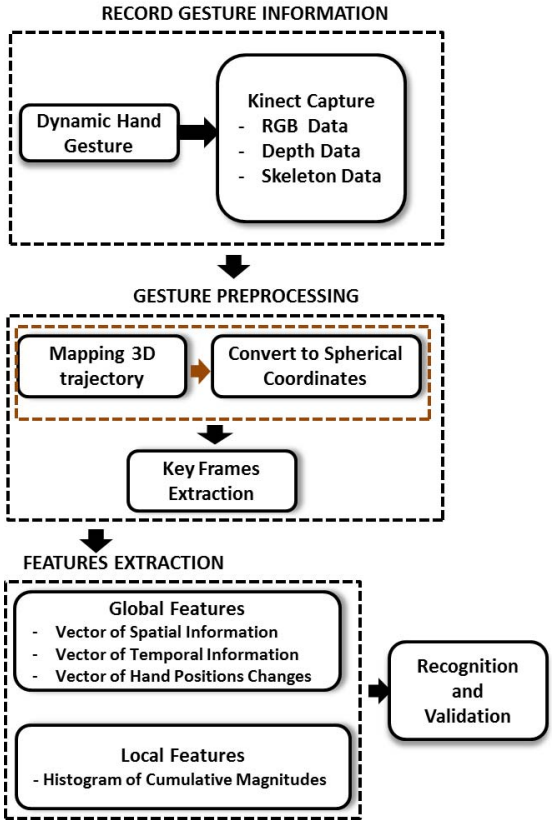

Fig. 1. Hand Gesture Recognition Proposed Model.

- *The azimuth (or azimuthal angle) φ, is the signed angle measured from the azimuth reference direction to the orthogonal projection of the line segment SC-P on the reference plane.*

$$\varphi = \arctan\left(\frac{y}{x}\right) \tag{3}$$

1) Keyframe Extraction: In dynamic gesture recognition exists the time variability problem that arises when a user makes a gesture with different speed. Work with all frames are inefficient and take a long time, so it is necessary to choose some frames. In this work, we propose a method to extract the most relevant frames in the trajectory, called keyframes.

To detect the keyframes, we convert the hand trajectories into spherical coordinates; in this work, we called the keyframe number as *sizeKF*. Then, we calculate the first derivative of the radial distance of all points to obtain the maximum and minimum difference between consecutive points, called P_{MMX}. To detect frames with the most significant difference, we select the P_{MMX} points belong to the convex hull trajectory.

Later, the P_{MMX} is used to segment the trajectory into segments, the points with the less Euclidean difference are discarded (Δr). Also, if Δr value is less than a threshold Tr, defined as the mean of the P_{MMX} points, the neighboring point is irrelevant and discarded. This process is repeated until

obtaining the *sizeKF* points.

Finally, for each keypoint, we extract its respective frame. Fig. 2 shows two examples of a hand gesture where the keyframes were detected by obtaining similar frames regardless the duration of each gesture.

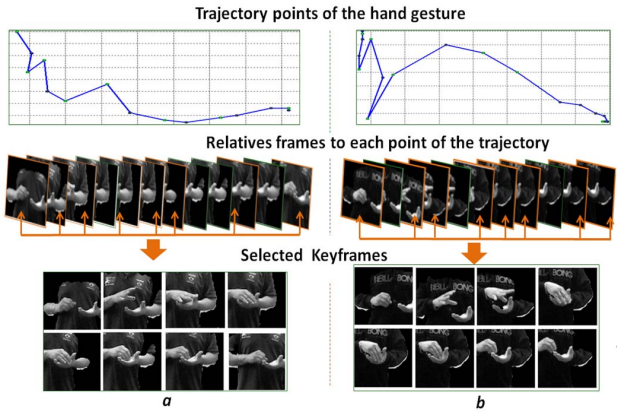

Fig. 2. Example of keyframe extraction for two videos of the same hand gesture where *sizeKF* = 8.

B. Feature Extraction

We combine global and local features to obtain a better characterization of hand gestures because there are some hand gestures with similar trajectories but different hand configuration. Otherwise, some signs have similar hand configuration but different trajectories. Thus, the combined local and global features are suitable for describing these gestures movements.

1) Global Features Extraction: The global feature extraction was designed to describe hand gestures that have structural movements, as *sign languages*. A structured movement has a well-defined trajectory while an unstructured movement can be characterized only by the final position of the gesture or by the configuration of the hand or body positions.

Trajectories of an unstructured movement may differ in its shape, but the final hand and body positions are always the same. Thus, global features (trajectories) are an important clue in structured movements

The global features are represented by three main vectors: spatial information vector V_{SI}, temporal data vector V_{TI} and hand position changes vector V_{HC}.

Firstly, we use the spherical coordinates (r, θ, φ) calculated in the new origin SC. Each keyframe contains seven spherical positions for each upper body joint: head (*h*), elbow right (*er*), wrist right (*wr*), hand right (*hr*), elbow left (***el***), wrist left (***wl***) and hand left (***hl***). These positions are concatenated to form the spatial information vector V_{SI} which is defined as:

$$V_{SI} = \bigcup_{k=1}^{k=sizeKF} \{SP_h^k, SP_{er}^k, SP_{wr}^k, SP_{hr}^k, SP_{el}^k, SP_{wl}^k, SP_{hl}^k\} \tag{4}$$

where *SP* represents a body joint spherical coordinate and $sizeKF$ is the keyframe number. The V_{SI} vector size is $3 \times 7 \times sizeKF$, where, 3 is the size of each spherical coordinate and 7 is the body joints number used.

Then, we include the relative trajectory concept, defined in [16], to incorporate the temporal relation from the wrist and elbow to the hand and generate the V_{TI} vector. First, we choose the hand trajectory T_{root} as the primary trajectory, the wrist and elbow trajectories (T_1, T_2) are considered secondaries. Each trajectory T_i is defined as:

$$L_i = \{(x_t, y_t, z_t) | t \in [1, N]\} \tag{5}$$

The relative trajectories ΔT_1 and ΔT_2 are obtained by computing the difference from T_{root} to T_1 and T_2, respectively.

$$\Delta T_i = T_{root} - T_i = \{(\Delta x_t, \Delta y_t, \Delta z_t) | t \in [1, N]\} \tag{6}$$

Thus, to represent the relative trajectory uniformly in both cases, we convert the Cartesian coordinates to the spherical coordinates to depict the orientation and distance changes between secondaries trajectories and the primary trajectory across the time.

$$\Delta T_i = \{T_{root} - T_i\} \stackrel{\triangle}{=} \{(r, \theta, \varphi) | t \in [1, N]\} \tag{7}$$

The V_{TI} vector is represented as the union of all relative trajectories in each keyframe for both hands. The vector size is: $4 \times 3 \times sizeKF$.

$$V_{TI} = \bigcup_{k=1}^{k=sizeKF} \{\Delta T_{1_{er}}^k, \Delta T_{2_{wr}}^k, \Delta T_{1_{el}}^k, \Delta T_{2_{wl}}^k\} \tag{8}$$

Finally, from the point SC, we divided the space into eight octants and recorded the octant where the hand was positioned at each keyframe. The octant number can be represented in a V_q vector of size 3 ($\log_2 8 = 3$). The V_{HC} vector is represented as the union of left and right-hand movements for each keyframe with $2 \times 3 \times N$ size. Fig. 3 shows the vectors generated for two hand gestures.

$$V_{HC} = \bigcup_{k=1}^{k=sizeKF} \{v_{q_{hr}}^k, v_{q_{hl}}^k\} \tag{9}$$

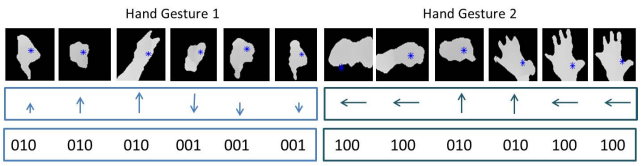

Fig. 3. Examples of hand position changes of two gestures: justice (right) and shine (left). The V_{HC} vector is different for both signs. Each binary code represents a octant position for each keyframe.

Therefore, the V_{GF} Global Features Vector is the concatenation of the three previous vectors, with a total size of $39 \times N$.

$$V_{GF} = \{V_{SI}, V_{TI}, V_{HC}\} \tag{10}$$

2) Local Features Extraction: The local features describe body positions and hand configurations when a user makes a gesture; we extract these features from depth data. Such as they are three-dimensional spatial data, they need a robust and fast descriptor to describe them. For that, we use the method proposed in [17], which was adapted to work with body positions of each keyframe. It is based on the *direction cosines concept*:

Definition 2 (Direction Cosines)**.** *The direction cosines of a* $\mathbf{V}$ *vector are the cosines of the angles between the vector and the coordinate axis. In three-dimensional Cartesian coordinates, if* $\mathbf{V}$ *is a vector in the Euclidean space,* $\mathbb{R}^3$*, then:*

$$V = v_x e_x + v_y e_y + v_z e_z \tag{11}$$

where e_x*,* e_y *and* e_z *are the standard basis in Cartesian notation and the scalars* v_x*,* v_y*,* v_z *are the scalar components of the* $\mathbf{V}$ *vector. Then, the direction cosines are:*

$$|V| = \sqrt{v_x^2 + v_y^2 + v_z^2} \tag{12}$$

$$\alpha = \cos a = \frac{V \cdot e_x}{|V|} = \frac{v_x}{|v|} \tag{13}$$

$$\beta = \cos b = \frac{V \cdot e_y}{|V|} = \frac{v_y}{|v|} \tag{14}$$

$$\gamma = \cos c = \frac{V \cdot e_z}{|V|} = \frac{v_z}{|v|} \tag{15}$$

Furthermore, $\cos a$*,* $\cos b$ *and* $\cos c$ *must meet the follow equality:*

$$\cos^2 a + \cos^2 b + \cos^2 c = 1 \tag{16}$$

Based on this concept, each keyframe is converted into a point-cloud (PC_{depth}) and we calculate the V_{depth} vector, which is the concatenation of several histograms of cumulative magnitudes (*HCM*). The steps to generate V_{depth} are the following.

- Divide the point-cloud PC_{depth} into $N_s \times N_s$ spatial subregions S_i.
- For each subregion S_i calculated the central point CP_{S_i}. Then, generate the directional VS_{P_d} vectors between CP_{S_i} and the points $P_d \in S_i$.

$$VS_{P_d} = \{CP_S - P_d | \forall P_d \in S_i\} \tag{17}$$

- For each directional VS_{P_d} vector in S_i, decompose it into its directional cosines for each Cartesian axis (α, β and γ) and calculate its magnitude $|VS_{P_d}|$. For that, use Equations 13, 14, 15 and 12, respectively. Then, we obtain the angles a, b and c by using an inverse function.
- In order to obtain spatial information of depth data, for each S_i subregion is calculated three *HCMs*: one for each axis (H_x, H_y and H_z). Each histogram is distributed into B *bins* and groups a number of angles from 0 to 180 degrees. Then, each VS_{P_d} vector casts a weighted vote for a *bin* histogram depending on the calculations of angles value in the previous steps.
  Thus, we obtain three different body projections which allow finding the highest local difference number between gestures, because the magnitude value varies depending on the body configuration.
- Finally, the V_{depth} vector is created by concatenating the cumulative histograms from each S_i subregion. We obtain a final vector which characterizes a particular body position. Fig 4, shows an example of HCM generation (H_x, H_y and H_z) for one S_1 subregion in a particular body position.

$$V_{depth} = \bigcup_{i=1}^{i=N_s x N_s} \{H_x^i, H_y^i, H_z^i\} \tag{18}$$

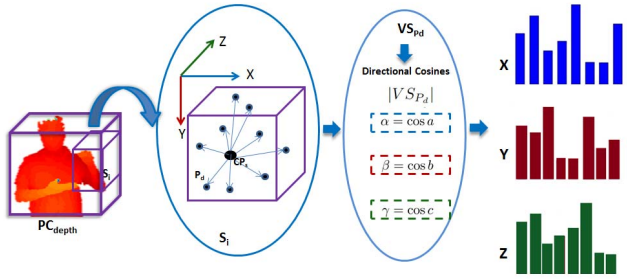

Fig. 4. Generation of Histograms of Cumulative Magnitudes for H_x, H_y and H_z in a S_1 sub region.

Therefore, the Local Feature Vector (V_{LF}) is defined as follow:

$$V_{LF} = \bigcup_{k=1}^{k=sizeKF} \{V_{depth}^k\} \tag{19}$$

Where $sizeKF$ is the keyframe number and V_{depth}^k the vector for each keyframe K.

C. Recognition and Validation

In this work, we use *Support Vector Machines* (SVM) to classify global and local features due to its excellent performance in time classification when compared to other classifiers like *KNN* [18] or *Random Forest* [19] .

Support Vector Machines (SVM) were proposed in [20], it is a useful classification method. Furthermore, SVMs have been successfully applied in many real-world problems and several areas: text categorization, handwritten digit recognition, object recognition, *etc*. An important characteristic of the SVM classifier is to allow a non-linear classification without requiring an explicit nonlinear algorithm. In kernel framework data, points may be mapped into a higher dimensional feature space, where a separating hyperplane can be found. Common kernel functions are: linear, polynomial, Radial Basis Function (RBF), *etc*.

III. Experiments

In order to measure the performance of this approach, we used different datasets to test and compare results between our method and other proposed works in the literature.

A. Datasets

We used different datasets with different structures to evaluate and validate our proposed method.

1) The LIBRAS Dataset: In this work, we present *The Brazilian Sign Language* dataset which consists of 20 different signals performed by two participants, each participant executes each sign 20 times recorded by a Kinect™ device. The dataset contains intensity, depth data and skeleton joints. This dataset is suitable for the evaluation of the robustness of our method. Our database signs presents the following characteristics:

- Signs with similar trajectories but different hands and body configurations,
- Signs with similar hand and body positions but different trajectories,
- Signs using one hand, and
- Signs using both hands,

Therefore, these characteristics make the dataset be challenging.

Fig. 5 shows the sign examples that belong to the two first type of characteristics described above.

(a) Similar trajectories but different hand configuration.

(b) Similar hand configuration but different trajectories.

Fig. 5. First type of characteristics in two sign examples in LIBRAS dataset.

2) The MSRC-12 Dataset: This is a large dataset for action/gesture recognition from 3D skeleton data recorded by Kinect sensor and proposed by Fothergill *et al.* [21]. In this paper, we used the MSRC-12 dataset to test the global feature performance. The dataset has 594 sequences and contains the performances of 12 gestures executed by 30 subjects. There are 6,244 annotated gesture instances in total. The gesture classes are divided into two groups: metaphoric gestures and iconic gestures.

3) The UTD-MHAD Dataset: The Multimodal Human Action Dataset [22] was collected by the Microsoft Kinect sensor and a wearable inertial sensor in an indoor environment. The dataset contains 27 actions performed by 8 subjects. Each subject repeated each action 4 times. Four data modalities of RGB videos, depth videos, skeleton joint positions, and the inertial sensor signals were recorded in three channels or threads. In this work, we only use the information registered by the Kinect™ sensor to measure global and local features performance.

4) CelebiGesture Dataset: Celebi *et al.* [23] provided a new gesture dataset, where only joint positions were recorded by a Microsoft Kinect, which is divided into train and test data, containing 8 gestures

5) The SDUSign Dataset: Geng *et al.* [5] presented a dataset for Chinese Sign Language Recognition. The dataset consists of 20 signs performed by four participants. Each person executes four times each sign; trajectories of right elbow, wrist and hand were collected using a Kinect™ device, recording a total of $20 \times 5 \times 4 = 400$ samples.

B. Experiments

To evaluate our proposed model, we conducted three experiments. In the first one, we assess the performance of global features (GF) while in the second one, we evaluate the Global and Local Features (GL+ LF) for gesture classification. Finally, in the third experiment, we define a protocol to work with our LIBRAS dataset. We define the following specifications for the experiments:

- We defined $sizeKF = 10$ keyframe number. This value was established as standard to all datasets after testing different N values (6, 8, 10, 12, 13, 15) in previous experiments.
- For local feature extraction process, we used the same parameters defined in [17], where N_s is equal to 5; and B (bins number) is equal to 8.
- After testing the three classifiers (KNN, RF, SVM) in the previous experiments, the SVM classifier obtained the best results, so we performed our experiment using the LIBSVM library [24] with a linear kernel. The choice of the classifier was validated using subsets of UTD-MHAD dataset; the results are shown in Table I.
- To obtain accurate results, we performed 20 times each experiment and showed the mean accuracy for each one. In other cases, we used the same protocol defined by the state-of-the-art methods, which are used to compare our results.

TABLE I
Results Obtained in Previous Experiment to Choose the Best Classifier.

Training	SVM	KNN	RF
25	**96.41**	91.86	55.78
40	**98.17**	96.13	65.59
60	**98.79**	97.65	73.16

First experiment: We evaluate our method using only global features. We used the datasets provided by Celebi *et al.* [23] and Geng *et al.* [5], and the MSCR-12 and UTD-MHAD datasets.

- For MSCR-12 dataset, we compared our results with the method proposed by Hussein *et al.* [25]. They presented a novel descriptor that uses a Covariance of 3D Joints (Cov3DJ) to encode the temporal dependency of joint locations and generate a fixed size vector that is independent of the sequence size. Moreover, we compared our results with Negin *et al.* [26], they extract spatiotemporal

features from joints in motion. Then, they used a discriminative RDF-based feature selection framework combined with a linear SVM classifier to improve the performance. Table II shows the results. We obtained an average rate of 98.58%, achieving better results than [26] and a little difference of 0.12% respect to [25]. Here we emphasize the capability of our method to be executed in real time.

- In the UTD-MHAD dataset, we compared our results by using only Kinect information, and we followed the same protocol defined by Chen *et al.* [22], we can see the results in Table III. Our method is based on global features and achieves an accuracy of 73.32 %. The results were not high as the first experiment due to the intrinsic characteristics of unstructured movements, as was explained in section II-B1, unstructured movements require local and global features to describe them better. Although this limitation, we outperform the results of the group that created the dataset, they achieved 66.10% in all information provided by the Kinect sensor.
- Finally, we used the dataset provided by Celebi *et al.* [23], they proposed a weighted Dynamic Time Warping method to boost the discrimination capability of DTW. For this, they suggested using a weighted distance in the cost computation based on the relevance of a body joint in a particular gesture class. When we compared our results by using the same protocol, we obtained similar results 97.50%; showing that our fixed vector V_{GF} can achieve as good results as methods based on time series. Geng *et al.* [5] also based their method on Dynamic Time Warping algorithm. We obtained in the same dataset used in [5] an accuracy of 82.97%. Table IV shows the results.

TABLE II
RESULTS WITH THE MSRC-12 GESTURE DATASET

Method	Accuracy
Global Features (GF)	*98.58*
Mohamed E., et al. [21]	98.70
Negin, Farhood, et al.[26]	93.00

TABLE III
RESULTS WITH THE UTD-MHAD DATASET

Method	Accuracy
Global Features (GF)	*73.32*
Local Features (LF)	*78.95*
GF + LF	***84.89***
Chen, C, et al. (only Kinect) [22]	66.10

TABLE IV
RESULTS OF USING GLOBAL FEATURES IN DIFFERENT DATASETS.

Method	Result	GF
Geng et al. [5]	69.32	82.97
Celebi, Sait, et al.[23]	97.50	97.50

Second experiment: In this experiment, our method was evaluated by using global, local and the combination of both features. Again, we used the dataset provided by Chen *et al.* [22]. The results obtained are shown in Table III. The local features that describe the body positions and hand configurations achieved 78.95% of accuracy, outperforming the results obtained by global features. However, the combination of both attributes produced the best results (84.89%). Therefore, it is possible to see the importance of combining both types of features.

Third experiment: Finally, in this experiment, our method was evaluated by using the LIBRAS dataset. First, we evaluated the performance of our method with global, local and a combination of both features. Also, we defined different training and testing sizes; our goal was to determine the robustness of our method. We defined the training sizes of $60\%, 45\%, 40\%, 35\%, 30\% and 25\%$. In Table V, we show the average accuracies for all the experiments.

- For global features, we obtained values from 89.70% to 95.63%. with a mean standard deviation value of 1.32.
- For local features, we obtained values from 89.70% to 95.63%. with a mean standard deviation value of 1.32.
- For global and local features, we obtained values from 97.58% to 99.84%. with a mean standard deviation value of 0.688.

Again, local features better describe the body positions and hand configurations. They have a better performance than global features and, the combination of both features achieved the best results.

For all experiments, as the number of keyframes ($sizeKF$) is constant, the processing time not depends of the video size, reducing significantly the processing time in general. In all our experiments, the maximum time recorded was 2.85 seconds for a video with more than 80 frames.

TABLE V
RESULTS WITH LIBRAS DATABASE FOR DIFFERENT TRAINING VALUES.

	Global		Local		Global + Local	
Training (%)	**Acc**	**SD**	**Acc**	**SD**	**Acc**	**SD**
25	89.70	1.75	97.28	1.17	97.58	0.78
30	91.67	1.31	98.02	0.44	97.86	0.89
40	92.87	1.19	98.48	0.57	98.54	0.52
45	93.13	1.24	98.66	0.61	98.66	0.66
60	95.63	1.22	99.08	0.69	99.84	0.59

As LIBRAS dataset has structured movements, it is expected that our approach achieves a higher accuracy. In this experiment, we are evaluating signs with characteristics presented in Section III-A1, *i.e.*, signs with the same trajectory and different hand configurations; signs with the same hand shape and different trajectories; signs performed with one hand; and signs carried out with both hands. In Table VI, we present the confusion matrix for global features, which has a small standard deviation. The signals *spread* and *employee* achieved the highest accuracy: 100%. The lowest recognition rates are achieved by signs *prison* similar to *truth* and just similar to *expert*. Even though they achieved the worst recognition rate, it is still high.

In Table VII, we present the confusion matrix of local features. We can observe that signals *shine, celebrate* have

the lowest average (90% and 96%), this is due to the hand configuration similarity with other signs. Finally, in Table VIII, we present the confusion matrix of the global and local features combination. The sign with the lowest recognition rate is about 91%, a high accuracy rate. These results show the efficiency of our method when we work with global and local features, besides present a greater stability.

IV. Conclusion

In this paper, we propose a method for hand gesture recognition by combining global and local features. The difference between our approach and others is the keyframe extraction; the trajectory is represented by three main vectors: spatial information vector V_{SI}, temporal data vector V_{TI} and hand position changes vector V_{HC}, which represent the global features. In addition, the depth data conversion to point-clouds allow generating histograms of cumulative magnitudes to represent body positions, the combined features (global and local) contribute to a better hand gesture description and stability. To evaluate our method, we used different hand gesture datasets such as UTD-MHAD, MSRC-12, among others. Also, we proposed a new challenging Brazilian Sign Language dataset (LIBRAS). For each experiment, our method achieved a good performance when used global, local and the combined features obtaining higher results.

Our approach provides a fast method for hand gesture recognition with a fixed-size feature vector. Our global and local features can be easily extracted, with a quick processing time. Based on the experiments, it is possible to demonstrate the robustness of our proposed approach.

As future work, we expect to improve the performance of our method by researching new global and local descriptors based on Kinect information and increasing the gestures number in the LIBRAS dataset.

Acknowledgment

References

[1] G. Murthy and R. Jadon, "A review of vision based hand gestures recognition," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 405–410, 2009.

[2] S. Mitra and T. Acharya, "Gesture recognition: A survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 3, pp. 311–324, 2007.

[3] Z. Zhang, "Microsoft kinect sensor and its effect," *MultiMedia, IEEE*, vol. 19, no. 2, pp. 4–10, 2012.

[4] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," 2013.

[5] L. Geng, X. Ma, B. Xue, H. Wu, J. Gu, and Y. Li, "Combining features for chinese sign language recognition with kinect," in *Control & Automation (ICCA), 11th IEEE International Conference on*. IEEE, 2014, pp. 1393–1398.

[6] M. Geetha, C. Manjusha, P. Unnikrishnan, and R. Harikrishnan, "A vision based dynamic gesture recognition of indian sign language on kinect based depth images," in *Emerging Trends in Communication, Control, Signal Processing & Computing Applications (C2SPCA), 2013 International Conference on*. IEEE, 2013, pp. 1–7.

[7] E. Rakun, M. Andriani, I. W. Wiprayoga, K. Danniswara, and A. Tjandra, "Combining depth image and skeleton data from kinect for recognizing words in the sign system for indonesian language (sibi [sistem isyarat bahasa indonesia])," in *Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on*. IEEE, 2013, pp. 387–392.

[8] E. Escobedo-Cardenas and G. Camara-Chavez, "A robust gesture recognition using hand local data and skeleton trajectory," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1240–1244.

[9] H. Takimoto, J. Lee, and A. Kanagawa, "A robust gesture recognition using depth data," *International Journal of Machine Learning and Computing*, vol. 3, no. 2, pp. 245–249, 2013.

[10] A. Hernández-Vela, M. Á. Bautista, X. Perez-Sala, V. Ponce-López, S. Escalera, X. Baró, O. Pujol, and C. Angulo, "Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in rgb-d," *Pattern Recognition Letters*, 2013.

[11] A. Budiman, M. I. Fanany, and C. Basaruddin, "Constructive, robust and adaptive os-elm in human action recognition," in *Industrial Automation, Information and Communications Technology (IAICT), 2014 International Conference on*. IEEE, 2014, pp. 39–45.

[12] X. Chen and M. Koskela, "Using appearance-based hand features for dynamic rgb-d gesture recognition," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 411–416.

[13] D. Wu, L. Pigou, P.-J. Kindermans, L. Nam, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," 2016.

[14] P. Blunsom, "Hidden markov models," *Lecture notes, August*, vol. 15, pp. 18–19, 2004.

[15] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[16] Z. Shao and Y. Li, "A new descriptor for multiple 3d motion trajectories recognition," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 4749–4754.

[17] E. J. Escobedo Cardenas and G. Camara Chavez, "Finger spelling recognition from depth data using direction cosines and histogram of cumulative magnitudes," in *Graphics, Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on*. IEEE, 2015, pp. 173–179.

[18] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[19] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1. IEEE, 1995, pp. 278–282.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[21] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1737–1746.

[22] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 168–172.

[23] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici, "Gesture recognition using skeleton data with weighted dynamic time warping." in *VISAPP (1)*, 2013, pp. 620–625.

[24] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[25] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations." in *IJCAI*, vol. 13, 2013, pp. 2466–2472.

[26] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, and A. Erçil, "A decision forest based feature selection framework for action recognition from rgb-depth cameras," in *Image Analysis and Recognition*. Springer, 2013, pp. 648–657.

TABLE VI
CONFUSION MATRIX OF 20 SIGNALS OF LIBRAS DATASET USING GLOBAL FEATURES.

	catch	love	shine	celebrate	compare	copy	employee	spread	expert	forget	scream	justice	just	look	person	prison	rancor	replace	television	truth
catch	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
love	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
shine	0.00	0.13	0.70	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
celebrate	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00
compare	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.03	0.00
copy	0.00	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.10
employee	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
spread	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
expert	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
forget	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.01	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00
scream	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
justice	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
just	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
look	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.01	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00
person	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00
prison	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.10
rancor	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00
replace	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.98	0.00	0.00
television	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00
truth	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.87

TABLE VII
CONFUSION MATRIX OF 20 SIGNALS OF LIBRAS DATASET USING LOCAL FEATURES.

	catch	love	shine	celebrate	compare	copy	employee	spread	expert	forget	scream	justice	just	look	person	prison	rancor	replace	television	truth
catch	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
love	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
shine	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
celebrate	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
compare	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
copy	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
employee	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
spread	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
expert	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
forget	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
scream	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
justice	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
just	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
look	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
person	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
prison	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
rancor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
replace	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
television	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
truth	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

TABLE VIII
CONFUSION MATRIX OF 20 SIGNALS OF LIBRAS DATASET USING GLOBAL AND LOCAL FEATURES.

	catch	love	shine	celebrate	compare	copy	employee	spread	expert	forget	scream	justice	just	look	person	prison	rancor	replace	television	truth
catch	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
love	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
shine	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
celebrate	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
compare	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
copy	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
employee	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
spread	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
expert	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
forget	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
scream	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
justice	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
just	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
look	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
person	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
prison	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
rancor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
replace	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
television	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
truth	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.99