



HAL
open science

3D Hand Gesture Recognition by Analysing Set-of-Joints Trajectories

Quentin de Smedt, Hazem Wannous, Jean-Philippe Vandeborre

► **To cite this version:**

Quentin de Smedt, Hazem Wannous, Jean-Philippe Vandeborre. 3D Hand Gesture Recognition by Analysing Set-of-Joints Trajectories. International Conference on Pattern Recognition (ICPR) / UHA3DS 2016 workshop, Dec 2016, Cancun, Mexico. hal-01535168

HAL Id: hal-01535168

<https://hal.science/hal-01535168v1>

Submitted on 8 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

3D Hand Gesture Recognition by Analysing Set-of-Joints Trajectories

Quentin De Smedt¹, Hazem Wannous², and Jean-Philippe Vandeborre¹

¹ Télécom Lille, CNRS, UMR 9189 - CRISStAL, F-59000 Lille, France
{desmedt,vandeborre}@telecom-lille.fr,

² Télécom Lille, Univ. Lille, CNRS, UMR 9189 - CRISStAL, F-59000 Lille, France
hazem.wannous@univ-lille1.fr

Abstract. Hand gesture recognition is recently becoming one of the most attractive field of research in Pattern Recognition. In this paper, a skeleton-based approach is proposed for 3D hand gesture recognition. Specifically, we consider the sequential data of hand geometric configuration to capture the hand shape variation, and explore the temporal character of hand motion. 3D Hand gesture are represented as a set of relevant spatiotemporal motion trajectories of hand-parts in an Euclidean space. Trajectories are then interpreted as elements lying on Riemannian manifold of shape space to capture their shape variations and achieve gesture recognition using a linear SVM classifier.

The proposed approach is evaluated on a challenging hand gesture dataset containing 14 gestures, performed by 20 participants performing the same gesture with two different numbers of fingers. Experimental results show that our skeleton-based approach consistently achieves superior performance over a depth-based approach.

Keywords: Gesture recognition, Riemannian Manifold, Hand skeleton, Depth image

1 Introduction

Among other human body parts, the hand is the most effective interaction tool in Human-Computer Interaction (HCI) applications. The essential types of sensors used to capture the hand gesture are multi-touch screen sensors, motion capture magnetic sensors and vision based sensors. Being less cumbersome and uncomfortable, without contact with the user, vision based sensors have some advantages compared to the other sensors.

Recently, thanks to the advance in information technologies, effective and inexpensive depth sensors, like Microsoft Kinect or Intel RealSense, are increasingly used in the domain of computer vision. The development of these sensors has brought new opportunities for the hand gesture recognition area. Compared to 2D cameras, these sensors are more robust to common low-level issues in RGB imagery like background subtraction and light variation.

3D Hand gesture recognition is becoming a central key for different application domains such as virtual game controller, sign language recognition, daily

assistance, palm verification, gaming and the emerging human-robot interaction. Consequently, the improvements in hand gesture interpretation can benefit a wide area of research domains.

In this paper, we present a novel hand gesture recognition solution, where the main advantage of our approach is the use of trajectories of a set of geometric configurations, extracted from 3D hand joint coordinates. These later can be returned by the Intel RealSense depth camera . Our ultimate goal is to develop an approach to recognize a variety challenging of dynamic hand gestures without needs to overly complex design of feature extraction.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 describes feature extraction from hand-part motion trajectory and the proposed learning algorithm on a Riemannian manifold. Section 4 introduces our new dynamic hand gesture dataset and discusses the experimental results of proposed approach. Section 5 concludes the paper.

2 Related Work

The interest in hand gesture recognition has increased in the past few years, motivated by the recent advances in acquisition technologies and the increase of societal needs in terms of multimedia and Human-Computer Interaction (HCI) applications. The hand gesture recognition is a challenging topic due to the hand motion’s complexity and diversity. We reviewed in this section the earlier works proposed for 3D hand gesture recognition, which can be divided into two approaches: static and dynamic.

Static approaches Mostly, 3D depth information can be used here to extract hand silhouettes or simply hand area and the focus will be on the feature extraction from segmented hand region. Features are usually based a global information as proposed by Kuznetsova et al. [1], where an ensemble of histograms is computed on random points in the hand point cloud. Other local descriptors are expressed as the distribution of points in the divided hand region into cells [2]. Instead of using the distribution of points in the region of the hand, Ren et al. [3] represented the hand shape as time-series curve and use distance metric called Finger-Earth Mover Distance to distinguish hand gestures from collected dataset of 10 different gestures. The time-series curve representation is also used by Cheng et al. [4], to generate a fingerlet ensemble representing the hand gesture. Sign language recognition with hand gestures has been widely investigated. Pugeault and Bowden [5] proposed a method using Gabor filter for hand shape representation and a Random Forest for gesture classification. They applied their method on a collected *ASL Finger Spelling* dataset, containing 48.000 samples of RGB-D images labelled following 24 static gestures of the American Sign Language. Recently, Dong et al. [6] outperformed the previous results on this database by going more deeply into the hand representation. They proposed a hierarchical mode-seeking method to localize hand joint positions under kinematic constraints, segmenting the hand region into 11 natural

parts (one for the palm and two for each finger). A Random Forest classifier is then built to recognize ASL signs using a feature vector of joint angles. Finally, Marin et al. [7] released a publicly database of 10 static hand gestures giving the depth image from a *Kinect* but also information about the hand using the hand pose recognition device *LeapMotion*. They also proposed a classification algorithm using fingertips distances, angles and elevations and also curvature and correlation features on the depth map.

Dynamic approaches Instead of relying on hand description on a single image, we exploit here the temporal character of hand motion, by considering the gesture as a sequence of hand shape. Kurakin et al. [8] presented the MSR-3D hand gesture database containing 12 dynamic *American Sign Language*. They recorded 360 sequences of depth images from a *Kinect*. Their recognition algorithm is based on a hand depth cell occupancy and a silhouette descriptor. They used an action graph to represent the dynamic part of the gesture. Using a histogram of 3D facets to encode 3D hand shape information from depth maps, Zhang et al. [9] outperformed the last results using a dynamic programming-based temporal segmentation. One of the track of the *Chalearn 2014* [10] was to use a multimodal database of 4,000 gestures drawn from a vocabulary of 20 dynamic Italian sign gesture categories. They provided sequences of depth images of the whole human body and body skeletons. On this database, Monnier et al. [11] employ both body skeleton-based and Histogram of Oriented Gradients (HOG) features on the depth around the hand to perform a gesture classification using a boosted cascade classifier. Recently, the use of deep learning has change the paradigm of many research field in computer vision. Recognition algorithms using specific neural network, like Convolutional Neural Network (ConvNet), obtain previously unattainable performance in many research field. Still on the *Chalearn 2014* [10], Neverova et al. [12] use stacked ConvNets on raw intensity and depth sequences around the hand and neural network on body skeletons. Recently, Ohn-Bar and Trivedi [13] made a publicly available database of 19 gestures performed in a car using the *Kinect* camera. The initial resolution obtained by such a sensor is 640x480 and the final region of interest is 115x250. Moreover, at some distance from the camera, the resulting depth is very noisy, making the challenge of gesture recognition tougher. They reported the accuracy results using several known features (HOG, HOG3D, HOG²) for gesture recognition.

This brief review highlights mainly the lack of publicly available skeleton-based dynamic hand gesture datasets for benchmarking, and progress to do to improve hand pose estimation using recent depth captures and approaches. In this context, very recently we have proposed a new public 3D hand gesture dataset [14], challenging in terms of gesture types and performing manner.

3 Approach

To express temporal character of hand gesture, our approach describes a motion trajectory of gesture as an element lying on a Riemannian manifold. A mani-

fold representation of gesture sequence can provide discriminating structure for dynamic gesture recognition. This leads to manifold-based analysis, which has been successfully used in many computer vision applications such as visual tracking [15] and action recognition in 2D video [16, 17] or 3D video [18, 19]. Such a representation allows to develop an effective machine learning process using some statistical properties of the manifold.

Generally, the motion trajectory-based approaches considers the sequential data of the movement and explores its temporal aspect of motion. To formulate trajectory features, hand positions, velocity, acceleration are the mostly used in the literature.

Besides, many approaches have been recently proposed to exploit a Riemannian manifold representation of feature data extracted from human pose and motion. Slama et al. [20] represent the 3D human body shape as a collection of geodesic 3D curves extracted from the body surface. The curves in \mathbb{R}^3 space are viewed as a point in the shape space of open curves, where an elastic metric can be calculated to estimate the similarity. Devanne et al. [19] extend this idea to represent a spatiotemporal motion characterized by full human skeleton trajectory. These motion trajectories are extracted from 3D joints and expressed in \mathbb{R}^{60} . The action recognition is performed using a K-NN classifier using geodesic distances obtained an open curve shape space.

Inspired by these approaches, we model in this work the dynamics of hand-part motion trajectories using shape analysis on a such manifold. The dynamic of the hand is then modeled as trajectories of nine n-tuples of joints according to the hand geometric shape and its cinematic.

3.1 Gesture space trajectories

The Software Development Kit (SDK) released for *Intel RealSense F200* provides a full 3D skeleton of the hand corresponding to 22 joints. The real-world 3D position of each joint J is represented by three coordinates expressed in the camera reference system $J_i(t) = (x_i(t), y_i(t), z_i(t))$ for each frame t of a sequence.

We choose to represent the hand pose as a collection of nine 5-tuples of joints according to the hand geometric structure, shown in Figure 1. The dynamic of the gesture is then analyzed and the temporal evolution of these tuples expressed as feature matrices. For a gesture sequence composed of N frames, columns in these matrices are regarded as a sample of a continuous trajectory in R^{15N} .

Practically, for each gesture sequences, motion trajectories are extracted as time-series of a set of tuples of joints according to the hand physical structure. Tuples are formed as a combinations of 3D finger joints and the relevant ones are selected. Each trajectory is then interpreted as a point on a Riemannian manifold, to allow in the first time the analysis of its shape and then to perform gesture recognition using the geometry of this manifold.

On the Riemannian manifold, each trajectory is characterized by a curve in \mathbb{R}^n and can be represented as $\beta : I \rightarrow \mathbb{R}^n$, with $I = [0, 1]$. The shape of β is mathematically represented through the square-root-velocity function (SRVF) [21], denoted by $q(t) = \dot{\beta}(t)/\|\dot{\beta}(t)\|$.

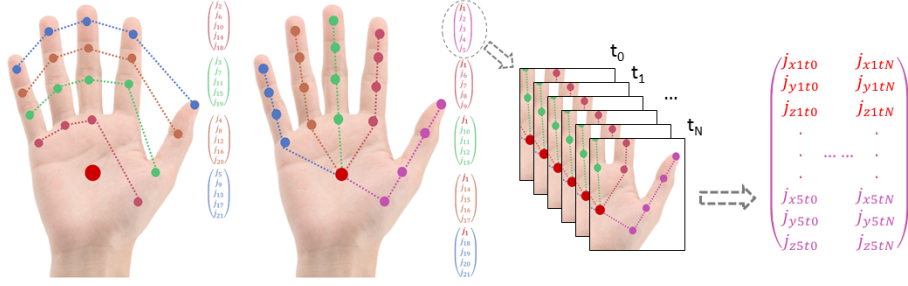


Fig. 1. An illustration of hand-part trajectory configuration. A 5-tuple are constructed using the thumb joints in \mathbb{R}^3 . The temporal evolution of these tuples over N frames is concatenated, resulting in feature matrices of size $15 * N$, where columns are regarded as a sample of a continuous trajectory in R^N .

The set of all unit-length curves in \mathbb{R}^n is given by $\mathcal{C} = \{q : I \rightarrow \mathbb{R}^n, \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^n)$, with $\|\cdot\|$ indicating the \mathbb{L}^2 norm. With the \mathbb{L}^2 metric on its tangent space, \mathcal{C} becomes a Riemannian manifold.

Since elements of \mathcal{C} have a unit \mathbb{L}^2 norm, \mathcal{C} is a hyper-sphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^n)$ and the distance between two elements q_1 and q_2 is defined as $d_{\mathcal{C}}(q_1, q_2) = \cos^{-1}(\langle q_1, q_2 \rangle)$. Such distance represents the similarity between the shape of two curves in \mathbb{R}^n . Basically, it quantifies the amount of deformation between two shapes.

3.2 Notion of average gesture

The Karcher mean is the intrinsic means which are defined as the point with minimal sum-of-squared distances to the data:

$$\mu = \arg \min \sum_{i=1}^n d_{\mathcal{C}}(\mu, q_i)^2. \quad (1)$$

Where $d_{\mathcal{C}}$ is the similarity metric between two elements on the manifold, knowing that a Riemannian metric on a manifold is an inner product in the Euclidean space that locally approximates the manifold at each point, which called tangent space. The exponential map ($T_{\mu}\mathcal{S} \rightarrow \mathcal{S}$) and inverse mapping, known as logarithmic map, ($\mathcal{S} \rightarrow T_{\mu}\mathcal{S}$) are a pair of operators mapping between the manifold \mathcal{S} and the tangent space at μ :

$$\exp_{\mu}(v) = \exp(\log(\mu) + v) = Q; . \quad (2)$$

$$\log_{\mu}(Q) = \log(Q) - \log(\mu) = v; . \quad (3)$$

where $v \in T_{\mu}\mathcal{S}$ is the tangent vector whose projected point on the manifold is Q .

For each gesture class k of the training data, a base element μ_k is computed as the center of mass of all elements of the class k , where an inner product can be defined in the associated tangent space. Karcher mean algorithm can be employed here for the computation of the center of mass [22]. For sample gesture elements belong to a same class, we can assume that they lie in a small neighborhood, where the exponential map can be defined correctly.

3.3 Learning on the manifold

The basic idea is to apply one-versus-all mapping given in Eq. 3 under the log-Euclidean metric, that maps manifold elements onto tangent spaces associated to all gesture classes. This incorporates an implicitly release properties in relation to all class clusters in the training data, and constitutes an ordered list of mapped tangent vectors representing the proximity of a gesture to all existing classes. The final representation of each gesture element is obtained by concatenating local coordinates in tangent spaces associated with different classes, and provides the input of a classifier algorithm. Slama et al. [19] proved that that training a linear SVM classifier over a representation of points provided by Grassmann manifold is more appropriate than the use of SVM with classical Kernel, like RBF, on original points on the manifold.

In our approach, we train a linear SVM on the obtained representation of points originally correspond to a set of joint trajectory in gesture space. We use a multi-class SVM classifier from the LibSVM library [23], where a penalty parameter C is tuned using a 5-fold cross-validation on the training dataset.

4 Experimental results

First, we present briefly the depth and skeleton-based dynamic hand gesture (DHG) dataset . Secondly, we analyze the impact of several sets of joints on the classification process and then conduct several tests of our approach presented in Section refsec:approach using respectively 14 and 28 different gesture classes. We finally compare our approach to recent skeleton-based approach presented in [14], and also to other existing depth-based approaches like HON4D [24] and HOG^2 [25].

4.1 Dynamic Hand Gesture dataset

In order to study these challenges, a RGB-D dataset (DHG ³) was collected. It provides sequences of hand skeleton in addition to the depth image. Such a dataset facilitates the analysis of hand gestures and open new scientific axes to consider.

The DHG dataset contains 14 gestures performed in two ways: using one finger and the whole hand (an example is shown in Figure 2). Each gesture is

³ <http://www-rech.telecom-lille.fr/DHGdataset>

performed 5 times by 20 participants in 2 ways, resulting in 2800 sequences. To increasing the challenges facing the recognition algorithm, we divide the gesture sequences in two categories: coarse-grained (*Tap* "T", *Swipe Right* "S-R", *Swipe Left* "S-L", *Swipe Up* "S-U", *Swipe Down* "S-D", *Swipe X* "S-X", *Swipe V* "S-V", *Swipe +* "S+", *Shake* "Sh") and fine-grained gestures (*Grab* "G", *Expand* "E", *Pinch* "P", *Rotation CW* "R-CW", *Rotation CCW* "R-CCW").

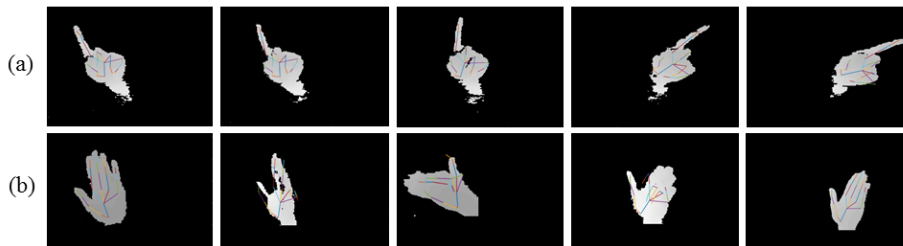


Fig. 2. Dynamic *Swipe* gesture performed (a) with one finger and (b) with the whole hand.

Sequences are labelled following their gesture, the number of fingers used, the performer and the trial. Each frame contains a depth image, the coordinates of 22 joints both in the 2D depth image space and in the 3D world space forming a full hand skeleton. The *Intel RealSense* short range depth camera is used to collect our dataset. The depth images and hand skeletons were captured at 30 frames per second, with a 640x480 resolution of the depth image. The length of sample gesture ranges goes from 20 to 50 frames. The Software Development Kit (SDK) released for *Intel RealSense* F200 provides a full 3D skeleton of the hand corresponding to 22 joints labelled as shown in Figure 3. However, the sensor stills has trouble to properly recognize the skeleton when the hand is closed, perpendicular to the camera, without a well initialization or when the user performs a quick gesture. The reader is referred to [14] for a more details about DHG dataset.

4.2 Gesture recognition

We evaluate our approach for hand gesture recognition in two cases, by considering 14 and 28 classes of gestures, thus taking account of the number of fingers used. Then, a comparison analysis on depth-vs-skeleton based descriptors is presented. For all following experiments, we use a *leave-one-subject-out cross-validation* protocol to evaluate our approach.

Choices of relevant set of joints The hand pose is represented as a collection of nine 5-tuple of joints. Five of these 5-tuples are constructed with the 4 joints

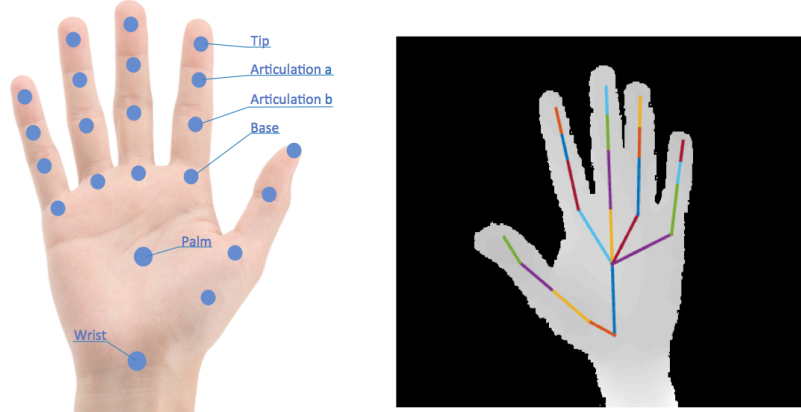


Fig. 3. Depth and hand skeleton of the DHG dataset. The 22 joints of the hand skeleton returned by the *Intel RealSense* camera. The joints include: 1 for the center of the palm, 1 for the position of the wrist and 4 joints for each finger represent the tip, the 2 articulations and the base. All joints are represented in \mathbb{R}^3 .

of each finger in addition to the palm one. The 4 remaining concern the 5 tips, the 5 first articulations, the 5 second articulations and the 5 bases. Notice that the points of each tuple follows the same order (see Table 1).

Before all tests, we analyze the redundancy and the impact of each tuple (set of joints) on the hand gesture recognition process. To do this, our approach is repeated for each set of joints separately. The Table 1 presents the obtained results of our approach by considering the trajectory of each set of joints independently and also combining them by simple concatenation.

# of the set	Joints in the set	Accuracy
1	4 Thumb's joint + palm joint	0.69
2	4 Index's joint + palm joint	0.75
3	4 Middle's joint + palm joint	0.71
4	4 Ring's joint + palm joint	0.69
5	4 Pinky's joint + palm joint	0.70
6	Tip of every fingers	0.79
7	First articulation of every fingers	0.78
8	Second articulation of every fingers	0.76
9	Base articulation of every fingers	0.69
10	combination of all sets	0.828
11	combination of set 1+2+7	0.825

Table 1. The list of all tuples configurations from five joints, tested in our approach.

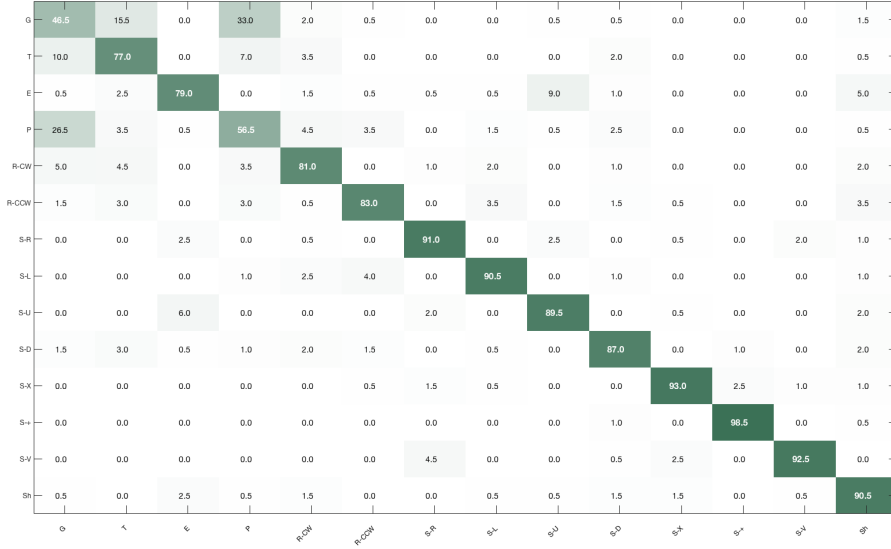


Fig. 4. The confusion matrix obtained by the proposed approach on DHG for 14 gesture classes.

According to the gesture motion, tuple trajectories exhibit different performance and some ones are more efficient to describe the dynamic of gesture shape. We first notice that best result is obtained using the set #6 as it gives the most information about the general hand shape. If their result are a bit lower, the set #7 and #8 as the same form that the later. They also carry the general hand shape leading to redundant information.

For hand fine-grained gesture as *Expands* or *Rotations*, the set #1 and #2 are primordial as the thumb and the index are the most relevant and used fingers for them.

14 gestures classes Experimental tests performed on all combinations show that adding more set of joints increase considerably the size of our final descriptor, making the feature extraction and classification step much longer. However, using only 3 sets, which divide the size of our descriptor by 3, decreases the accuracy only by 0.3%. The confusion matrix obtained by our approach using the selected combination is illustrated in Figure 4.

The accuracy of our approach when considering 14 gesture classes is 82.5% (more than 90% for only the coarse gestures and 68% for the fine ones).

28 gestures classes To increase the challenge of the gesture recognition, we take into consideration not only the gesture but also the way it have been performed, i.e. with one finger or the whole hand. In this case, we consider the

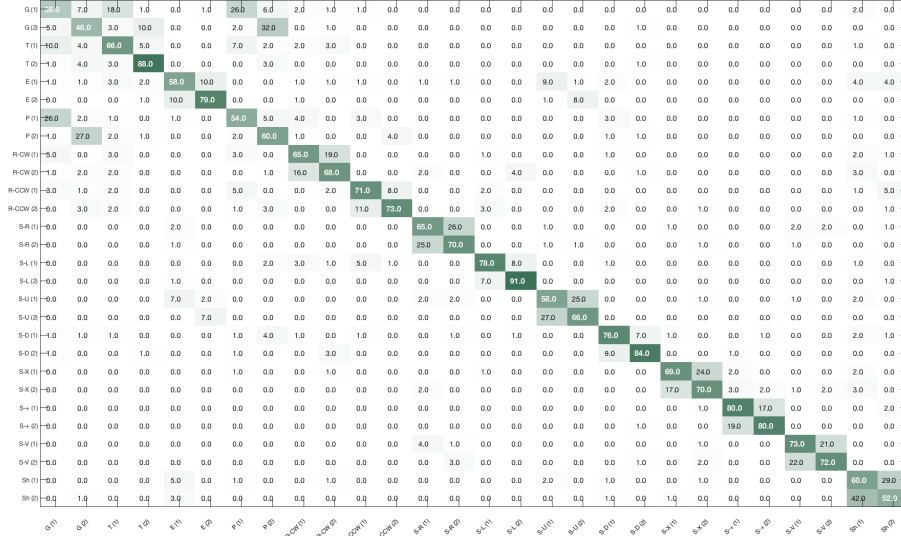


Fig. 5. The confusion matrix obtained by the proposed approach on DHG for 28 gesture classes.

sequences of the DHG dataset as belonging to 28 classes. The obtained confusion matrix is shown in Figure 5.

To analyze these results, let see the example of *R-CW* gesture; we obtained an accuracy of 81% when considering 14 classes. However, a loss of accuracy of more than 14% is observed when considering 28 classes (65% with one finger and 68% with the whole hand). More loss of accuracy can be observed for *S-R*, *S-U* and *S-H* gestures.

The decreased accuracy of 68.11% obtained in this case due to the intra-gestures confusion. This can be explained by the limitation of part-hand trajectory to discriminate gesture motions when performing with one or more fingers.

Comparing to existing approaches We compare our approach to a recent skeleton-based approach presented in [14], and also to other existing depth-based approaches like HON4D [24] and *HOG²* [25].

In the absence of any other dynamic gesture dataset providing hand joints, we propose to compare our approach to a recent skeleton-based approach presented in [14], and also to other existing depth-based approaches: HON4D [24] and *HOG²* [25]. As shown in Table 2, skeleton-based approaches outperform overly depth-based approaches. We note here that the two last descriptors are computed on cropped hand zone motion on the image, which can be considered as drawback of depth based approach and without it the rate are very low.

We can also observed that our descriptor achieves a comparable accuracy to the SoCJ+HoHD+HoWR combined descriptors for 14 gesture classes. However,

Approach	14 gesture classes	28 gesture classes
HON4D [24]	75.53%	74.03%
HOG^2 [25]	80.85%	76.53%
SoCJ+HoHD+HoWR [14]	83.00%	79.14%
Our approach	82.50%	68.11%

Table 2. Accuracies of approach to recognize 14 and 28 gesture classes on the DHG dataset, compared to a skeleton-based approach (SoCJ+HoHD+HoWR) [14], and two depth-based approaches: HON4D [24] and HOG^2 [25].

it fails for 28 gesture classes due to the the limitation of that hand-part trajectories to distinguish between gestures performed with one and more fingers. Indeed, the intra-gesture confusion when passing from 14 to 28 gesture classes is more important in this work than that presented recently [14], as the later include more complex hand shape description. Part of this confusion may be related to the failing of the hand pose estimation algorithm used by the camera. Using more efficient pose estimator or improve the current skeleton could lead to a better accuracy of our gesture recognition method.

5 Conclusion

An effective approach with compact representation is proposed to recognize dynamic hand gesture as time-series of 3D hand joints returned by the Intel RealSense depth camera. We extracted relevant hand-part trajectories to model their dynamics by exploiting the geometric properties of Riemannian manifold to express the motion character of gesture. Finally, recognition is achieved by introducing an efficient learning process on manifold elements. Experimental tests of our proposed approach on challenging dataset, shows promising results with high accuracies.

As future work, a combination of skeleton and depth based features could be investigated, to face expected limitations and provide more informative description for more challenging contexts like hand-object interaction.

References

1. Kuznetsova, A., Leal-Taix, L., Rosenhahn, B.: Real-time sign language recognition using a consumer depth camera. In: IEEE International Conference on Computer Vision Workshops (ICCVW). (Dec 2013) 83–90
2. Wang, H., Wang, Q., Chen, X.: Hand posture recognition from disparity cost map. In: ACCV (2). Volume 7725 of Lecture Notes in Computer Science., Springer (2012) 722–733
3. Ren, Z., Yuan, J., Zhang, Z.: Robust hand gesture recognition based on finger-earth mover’s distance with a commodity depth camera. In: ACM International Conference on Multimedia. MM ’11, New York, NY, USA, ACM (2011) 1093–1096

4. Cheng, H., Dai, Z., Liu, Z.: Image-to-class dynamic time warping for 3d hand gesture recognition. In: 2013 IEEE International Conference on Multimedia and Expo (ICME). (July 2013) 1–6
5. Pugeault, N., Bowden, R.: Spelling it out: Real-time asl fingerspelling recognition. In: IEEE computer Vision Workshops (ICCV Workshops). (Nov 2011) 1114–1119
6. Dong, C., Leu, M.C., Yin, Z.: American sign language alphabet recognition using microsoft kinect. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2015) 44–52
7. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with leap motion and kinect devices. In: IEEE International Conference on Image Processing (ICIP). (2014) 1565–1569
8. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: 20th European Signal Processing Conference (EUSIPCO). (Aug 2012) 1975–1979
9. Zhang, C., Yang, X., Tian, Y.: Histogram of 3d facets: A characteristic descriptor for hand gesture recognition. In: IEEE Int. Conference and Workshops on Automatic Face and Gesture Recognition (FG),. (April 2013) 1–8
10. Escalera, S., Baró, X., Gonzalez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. In: Computer Vision-ECCV 2014 Workshops, Springer (2014) 459–473
11. Monnier, C., German, S., Ost, A.: A multi-scale boosted detector for efficient and robust gesture recognition. In: Computer Vision-ECCV 2014 Workshops, Springer (2014) 491–502
12. Neverova, N., Wolf, C., Taylor, G.W., Nebout, F.: ModDrop: adaptive multimodal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (April 2016)
13. Ohn-Bar, E., Trivedi, M.M.: Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. IEEE Trans. on Intelligent Transportation Systems **15**(6) (2014) 2368–2377
14. De Smedt, Q., Wannous, H., Vandeborre, J.P.: Skeleton-based dynamic hand gesture recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (June 2016)
15. Lee, C.S., Elgammal, A.M.: Modeling view and posture manifolds for tracking. In: IEEE International Conference on Computer Vision. (2007) 1–8
16. Lui, Y.M.: Advances in matrix manifolds for computer vision. In: Image and Vision Computing. Volume 30. (2012) 380 – 388
17. Harandi, M.T., Sanderson, C., Shirazi, S., Lovell, B.C.: Kernel analysis on grassmann manifolds for action recognition. In: Pattern Recognition Letters. Volume 34. (2013) 1906 – 1915
18. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. CVPR '14, Washington, DC, USA, IEEE Computer Society (2014) 588–595
19. Slama, R., Wannous, H., Daoudi, M., Srivastava, A.: Accurate 3d action recognition using learning on the grassmann manifold. Pattern Recognition **48**(2) (2015) 556 – 567
20. Slama, R., Wannous, H., Daoudi, M.: 3d human motion analysis framework for shape similarity and retrieval. Image Vision Comput. **32**(2) (February 2014) 131–154

21. Joshi, S.H., Klassen, E., Srivastava, A., Jermyn, I.: A novel representation for Riemannian analysis of elastic curves in R^n . In: Proc IEEE Int. Conf. on Computer Vision and Pattern Recognition, Minneapolis, MN, USA (June 2007) 1–7
22. Karcher, H.: Riemannian center of mass and mollifier smoothing. *Comm. on Pure and Applied Math.* **30** (1977) 509–541
23. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**(3) (May 2011) 27:1–27:27
24. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA (2013) 716–723
25. Ohn-Bar, E., Trivedi, M.M.: Joint angles similarities and HOG2 for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013, Portland, OR, USA, June 23-28, 2013. (2013) 465–470