SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset

Quentin De Smedt, Hazem Wannous, Jean-Philippe Vandeborre, Joris Guerry, Bertrand Le Saux, David Filliat

Eurographics Workshop on 3D Object Retrieval (2017)
I. Pratikakis, F. Dupont, and M. Ovsjanikov (Editors)

SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset

Q. De Smedt[1], H. Wannous[2], J.-P. Vandeborre[1], J. Guerry[3], B. Le Saux[3], D. Filliat[4]

[1]IMT Lille Douai, Univ. Lille, CNRS, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France.
e-mail: {quentin.desmedt,jean-philippe.vandeborre}@imt-lille-douai.fr
[2]Univ. Lille, CNRS, Centrale Lille, IMT Lille Douai, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France. e-mail: hazem.wannous@univ-lille1.fr
[3]ONERA The French Aerospace Lab, DTIM, F-91761 Palaiseau, France
e-mail: joris.guerry@onera.fr
[4]ENSTA Paristech, U2IS, F-91762 Palaiseau, France
e-mail: david.filliat@ensta-paristech.fr

Abstract

Hand gesture recognition is recently becoming one of the most attractive field of research in pattern recognition. The objective of this track is to evaluate the performance of recent recognition approaches using a challenging hand gesture dataset containing 14 gestures, performed by 28 participants executing the same gesture with two different numbers of fingers. Two research groups have participated to this track, the accuracy of their recognition algorithms have been evaluated and compared to three other state-of-the-art approaches.

Categories and Subject Descriptors (according to ACM CCS): I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Shape

1. Introduction

Among human body parts, the hand is an effective and intuitive interaction tool in most Human-Computer Interaction (HCI) applications. Using the hand gesture as a HCI modality introduces intuitive and easy-to-use interfaces for a wide range of applications in virtual and augmented reality systems, offer helps for hearing-impaired support and provides solutions for all environments using touch-less interfaces. However, the hand is an object with a complex topology and has many possibilities to perform the same gesture. For example, Feix *et al.* [FPS*09] summarize the grasping taxonomies and found 17 different hand shapes to perform a grasp. Other gestures, as *swipes*, which are more defined by the hand motion than its shape, are already commonly used in tactile HCI. This difference between useful gestures have to be taken into account in a hand gesture recognition algorithm. 3D hand gesture recognition has been an active research field for the past 20 years, where various different approaches have been proposed.

Over the few past years, advances in commercial 3D depth sensors have substantially promoted the research of hand gesture detection and recognition. According to recent state-of-the-art, the approaches focus on 3D hand gesture recognition can be gathered into two main categories: **static** and **dynamic** hand gesture recognition using **depth images** and/or hand **skeletal data**. In most of the **static** approaches, 3D depth information is used to extract hand silhouettes or simply hand areas in order to extract features from segmented hand region [KLTR13, RYMZ13, CDL13, WLC15, PB11]. However, **dynamic** methods exploit the temporal aspect of hand motion, by considering the gesture as a sequence of hand shapes [KZL12, ZYT13, MGO14]. Recently, the use of deep learning has changed the paradigm of many research fields in computer vision. Recognition algorithms using specific neural network — like Convolutional Neural Network (CNN) – obtained previously unattainable performance [EBG*14, NWTN16].

In this track, we present a new 3D dynamic hand gesture dataset which provides sequences of hand skeletal data in addition to the depth images. Such a dataset will facilitate the analysis of hand gestures and open new scientific axes to consider. Two methods have been registered for this track and will be presented in the next sections. These two methods are also compared to three state-of-the-art methods in terms of results.

2. Dataset

The dataset contains sequences of 14 hand gestures performed in two ways: using one finger and the whole hand. Each gesture is performed between 1 and 10 times by 28 participants in 2 ways, resulting in 2800 sequences. All participants are right handed. Sequences are labeled following their gesture, the number of fingers

DOI: 10.2312/3dor.20171049

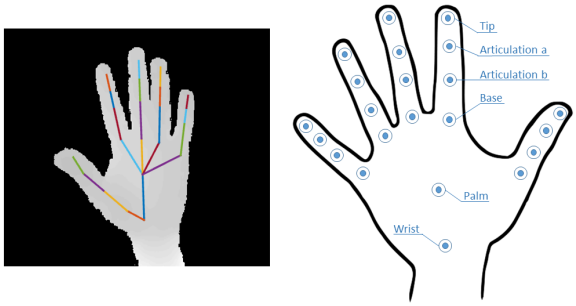

Figure 1: *Depth and hand skeleton of the dataset. The hand skeleton returned by the* Intel RealSense *camera contains 22 joints. The joints include: 1 for the center of the palm, 1 for the position of the wrist and 4 joints for each finger represent the tip, the 2 articulations and the base. All joints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

used, the performer and the trial. Each frame of sequences 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 (Fig. 1). 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 resolution of the depth image of 640x480. The length of sample gestures ranges from 20 to 50 frames.

The list of the proposed gestures can be found in Table 1 and an example is given in Fig. 2. Most of them have been chosen to be close to the state-of-the-art, like the VIVA challenge's dataset [OBT14]. Nevertheless, we removed the differentiation between normal and scroll swipe as you can find it in our number-of-fingers approach. The same thing appears with the pair of gesture *Pinch/Expand* and *Open/Close*. In addition, we supplement this base with the gesture *Grab* because of its usefulness in the augmented reality applications, but also for its scientific challenges related to the high potentially variation among performers. We also add the gesture *Shake*, as it can be interesting for recognition algorithm to be able to differentiate gesture composed of other gestures (a shake gesture can be seen as a repetition of opposed swipe gestures). The dataset is available at http://www-rech.telecom-lille.fr/shrec2017-hand/.

Figure 2: *Example of a swipe left gesture represented in color (top), depth map (middle) and skeletal data (bottom).*

Table 1: *List of the gestures included in the dataset.*

Gesture	Label	Tag name
Grab	Fine	G
Expand	Fine	E
Pinch	Fine	P
Rotation CW	Fine	R-CW
Rotation CCW	Fine	R-CCW
Tap	Coarse	T
Swipe Right	Coarse	S-R
Swipe Left	Coarse	S-L
Swipe Up	Coarse	S-U
Swipe Down	Coarse	S-D
Swipe X	Coarse	S-X
Swipe V	Coarse	S-V
Swipe +	Coarse	S-+
Shake	Coarse	Sh

3. Evaluation methodology

We emphasized the main challenges compared to existing hand gesture datasets: (1) Study the dynamic hand gesture recognition using depth and full hand skeleton; (2) Evaluate the effectiveness of recognition process in terms of coverage of the hand shape that depend on the number of fingers used. The same movement is performed with one or more fingers, and the sequence can be labeled according to 14 or 28 classes, depending on the gesture represented and the number of fingers used.

Indeed, labeling the sequences using the 14 gesture during the recognition process allows to judge if an algorithm can face the high coverage of hand shape while performing the same gesture with different number of fingers. In the other hand, we can use 28 gesture classes by grouping sequences following their type and the number of finger used to perform the gesture. In this manner, we will be able to evaluate the different methods on the task of fine-grained hand gesture recognition task.

The recognition accuracy will be computed on the proposed methods following 14 or 28 gesture classes as described above.

4. Methods

Two recognition methods have been registered to the track:

- *Skeleton-based Dynamic hand gesture recognition* [DSWV16] from IMT Lille Douai / University of Lille, France.
- *Classify sequence by key frames with convolutional neural network* from ONERA and ENSTA ParisTech, France.

These algorithms are detailed in the next subsection.

4.1. Skeleton-based Dynamic hand gesture recognition

This method [DSWV16] is proposed by Quentin De Smedt, Hazem Wannous and Jean-Philippe Vandeborre, from IMT Lille Douai / University of Lille, France.

Using 3D hand skeletal data, as shown in Fig. 1, a dynamic gesture can be seen as a time series of hand skeleton. It describes the

motion and the hand shapes along the gesture. For each frame t of the sequence, the position in the camera space of each joint i is represented by three coordinates i.e., $j_i(t) = [\, x_i(t)\ y_i(t)\ z_i(t)\,]$.

In order to modelize the hand gesture, the approach uses different features computed from skeletal data.

Some gestures are defined almost only by the way the hand moves in space (e.g. *swipes*). To take this characteristic into account, we compute a direction vector for each frame t of our sequence using the position of the palm joint noted j_{palm}:

$$\overrightarrow{d}_{dir}(t) = \frac{j_{palm}(t) - j_{palm}(t-c)}{\left\| j_{palm}(t) - j_{palm}(t-c) \right\|} \quad (1)$$

where c a constant value chosen experimentally. We normalize the direction vector by dividing it by its norm. Finally, we create the direction matrix M_{dir} describing the hand motion in the sequence. It is of size $N_f \times 3$ and each line t is the row vector $\overrightarrow{d}_{dir}(t)$.

The rotation of the wrist along the gesture describes also how the hand is moving into space. For each frame t, we compute the vector from the wrist node to the palm node to get the rotational information of the hand:

$$\overrightarrow{d}_{rot}(t) = \frac{j_{palm}(t) - j_{wrist}(t)}{\left\| j_{palm}(t) - j_{wrist}(t) \right\|} \quad (2)$$

Finally, we create the rotation matrix M_{rot} describing the rotation of the wrist along the sequence. It is of size $N_f \times 3$ and each line t is the row vector $\overrightarrow{d}_{rot}(t)$.

To represent the hand shape, we divide the hand skeleton into nine tuples of five joints, named Shape of Connected Joints (SoCJ) according to the hand physical structure, presented in Figure 3.

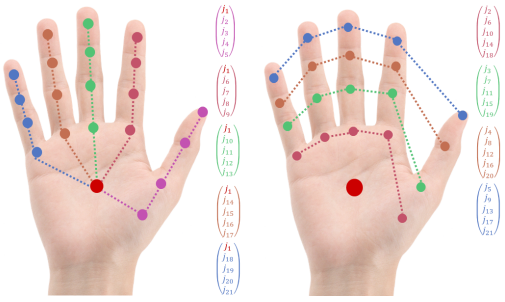

Figure 3: *The nine tuples used to construct the SoCJ descriptors.*

Fisher vector representation and classification. After feature extraction, we represent a hand gesture by three matrices of different descriptors, describing the direction of the movement (M_{dir}), the rotation (M_{rot}) and the shape of the hand (M_{SoCJ}) along the sequence. We use the statistical representation named *Fisher Vector* (FV) in order to get our final descriptor. FV coding method was firstly introduced for large-scale image classification. It can be considered as an extension of the *Bag-Of-Word* (BOW) method by going beyond count analysis.

For gesture classification, we use a supervised learning classifier SVM with a linear kernel as it easily deals with our high-dimensional representation. Further details are available in [DSWV16].

4.2. Classify sequence by key frames with convolutional neural network

This method is proposed by Joris Guerry, Bertrand Le Saux and David Filliat, from ONERA and ENSTA ParisTech, France.

The approach is based on Convolutional Neural Networks (CNNs). This is why we sought an encoding that could adapt to the varying length of the sequences and that could be stored as a tensor. Intuitively, one can note it is possible to guess the gesture by just looking at subsets of images of the sequence. Based on this observation we thought to simply concatenate depth images on each other: each channel representing a keyframe. An example is shown for three-keyframe concatenation in Figure 4. The keyframes are picked regularly with a random variation up to three neighboring frames.

Figure 4: *Three keyframe concatenation of a grab gesture.*

The tensor previously defined is then processed by a CNN. Namely, we took a VGG11 network [SLJ*15] whose parameters were initiated with weights resulting from ImageNet [KSH12] training.

1. We first learn to classify the 14 classes (giving much better results than starting by 28 classes classification)
2. Then we use the same features until "conv5" layer to learn a second binary task : is the hand open or not ? (Which is directly related to the 28 classes problem : $class_{28} = 2 * class_{14} - \delta$, where δ corresponds to 1 if the hand is closed and 0 otherwise.)
3. We could then use the 14 classes predictions ("fc8_14" layer) and the binary predictions ("fc8_δ" layer) to predict the 28 classes with a last fully connected layer.
4. Lastly, we get better results by removing the multi-task training (14 and δ then 28) and just training on the 28 classes prediction.

5. Evaluation results

The proposed approaches are evaluated and compared to three state-of-the-art methods using depth images and skeletal data. We chose two depth-based descriptors: HOG^2 proposed by Ohn-Bar *et*

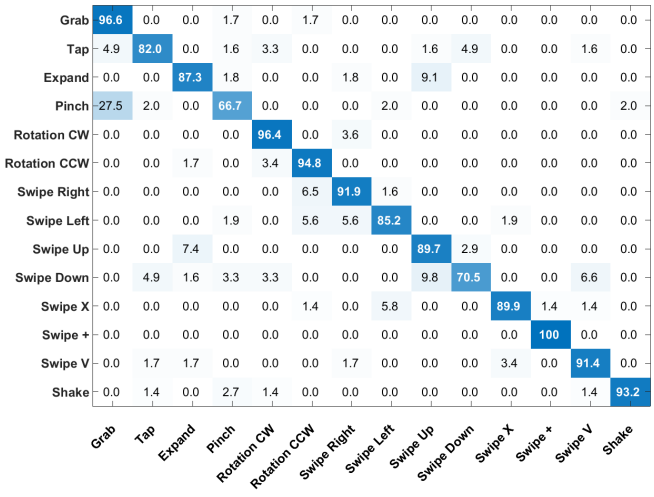

Figure 5: *Confusion matrix (14 gesture classes) obtained by De Smedt* et al. *[DSWV16]*

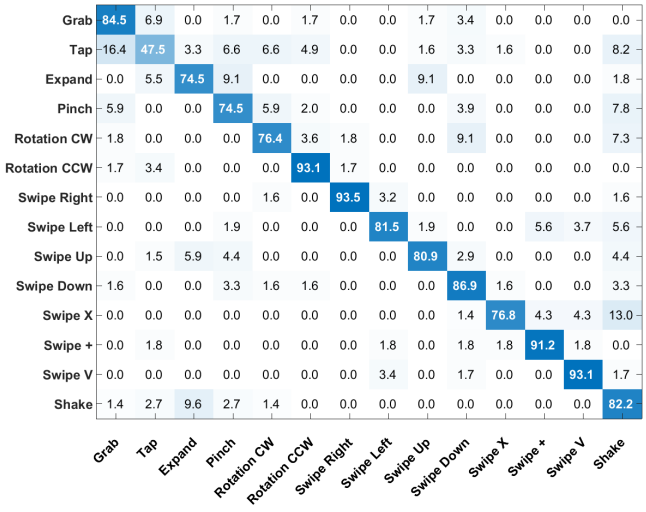

Figure 6: *Confusion matrix (14 gesture classes) obtained by Guerry* et al.

al. [OBT13] and HON4D proposed by Oreifej *et al.* [OL13]; and a skeleton-based method proposed by Devanne *et al.* [DWB*15] originally presented for human action recognition.

The Table 2 analyzes the results obtained by the methods cited below using 14 and 28 gestures.

Table 2: *Accuracy comparison 14 versus 28 gestures.*

Method	14 gestures (%)	28 gestures (%)
Guerry *et al.*	82.90	71.90
De Smedt *et al.* [DSWV16]	88.24	81.90
Ohn-Bar *et al.* [OBT13]	83.85	76.53
Oreifej *et al.* [OL13]	78.53	74.03
Devanne *et al.* [DWB*15]	79.61	62.00

To better understand the behavior of the approaches according to the recognition per class, confusion matrices are also given.

6. Conclusion

In this paper, we have presented a dynamic hand gesture dataset, an evaluation methodology, proposed approaches and results for the SHREC 2017 track "3D Hand Gesture Recognition Using a Depth and Skeletal Dataset". The evaluation of proposed methods shows the promising way to perform hand gesture recognition with skeleton-based approach. Hand skeleton-based method has demonstrated superior results of 88.24% and 81.90% of accuracy respectively for 14 and 28 different gestures.

References

[CDL13] 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), pp. 1–6. doi:10.1109/ICME.2013.6607524. 1

[DSWV16] DE SMEDT Q., WANNOUS H., VANDEBORRE J.-P.: Skeleton-based dynamic hand gesture recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (June 2016). 2, 3, 4, 5

[DWB*15] DEVANNE M., WANNOUS H., BERRETTI S., PALA P., DAOUDI M., BIMBO A. D.: 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Transactions on Cybernetics 45*, 7 (July 2015), 1340–1352. doi:10.1109/TCYB.2014.2350774. 4

[EBG*14] 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 Workshops* (2014), Springer, pp. 459–473. 1

[FPS*09] FEIX T., PAWLIK R., SCHMIEDMAYER H.-B., ROMERO J., KRAGIC D.: A comprehensive grasp taxonomy. In *Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation* (2009), pp. 2–3. 1

[KLTR13] KUZNETSOVA A., LEAL-TAIXÃL' L., ROSENHAHN B.: Real-time sign language recognition using a consumer depth camera. In *IEEE International Conference on Computer Vision Workshops (ICCVW)* (Dec 2013), pp. 83–90. doi:10.1109/ICCVW.2013.18. 1

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105. 3

[KZL12] 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), pp. 1975–1979. 1

[MGO14] MONNIER C., GERMAN S., OST A.: A multi-scale boosted detector for efficient and robust gesture recognition. In *Computer Vision - ECCV Workshops* (2014), Springer, pp. 491–502. 1

[NWTN16] NEVEROVA N., WOLF C., TAYLOR G. W., NEBOUT F.: ModDrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Apr. 2016). 1

[OBT13] OHN-BAR E., TRIVEDI M. M.: Joint angles similarities and hog2 for action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops - HAU3D* (2013). 4

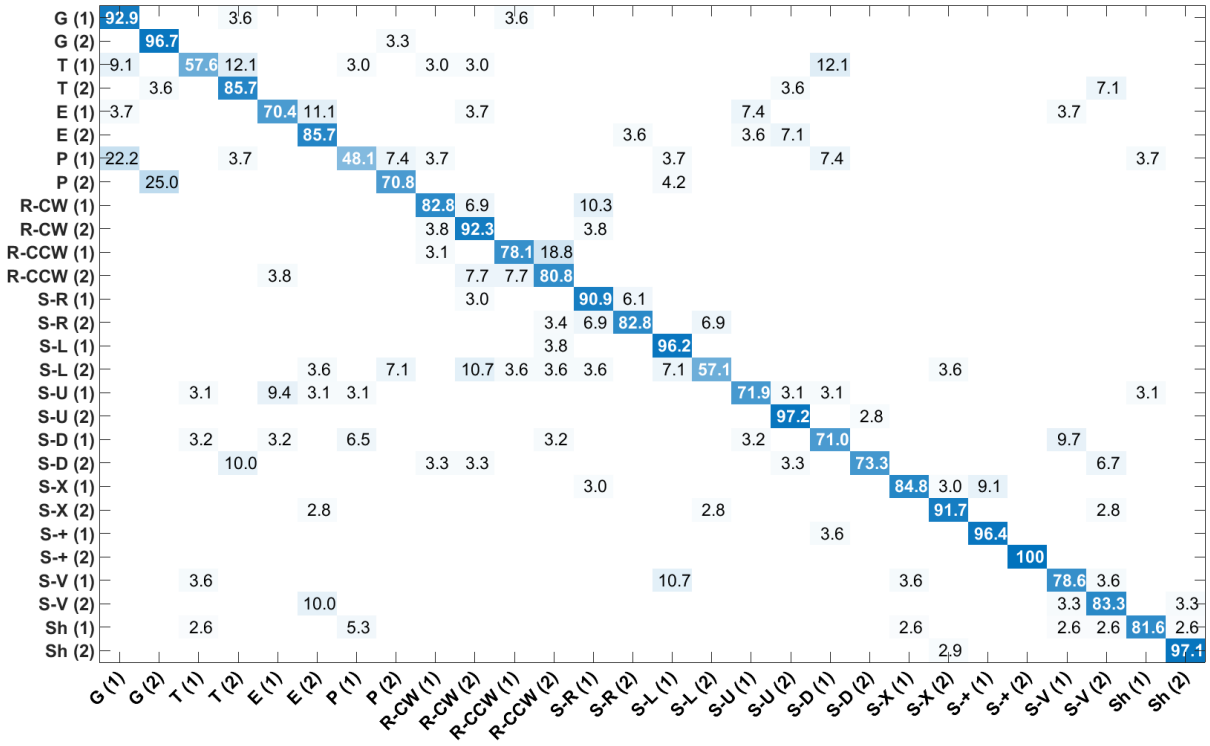

Figure 7: *Confusion matrix (28 gesture classes) obtained by De Smedt* et al. *[DSWV16].*

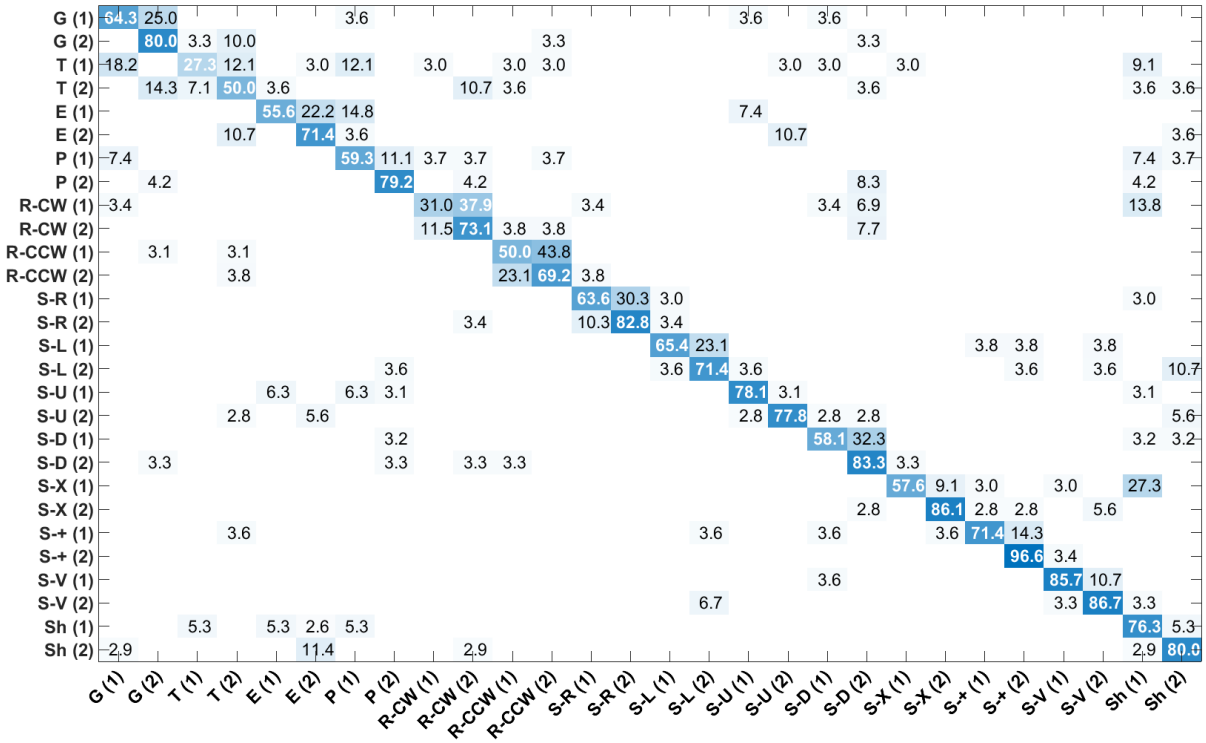

Figure 8: *Confusion matrix (28 gesture classes) obtained by Guerry* et al.

[OBT14] OHN-BAR E., TRIVEDI M. M.: Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Transactions on Intelligent Transportation Systems 15*, 6 (2014), 2368–2377. 2

[OL13] OREIFEJ O., LIU Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2013). 4

[PB11] PUGEAULT N., BOWDEN R.: Spelling it out: Real-time asl fingerspelling recognition. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (Nov 2011), pp. 1114–1119. doi:10.1109/ICCVW.2011.6130290. 1

[RYMZ13] REN Z., YUAN J., MENG J., ZHANG Z.: Robust part-based hand gesture recognition using kinect sensor. *IEEE Transactions on Multimedia 15*, 5 (2013), 1110–1120. 1

[SLJ*15] SZEGEDY C., LIU W., JIA Y., SERMANET P., REED S., ANGUELOV D., ERHAN D., VANHOUCKE V., RABINOVICH A.: Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015). 3

[WLC15] WANG C., LIU Z., CHAN S.-C.: Superpixel-based hand gesture recognition with kinect depth camera. *IEEE Transactions on Multimedia 17*, 1 (2015), 29–39. 1

[ZYT13] ZHANG C., YANG X., TIAN Y.: Histogram of 3d facets: A characteristic descriptor for hand gesture recognition. In *IEEE Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (April 2013), pp. 1–8. doi:10.1109/FG.2013.6553754. 1