Article Title

A Dynamic Hand Gesture Recognition Dataset for Human-Computer Interfaces

Authors

Graziano Fronteddu[1], Simone Porcu[1,2], Alessandro Floris[1,2], Luigi Atzori[1,2]

Affiliations

1. DIEE, University of Cagliari, 09123 Cagliari, Italy
2. CNIT, University of Cagliari, 09123 Cagliari, Italy

Corresponding author(s)

Alessandro Floris (alessandro.floris84@unica.it)

Abstract

Computer vision systems are commonly used to design touch-less human-computer interfaces (HCI) based on dynamic hand gesture recognition (HGR) systems, which have a wide range of applications in several domains, such as, gaming, multimedia, automotive, home automation. However, automatic HGR is still a challenging task, mostly because of the diversity in how people perform the gestures. In addition, the number of publicly available hand gesture datasets is scarce, often the gestures are not acquired with sufficient image quality, and the gestures are not correctly performed. In this data article, we propose a dataset of 27 dynamic hand gesture types acquired at full HD resolution from 21 different subjects, which were carefully instructed before performing the gestures and monitored when performing the gesture; the subjects had to repeat the movement in case the performed hand gesture was not correct, i.e., the authors of this paper that were observing the gesture found that it did not correspond to the exact expected movement and/or the camera recorded a viewpoint did not allow for a plain visualizing of the gesture. Each subject performed 3 times the 27 hand gestures for a total of 1701 videos collected and corresponding 204120 video frames.

Keywords

Dynamic hand gesture recognition, Hand gesture dataset, Human-Computer Interface, Touch-less screen, Computer vision.

Specifications Table

Subject	Computer Vision and Pattern Recognition
Specific subject area	Hand gesture recognition
Type of data	Video Image Table
How data were acquired	Hardware: Logitech Brio Stream 4K HDR video camera. Software: a dedicated program implemented in Python.
Data format	Raw: .AVI, .PNG Analysed: .CSV
Parameters for data collection	The resolution for the video camera is 1920 x 1080 at 30 fps. A dataset of 27 dynamic hand gesture types was acquired. A total of 21 subjects participated in data collection. Each subject performed each hand gesture for 3 times for a total of 1701 videos collected (63 videos for each different hand gesture).
Description of data collection	The subjects were carefully instructed before performing the gestures and monitored when performing the gesture. The subjects were prompted to perform each hand gesture with a 3 second sample video of the hand gesture followed by a 3 second countdown. The subjects performed the hand gestures with their right hand in front of the video camera. The subjects had to repeat the movement in case the performed hand gesture was not correct. The starting time and ending time to perform each hand gesture were extracted from the videos and reported in a table for each of the subjects.
Data source location	Institution: University of Cagliari City: Cagliari Country: Italy
Data accessibility	Repository name: Dataset for Dynamic Hand Gesture Recognition Systems Direct URL to data: https://dx.doi.org/10.21227/43mn-bb52

Related research article	-
Related project(s)	This work has been partially funded by the Italian Ministry for the Economic Development (MISE), under the framework "Asse II del programma di supporto tecnologie emergenti (FSC 2014-2020)", project Monifive.

Value of the Data

- Automatic hand gesture recognition (HGR) is still a challenging task for the following reasons that have motivated the creation of the proposed dataset:
    - Scarcity of publicly available hand gesture datasets: public hand gesture datasets are limited given the complexity of the HGR problem when considering different environments. To the authors' knowledge the following is the complete list of existing relevant datasets. The NVIDIA Dynamic Hand Gesture Dataset [1] includes dynamic hand gestures captured with depth, color and stereo-IR sensors. A total of 20 subjects participated in the dataset collection, which included 25 gesture classes intended to be used while driving to control in-car automotive devices. The IPN Hand dataset [2] includes 13 classes of static and dynamic gesture for interaction with touch-less screens acquired from 50 distinct subjects. The Jester Dataset [3] is a large-scale gesture recognition real-world video dataset. It is the result of a crowdworking-based data collection, which involved 1,376 subjects performing a set of 27 gestures that encompass the most commonly performed human gestures in the context of visual HCIs. Further datasets, such as [4-5] focused on hand gestures not intended for HCIs. In [4], the ChaLearn LAP Continuous Gesture Dataset (ConGD) contains 249 gesture classes regarding different domains, including sign language, body language, symbolic gestures and Italian gestures. In [5], the EgoGesture dataset consists of 83 static and dynamic gesture classes acquired from 50 distinct subjects. In this case, the gesture recognition domain is the first-person, i.e., the considered gestures were specifically designed for interaction with wearable devices.
    - Hand gestures acquired with poor image quality: in [1], the videos were captured at 320×240 resolution at 30 fps. In [2], the videos were captured at 640×480 resolution at 30 fps. In [3], the videos are captured with a height of 100 pixels and variable width (due to crowdworking-based data collection).

    - Diversity in how people perform the gestures and hand gestures not correctly performed by the subjects: the subjects who performed the gestures for datasets [1-3] were not monitored when performing the gesture. This may result in the same gesture performed differently by diverse people or gestures not properly performed by the subjects.
- Motivated by the aforementioned limitations of existing datasets, with this paper we aim to provide a public dataset including hand gestures properly performed and captured at high quality. In particular, this dataset provides 1701 total videos of 27 dynamic hand gestures performed 3 times by 21 different subjects (63 videos for each different gesture). The videos are acquired at 1920 x 1080 resolution at 30 fps. The subjects were carefully instructed before performing the gestures and were monitored when performing the gesture; in case the performed hand gesture was not correct, the subjects had to repeat the movement.
- The proposed dataset motivates researchers to implement accurate HGR systems since it includes high quality videos of subjects performing 27 different hand gestures properly. These characteristics are not easy to find in current available public hand gesture datasets. Also, 25 of the considered gestures are those considered in [1], which are commonly used for online HGR evaluation. We have proposed 2 additional novel hand gestures which have been designed to be used to command the playback of previous/next video sequence when controlling multimedia systems.
- The proposed dataset can be used to implement HGR systems, which are commonly utilized to design touch-less human-computer interfaces (HCI) for a wide range of applications, such as, gaming, multimedia, automotive, home automation.

1. Data

1.1 hand_gestures_dataset_videos.zip

This dataset contains the videos of the recorded hand gestures. The zip contains 27 main folders. Each main folder refers to a hand gesture class, for a total of 27 main folders named “class_xx”, where “xx” identifies the class from 01 to 27. Within each of the class folders there are 21 sub-folders, one folder for each of the subjects that performed the hand gestures. These folders are named “Useryy_”, where “yy” identifies the user from 01 to 21. Each of the user folders contains three videos (.avi) corresponding to the three hand gestures performed by the user for each hand gesture class. The size of the full dataset is 21.34 GB.

1.2 HGD_VideoFrames_class_XX.zip

These datasets contain the video frames extracted from the videos of the recorded hand gestures. Each zip file contains the video frames of a hand gesture class, for a total of 27 zip files named "*HGD_VideoFrames_class_XX.zip*", where "xx" identifies the class from 01 to 27. Therefore, each zip file contains one of the 27 class folders. Within each of the class folders there are 21 sub-folders, one folder for each of the subjects that performed the hand gestures. These folders are named "Useryy_", where "yy" identifies the user from 01 to 21. Each of the user folders contains, in turn, 3 sub-folders, one folder for each of the three hand gestures performed for each hand gesture class. These sub-folders are named, respectively, "Useryy_1", "Useryy_2", and "Useryy_3", and contain 120 video frames (.png) extracted from the corresponding video. The size of each zip file is about 9 GB.

1.3 hand_gesture_timing_stats.csv

This dataset contains timing information regarding the gestures performed by the subjects. The size of this dataset is 36 KB. It has 567 records plus the header.
The structure of the table is shown in Fig. 1. The meaning of the columns is as follows:

- *class*: hand gesture class, from 01 to 27.
- *user*: user who performed the hand gestures, from 01 to 21.
- *start_frame_1*: starting frame related to the first performed hand gesture. It is a number between 0 and 119.
- *end_frame_1*: ending frame related to the first performed hand gesture. It is a number between 0 and 119.
- *exec_time_1*: execution time (in seconds) related to the first performed hand gesture. It is computed as the difference between the ending frame and the starting frame divided by the 30 fps set for video recording, i.e., $\frac{end\_frame\_1 - start\_frame\_1 + 1}{30}$.
- *start_frame_2*: starting frame related to the second performed hand gesture. It is a number between 0 and 119.
- *end_frame_2*: ending frame related to the second performed hand gesture. It is a number between 0 and 119.
- *exec_time_2*: execution time (in seconds) related to the second performed hand gesture. It is computed as the difference between the ending frame and the starting frame divided by the 30 fps set for video recording, i.e., $\frac{end\_frame\_2 - start\_frame\_2 + 1}{30}$.
- *start_frame_3*: starting frame related to the third performed hand gesture. It is a number between 0 and 119.

- *end_frame_3*: ending frame related to the third performed hand gesture. It is a number between 0 and 119.
- *exec_time_3*: execution time (in seconds) related to the third performed hand gesture. It is computed as the difference between the ending frame and the starting frame divided by the 30 fps set for video recording, i.e., $\frac{end\_frame\_3 - start\_frame\_3 + 1}{30}$.
- *mean_exec_time*: mean execution time (in seconds) for that related user and hand gesture. It is computed as the mean of the execution times measured for the three hand gestures performed by that user for that class, i.e., $\frac{exec\_time\_1 + exec\_time\_2 + exec\_time\_3}{3}$.
- *std_dev_exec_time*: standard deviation of the three execution times (in seconds) measured for the three hand gestures performed by that user for that class. It is computed as $\sqrt{\frac{\sum_{i=1}^{3} (exec\_time\_i - mean\_exec\_time)^2}{3}}$.
- *total_mean_exec_time*: total mean execution time (in seconds) for that class. It is computed as the mean of all the execution times measured for the three hand gestures performed by all the users for that class.
- *total_std_dev_exec_time*: total standard deviation of all the execution times (in seconds) for that class. It is computed as the standard deviation of all the execution times measured for the three hand gestures performed by all the users for that class. Note that in this case the standard deviation has been computed dividing by (N-1) as the entire population is considered.

Note that there is only one value of "total_mean_exec_time" and "total_std_dev_exec_time" for each class, which are reported at the first row of each class table.

class	user	start_frame_1	end_frame_1	exec_time_1	start_frame_2	end_frame_2	exec_time_2	start_frame_3	end_frame_3	exec_time_3	mean_exec_time	std_dev_exec_time	total_mean_exec_time	total_std_dev_exec_time
1	1	36	69	1.1333	35	64	1	35	65	1.0333	1.0556	0.069389	0.80265	0.23294
1	2	70	113	1.4667	20	44	0.83333	24	41	0.6	0.96667	0.44845	0	0
1	3	54	81	0.93333	29	55	0.9	43	85	1.4333	1.0889	0.29876	0	0
1	4	22	36	0.5	20	40	0.7	38	82	1.5	0.9	0.52915	0	0
1	5	33	52	0.66667	33	49	0.56667	31	51	0.7	0.64444	0.069389	0	0
1	6	25	44	0.66667	30	55	0.86667	22	43	0.73333	0.75556	0.10184	0	0
1	7	35	57	0.76667	35	62	0.93333	26	53	0.93333	0.87778	0.096225	0	0
1	8	28	50	0.76667	23	52	1	18	45	0.93333	0.9	0.12019	0	0
1	9	9	34	0.86667	27	48	0.73333	31	47	0.56667	0.72222	0.15031	0	0
1	10	30	55	0.86667	30	62	1.1	30	48	0.63333	0.86667	0.23333	0	0
1	11	1	29	0.96667	12	29	0.6	10	32	0.76667	0.77778	0.18359	0	0
1	12	22	40	0.63333	19	37	0.63333	20	38	0.63333	0.63333	0	0	0
1	13	26	52	0.9	25	40	0.53333	20	44	0.83333	0.75556	0.19532	0	0
1	14	60	83	0.8	35	54	0.66667	34	54	0.7	0.72222	0.069389	0	0
1	15	62	85	0.8	34	54	0.7	35	59	0.83333	0.77778	0.069389	0	0
1	16	35	79	1.5	30	54	0.83333	48	80	1.1	1.1444	0.33555	0	0
1	17	36	58	0.76667	32	54	0.76667	34	60	0.9	0.81111	0.07698	0	0
1	18	37	54	0.6	33	48	0.53333	38	56	0.63333	0.58889	0.050918	0	0
1	19	49	68	0.66667	39	58	0.66667	35	55	0.7	0.67778	0.019245	0	0
1	20	47	61	0.5	39	57	0.63333	37	54	0.6	0.57778	0.069389	0	0
1	21	87	105	0.63333	14	32	0.63333	17	33	0.56667	0.61111	0.03849	0	0
2	1	31	57	0.9	34	67	1.1333	35	60	0.86667	0.96667	0.1453	0.7037	0.18839
2	2	10	30	0.7	22	41	0.66667	21	38	0.6	0.65556	0.050918	0	0
2	3	45	64	0.66667	30	50	0.7	43	60	0.6	0.65556	0.050918	0	0

Fig. 1 Structure of the table *hand_gesture_timing_stats.csv*.

Experimental Design, Materials, and Methods

We considered 27 dynamic hand gestures commonly used for online HGR evaluation. Most of these gestures (1-25) were adopted by the NVIDIA popular dataset [1], which in turn had already adopted some hand gestures from existing commercial systems or popular datasets. We have proposed 2 additional novel hand gestures (26-27) which have been designed to be used to command the playback of previous/next video sequence when controlling multimedia systems. These 2 dynamic hand gestures consist in closing the open hand, except for the thumb, and then moving the closed hand to the left (to command the playback of the previous video) or to the right (to command the playback of the next video). The thumb must be pointed to the chosen direction (left or right). The 27 hand gestures are shown in Fig. 2.

The subjects who participated in the data collection were provided with written and oral instructions describing the tasks to be performed. Each subject signed an informed consent, where they were informed that their actions would be recorded during the experiment and that the dataset would be made publicly available online.

The subject had to sit on a chair in front of a desk, where sample videos of the hand gestures to be performed were shown on a monitor. The camera was fixed at the desk and the subject had to perform the hand gestures in front of the camera. The subjects performed the gestures with their right hand in front of the video camera. Only the hand and the arm of the subjects are visible in the video so that they are not recognizable and their privacy is preserved. The subjects were prompted to perform each gesture with a 3 second sample video of the gesture followed by a 3 second countdown. Then, the subject had to perform the hand gesture. The subjects had to repeat the hand gesture in case the movement was not correct, i.e., the authors of this paper that were observing the gesture found that it did not correspond to the exact expected movement and/or the camera recorded a viewpoint did not allow for a plain visualizing of the gesture. Indeed, the subjects were monitored when performing the gesture and corrected when needed. Each recorded video of the performed hand gestures lasted for a total of 3 seconds, i.e, 120 video frames. The subjects had to perform 3 times each of the 27 hand gestures, for a total of 1701 videos collected and corresponding 204120 video frames.

The software used to collect the data was a dedicated program implemented in Python 3.7, mostly based on the OpenCV library that comprises several functions for video capture, editing and analysis. The Python script in charge of collecting the data acquires in 4 seconds 120 frames of the same size, showing the test subject their reproduction in real-time. This approach helps the subjects to centre the hand into the frames. Moreover, to further support the test subjects, a video of the gesture they have to perform was shown 3 seconds before the performed gesture was recorded. The same script was used to export all the collected videos into video frames. This

step has been accomplished in a second moment because it leads the used workstation to resources starvation.

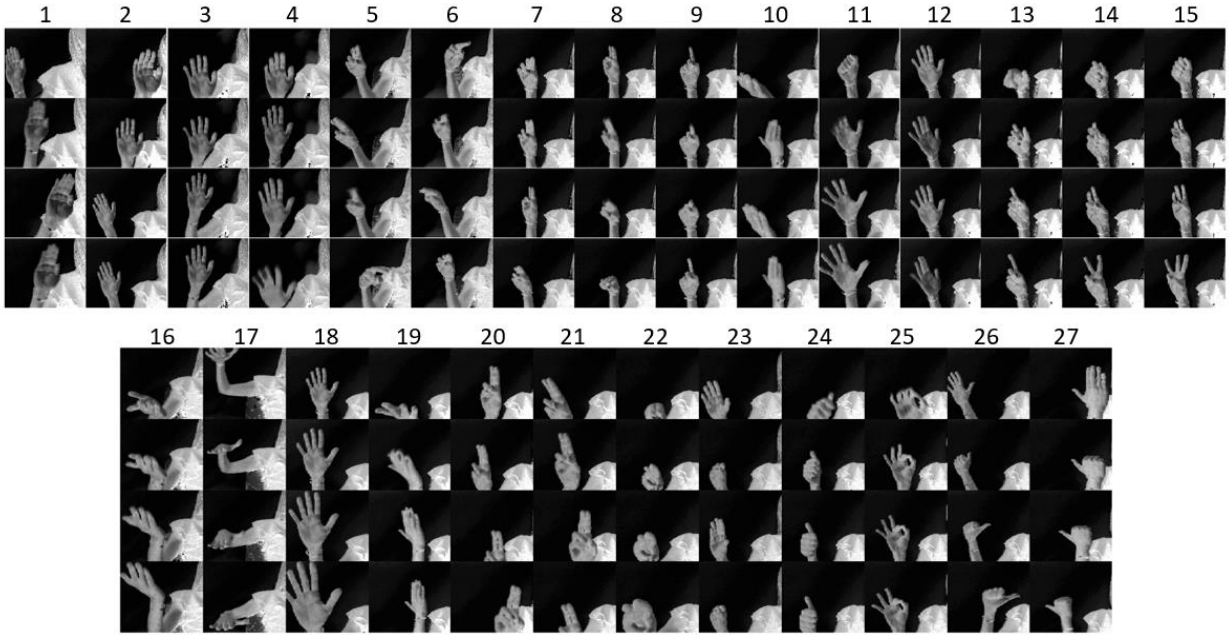

Fig. 2 Twenty-seven dynamic hand gesture classes. The first 25 gestures were adopted by the NVIDIA popular dataset [1]. We have proposed 2 additional novel hand gestures (26-27) which have been designed to be used to command the playback of previous/next video sequence when controlling multimedia systems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially funded by the Italian Ministry for the Economic Development (MISE), under the framework “Asse II del programma di supporto tecnologie emergenti (FSC 2014-2020)”, project Monifive.

References

[1] P. Molchanov, X. Yang, S. Gupta K. Kim, S. Tyree, J. Kautz, “Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks,” in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4207–4215, doi:10.1109/CVPR.2016.456.

[2] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez and K. Yanai, "IPN Hand: A Video Dataset and Benchmark for Real-Time Continuous Hand Gesture Recognition," in 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 4340-4347, doi: 10.1109/ICPR48806.2021.9412317.

[3] J. Materzynska, G. Berger, I. Bax and R. Memisevic, "The Jester Dataset: A Large-Scale Video Dataset of Human Gestures," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 2874-2882, doi: 10.1109/ICCVW.2019.00349.

[4] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon and S. Escalera, "ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016, pp. 761-769, doi: 10.1109/CVPRW.2016.100.

[5] Y. Zhang, C. Cao, J. Cheng and H. Lu, "EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition," in IEEE Transactions on Multimedia, vol. 20, no. 5, pp. 1038-1050, May 2018, doi: 10.1109/TMM.2018.2808769.