A Survey of Datasets for Human Gesture Recognition

Simon Ruffieux[1], Denis Lalanne[2], Elena Mugellini[1], and Omar Abou Khaled[1]

[1] University of Applied Sciences and Arts of Western Switzerland, Fribourg
{Simon.Ruffieux,Elena.Mugellini,Omar.AbouKhaled}@Hefr.ch
[2] University of Fribourg
Denis.Lalanne@unifr.ch

Abstract. This paper presents a survey on datasets created for the field of gesture recognition. The main characteristics of the datasets are presented on two tables to provide researchers a clear and rapid access to the information. This paper also provides a comprehensive description of the datasets and discusses their general strengths and limitations. Guidelines for creation and selection of datasets for gesture recognition are proposed. This survey should be a key-access point for researchers looking to create or use datasets in the field of human gesture recognition.

Keywords: human-computer interaction, gesture recognition, datasets, survey.

1 Introduction

The fields of human activity, action and gesture recognition gained more and more attention these last years, notably due to the numerous affordable sensors commercially released. In recent years, more and more datasets have been created by researchers in order to develop, train, optimize and evaluate algorithms; several of them have been made publicly available to developers and researchers. Several articles have already addressed the topic of datasets for the general field of human activity and action recognition [1,2] and a couple of websites already list publicly available datasets [3,4]. However the topic of datasets for the specific field of gesture recognition has not been addressed yet. Gesture recognition is defined as a subset of human action and activity recognition and generally requires its own specific datasets for the development of algorithms. The devices and sensors employed are often similar in both fields however they are generally used with different setups in gesture recognition: the sensors tend to be closer to the user in order to augment the granularity and the users are generally aware of the presence and position of the sensor thus interacting towards it. Therefore, most datasets acquired for activity and action recognition cannot be directly used for gesture recognition; the same asset is, in most cases, also applicable with algorithms.

The goal of the present survey is two-fold: provide an overview and a discussion about the available datasets and provide brief guidelines to help researchers when selecting or creating datasets.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 337–348, 2014.

This work takes place in the context of the FEOGARM project [5]. The goal of FEOGARM is to provide a comprehensive framework for facilitating gesture evaluation and recognition methods. A dataset for gesture recognition has been publicly released in this context [6].

2 Related Works

Several surveys have already addressed topics related to datasets although most of them have mostly considered the field of human action and activity recognition; only short sub-sections were addressing the gesture recognition domain. A recent and informative survey addressed the topic of datasets for activity recognition but explicitly omitted datasets focusing solely on gesture recognition in order to narrow the survey [3]. Another survey addressed the methods, systems and evaluation metrics for vision based human-activity recognition to detect abnormal behaviors in videos streams, a subset of activity recognition called surveillance systems [7]. Large surveys of the activity recognition domain are also available [8,9], resuming the taxonomies, techniques, challenges and listing the datasets for full-body activity recognition. In [10], a survey of the datasets for action recognition are presented, a domain at the frontier between activity and gestures. In [11], the datasets available for pose estimation and tracking are listed and discussed; the need for common standards in the domain is strongly highlighted. These surveys provide a good overview of the human activity and action recognition field, although they do not address directly gesture recognition.

The surveys that specifically addressed the field of gesture recognition have mostly considered three perspectives: the topic of gesture recognition in general [12,13], the specific topic of hand gestures for human-computer interaction [14,15] and the topic of sign language [16]. None of these surveys focused on the specific topic of the existing datasets for gesture recognition. A few research papers have addressed topics such as modeling, building and using datasets in the context of gesture recognition. In [17], they presented a framework based on databases for gesture recognition. They developed an ASL and hand shape real-time recognition systems based on comparisons with examples of images stored in their databases. The developed method demonstrated the ability to search a gesture database fast enough for real-time gesture recognition applications. However the low accuracy rate of the recognition system was not satisfactory and required some additional work. In [18], they discussed and highlighted some important modeling considerations when creating a database for hand gesture recognition in the context of natural interfaces. They identified the required assumptions to create an effective database: naturality of the gesture set, size of the set, a precise analysis of the potential effects of the recording conditions and a precise description of the acquisition process. They also stated the importance of recording the data with multiple sensors as a way to achieve independence from the acquisition conditions; they notably promote motion capture systems and video cameras. Finally, in [19], they study the impact of the semiotic modalities such as text, images or videos, which are used to instruct the subjects, on the quality of the performed gestures. They also illustrate the importance to balance correctness and

coverage properties of a gesture dataset in order to obtain the best recognition performances with machine learning algorithms. The study demonstrated that video instructions promote correctness while texts and images together are best for coverage; the latter also giving a strong sense of freedom to the subjects. Gesture datasets are also slowly moving away from research and spread to the commercial market; for example, ARB Labs [20] has recently started a company based on a gesture dataset and the related acquisition software.

3 Survey

This section presents the main datasets that have been employed or developed for the field of gesture recognition these last years. The datasets are presented through two chronologically ordered tables: Table 1contains the general information and a short description for each datasets. Then Table 2 resumes the main technical characteristics and categorizes the datasets according to the three main types of ground truth annotations. Note that older datasets have been omitted due to the important changes in data quality and on the types of sensors employed. This survey has also been limited to datasets containing gestures mostly involving hand(s) and arm(s) motion.

The Table 1 provides an overview of the 15 reviewed datasets. The table presents the *name* or acronym of the datasets and their reference paper. The number of *citations* for the reference papers, which have been retrieved from Google Scholar the 03.02.2014. Two of the papers have more than one hundred citations. The *placement of the sensor(s)* indicates if the sensor was placed in the environment or on the user. The sensors and their placement are rather constant amongst reviewed datasets. Most datasets rely on a single video camera at a fixed location in the environment. Only a couple of datasets used alternative setups such as multiple video cameras or a combination of environmental and wearable sensors. Only two datasets are based on environmental and wearable data. The ChAirGest dataset uses a combination of RGB-D camera fixed in the environment and inertial motion units (IMU) located on the arm of the user. The 6DMG dataset uses a combination of hand-held controller and optical tracker to obtain both the motion of the hand and its position in the space. Such setup enables the comparison or the fusion of both approaches on common material. *The quality of information* depicts the amount of documents, description and information which have been provided with a dataset. Such documentation can be very important to understand and use a dataset. Large variations can be observed between datasets. The *types of gestures* distinguish the gesture vocabularies present in the datasets. Datasets are either taking their vocabulary from existing ones such as sign language [21], cultural signs [22] or military gestures [23] or creating original vocabularies. Numerous datasets uses their own vocabulary of gestures which are thought for specific applications or domains such as gaming [24] or human-computer interaction [6]. These vocabularies usually rely on iconic gestures which imbue a correspondence between the gesture and the reference, symbolic gestures which are highly lexicalized and metaphoric gestures which correspond to an abstract representation.

Table 1. This table provides a general description of the most recent datasets for human gesture recognition. Notation for; for the information: from poor (1) to very good (5); and finally for the availability: Public, Public on Request or Not Yet.

Name	Year	Citations (03.02.2014)	Sensors placement	Information	Types of gestures	Purposes & Description	Availability
3DIG [25]	'13	1	Environment	2	Iconic	**Recognition** of iconic gestures where subjects were free to perform their own gesture to depict each object	P
ASL Dataset [21]	'13	-	Environment	5	Sign language	American sign **recognition**. Evaluation of hands **detection** & **tracking.** Acquisition still on-going.	NY
CGD2013 [22] **ChaLearn Dataset**	'13	2	Environment	5	Metaphoric	Multimodal gesture **recognition** of cultural Italian gestures accompanying speech. **Challenge**-related dataset	P
ChAirGest [6]	'13	1	Env. & wear.	5	Iconic & metaphoric	Gesture **spotting** & **recognition** from multimodal data in the context of close HCI. **Challenge**-related dataset	PR
SKIG [26]	'13	5	Environment	3	Iconic & metaphoric	Improve gesture recognition from RGB-D data, notably with different illuminations. Hand gesture **recognition** seen from above	P
6DMG [27]	'12	2	Env. & wear.	5	Iconic & metaphoric	Explore gesture **recognition** from implicit & explicit data. Subjects performed the gestures with a Wii-mote in their right hand	P
MSRC-12 [19]	'12	21	Environment	5	Iconic & metaphoric	Gesture **recognition** from the skeleton data. Study the motion variation across users with skeleton data	P
G3D [24]	'12	7	Environment	5	Iconic	Gaming actions and gestures **recognition** & **spotting.** Specifically designed to improve gaming without controller	PR
MSRGesture3D [28]	'12	17	Environment	3	Sign language	Sign language **recognition** from hand depth data. Only the segmented hand sections of the images are provided	P
CGD2011 [29] **ChaLearn Dataset**	'11	17	Environment	5	Iconic & metaphoric	Improve one-shot learning for **recognition**. **Challenge**-related dataset. The competition had a large success.	P
NATOPS Aircraft Handling Signals Database [30]	'11	26	Environment	5	Metaphoric & symbolic (Real vocabulary)	Body-and-hand **tracking** & gesture **recognition** requiring both body and hand information to distinguish gestures	PR
NTU Dataset [31]	'11	68	Environment	2	Metaphoric & symbolic poses	Hand pose & shape **recognition** in cluttered conditions. Only contains static images, no motion.	P
Keck Gesture Dataset [23]	'09	153	Environment	4	Metaphoric & symbolic (Real vocabulary)	Military gestures performed with perturbations in the background. Designed to evaluate gesture **recognition** and **spotting** in harsh conditions.	P
ASLLVD [32]	'08	17	Environment	4	Sign language	A reference database in automatic sign language **recognition** and **spotting** with data captured from several viewpoints.	PR
CHGD [33] **(Cambridge Hand Gesture Dataset)**	'07	136	Environment	4	Metaphoric	Hand segmentation & gesture **recognition** in varying illuminations conditions. It only contains sequences of images.	P

Although most datasets span on multiple gestures types, some dataset focus on a specific type. For example the approach of 3DIG dataset focusing on iconic gestures is interesting: the subjects were free to perform the gesture of their choice to depict a specific object; the classification goal being to recognize the depicted object. Such approach generates large variations within a class which complexifies the recognition. Then the *purposes* and a short *description* of the datasets are provided to better characterize each dataset. Finally the last column shows the current *availability* of each dataset. In this survey, all the presented datasets are available online either publicly or on request, except one which was not yet available. Generally datasets are available on request due to image rights of the recorded subjects; researchers have to sign an End-User License Agreement (EULA) to obtain a dataset. This EULA ensures that researchers will preserve the data of the subjects. Only one of the reviewed datasets is available commercially and has not been listed in the tables due to the lack of information about it [20]. Note that for some datasets, notably the ones used in challenges, only around 75% percent of the instances are publicly available, the remaining is kept private to safely evaluate the performances of the algorithms developed by the challengers.

The Table 2 resumes the main technical characteristics of the reviewed datasets. It resumes the *body-parts* that are involved in the gestures to recognize. The reviewed datasets are quite heterogeneous in that respect, spanning from single hand to full-body. For example, the gestures from the CGD2011 dataset could be recognized only by having the information from the two hands and arms. The *sensor view-point* indicates the position(s) of the video sensor(s), when applicable, with respect to the subjects. Most datasets use a front-view, with the sensor in front of the subject. However a couple of datasets use a top-view, with the sensor above the user and facing downward, which greatly simplifies hand recognition from the images. A few other datasets use different approaches: a trade-off between top and front view for the ChAirGest dataset which uses a sensor inclined at 45° or multiple simultaneous view-points for the ASSLVD dataset. A single dataset contains a moving camera in order to evaluate algorithms in difficult conditions. The *subject stance* corresponds to the position of the user during the recording. For most datasets, subjects were standing in front of the camera, although in a few datasets, subjects were sitting on a chair which implies interaction with whole or part of the upper-body. The Keck Gesture Dataset is the only reviewed dataset containing subjects who are moving during the interaction; a very challenging recognition task. Finally the more classical characteristics: the *number of subjects* who are available in the data, the number of distinct classes (gestures) and the total number of instances. In general, the more subjects, classes and instances, the better. However, it is usually important to have a high ratio between the number of instances and the number of classes to properly train machine learning algorithms. Then the *sensors* used to acquire the data are described. The Kinect-based dataset have not all recorded each of the streams from the sensor; Kinect being a multimodal sensor, it provides color and depth stream, the approximate position of the subject's body-parts through a skeleton representation and the sound. Non video-based sensors include inertial motion units (similar to motion sensors embedded in phones, smartwatches and smart-bands), Optical tracker or Vicon system for motion capture or a Wiimote+ controller from Nintendo. The next column indicates the *resolution f*or the

sensors based on videos. An increase of the resolution through the years is clearly observable. Higher resolution implies more information in the image but also more processing time when processing an image. The *frequency* is indicated for all mentioned sensors, when applicable. Similarly to the resolution, a higher frequency means more information but increases processing time and data storage size; many algorithm implementations artificially down sample the frequency for real-time applications. However, a high frequency is important in order to capture all the information during rapid movements. Finally the *size* of the datasets in Gigabytes (GB) usually results from the previous choices and can largely vary across datasets.

Table 2. This table provides technical information about the most recent datasets for human gesture recognition. Notation for *body-parts*: Full-Body, Upper-Body, Hand and Arm; for the *sensor-view*: Front-View, Top-View, Lateral-View and Moving-View; for the *user stance*: Standing, Sitting and Moving; for the *Kinect sensor*: Color, Depth, Skeleton and Sound.

Name	Body-parts	Sensor view	Subject stance	Subjects	Classes	Instances	Sensors	Resolution	Frequency [Hz]	Size [GB]	Ground truth		
											Label	Temporal	Spatial
3DIG [25]	HA	FV	St.	29	20	1739	Kinect^{CDS}	640x480	30	85[2]	X		
ASL Dataset [21]	UB	FV	St.	2	1300+	1300+	Kinect^{CDS}	640x480	25	?	X	X	X[1]
CGD2013 [22] **ChaLearn Dataset**	UB	FV	St.	27	20	13000	Kinect^{CDSSo}	640x480	20	27[2]	X	X	
ChAirGest [6]	HA	FV 45°	Si.	10	10	1200	Kinect^{CDS} 4 IMU	640x480 -	30 50	1000 3[2]	X	X	
SKIG [26]	H	TV	Si.	6	10	1080	Kinect^{CD}	320x240	10	1.2[2]	X		
6DMG [27]	H	-	-	28	20	5600	Wiimote+ Optical tracker	-	60	0.02	X	X	X
MSRC-12 [19]	FB	FV	St.	30	12	6244	Kinect^{S}	-	30	0.2	X	X[3]	
G3D [24]	FB	FV	St.	10	20	600	Kinect^{CDS}	640x480	30	47[2]	X	X	
MSRGesture3D [28]	H	FV	-	10	12	336	Kinect^{D}	130x130	20	0.03	X		
CGD2011 [29] **ChaLearn Dataset**	2HA	FV	St.	20	30	50'000	Kinect^{CD}	320x240	10	30 5[2]	X	X	X[1]
NATOPS Aircraft Handling Signals Database [30]	UB	FV	St.	20	24	9600	Stereo Cam. Vicon[1]	320x240	20	19	X	X	X[1]
NTU Dataset [31]	H	FV	Si.	10	10	1000	Kinect^{CD}	640x480	-	0.1	X		
Keck Gesture Dataset [23]	2HA	FV MV	St. Mo	3	14	294	Color Cam.	640x480	15	0.15	X	X	
ASLLVD [32]	UB	3FV LV	St.	6	2700	3300	4 Color Cameras	640x480	60	1.6[2]	X	X	
CHGD [33] **(Cambridge Hand Gesture Dataset)**	H	TV	Si.	2	9	900	Color camera	320x240	?	1	X		

When working with video, many datasets offer a couple of data qualities: raw or compressed/encoded qualities. Encoding video dramatically reduces the size of the data with only a partial loss of information but a large gain in download, loading and processing times. The *types of ground truth* present in the datasets have strong impli-

1 Only for part of the data.
2 The data has been encoded or compressed.
3 Only the start event of gestures has been temporally labeled.

cations on the type of algorithms that may be trained and evaluated. Therefore this information has been used as a way to categorize the datasets. In this work, datasets are grouped in three non-exclusive incremental categories: recognition, spotting and tracking. This categorization allows the definition of the potential usage(s) of the dataset. Gesture labels are normally always provided because they allow recognition algorithms to be trained and evaluated. Spotting algorithms require temporal segmentation which corresponds to annotate the time at which gestures occur. Finally tracking algorithms require the labeling of the positions of the body-parts of interest in all frames, also called spatial segmentation. Temporal and spatial segmentation may involve several levels of accuracy. Temporal segmentation can be provided as an ordered list of appearance of the gestures or as accurate start and stop timestamps. Similarly, spatial segmentation can be provided as an approximate position of body-parts using bounding boxes or as an accurate position in the 2d/3d space. Bounding boxes are generally used for body-parts detection and segmentation while accurate positions are used to evaluate tracking algorithms. This categorization appears on both tables; it is represented in Table 1 by the bolded terms in the description of the main purposes of the datasets and can be inferred from the three types of ground truth shown in Table 2. Temporal segmentation is provided for most of the datasets; although several datasets only provide the gesture ordering. Spatial segmentation is rarely provided and when provided, it is generally only for a small percentage of the data. The 6DMG dataset provides an accurate spatial segmentation which has been acquired using an optical tracker. This approach is generally not considered valid when acquired concurrently with video streams due to visual artifacts on the images resulting from markers attached to the subject.

4 Discussion

The number of datasets released in the domain of gesture recognition has largely increased these last years, simultaneously with the regain of interest for human gesture recognition. The transition from color cameras and stereo cameras to single multi-modal sensors capable of providing color and depth images and body-joint position is clearly visible in Table 2. Although the number of citations may seem a good indication of the popularity of a dataset most of the reviewed papers introducing a dataset are focused on novel recognition algorithms rather than the dataset itself. This tends to bias the number of citations about the dataset itself. Most of the reviewed datasets have been developed to explore one or several specific contexts; general-interest datasets have currently not been explored. These contexts can concern the type of gestures involved: gaming, iconic, metaphoric, deictic or sign language; the types of algorithms that can be applied: static or dynamic gesture recognition, one-shot learning, spotting, body-part segmentation or tracking or the type of input data: implicit or explicit, depth, color, body-joint position or acceleration data. The type of ground truth available for each dataset is related to the intended algorithm(s) and on the available “man-power” dedicated to manage the dataset. Indeed, ground truthing of datasets remains problematic. In theory and practice, a dataset is considered better if it contains more annotations. However, the ground truthing task is generally performed

manually by one or more expert annotators and may consume a lot of time and/or money depending on the amount of data to annotate and the precision level of the desired annotations. Some automatic, semi-automatic and crowd-sourced systems and methods are being explored to solve this problem; however first results tend to show problems in accuracy [34]. Notably, temporal segmentation and spatial segmentation of body-parts can be particularly costly to provide. Note that accurate spatial segmentation can be provided automatically using expensive and cumbersome motion capture systems at the cost of visual artifacts in video streams. The Skeleton data from the Kinect has been used and considered as a marker-free tracking system in a research paper based on MSRC-12. Although this can be valid for an approximate study of motion [19], the problems of accuracy and lost-of-tracking should not be neglected when evaluating tracking algorithms.

Another interesting and surprising information than can be observed from the reviewed datasets is the limited number of multi-sensors datasets; only two datasets contains multiple sensors: 6DMG contains inertial and motion capture data thus providing both implicit and explicit data. Similarly, the ChAirGest contains data from two popular sensors (Kinect and IMU). Although having multiple sensors may require more development on the acquisition software and complexify the acquisition procedure, the added value to the dataset can be worth it and may lead to innovative research directions [35]. The Kinect sensor is a multimodal device in itself as it provides image, depth, approximate body-joints positions and sound which greatly reduces problems of synchronization between sensors. Additionally, comparison methods for the performance of algorithms based on multimodal data must be carefully designed and defined. A discussable example is the ChaLearn 2013 challenge, which was relying on all modalities provided by a Kinect sensor. The best results of the challenge have been obtained by algorithms relying mostly on speech although the task was to recognize the gestures [22]. Even if this is not incorrect, it illustrates the importance of producing well designed vocabularies, datasets and tasks in order to prevent such shortcuts. Multimodal datasets also provide a way to prove quantitatively that some technologies, sensors, data or algorithms may be better suited for recognition than others depending on the conditions. Multimodal datasets for gesture recognition enable researchers to perform quantitative comparisons of modalities and combination of modalities on common data.

Most of the reviewed datasets have been first developed for internal projects and then released publicly. However datasets specifically and carefully designed for benchmarking and comparisons purposes gain more and more interest in the research community. This interest promotes challenges and workshops organized around datasets. Indeed, challenges provide a few advantages such as ensuring that participants can compare their results with a guarantee of validity and fairness and incentive for researchers to compete on similar data and goals.

5 Guidelines

This section contains the guidelines that have been developed to help researchers during the task of selecting or creating datasets.

Selecting a dataset that fits perfectly your needs is not a trivial task and often implies several considerations. Two approaches are distinguished in this paper: researcher and developer .A researcher usually needs a dataset for the evaluation of a new algorithm in order to prove its validity and performances compared to others. The developer usually needs data to provide a rapid and constant solution for testing and optimizing his platform and existing algorithms during the development phase, before starting the tests in real conditions. The following guidelines have been devised for researchers desiring to find a dataset suiting their needs.

- **Task**: The first selection depends on the task of the intended algorithm (recognition, spotting or tracking). Note that adding the missing ground truth information to a dataset might be feasible in certain cases and would probably be welcomed by any dataset author.
- **Requirements of algorithm**: an algorithm implementation generally relies on specific data and features which may be related to certain types of sensors or data types (body joint, depth information, acceleration, etc.).
- **Situation and interaction setup:** the interaction setup such as the position of video-based (front-view, top-view, etc.) and user conditions (standing, sitting, moving, etc.) must be clearly defined.
- **Types of gestures:** some gestures vocabularies may not be suited for all algorithms. Subtle gestures involving limited motion of hands and finger might yield problems for an algorithm initially intended for full-body gesture recognition.
- **Classes and instances:** a dataset with more classes is usually more interesting at the condition that it has enough instances of each class to train and validate the your algorithms. A dataset with many classes and very few instances is generally not usable for most machine learning algorithms.
- **Practical tests:** Researchers should download, when possible, small portion of the selected datasets and then visualize and test the data to take their final decision.

Once the selection finished, researchers should try to take advantage of all the potential of the dataset. When multiple recording conditions are available, the performances of the algorithm for each available condition should be evaluated. Specific evaluation metrics are often imposed, specifically in challenges; researchers should take this into account during the optimization of their performances. Similarly, challenges generally impose specific recognition task(s), if a developed algorithm does not fit exactly the task; researchers should not hesitate to contact the organizers as some alternative solutions can often be found.

Creating a dataset is also a complex task which involves many hours of work. The researcher creating a dataset should always keep in mind the possibility of releasing the dataset publicly at the end of his work. Indeed the time spent to record a dataset may quickly become very long and the dataset could be valuable to other researchers. The following brief guidelines should give an insight of the main tasks when creating a dataset.

- **Careful design**: The initial design of the dataset is very important. All the desired characteristics and recording conditions should be well defined and thought before

starting the implementation. The dataset should aim for novelties compared to existing datasets as previously outlined in this paper.

- **Software development**: Several frameworks provide tools to record simple datasets with standard sensors. For more complex scenarios, specific development is usually required. Several frameworks accept the addition of custom plugins.
- **Acquisition methodology**: the acquisition methodology should be accurately defined simultaneously with the software development. A well-defined methodology simplifies the acquisition process. Consider automatizing all the possible processes such as gathering of subjects data, labeling of conditions or ground truthing.
- **Acquisition**: The acquisition data is a time-consuming process. Before starting real acquisition with subjects, the setup should have been thoroughly tested several times in real conditions to ensure the validity of the final recordings. When possible, acquire the data with the highest possible quality and then convert it to lower quality for public release.
- **Annotation and Verification**: Once the dataset has been recorded, perform manual or automatic annotation and verifications on the data to ensure absence of errors. Finally apply a few well-known algorithms on the dataset before release it in order to provide a baseline to researchers.
- **Documentation**: A good documentation and description of the dataset is important for a public release of the dataset. The acquisition setup and the data should be precisely described.

6 Conclusions

This paper filled a void in the literature by providing a survey of the available datasets for the field of gesture recognition, a sub-domain of human actions and activity recognition. The survey provided a comprehensive description of the main publicly available datasets, exhibiting their characteristics, potential usage and highlighting their strengths and weaknesses through two tables. The categorization of the datasets provided a clear distinction between them. This distinction has been based on the usability of the datasets for the different algorithms involved in the gesture recognition. The survey and discussion also highlighted the current design space of the existing datasets and hinted at potential perspectives and challenges for the future datasets such as multimodal and multi-sensors approaches, automatic ground truthing methods and common standards. The discussion outlined the evolution of gesture recognition datasets and highlighted the importance of the presented characteristics through examples. The lack of documentation and information has also been highlighted as a major problem in most reviewed datasets. Finally, brief guidelines have been provided on the main notions and facts researchers should keep in mind when selecting or creating datasets for research.

Acknowledgements. This research has been supported by the HASLER foundation within the framework of “Living in Smart Environment” project.

References

1. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: 2011 Int. Conf. Comput. Vis., pp. 2556–2563 (2011)
2. Chaquet, J.M., Carmona, E.J., Fernández-Caballero, A.: A survey of video datasets for human action and activity recognition. Comput. Vis. Image Underst. 117, 633–659 (2013)
3. Computer Vision Index Dataset: http://riemenschneider.hayko.at/vision/dataset/index.php (accessed: October 23, 2013)
4. CV Datasets on the web: http://www.cvpapers.com/datasets.html (accessed: October 1, 2014)
5. Ruffieux, S., Mugellini, E., Lalanne, D., Khaled, O.A.: FEOGARM : A Framework to Evaluate and Optimize Gesture Acquisition and Recognition Methods. In: Work. Robust Mach. Learn. Tech. Hum. Act. Recognition; Syst. Man Cybern., Anchorage (2011)
6. Ruffieux, S., Lalanne, D., Mugellini, E.: ChAirGest: A Challenge for Multimodal Mid-Air Gesture Recognition for Close HCI. In: Proc. 15th ACM Int. Conf. Multimodal Interact. - ICMI 2013, pp. 483–488. ACM Press, Sydney (2013)
7. Xu, X., Tang, J., Zhang, X., Liu, X., Zhang, H., Qiu, Y.: Exploring techniques for vision based human activity recognition: methods, systems, and evaluation. Sensors (Basel) 13, 1635–1650 (2013)
8. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. 28, 976–990 (2010)
9. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis. ACM Comput. Surv. 43, 1–43 (2011)
10. Ahad, S., Tan, M.A.R., Kim, J., Ishikawa, H.: Action dataset—A survey. In: 2011 Proc., SICE Annu. Conf. (SICE), pp. 1650–1655 (2011)
11. Andriluka, M., Sigal, L., Black, M.J.: Benchmark Datasets for Pose Estimation and Tracking. In: Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (eds.) Vis. Anal. Humans. Springer, London (2011)
12. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Trans. Syst. Man, Cybern. Part C Appl. Rev. 37, 311–324 (2007)
13. Wachs, J.P., Kölsch, M., Stern, H., Edan, Y.: Vision-based hand-gesture applications. Commun. ACM. 54, 60 (2011)
14. Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Trans. on Pattern Anal. Mach. Intell. 19, 677–695 (1997)
15. Hasan, H., Abdul-Kareem, S.: Human–computer interaction using vision-based hand gesture recognition systems: A survey. Neural Comput. Appl. (2013)
16. Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., Ney, H.: Benchmark Databases for Video-Based Automatic Sign Language Recognition. In: Int. Conf. Lang. Resour. Eval., Marrakech, Morocco, pp. 1–6 (2008)
17. Athitsos, V., Wang, H., Stefan, A.: A database-based framework for gesture recognition. Pers. Ubiquitous Comput. 14, 511–526 (2010)
18. Glomb, P., Romaszewski, M., Opozda, S., Sochan, A.: Choosing and Modeling Hand Gesture Database for Natural User Interface. In: Proc. 9th Int. Gesture Work., Athens, Greece, pp. 72–75 (2011)
19. Fothergill, S., Mentis, H., Kohli, P., Nowozin, S.: Instructing people for training gestural interactive systems. In: Proc. 2012 ACM Annu. Conf. Hum. Factors Comput. Syst. - CHI 2012, p. 1737 (2012)

20. ARB Labs: http://www.arblabs.com/ (accessed: October 23, 2013)
21. Conly, C., Doliotis, P., Jangyodsuk, P., Alonzo, R., Athitsos, V.: Toward a 3D Body Part Detection Video Dataset and Hand Tracking Benchmark Categories and Subject Descriptors. In: Pervasive Technol. Relat. to Assist. Environ. (2013)
22. Escalera, S., Sminchisescu, C., Bowden, R., Sclaroff, S., Gonzàlez, J., Baró, X., et al.: ChaLearn multi-modal gesture recognition 2013. In: Proc. 15th ACM Int. Conf. Multimodal Interact. - ICMI 2013, pp. 365–368. ACM Press, New York (2013)
23. Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: 2009 IEEE 12th Int. Conf. Comput. Vis., pp. 444–451. IEEE (2009)
24. Bloom, V., Makris, D., Argyriou, V.: G3D: A gaming action dataset and real time action recognition evaluation framework. In: 2012 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., pp. 7–12 (2012)
25. Sadeghipour, A., Morency, L., Kopp, S.: Gesture-based Object Recognition using Histograms of Guiding Strokes. In: Procedings Br. Mach. Vis. Conf. 2012, British Machine Vision Association, pp. 44.1–44.11 (2012)
26. Liu, L., Shao, L.: Learning Discriminative Representations from RGB-D Video Data. In: Proc. Int. Jt. Conf. Artif. Intell. (2013)
27. Chen, M., AlRegib, G.: A new 6d motion gesture database and the benchmark results of feature-based statistical recognition. Emerg. Signal Process, 131–134 (2012)
28. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: Signal Process. Conf., pp. 1975–1979 (2012)
29. Guyon, I., Athitsos, V., Jangyodsuk, P., Hamner, B., Escalante, H.J.: ChaLearn Gesture Challenge: Design and First Results. In: IEEE Conf. Comput. Vis. Pattern Recognit. Work., pp. 1–6. IEEE (2012)
30. Song, Y., Demirdjian, D., Davis, R.: Tracking Body and Hands for Gesture Recognition: NATOPS Aircraft Handling Signals Database. In: Proc. IEEE Int. Conf. Autom. Face Gesture Recognit., pp. 500–506. IEEE, Santa Barbara (2011)
31. Ren, Z., Meng, J., Yuan, J., Zhang, Z.: Robust hand gesture recognition with kinect sensor. In: Proc. 19th ACM Int. Conf. Multimed. - MM 2011, p. 759. ACM Press, New York (2011)
32. Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Thangali, A.: The American Sign Language Lexicon Video Dataset. In: 2008 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work, pp. 1–8. IEEE (2008)
33. Kim, T.-K., Wong, S.-F., Cipolla, R.: Tensor Canonical Correlation Analysis for Action Classification. In: 2007 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1–8 (2007)
34. Ruffieux, S., Lalanne, D., Mugellini, E., Abou Khaled, O.: Gesture Recognition Corpora and Tools: A Scripted Ground Truthing Method, Publ. Submitt. to J. Comput. Vis. Image Underst. (2014)
35. Banos, O., Calatroni, A., Damas, M., Pomares, H., Rojas, I., Sagha, H., et al.: Kinect=IMU? Learning MIMO Signal Mappings to Automatically Translate Activity Recognition Systems across Sensor Modalities. In: IEEE 2012 16th Int. Symp. Wearable Comput., pp. 92–99 (2012)