When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets

Conference Paper · October 2017

DOI: 10.1109/ITSC.2017.8317828

CITATIONS
5

READS
3,270

2 authors, including:

Hang Yin
Zenuity
25 PUBLICATIONS **117** CITATIONS

SEE PROFILE

When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets

Hang Yin
Department of Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden
Email: yhang@chalmers.se

Christian Berger
Department of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden
Email: christian.berger@gu.se

***Abstract*—Data collection on public roads has been deemed a valuable activity along with the development of self-driving vehicles. The vehicle for data collection is typically equipped with a variety of sensors such as camera, LiDAR, radar, GPS, and IMU. The raw data of all sensors is logged on a disk while the vehicle is manually driven. The logged data can be subsequently used for training and testing different algorithms for autonomous driving, e.g., vehicle/pedestrian detection and tracking, SLAM, and motion estimation. Data collection is time-consuming and can sometimes be avoided by directly using existing datasets including sensor data collected by other researchers. A multitude of openly available datasets have been released to foster the research on automated driving. These datasets vary a lot in terms of traffic conditions, application focus, sensor setup, data format, size, tool support, and many other aspects. This paper presents an overview of 27 existing publicly available datasets containing data collected on public roads, compares each other from different perspectives, and provides guidelines for selecting the most suitable dataset for different purposes.**

I. Introduction

Research and development of self-driving vehicles has been gaining momentum in the automotive industry over the last decade. There are a multitude of existing algorithms covering various aspects of automated driving, such as lane following, object detection, semantic understanding, localization, SLAM (simultaneous localization and mapping), motion control, and V2X communication. Furthermore, these algorithms are frequently revised, extended, and replaced with new ones that roll out at a remarkably high pace. Testing such algorithms on real vehicles is often a costly (and sometimes even dangerous) task. As machine learning based algorithms are prevailing for automated driving in the recent years, enormous amount of driving data is in demand for training and testing purposes, making data collection on public roads a valuable activity.

Data collection is typically conducted by a manually driven vehicle equipped with a set of sensors and other automotive devices such as camera, LiDAR (Light Detection and Ranging), radar, GPS, and IMU (inertial measurement unit). The sensor/device configuration (e.g., selected sensors, number and placement of each sensor, parameter configuration of each sensor) may vary depending on how the collected data is planned to be used. For instance, a camera can be monocular, stereo, or omnidirectional, grayscale, color, and HDR (high dynamic range). Viewing perspective, image resolution, and frame rates are the key parameters of a camera to be configured before a data collection. During data collection, the raw data of different sensors is logged on disk with time synchronization and timestamps. Data collected on public roads is of particular interest for automated driving by virtue of realistic traffic scenarios reflected from the data. LiDARs and vision-based automotive sensors can produce large amounts of data with high frequency. Hence, the resultant data size can reach the magnitude of tens or even hundreds of gigabytes.

Data collection on public roads is extremely time-consuming and tedious. Fortunately, there are a lot of existing driving datasets containing data collected by other researchers and engineers. Many of those datasets are openly accessible and well documented. The training and testing of new self-driving algorithms can be facilitated by directly using suitable existing datasets. These datasets differ from each other in various aspects such as data size, sensor setup, and data format. This raises a key pertinent question: Which datasets are suitable for what self-driving algorithms? When a new self-driving algorithm is ready to be tested using an existing dataset, one needs to know not only the list of datasets available for use but also which dataset is suitable for the algorithm. To the best of our knowledge, there are still no scientific publications which provide a comprehensive summary of existing driving datasets on public roads. Very recently Janai et al. [1] published a survey paper on a broad spectrum of datasets. However, those datasets are related to computer vision in general, while only a few fit the scope of self-driving. With a particular focus on training and testing self-driving algorithms, this paper presents an overview of 27 existing publicly available datasets containing data collected on public roads, compares each other from different perspectives, and provides guidelines for selecting the most suitable dataset(s) for different purposes.

The rest of the paper is organized as follows: Section II describes the applied methodology to collect the 27 existing datasets. Section III briefly summarizes the collected datasets,

while Section IV analyzes and compares these datasets. Finally, Section V summarizes the contribution of the paper and envisions future work.

II. Methodology on dataset collection

Our methodology to find and collect existing datasets is based on Google search and snowballing. Google does not guarantee a systematic search that can be achieved by scientific databases such as IEEExplore and ACM Digital Library. However, our prime concern is dataset web pages whereas not all datasets provide associated scientific publications. Datasets were collected in four sequential phases:

1) Direct Google search: The search string "driving dataset" was used in the Google search engine to find the most popular dataset web pages. Since it is unrealistic to traverse all the search results, we only considered the top 200 results, i.e., the first 20 pages. Actually, most dataset web pages were found among the top 100 results, while no new datasets were discovered already after 150 results.
2) Snowballing among dataset web pages: The dataset web pages found in the first phase were thoroughly examined. Some dataset web pages include a reference to other related dataset pages, thus allowing us to identify additional datasets not covered in the first phase.
3) Publication collection: Most datasets have one or multiple associated scientific publications listed on their web pages. In this phase, we collected the major scientific publications associated with the datasets found in the first two phases.
4) Backward snowballing with publications: The related work of each collected publication was explored to find more relevant publications associated with new datasets.

Throughout the four phases above, we came across a variety of datasets irrelevant to self-driving. We set the following inclusion criteria to select only relevant driving datasets:

- The dataset must contain data from on-board sensors collected by a vehicle running on public roads.
- The dataset must contain camera, LiDAR, or radar data. Otherwise, it is considered to be of insignificant value for training and testing self-driving algorithms.
- The dataset must allow full/partial open access.

The inclusion criteria excluded a number of datasets with simulated data in virtual world, data collected indoors, or data collected in confined areas such as parks and campuses. One dataset was excluded because it contained GPS data only. The first two phases resulted in 11 datasets, which were subsequently extended to 26 datasets after phase 4 was completed; in addition, the Stanford track collection [2] was also included. Eventually, 27 datasets have been collected and will be presented in the next section.

III. Overview of the collected datasets

In this section, a concise introduction highlighting the most distinctive characteristics of the 27 datasets collected by us is given in alphabetic order. A short alias based on the full name of each dataset is indicated in parentheses. In the rest of the paper, the alias will be used whenever a dataset is referenced.

- *Dataset 1: Automotive multi-sensor dataset (AMUSE) [3]* (https://goo.gl/1YbD5E)
  Provider: Linköping University, Sweden
  Highlight: omnidirectional visual data for full surround sensing; include winter conditions with snow
- *Dataset 2: Caltech Pedestrian Detection Benchmark (Caltech) [4]* (https://goo.gl/07Us6n)
  Provider: California Institute of Technology, US
  Highlight: the largest pedestrian dataset on public roads; pedestrian annotation; the first dataset with temporal correspondence between bounding boxes and occlusion labels
- *Dataset 3: Cambridge-driving Labeled Video Database (CamVid) [5]* (https://goo.gl/I2pbdP)
  Provider: University of Cambridge, UK
  Highlight: the first collection of videos with object class semantic labels; pixel-level annotation
- *Dataset 4: CCSAD dataset [6]* (https://goo.gl/pxr3Yc)
  Provider: Centro de Investigacin en Matemticas, Mexico
  Highlight: stereo video captured in developing countries
- *Dataset 5: Cheddar Gorge Dataset [7]*
  Provider: BAE Systems (Operations) Limited, UK
  Highlight: diversified sensor setup with stereo, monocular, and infrared cameras, Velodyne 64 LiDAR, GPS/IMU etc.
- *Dataset 6: Cityscapes dataset [8], [9]* (https://goo.gl/qLM3V4)
  Provider: Daimler AG R&D, Germany; Max Planck Institute for Informatics (MPI-IS), Germany; TU Darmstadt Visual Inference Group, Germany
  Highlight: stereo sequences from 50 cities; pixel-level annotation for semantic urban scene understanding; benchmark suite with an evaluation server
- *Dataset 7: CMU Visual Localization Dataset (CMU)* (https://goo.gl/0R8XX6)
  Provider: Carnegie Mellon University, US
  Highlight: various weather and light conditions
- *Dataset 8: comma.ai driving dataset (comma.ai) [10]* (https://goo.gl/B3TWf2)
  Provider: comma.ai
  Highlight: highway traffic scenarios
- *Dataset 9: Daimler Pedestrian Benchmarks (Daimler pedestrian)* (https://goo.gl/l3U2Wc)
  Provider: Daimler AG R&D, Germany; University of Amsterdam, the Netherlands
  Highlight: encompass multiple benchmark datasets for pedestrian detection, classification, segmentation, and path prediction based on monocular and stereo images; the first dataset with partially occluded pedestrians; include the only cyclist dataset [11] that we have found so far
- *Dataset 10: Daimler Urban segmentation (Daimler urban) [12]* (https://goo.gl/KRBCLa)
  Provider: 6D-Vision, Germany
  Highlight: stereo vision sequences in urban traffic; pixel-level semantic class annotation
- *Dataset 11: DIPLECS Autonomous Driving Datasets (DIPLECS) [13]* (https://goo.gl/8isjeJ)
  Provider: University of Surrey, UK

Highlight: include two datasets on public roads, one in UK, the other in Sweden; labeled frame by frame with speed and steering data (the first dataset) and driving environments and driver actions (the second dataset)

- *Dataset 12: Dr(eye)ve [14]* (https://goo.gl/45bwXr)
  Provider: ImageLab, Italy
  Highlight: the first dataset for researching driver attention: human attention, eye fixation, and visual saliency
- *Dataset 13: EISATS [15]* (https://goo.gl/ausKsL)
  Provider: University of Auckland, New Zealand; Daimler AG, Germany; Hella Aglaia Mobile Vision GmbH, Germany; HU Berlin, Germany
  Highlight: include multiple datasets with stereo vision sequences for comparative performance evaluation of stereo vision, optic flow, motion analysis etc.
- *Dataset 14: Elektra* (https://goo.gl/GNNq0f)
  Provider: Autonomous University of Barcelona, Spain; Polytechnic University of Catalonia, Spain
  Highlight: various types of images with annotated pedestrians; include far infrared images
- *Dataset 15: ETH pedestrian dataset [16]* (https://goo.gl/xXDTwI)
  Provider: ETH Zürich, Switzerland
  Highlight: stereo images captured in a crowed city center with many pedestrians
- *Dataset 16: Ford Campus Vision and Lidar Data Set (Ford) [17]* (https://goo.gl/6ZkCpc)
  Provider: University of Michigan, US
  Highlight: diversified sensor setup, including high precision localization devices, multiple LiDARs, omnidirectional camera etc.; full software support
- *Dataset 17: German Traffic Sign Detection Benchmark (German traffic sign)* (https://goo.gl/FqaCJQ)
  Provider: Ruhr University Bochum, Germany
  Highlight: still images with traffic signs in Germany
- *Dataset 18: Heidelberg benchmarks (Heidelberg) [18]* (https://goo.gl/6c2lAs)
  Provider: Heidelberg University, Germany
  Highlight: associated with an event called Robust Vision Challenge; provide challenging data for stereo and optical flow, e.g., rain flares and flying snow
- *Dataset 19: Joint Attention for Autonomous Driving Dataset (JAAD) [19]* (https://goo.gl/cXoPnp)
  Provider: York University, Canada
  Highlight: focus on joint attention between pedestrians and drivers for autonomous driving; provide both textual and behavioral annotations for pedestrians and vehicles
- *Dataset 20: Karlsruhe Dataset: Labeled Objects (Karlsruhe labeled objects) [20]* (https://goo.gl/5fk0js)
  Provider: MPI-IS
  Highlight: images with object bounding boxes for cars and pedestrians; include even object orientation
- *Dataset 21: Karlsruhe Dataset: Stereo Video Sequences + rough GPS Poses (Karlsruhe stereo) [21]* (https://goo.gl/V6Q7Vx)
  Provider: MPI-IS
  Highlight: high-quality stereo sequences in Karlsruhe
- *Dataset 22: KITTI Vision Benchmark Suite (KITTI) [22], [23]* (https://goo.gl/cvSbGl)
  Provider: Karlsruhe Institute of Technology, Germany; Toyota Technological Institute, US
  Highlight: the current most prestigious dataset for self-driving; provide a number of excellent benchmarks for the evaluation of stereo vision, optical flow, scene flow, visual odometry, SLAM, object detection and tracking, road/lane detection, semantic segmentation
- *Dataset 23: Málaga Stereo and Laser Urban Data Set (Málaga) [24]* (https://goo.gl/EdLHtW)
  Provider: University of Málaga, Spain
  Highlight: well-documented; full tool support; message board embedded on its web page for convenient communication
- *Dataset 24: Oxford robotcar dataset (Oxford) [25]* (https://goo.gl/nJOQkq)
  Provider: Oxford University, UK
  Highlight: the first dataset stressing periodic long-term data collection (over a year) following predefined routes to cover long-term changes of road conditions
- *Dataset 25: Stanford track collection (Stanford) [2]* (https://goo.gl/KNOYpX)
  Provider: Stanford University, US
  Highlight: Velodyne 64 point cloud with object labels and GPS/IMU data
- *Dataset 26: Ground Truth Stixel Dataset (Stixel) [26]* (https://goo.gl/rf12z6)
  Provider: 6D-Vision, Germany
  Highlight: heavy rain on highways; stixel annotation
- *Dataset 27: Udacity dataset* (https://goo.gl/AoxEt1)
  Provider: Udacity
  Highlight: open source project; driving data with or without annotation: annotated objects (even including traffic lights)

IV. Dataset comparison and selection guidelines

After studying the 27 datasets introduced in Section III with thorough attention, we extract the most important information of each dataset and summarize them in Table I and II, which can be conveniently consulted to get a quick overview of each dataset and compare different datasets from various perspectives. The summary tables jointly present the following aspects of each dataset: (1) time and venue: when and where was the data collected? (2) data size; (3) traffic conditions during the data collection; (4) sensor setup; (5) data format; (6) provided resources (e.g., raw data, annotation, benchmark, open source code, and tool support). Due to limited space, the tables do not include the dataset URLs and dataset providers, which are already given in Section III.

We exerted ourselves to make the summary tables as complete as possible, despite a few empty cells due to insufficient available information of certain datasets. Later in this section, we shall supplement Table I and II by discussing the license, accessibility, and popularity of these datasets.

The summary tables indicate that most datasets have their data collected after 2009. This implies that data collection on

public roads became active soon after research and development of self-driving was significantly fostered by the 2004 & 2005 DARPA Grand Challenge [27] and the 2007 DARPA Urban Challenge. We also witness a growing trend in running data collections, as 7 out of 27 datasets were released in 2016.

As shown in the *Time & Venue* column, the data collection venue is consistent with the location of the dataset provider for most datasets. In other words, most dataset providers collected data in their own cities. Nevertheless, some datasets do not explicitly tell where the data was collected and sometimes the data collection venue can be different from the provider's city. We thus made strenuous efforts to figure out the actual data collection venue of certain datasets. For instance, the venues of CamVid and ETH pedestrian were eventually confirmed by downloading sample video sequences from their web pages and comparing the recorded scenes with Google Maps street views. Most datasets have a single city as the only venue, while a few datasets such as Cityscapes, DIPLECS, ESATS, Elektra, and JAAD contain data collected from multiple places.

The geographical distribution of data collection venues among the 27 datasets is illustrated in Fig. 1. Apparently, most data were collected in Europe and the US. Germany is the most active country for running data collection. These data collection venues only cover a tiny portion of the world map. We strongly urge the future release of new datasets from other continents outside Europe and the US. The discrepancy between traffic conditions in different continents and countries necessitates global driving data on public roads, which would contribute to making future self-driving algorithms more robust and less dependent on geographical regions. Moreover, most datasets were attributed to single organizations, reflecting a lack of global coordination in terms of data collection. We believe that the synergy between multiple organizations would bring higher quality datasets.

In terms of dataset size, the Oxford robotcar dataset stands out with over 23TB. Even the average size among these datasets is still rather large. Most datasets partition their data into different files and categories for separate download, as a dataset user may be interested in only a subset of the provided data. The comma.ai dataset is an exception in the sense that its data (11 sequences) is compressed into a single zip file which must be downloaded as a complete chunk.

For each dataset, we also investigated the traffic condition reflected from the collected data, including the type of traffic (e.g., urban traffic, rural road, highway), light condition (e.g., daylight, night), and weather condition (e.g., sunny, overcast, rainy). A majority of the 27 datasets focuses on urban traffic, daylight, and sunny weather. Driving data collected under perfect light and weather conditions are ideal for training and testing self-driving algorithms. However, sometimes driving data collected under adverse conditions is desired for more robust algorithms. Such data can be accessed from a number of datasets: AMUSE, CCSAD, CMU, Dr(eye)ve, ESATS, Elektra, Heidelberg, JAAD, Oxford, and Stixel. Datasets such as Caltech and Daimler pedestrian contain only urban traffic on account of their inherent nature as there are more pedestrians and cyclists in urban traffic. In contrast, comma.ai and Stixel focus more on highway traffic.

The sensor setup varies among these datasets. Almost all datasets include at least one type of camera for the data collection except for Stanford track collection, which contains LiDAR data instead complemented with GPS/IMU data. We observe a balanced use of monocular and stereo cameras while the color option is slightly preferred to grayscale. Omnidirectional cameras are only used in AMUSE and Ford. In addition to visual data, location data and vehicle motion data are also deemed important for self-driving vehicles. 12 datasets include GPS in their sensor setup, mostly in combination with IMU. LiDAR sensors are used in 8 datasets, with Velodyne 64 and Sick as the most popular models. It is interesting to note that certain types of sensors only exist in a single dataset, including the monocular infrared camera in Cheddar Gorge, the driver's eye tracking device to capture driver fixation in Dr(eye)ve, and the far infrared sensor in one subset of Elektra. No radar sensors have been reported in the 27 datasets, though radar sensors are still essential to self-driving vehicles. Our conjecture is that many existing radar sensors adopt proprietary data formats which cannot be shared easily.

These datasets exhibit miscellaneous file types resulted from raw data, annotations, calibration files, labels, etc. Most datasets share data in standard formats. AMUSE, one subset of Elektra, Málaga, and Stanford contain own data formats which can be parsed by the example code or tools that were released alongside the data. It seems that dataset providers are more inclined to share vision data from the cameras as separate and consecutive image files instead of video files.

The resources provided by most datasets are much more than the raw data from on-board sensors, which are often accompanied with lots of additional supplements to maximize the usability of these datasets, such as annotations and labels (e.g., object bounding boxes), benchmark suites, open source code, tools and scripts, scientific publications and demo videos. It is common to post-process the raw data, in particular visual data, and classify them into different sets. For instance, some datasets (KITTI, Málaga, and Heidelberg) provide both the original and rectified images while some datasets (Caltech, Cityscapes, Daimler pedestrian, Elektra, German traffic sign, Stanford) classify images into training, testing, and sometimes even validation sets. Caltech, Cityscapes, German traffic sign, and KITTI offer benchmarks that serve as an open platform where other people can upload the evaluation results of their own algorithms so as to get a ranking in comparison with other algorithms of similar types. This feature is extremely rewarding and appreciated, as it gives the opportunity to compare the performance of different algorithms. Another appealing feature that we notice is unique for Málaga. The web page of Málaga embeds a message board for visitors to leave messages, thereby enabling more efficient communication between the dataset provider and potential users.

As mentioned in Section II, we acquired a large number of publications in the third phase of dataset collection. These publications indicate that numerous prospective dataset users

TABLE I
OVERVIEW OF EXISTING DRIVING DATASETS ON PUBLIC ROADS — PART 1)

Dataset	Time & Venue	Size	Traffic condition	Sensors	Data format	What is provided
AMUSE	Feb-March, 2013 Linköping (Sweden)	1,169GB (7 sequences)	loop, closing, (nearly) static scene, snow, suburb, urban, low altitude of sun, water and snow on lens	omnidirectional camera; GPS+IMU; velocity sensor; 3 height sensors	png: image; liu: own format	partial raw data; API for C/C++, Python, Matlab; ROS support
Caltech	before May, 2009 Los Angeles (US)	ca 11GB	urban	monocular color camera	seq: video; vbb: bounding box	videos(training/ testing sets); annotation; benchmark results; Matlab code
CamVid	before 2009 Cambridge (UK)	ca 8GB (4 videos)	urban	monocular color camera	png: labeled image; mxf: video; avi: video	video+label; png image extraction tool; Matlab code; paint strokes during labeling
CCSAD	May-Jul, 2014 Guanajuato (Mexico)	ca 500GB (42 sequences, total 1h 20min)	urban, small roads, tunnel at night, varying light conditions	stereo vision, grayscale; attitude and heading reference system; GPS-enabled smartphone	png: image; txt: timestamp, GPS, IMU, and vehicle data; xml: calibration	raw data
Cheddar Gorge	2010-03-05 Cheddar Gorge (UK)	329GB (57min)	dry, sunny, clear, cold	stereo vision, color; monocular color camera; monocular infrared camera; Velodyne 64 LiDAR; professional GPS/IMU; low cost IMU; 4 wheel distance encoders ; laser tracker for sensor pose measurements	standard formats (no description)	raw data
Cityscapes	before 2016-02-20 Germany, Zürich (Switzerland), Strasbourg (France)	63.141GB, 5 files, more data available upon request	daytime, no adverse weather conditions	stereo vision, color	png: image; json: annotation	images (training/ validation/test sets); pixel-level annotation; coarse annotation; benchmark suite; evaluation server; scripts
CMU	from 2010-09-01 to 2011-09-02 Pittsburgh (US)	16 sequences, 11-22GB each	various weather/light conditions	3 monocular color cameras; 4 Sick LiDARs; GPS+IMU	jpg: image; txt: LiDAR, GPS, and vehicle data	raw data
comma.ai	from 2016-01-30 to 2016-06-08 San Francisco (US)	80GB, 11 sequences, 7.25h	daylight, mostly highway	monocular color camera; GPS+IMU; gyroscope	HDF5	raw data; open source code
Daimler pedestrian	2006-2016 Beijing(China), others unknown	8 datasets: 53MB, 10-15GB, 10GB, 12GB, 6GB, 300MB, 2.5MB, 45GB	urban	monocular grayscale camera; stereo vision, color or grayscale	png: image; pgm: image; mat: image; ...	raw/processed data (training/ testing/validation sets); annotation;
Daimler urban	2014	7.55GB, 5000 stereo image pairs, 1024*440	urban	stereo vision, grayscale	pgm: image, ground truth label, disparity map; xml: camera calibration, vehicle data	video; pixel-level annotation; ego-motion data; disparity map; development kit
DIPLECS	2015 Surrey (UK), Stockholm (Sweden)	4.29GB+1.06GB	country road, highway, urban	Surrey: monocular color camera Stockholm: 3 monocular grayscale cameras	Surrey: mp4: video; txt: vehicle data Stockholm: avi: video; dat: label	Surrey: raw data Stockholm: video; frame label
Dr(eye)ve	2016 Modena (Italy)	74 sequences (5min each)	urban, countryside, highway; sunny, cloudy, rainy; morning, evening, night	monocular color camera; driver's eye tracking device	avi: video; txt: driver fixation, vehicle data, annotation; png: image	raw data; annotation
ESATS	2007-2010 Stuttgart+Lippstadt (Germany), Auckland (New Zealand)	8 relevant subsets: Set 1, 3, 4, 5, 6, 7, 9, 10: 525MB, 12GB, 705MB, 200MB, 1GB, 3MB, 5.7GB, 1.54GB (estimated)	highway, rural, urban; various weather/light conditions, adverse conditions	2-3 monocular grayscale/color cameras	pgm: grayscale image; ppm: color image; jpg/bmp: image; binary data	raw data
Elektra	Apr-Jun, 2016 Barcelona (Spain)	6 relevant subsets: CVC 01, 02, 05, 08, 09, 14: 86.9MB, 2.44GB, 280MB, 2.11GB, 1.92GB, 3.48GB	urban, mostly daylight; night sequences in one dataset	monocular grayscale or color camera; stereo vision, color; far infrared sensor	png: image; pts: 3D points (own format)	raw data (training/testing sets); annotation

TABLE II
OVERVIEW OF EXISTING DRIVING DATASETS ON PUBLIC ROADS — PART 2)

Dataset	Time & Venue	Size	Traffic condition	Sensors	Data format	What is provided
ETH pedestrian	2009 Zürich (Switzerland)	660MB	downtown	2 monocular color cameras	png: image; cal: calibration; idl: annotation	raw images; calibration; annotation; demo videos
Ford	Nov-Dec, 2009 Michigan (US)	ca 100GB	downtown, loop closure, campus	Velodyne 64 LiDAR; omnidirectional camera; 2 Riegl LMS-Q120 LiDARs; Applanix+Trimble GPS; Xsens consumer IMU	mat: Velodyne scan; ppm: image; log: sensor data and timestamp; pcap: Velodyne stream; mat: calibration;	raw data; C and Matlab code
German traffic sign	available from 2012-12-01 Germany	1.6GB		monocular color camera	ppm: image; csv: annotation	images(training/ evaluation sets); annotation; C++/Matlab code; benchmark
Heidelberg	2011-2012 Hildesheim (Germany)	10TB in total; ca 12.6GB available	city/avenue/bend/ parking/village, various lighting/ weather conditions	2 monocular grayscale cameras (stereo vision)	pgm: image; png and h5: image; avi: demo video	raw data
JAAD	Before Nov, 2016 Toronto(Canada) Kremenchuk+ Lviv(Ukraine), Hamburg (Germany), New York(US)	347 video clips (5-15s each)	mainly urban, a few rural roads, most daytime, occasional night, sunset and sunrise, various weather conditions	monocular color camera	mp4: video; seq: video; vbb and tsv: textual annotation; xml: bounding box annotation	videos; textual and bounding box annotations; bash script for splitting videos
Karlsruhe labeled objects	2011 Karlsruhe (Germany)	631.2MB; ca 1800 images with labels	urban, daylight	monocular grayscale camera	png: image; mat: label	images; object labels; object orientation
Karlsruhe stereo	2009-2010 Karlsruhe (Germany)	20 sequences; 02-1.4GB each	urban, rural, daylight	stereo vision, grayscale; GPS+IMU	png: image; txt: GPS/IMU data	raw data; camera calibration
KITTI	Sep-Oct, 2011 Karlsruhe (Germany)	180GB	urban, rural, highway	2 monocular grayscale cameras; 2 monocular color cameras; Velodyne 64 LiDAR; GPS+IMU	png: image; txt: Velodyne and GPS/IMU data, calibration; xml: bounding box label	raw data; object annotation (3D bounding box); calibration; various benchmarks: stereo, optical flow, visual odometer, SLAM, 3D object detection/ tracking; development kit; Matlab/C++ code
Málaga	Before Feb, 2014 Málaga (Spain)	>70GB; 15 sequences; 93min	urban, highway, loop closure, direct sun etc.	stereo vision, color; 3 Hokuyo UTM-30LX laser scanners; 2 Sick LiDARs; GPS+IMU	txt: raw laser scan, GPS/IMU data, camera calibration; jpg: image; rawlog: binary log (own format); kml: Google earth file to represent path	raw data; C++ example code for parsing rawlog files; demo videos; support for posting public messages by users
Oxford	from 2014-05-06 to 2015-12-13 Oxford(UK)	23.15TB 133 sequences	various light/ weather conditions	stereo vision, color; 3 monocular color cameras; 2 Sick 2D LiDARs; Sick 3D LiDAR; GPS/INS	png: image; bin: LiDAR data; csv: GPS/INS data	raw data; calibration; Matlab/Python tools
Stanford	2009-2010 San Francisco (US)	5.72GB; 33 files	urban, campus, intersections	Velodyne 64 LiDAR; Applanix (GPS/IMU)	tm: Velodyne and Applanix data (own format)	raw data; background data without objects(training and testing sets); object labels; code in ROS package
Stixel	2013	3.1GB (zip); 12 sequences	highway, good weather, heavy rain	stereo vision, grayscale	pgm: image; xml: ground truth stixel	videos; ground truth stixel (own novel concept); vehicle data including timestamps
Udacity	Sep-Oct, 2016 Mountain View (US)+around	223GB; >10h	sunny, overcast, daylight	monocular color camera; Velodyne 32 LiDAR; GPS+IMU	png or jpg: image; log: GPS and vehicle motion; csv: label; ROSBAG	videos; labels: vehicle, pedestrian, traffic lights; open source code; tools for ROSBAG files

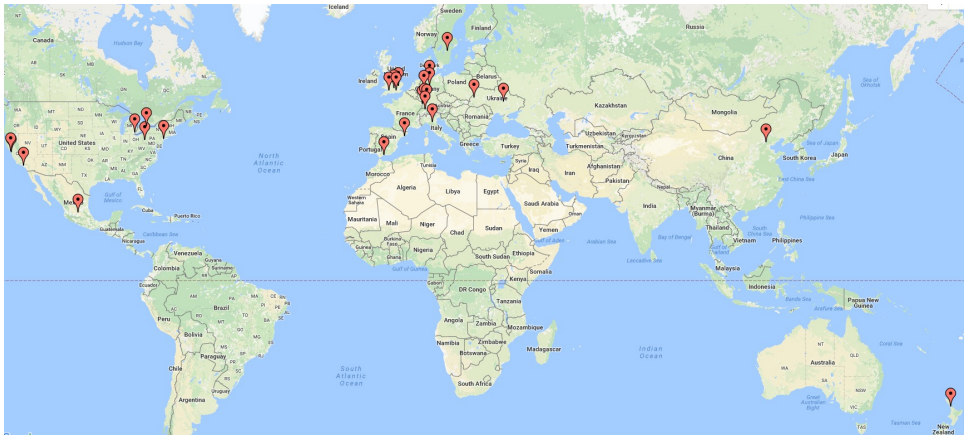

Fig. 1. Map distribution of venues for existing datasets acquisition (Image courtesy: ZeeMaps)

are from academia, thus underlining the importance of disseminating datasets via scientific publications. Publications allow a more technical and detailed description of any dataset. Furthermore, the fame of a dataset grows rapidly along with direct and indirect publication citations. We use number of citations as a metric to rank popularity among the 27 datasets. The number of citations of each dataset is computed as follows: First, we identify all the publications stated on the web page of each dataset. Then for each of these publications, we retrieve the number of citations. The number of citations of the dataset amounts to the total number of citations of all the publications. The popularity of the 27 datasets ranked by number of citations is illustrated in Fig. 2. KITTI, Caltech, and Daimler pedestrian are the top three most popular datasets that clearly outrank the remaining datasets. Note that German traffic sign and Udacity get no citation simply because we have not found any associated publications, whereas we do not exclude the possibility that these two datasets are already referenced in several existing publications. It is also worth to remark on the datasets with low citation numbers. Oxford, Dr(eye)ve, JAAD, and comma.ai were all released in 2016, which explains their low citation numbers. However, they all have the potential to become one of the most influential datasets in future.

Regarding the legal constraint of using these datasets, one third of these datasets have declared the licenses under which they were published. Creative Commons Attribution-NonCommercial-ShareAlike 3.0 is the most adopted license used by CMU, comma.ai, Karlsruhe labeled objects, Karlsruhe stereo, and KITTI. Elektra and Oxford go for Creative Commons Attribution-NonCommercial 4.0 License. AMUSE is licensed under Creative CommonsAttribution-NonCommercial-NoDerivs 3.0 Unported License. Udacity applies the MIT license for its data while everything else (e.g., code and tool) is licensed under GPLv3. The licenses of the other remaining datasets are still not indicated on their web pages.

Last but not least, data accessibility is a fundamental requirement. According to our last inclusion criterion specified in Section II for selecting datasets, any dataset presented in this paper must be fully or partially available for download. AMUSE is the only dataset which is partially available, while all the other 26 datasets are fully open. AMUSE prepares 6 short samples, with the total size of 39GB, for free direct download. The entire dataset (1,169GB) can be bought together with a hard disk for 4,000 SEK, including international shipping. Among the 26 fully open datasets, Cheddar Gorge is the least accessible (free) dataset, which is also the only dataset without a web page. The only information source is a technical report, which requires a dataset user to send email to james.revell@baesystems.com including the name of the researcher, organization, address and brief statement on how the data will be used. Once it is approved, one needs to send a signed copy of the license and a hard disk drive (at least 350GB) for them to copy the data. Cityscapes, Elektra, Heidelberg, and Dr(eye)ve require a valid email address to obtain the download links. Cityscapes is relatively more stringent in the sense that a data user must register an account with work email (private email is not accepted). Any new registration will be manually inspected and it takes a few days to get it approved. All the remaining datasets provide direct links for data access.

V. Conclusion

Driving data collected on public roads is a valuable resource for training and testing algorithms in the area of self-driving vehicles. This paper gathers 27 existing driving datasets on public roads that are openly accessible. These datasets are comprehensively analyzed and compared with regard to data collection time and venue, data size, traffic condition, sensor setup, data format, provided resource, popularity based on number of citations, license, and data accessibility. The comparison result serves as a guideline for dataset users to select the most suitable datasets for training and testing their automated driving algorithms based on their actual demand.

A threat to validity of this work is the completeness of the collected datasets. There may exist other relevant datasets that are overlooked. Nonetheless, the combination of Google search and snowballing definitely covers the most influential datasets. The web page and associated publications of each dataset contain a vast amount of information, while in this

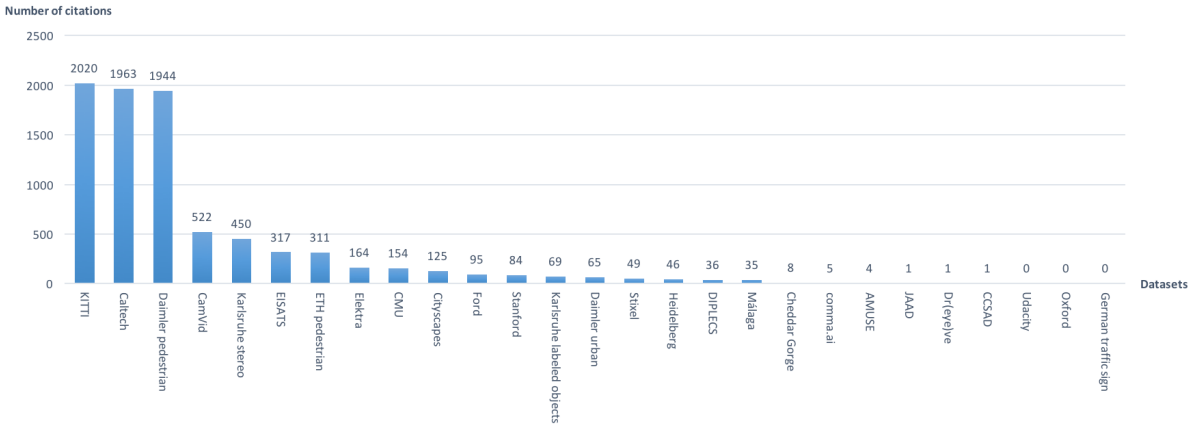

Fig. 2. Dataset popularity ranked by number of citations (retrieved on 2017-04-04)

paper we only present what we consider is most essential due to space limitation. However, other data users may be interested in other aspects of these datasets excluded in the paper, such as performance evaluation. An empirical study targeting prospective dataset users would complement our work to identify what is really wanted from a dataset. Another future work is to gain a deeper understanding of the data provided by different datasets, thereby rendering the dataset summary tables more complete and accurate.

ACKNOWLEDGMENT

This work has been supported by the COPPLAR Project, funded by Vinnova FFI, Diarienr: 2015-04849.

REFERENCES

[1] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art," *Journal of Photogrammetry and Remote Sensing*, Apr. 2017.

[2] A. Teichman, J. Levinson, and S. Thrun, "Towards 3D object recognition via classification of arbitrary object tracks," in *2011 IEEE International Conference on Robotics and Automation*, May 2011, pp. 4034–4041.

[3] P. Koschorrek, T. Piccini, P. berg, M. Felsberg, L. Nielsen, and R. Mester, "A Multi-sensor Traffic Scene Dataset with Omnidirectional Video," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 727–734.

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: A Benchmark," in *CVPR*, June 2009.

[5] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, pp. 88–97, 2009.

[6] R. Guzmán, J.-B. Hayet, and R. Klette, "Towards ubiquitous autonomous driving: The CCSAD dataset," in *16th International Conference on Computer Analysis of Images and Patterns*, Sep 2015, pp. 582–593.

[7] R. Simpson, J. Cullip, and J. Revell, "The cheddar gorge data set," BAE Systems (Operations) Limited, UK, Tech. Rep., 2011.

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," *CoRR*, vol. abs/1604.01685, 2016.

[9] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset," in *CVPR Workshop on The Future of Datasets in Vision*, 2015.

[10] E. Santana and G. Hotz, "Learning a Driving Simulator," *CoRR*, vol. abs/1608.01230, 2016.

[11] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila, "A new benchmark for vision-based cyclist detection," in *2016 IEEE Intelligent Vehicles Symposium (IV)*, June 2016, pp. 1028–1033.

[12] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth, *Stixmantics: A Medium-Level Model for Real-Time Semantic Scene Understanding*. Springer International Publishing, 2014, pp. 533–548.

[13] N. Pugeault and R. Bowden, "How Much of Driving Is Preattentive?" *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5424–5438, Dec 2015.

[14] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR(eye)VE: A Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2016, pp. 54–60.

[15] Z. Liu and R. Klette, "Performance evaluation of stereo and motion analysis on rectified image sequences," Computer Science Department, The University of Auckland, New Zealand, Tech. Rep., 2007.

[16] A. Ess, B. Leibe, K. Schindler, and L. van Gool, "Robust Multiperson Tracking from a Mobile Platform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1831–1846, Oct 2009.

[17] G. Pandey, J. R. Mcbride, and R. M. Eustice, "Ford Campus Vision and Lidar Data Set," *Int. J. Rob. Res.*, vol. 30, no. 13, pp. 1543–1552, Nov. 2011.

[18] S. Meister, B. Jähne, and D. Kondermann, "Outdoor stereo camera system for the generation of real-world benchmark data sets," *Optical Engineering*, vol. 51, no. 02, p. 021107, 2012.

[19] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint Attention in Autonomous Driving (JAAD)," *CoRR*, vol. abs/1609.04741, 2016.

[20] A. Geiger, C. Wojek, and R. Urtasun, "Joint 3D Estimation of Objects and Scene Layout," in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 2011, pp. 1467–1475.

[21] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d reconstruction in real-time," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, June 2011, pp. 963–968.

[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision Meets Robotics: The KITTI Dataset," *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[23] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 3354–3361.

[24] J.-L. Blanco, F.-A. Moreno, and J. González-Jiménez, "The Málaga Urban Dataset: High-rate Stereo and Lidars in a realistic urban scenario," *International Journal of Robotics Research*, vol. 33, no. 2, pp. 207–214, 2014.

[25] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017.

[26] D. Pfeiffer, S. Gehrig, and N. Schneider, "Exploiting the Power of Stereo Confidences," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 297–304.

[27] M. Buehler, K. Iagnemma, and S. Singh, *The 2005 DARPA Grand Challenge: The Great Robot Race*, 1st ed. Springer Publishing Company, Incorporated, 2007.