A Public Video Dataset for Road Transportation Applications

Nicolas Saunier (corresponding author)
Department of Civil, Geological and Mining Engineering, Polytechnique Montréal
nicolas.saunier@polymtl.ca

Håkan Ardö
Centre for Mathematical Sciences, Lund University
ardo@maths.lth.se

Jean-Philippe Jodoin
Department of Computer and Software Engineering, Polytechnique Montréal
jean-philippe.jodoin@polymtl.ca

Aliaksei Laureshyn
Transport and Roads, Lund University
Aliaksei.Laureshyn@tft.lth.se

Mikael Nilsson
Centre for Mathematical Sciences, Lund University
micken@maths.lth.se

Åse Svensson
Transport and Roads, Lund University
Ase.Svensson@tft.lth.se

Luis Miranda-Moreno
Department of Civil Engineering and Applied Mechanics, McGill University
luis.miranda-moreno@mcgill.ca

Guillaume-Alexandre Bilodeau
Department of Computer and Software Engineering, Polytechnique Montréal
guillaume-alexandre.bilodeau@polymtl.ca

Kalle Åström
Centre for Mathematical Sciences, Lund University
kalle@maths.lth.se

5293 words + 5 figures + 2 tables
November 15, 2013

ABSTRACT
Video data and the tools for automated analysis have a great potential to be used in road traffic research, particularly road safety. In this project a video dataset is built and made public so that researchers can evaluate their algorithms on it. The dataset focuses on the traffic research applications (data from real research projects) and provides recordings of the traffic scenes, meta-data, camera calibration, ground truth, protocols for comparing algorithms and software tools and libraries for reading/presenting the data. To the authors' knowledge, this public dataset is the first of its kind. With the proposed dataset, researchers get access to a large variety of recordings representing different traffic, weather and lighting conditions to evaluate and compare different tools and applications. As a consequence, discussions between computer vision and transportation researchers are expected to increase, contributing to more collaborations and better tools, more accurate and user-friendly, to obtain automatically rich traffic data from video.

INTRODUCTION

Video sensors have distinct advantages for road transportation applications. Compared to traditional road sensors, such as inductance loops, which can collect data only at a point along a roadway, video sensors cover relatively large areas. They can therefore be called spatial sensors. The main obstacle to their more widespread use in transportation applications is the availability of automated analysis methods to interpret video data. After the beginning of computer vision research in the 1960's, transportation applications have been developed since the 1980's for detection (*1*) and the 1990's for tracking (*2*). Commercial systems such as Autoscope (from companies like Econolite and Citilog) exist to replace traditional road traffic sensors to count and classify vehicles at specific locations as well as to detect incidents for simple environments such as highways. However, automatically detecting and tracking all road users in complex urban road environments such as intersections in all conditions is still an open problem. These urban road environments involve various types of road users (cars, buses, trucks, cyclists, pedestrians, etc.) at varying levels of density that may enter and exit the camera field of view through several zones, may turn, stop and park for varying amounts of time. Being able to detect, classify and track all road users in all environments would provide large amounts of trajectory data which is one of the most useful type of transportation data. Based on a literature review of 45 relatively recent traffic research articles in (*3*), 119 unique indicators were found to describe road users' behaviour. 86 % of the indicators can be calculated from road users' trajectories. The remaining 14 % describe the road users' personal characteristics (e.g. age, gender) and actions like head movements, eye contact or some informal signals. Automated video analysis may thus support a considerable share of all transportation applications, including the calibration and validation of microscopic models, studies of accessibility and livability of public spaces, and surrogate safety analysis, i.e. safety diagnosis based on the analysis of interactions without a collision such as conflicts (*4, 5*).

These applications require and benefit from large amounts of data: for example, since conflicts and collisions are rare events, data needs to be collected for various environmental and traffic conditions (e.g. weather and traffic flow, density and speed) in large enough quantities to allow statistical inference. The collection of large amounts of data is made possible, and is actually getting easier thanks to the falling price of video sensors, data collection devices and computer storage. Despite the interest for these applications and several research projects involving large amounts of data, no dedicated large public dataset of video data for transportation applications is available (large refers to several hours of video). This is a significant issue because there are few comparisons of different methods on the same datasets: comparisons are made on small datasets that cannot represent a wide variety of conditions and consequently few guidelines exist to choose, replicate, adapt and apply these methods. Hence, the general progress in this area is not clearly benchmarked, which limits future progress. The likely reasons include real and perceived privacy issues, the ownership of the video data (e.g. obtained through a third party like a transportation agency or a company), the overall lack of incentives for sharing data and the costs associated with creating a dataset, in particular the manual annotations relevant to the various applications that are needed to assess the performance of automated video analysis methods.

The goal of this paper is to remedy this situation by introducing the first public dataset of traffic video (PDTV) with annotations for transportation applications. In addition to the dataset, the paper presents an effort to standardize the description, or meta-data, and the annotations of the dataset and tools to facilitate the use of the data. Sample applications such as tracking are also demonstrated.

This paper also serves as a companion to the workshop organized at the 2014 Transportation Research Board Annual Meeting on the "Comparison of Surrogate Measures of Safety Extracted from Video Data" (see call for participation and details at `http://nicolas.saunier.confins.net/trb14workshop.html`). The objectives of the workshop are to promote this dataset and the habit of comparing methods on common datasets, tasks and metrics. Researchers and practitioners are invited to test their methods for video and safety analysis on the public datasets and report their results at the workshop. It is hoped that this dataset and the workshop will bring together researchers and practitioners from the fields of transportation and computer vision.

BACKGROUND

Among the various sensing technologies available for road transportation data collection (*6*), video sensors have several advantages: 1. the relative ease of installation; 2. the availability of sensors already installed by road management organizations; 3. the moderate cost; 4. the rich description of traffic; 5. the spatial coverage; 6. the automated analysis using computer vision techniques; 7. the ability to verify the data manually.

Video data has first been and is still often manually analyzed, but can also be processed automatically using methods from the field of computer vision. Various types of transportation data can be extracted from traffic video data, from the simple emulation of traditional traffic sensors at a specific location to the detection and identification of all objects in the scene and their tracking from one image to the next to reconstitute trajectories, all the way to higher semantic interpretation of activities occurring in the video. Solutions have been available for about two decades for the simplest types of data (classified counts and speeds) for simple environments, e.g. highways. However, complete and generic solutions for higher level interpretations of video data, starting with all the road users' trajectories, still elude researchers for complex environments such as urban intersections with mixed traffic of medium to high density.

The readers are referred to (*7*) and (*8*) for complete surveys of the field of object detection and tracking. There are three main categories of methods:

1. tracking by detection: objects are detected through background modelling and subtraction, edge detection, (3D) model fitting and image classifiers (*9, 10, 11, 12, 13, 14*);

2. tracking using flow: distinctive image points are detected and tracked in successive images (*15, 16, 17*);

3. tracking with probability: tracking is handled as a probabilistic inference problem, which aims to solve the data association problem in successive images (*18, 19*).

Methods for the surrogate analysis of safety belong to more difficult interpretation tasks as they first typically require sufficient tracking for all road users. There have been a few attempts at building complete video analysis systems for surrogate safety analysis. Most methods rely on indicators of spatio-temporal proximity with the simple motion prediction method at constant velocity (*4*). Faced with the complexity of the task, state of the art methods rely on improved probabilistic frameworks. They take into account various paths that may lead road users to collide (*5*), using supervised and unsupervised machine learning methods, e.g. hidden Markov models (*20*) and clustering (*21, 22*).

Despite the recent progress and all the activity in the development of computer vision techniques for transportation applications, the performance of tracking and higher level analysis,

such as surrogate safety analysis, is difficult to report and compare, especially when the systems are not publicly available and when benchmarks are rare and not systematically used. Most authors use small public datasets and their own non-public datasets of varying complexity. This situations makes it hard to evaluate the actual performance of video analysis methods for transportation applications.

Public datasets and benchmarking are common in several scientific fields, most notably in computer vision. The IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS) (*23*) has been held 15 times from 2000 to 2013 (*24*) with at least 8 different public datasets used to support the workshops (downloadable from `http://ftp.pets.rdg.ac.uk`). Many PETS datasets are relevant for transportation, e.g. scenes of parking with cars and pedestrians (2001) and crowds of pedestrians (2009). Another well-known example is the CAVIAR dataset (Context Aware Vision using Image-based Active Recognition) (*25*) containing videos of people walking indoors, e.g. in offices and malls. Datasets with annotated events also exist that include traffic scenes like the VIRAT Video Dataset (*26*) (downloadable from `http://www.viratdata.org`). Several other small datasets of video data can be found: for example, a dataset of pedestrians crossing in downtown Zurich used in (*18*) (downloadable from `https://www.vision.ee.ethz.ch/datasets/downloads/iccv07-data.tar.gz`, as well as datasets of single images for object detection and classification, e.g. from the PASCAL Visual Object Classes challenges from 2005 to 2012 (*27*). These have many applications for example to pedestrian detection, e.g. the "Daimler Pedestrian Benchmark Data Sets" (*28*) used in (*29*) and the "INRIA Person dataset" used in (*13*) (downloadable from `http://pascal.inrialpes.fr/data/human/`). A list of these datasets is maintained on a wiki at Polytechnique Montreal (*30*). While these examples are relevant for transportation, the video datasets are small (rarely longer than a few minutes) and insufficient for transportation applications, where surrogate safety analysis requires the analysis of several hours of video data, and cover few real life transportation applications.

There exist no standards and few tools for organizing the data that makes up an annotated video dataset, i.e. at least the video itself, its meta-data and the annotations. The most cited effort seems to be the Video Performance Evaluation Resource (ViPER) with its tools for ground truth authoring and performance evaluation (*31*). A more recent tool has been developed for interactive video annotation for computer vision research that crowdsources work to Amazon's Mechanical Turk (*32*). All the other datasets mentioned above also use ad hoc formats and organizations.

Although performance metrics exist for object detection in static images, there is a lack of agreed standards for tracking and higher semantic interpretation such as event detection. Defining tracking performance is difficult because there are different annotation methods (e.g. image bounding boxes, centroid trajectories or ground footprints), because it requires to define what a match between the tracker results and the ground truth is, spatially and temporally, and because there are often multiple solutions to the problem of finding the correspondence between the tracker output and the ground truth. The most used tracking performance metrics may be the CLEAR MOT metrics (*33*).

DATA ORGANIZATION AND TOOLS

A typical dataset includes the following data:

- The meta-data, describing when, where, under which conditions and for what purpose

the filming was done.

- The video recordings from one or several cameras in some commonly used format (either as one video file or as a sequence of images). The current dataset is focused primarily on fixed site-based recordings, i.e. the cameras are not located in vehicles but outside, attached to road signs, lamp posts or buildings nearby using mobile video data collection units such as the one described in (*34*).
- The data necessary to project the object positions in each camera view to the real world coordinate system. This is typically in the form of a homography matrix and normally includes some kind of road plane map or drawing, scale and a set of corresponding points between each camera view and the drawing. This may also include camera calibration parameters with the description of the technique used to produce the values.
- The ground truth, that can typically be:
  - The counts of road users passing given lines or in given areas during a certain time interval;
  - Sequences of points or rectangles (bounding boxes) defining the position of a road user in the camera view (image coordinate system) as it passes the scene;
  - The trajectories (tracks) of the road users extracted from video and projected on the road plane (the world coordinate system);
  - Description of the road user type classification and the 3D-models used for each type;
  - The event database - short video clips containing events of interest such as traffic conflicts with detailed description of the event, involved road users, their trajectories, speed profiles and other safety indicators, etc.
- Detailed documentation explaining how the data is to be used, specification of the file formats, code samples for reading the data, presentation of the automated analysis results, etc.

Some elements in the above list are optional, such as multiple cameras, their full calibration, or various types of annotation, which will depend on the task. Obviously, more data (e.g. from several cameras) has the potential to improve the result quality at the cost of the development of more complicated methods and higher computation.

The dataset is composed of two subsets collected in Europe and North America by different research teams with different means and for different case studies. The principles of data organization, and the tools, have been developed separately to meet the researchers' needs. They correspond in particular to two different setups, depending if only one or more cameras are used.

Single Camera Setup

The simplest necessary setup for video-based transportation data collection is one stationary camera. The following is a description of the meta-data, data requirements and tools developed at Polytechnique Montreal for such a setup. The philosophy of the development has been simplicity and incrementality, adding functionality as needed. The formats and software used to manage the

```
[iberville-10]
sitename = Iberville-Sherbrooke
data-filename = iberville-10.sqlite
homography-filename = iberville-10-homography.txt
calibration-filename = none
video-filename = iberville-10.avi
framerate = 30.
date = 2011-06-28 10:00:39
translation = [0.0, 0.0]
rotation = 0.0
duration = 60
```

FIGURE 1 Sample meta-data for the Sherbrooke-Iberville dataset

data are available in the open source "Traffic Intelligence" project (*34, 35*). The project therefore includes a reference implementation of the single camera setup presented here.

The meta-data is stored in text files in the INI format (with the .cfg extension). Text files have the advantage of being human-readable and easy to edit with any text editor. The INI format is widespread and benefits from implementation in all common programming languages. It contains one section per video file with several pieces of information related to it. An example from the public dataset is provided in FIGURE 1.

The meta-data references the names of the most important files, which should be in the same directory as the meta-data file. The most important data is the video data. It is stored in its original format in the video file as recorded by the camera if the video analysis tools can handle it, usually through the OpenCV library (*36*), which can read standard containers like AVI and MP4, and the codecs installed on the host computer. The data output by the video analysis tools in Traffic Intelligence is stored in databases managed by the public domain SQLite relational database management system (`http://www.sqlite.org/`). Several tables are used to store trajectories and velocities, object information, as well as safety indicator values for road user interactions. The advantage of using a database is to be able to perform simple operations and data transforms efficiently using the SQL language.

The next important pieces of information are the camera calibration and homography files used to project from the image space to the ground plane (and vice versa) in the real world. The simplest setup relies only on the 3 by 3 homography matrix estimated from at least 4 non-colinear point correspondences in image and world space (a simple tool is available in Traffic Intelligence for that purpose). The homography is saved to a separate text file. Better, but more involved, methods exist to obtain the full camera calibration (*37*). Note that the feature-based tracker provided in Traffic Intelligence can function in image space without any homography.

The frame rate is also provided for interpretation of the data, in particular for the conversion of time in standard units to time in number of video frames. An important meta-data for further analysis and interpretation is the data and time of the beginning of the video. The duration is also provided, even if it can be deducted from the number of frames and the frame rate. Finally, translation and rotation information can be added to transform the world coordinate systems accordingly.

Another configuration file in the INI format is used for the video analysis software. The tracking annotations, i.e. the complete trajectories of a sample of road users are created using a graphical user interface developed at Polytechnique (to be released as open source later) and stored in a SQLite database. It contains a table of the image bounding boxes of each road user in each frame and a table for each road user information, in particular its type and a free description.

FIGURE 2 A frame from a Montreal video with overlaid trajectories as replayed by Traffic Intelligence

The project Traffic Intelligence contains all the tools and functions necessary to carry out video tracking, storing the resulting data in a SQLite database, loading the trajectories from it, and processing the trajectory data, in particular for surrogate safety analysis. A sample video frame with the overlaid output of the Traffic Intelligence tracker from the tool to replay trajectories is presented in FIGURE 2. The presented framework has been successfully used for several video analyses and is under ongoing development to meet new needs. The reader is referred to the Traffic Intelligence website (*35*) for more documentation and tutorials designed to help new users get started.

Multi-Camera Setup

Using several cameras requires more work and data description, in particular to synchronize the time and identify the spatial correspondences. The following is a description of the meta-data formats and tools developed at Lund University for such a setup. The formats and software used to manage the data are available as open source (see the documentation at `ftp://barbapappa.tft.lth.se/pdtv/python/index.html`). As a single camera setup is a special case of the multi-camera setup, that data (as describe above) will also be available in the format described below. This gives a single common interface to work with all the datasets provided, that is slightly more complex as compared to the above section.

The dataset consists of a set of data files describing different kind of data objects such as a camera calibration, a video recording, or a set of ground truth tracks. Each data file comes in three equivalent formats; .json, .bson and .yaml. All of them contain the exact same information and can be automatically generated from each other. The yaml format is human-readable and can be edited with any text editor. However, the PDTV tools described below provide a more convenient interface for creating those files. The intention of providing the files in several formats is to allow whichever is most convenient to be used. The formats are all widespread and benefit from implementation in all common programming languages. To get acquainted with the dataset, it can also be browsed using an html interface (`ftp://barbapappa.tft.lth.se/events.html`).

The PDTV documentation describes the dataset as a set of python objects with attributes. Those objects are stored in the files in form of dictionaries with the attribute names as string keys. Each dataset consists of one recording and one calibration from each camera, and one or several synchronizations matching the frames between the cameras as well as the trajectories of the road users. An overview of each of those components are given in the sections below: for addition details please refer to the PDTV documentation.

The dataset files are typically created by instantiating those objects and then saving them using their save methods. This is similar to how the pickle module works in python, but by saving them in a common format such as json or yaml, they can later be loaded from other languages as well.

Video Recordings

The video is stored on disk as a zip file containing jpeg images together with a yaml/json/bson file describing it. The file refers to the camera calibration, the scene it belongs to and how its timestamps relates to the global timestamps. The name of each jpeg frame in the archive should be its timestamp in the canonical format YYYYMMDD-hhmmss.iii.jpg, where YYYY is the year, MM the month, DD then day, hh the hour, mm the minutes, ss the seconds, and iii the milliseconds.

Video Synchronization

The video synchronization represents a set of synchronized video streams. The file contains a stream of synchronized frames that each represents one frame from each video stream that is exposed as close as possible in time. It is constructed by re-sampling the original videos based on their timestamps, using a common master clock with constant frame-rate. The dataset can contain multiple video synchronization files synchronized to different frame rates or using different subsets of the available cameras. The synchronization files do not store the video frames again, but refer to the video recordings describe above.

Camera Calibration

Two different types of calibrations are supported. The first is a simple homography, which allows image coordinates to be projected onto the ground plane (and back) under the assumption of a perfect projective camera. The second is the Tsai camera model (*38*) that also models lens distortion. The camera model can project any 3D world coordinate onto the image plane, as opposed to the homography which is restricted to the ground plane only. For more advanced traffic analyses, an exact calibration becomes important as, for example, distances between road user borders need to be calculated very accurately to compute safety indicators such as time to collision.

Road User Trajectories

The road users are represented as 3D boxes, that are specified by their width, length, and height in meters. The center of the ground footprint of this box is used to represent each road user position, and a unit vector is used to represent their orientation. Using the camera calibration, those 3D boxes can be projected into each camera view to provide the more classical image bounding boxes. This is performed by the PDTV tools when needed (sample demonstration videos can be viewed in the dataset html interface at `ftp://barbapappa.tft.lth.se/events.html` and in FIGURE 3). It is also possible to store image bounding boxes directly in cases where 3D ground truth is not available.

FIGURE 3 Frame from the Minsk videos with overlaid trajectories and 3D boxes as replayed by PDTV

Ground Truth Comparison for Tracking Performance

The PDTV tools also has the functionality to compare the output from a tracking algorithms with the ground truth data. The same kind of objects and files that are used to represent the ground truth is also used to represent the output from the tracking algorithm. This means that the same tools in PDTV can be used to produce both of them.

The comparison problem then becomes a matching problem between two (unordered) sets of tracks. Each track consists of sequence of states or detections that gives the position and size of the road users for some frames. Those states are matched by considering their amount of overlap. The preferred method is to consider the overlap between their ground footprints. However in cases where 3D data is not available it is possible to instead use their image bounding boxes in the camera views.

The amount of overlap between two rectangles (ground footprints or image bounding boxes), A and B, is defined as the size of their intersecting area divided by the size of their combined area,

$$\frac{|A \cap B|}{|A \cup B|}.$$

This gives a number between 0 (no overlap) and 1 (complete overlap). If this ratio is large enough the states are considered to match. By default we use a threshold of $1/3$. That would correspond

to a 50% overlap when the two rectangles have the same area, i.e. $|A| = |B|$.

To match the tracks, the Hungarian (*39*) method is used to find the assignment that maximizes the total number of matched states between matched tracks. Assigned tracker and ground truth tracks whose states match for less than $1/3$ of their combined set of states are considered partial and discarded. Statistics about the number of matched/missed/extra tracks/states are then provided (see TABLE 1).

Finally, to assess the precision of the tracking algorithm, the distance between the centers of each of the matched states are calculated. A cumulative distribution of those distances is formed, normalized by the total number of states in the ground truth. For each distance d, this distribution gives the number of ground truth states that the tracker located within d meters from its ground truth position. This is then illustrated with a plot (see FIGURE 4 for an example).

THE DATASET

Description

The dataset made available for the 2014 TRB workshop currently contains video data collected at three sites; one in Minsk, Belarus, and two in Montréal, Canada. The intention is to allow this dataset to grow as we make new recordings and/or annotations in future studies. We also hope that third parties would be interesting in contributing with their data and annotations.

The Minsk videos were collected in June 2010 by the research team at Lund University. It consists of 3 months of recordings, from four cameras, observing one intersection with mixed traffic (including pedestrians) and one of its legs from different angles (see the views in FIGURE 3). Only a small portion of it is currently annotated and publicly available, the ground truth is expected to grow in the near future, including with extra information such as traffic counts per direction. 8 conflicts have been identified and the annotated tracks are available for the road users involved in the conflict.

The Montreal videos were collected by the research teams at Polytechnique Montreal and McGill University in June 2011 at two intersections on the major arterial Avenue Sherbrooke for a study of pedestrian infractions and safety at crossings (*40*) (see one view in FIGURE 2). Traffic is mixed, including pedestrians and cyclists. Each site was recorded by one camera during the morning and 2 hours are made available for each. A small section of a 1000 frames (33 s) containing 21 road users is currently annotated using image bounding boxes. The instants of conflicts and other traffic events that may be relevant for safety were manually recorded.

Traffic is mixed at the 3 intersections where data was collected, including vulnerable road users and motorized vehicles of varying sizes. Camera angles vary also widely, with tall vehicles causing significant occlusions in the Montreal dataset. These scenes correspond to the complex urban environments highlighted in the introduction for which tracking all road users is an open problem.

Sample Transportation Application: Tracking Road Users

A typical use of video data for transportation is road user tracking. To present a bit further the dataset and the annotations, and to demonstrate the tools provided with it, the Montreal and Lund research teams report the performance of their respective video trackers on some videos. The performance is computed using the PDTV tool for ground truth comparison presented previously.

A classical approach to tracking, is to use a single frame detector followed by a Kalman-filter and some data association. The tracker uses a detector that slides a 3D box along the ground

TABLE 1 Summary report of tracking performance of the tracker described in (*41*) on the Minsk dataset

	Bus	Car	Minivan	Pedestrian	Unknown
True tracks	1	33	3	12	1
Detected tracks	0	27	1	0	0
Missed tracks	1	6	2	12	1
Extra tracks	1	51	2	12	1
True states	79	3315	381	1448	435
Detected states	0	1579	22	0	0
Missed states	79	1736	359	1448	435
Extra states	78	5015	260	718	247

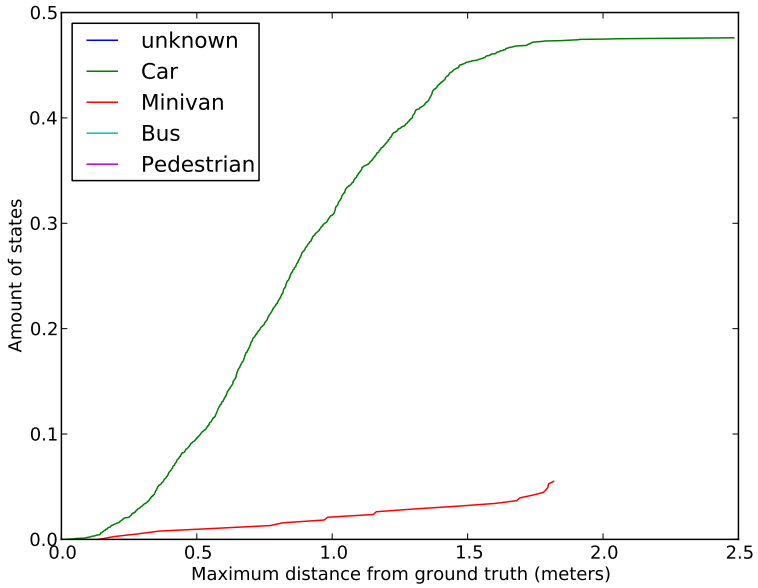

FIGURE 4 Proportion of ground truth states that the tracker described in (*41*) located within *d* meters from its ground truth position (on the ground plane) on the Minsk dataset

plane. This box is projected into each of the camera views and the amount of foreground within those projections is summed up. One such approach (*41*) was applied to the Minsk tracking dataset. The box size chosen was set to match the cars, which means that the cars are tracked nicely while larger vehicles are typically detected as multiple cars, while the smaller pedestrians are missed. The statistics produced by the comparison described above are provided in TABLE 1. In FIGURE 4, the precision assessment described above shows that the position of cars is much better than the positioning of vans, which is not surprising as the box size was tuned for cars.

Another classical approach to tracking is feature-based tracking (*16*), which is available in the Traffic Intelligence project. Distinctive points, or features, are detected, tracked from frame to frame and recorded as feature trajectories using the OpenCV library (*36*). A moving road user

TABLE 2 Summary report of tracking performance of the feature-based tracker available in Traffic Intelligence on the Amherst Montreal dataset

	Car	Pedestrian	Unknown
True tracks	15	4	2
Detected tracks	14	0	0
Missed tracks	1	4	2
Extra tracks	0	1	0
True states	2965	1987	568
Detected states	1872	0	0
Missed states	1093	1987	568
Extra states	164	34	0

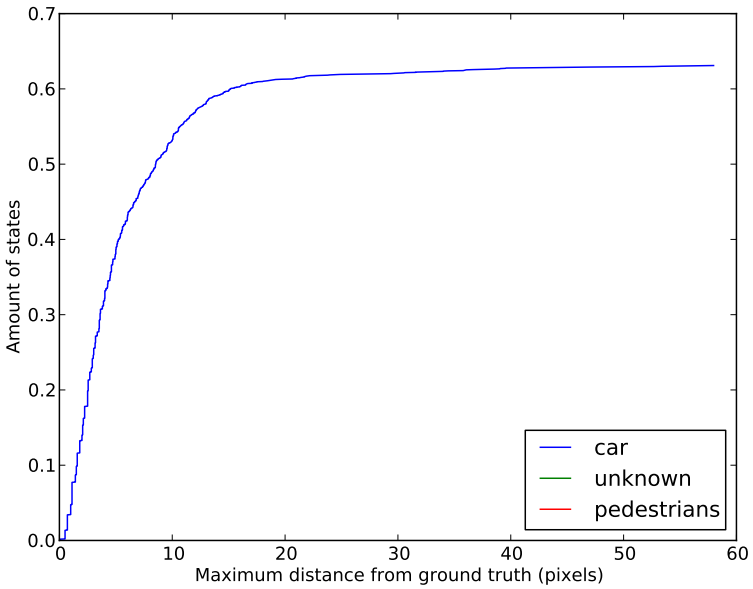

FIGURE 5 Proportion of ground truth states that the Traffic Intelligence tracker located within d pixels from its ground truth position (in image space) on the Amherst Montreal dataset

will have multiple features on it. They are grouped based on consistent common motion. The parameters of this algorithm are tuned through trial and error, leading to a trade-off between over-segmentation (one object being tracked as many) and over-grouping (many objects tracked as one). Because feature trajectories are interrupted when stationary, the motion of an object that stops tends to be represented by more than one trajectory. The statistics of tracking are reported in TABLE 2 and show that the tracker performs well for cars, but not for pedestrians or other road users. A confounding issue is that the tracker output, a set of feature trajectories grouped for each object, is not well suited to comparisons based on object volume (here image bounding boxes). The accuracy is reported in FIGURE 5, with distances in pixels. The tracker seems less accurate than

the previous one, although it is difficult to compare distances in the world and image spaces, which is related to its output type.

CONCLUSION

To the authors' knowledge, the existence and the content of the proposed dataset are unique. No video dataset could be found that includes as much data (hours of video) with such detailed information and that is in addition accessible to a whole research community. An important goal with the dataset is not only to make it accessible, but also to make it user-friendly.

The availability of a common dataset of this kind provides computer vision researchers with massive data to test the performance and robustness of the developed algorithms without spending time on making their own recordings. Using the same video input and having the manual ground truth available, the performance of different systems can be objectively measured and directly compared. This is a clear move forward for computer vision research.

The proposed dataset is unique with its clear focus on traffic applications and problems related to traffic research. The dataset is an important prerequisite for progress within the area of traffic research. For instance in road safety, it is actually possible to jointly assess the great variety of surrogate safety indicators proposed by different research groups and thereby agree on "more valid" and "less valid" indicators. It is for the first time possible for the different research groups to work with the very same video data and same ground truths. As the dataset, with time, will contain different types of road environments, video recorded for different purposes, indicators will be broadened besides safety to contain indicators on a location attractiveness, level of service, accessibility, etc. The dataset will thus provide opportunities to elaborate with completely new indicators describing the road environment and the users' relation to it.

ACKNOWLEDGEMENT

The authors wish to acknowledge the help of Marilyne Brosseau and Jean-Simon Bourdeau who collected the data in Montréal, and the Montreal borough of Plateau Mont-Royal for the opportunity to collect count and video data.

REFERENCES

[1] Michalopoulos, P. G., Vehicle detection video through image processing: the Autoscope system. *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1, 1991, pp. 21–29.

[2] Koller, D., K. Daniilidis, T. Thórhallson, and H.-H. Nagel, Model-based object tracking in traffic scenes. In *Computer Vision - ECCV'92* (G. Sandini, ed.), Springer Berlin Heidelberg, 1992, Vol. 588 of *Lecture Notes in Computer Science*, pp. 437–452.

[3] Laureshyn, A., *Application of automated video analysis to road user behaviour*. Ph.D. thesis, Lund University, 2010.

[4] Laureshyn, A., Å. Svensson, and C. Hydén, Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation. *Accident Analysis & Prevention*, Vol. 42, No. 6, 2010, pp. 1637–1646.

[5] Saunier, N., T. Sayed, and K. Ismail, Large Scale Automated Analysis of Vehicle Interactions and Collisions. *Transportation Research Record: Journal of the Transportation Research*

Board, Vol. 2147, 2010, pp. 42–50, presented at the 2010 Transportation Research Board Annual Meeting.

[6] Klein, L. A., M. K. Mills, and D. R. Gibson, *Traffic Detector Handbook: Third Edition - Volume I and II*. FHWA, 2006.

[7] Yilmaz, A., O. Javed, and M. Shah, Object tracking: A survey. *ACM Computing Surveys*, Vol. 38, No. 4, 2006, p. 13.

[8] Forsyth, D., O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis. *Foundations and Trends in Computer Graphics and Vision*, Vol. 1, No. 2-3, 2005, pp. 77–254.

[9] Gupte, S., O. Masoud, R. Martin, and N. Papanikolopoulos, Detection and classification of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 3, No. 1, 2002, pp. 37–47.

[10] Messelodi, S., C. M. Modena, and M. Zanin, A computer vision system for the detection and classification of vehicles at urban road intersections. *Pattern Analysis & Applications*, Vol. 8, No. 1-2, 2005, pp. 17–31.

[11] Leibe, B., E. Seemann, and B. Schiele, Pedestrian detection in crowded scenes. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, Vol. 1, pp. 878–885.

[12] Wu, B. and R. Nevatia, Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors. *International Journal of Computer Vision*, Vol. 75, No. 2, 2007, pp. 247–266.

[13] Dalal, N. and B. Triggs, Histograms of Oriented Gradients for Human Detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (C. Schmid, S. Soatto, and C. Tomasi, eds.), INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, 2005, Vol. 2, pp. 886–893.

[14] Kanhere, N. K., S. T. Birchfield, and W. A. Sarasua, Vehicle Segmentation and Tracking in the Presence of Occlusions. In *Transportation Research Board Annual Meeting Compendium of Papers*, Washington, D.C., 2006.

[15] Beymer, D., P. McLauchlan, B. Coifman, and J. Malik, A Real-time Computer Vision System for Measuring Traffic Parameters. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, IEEE Computer Society, Washington, DC, USA, 1997, pp. 495–501.

[16] Saunier, N. and T. Sayed, A feature-based tracking algorithm for vehicles in intersections. In *Canadian Conference on Computer and Robot Vision*, IEEE, Québec, 2006.

[17] Hsieh, J.-W., S.-H. Yu, Y.-S. Chen, and W.-F. Hu, Automatic traffic surveillance system for vehicle tracking and classification. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 7, No. 2, 2006, pp. 175–187.

[18] Leibe, B., K. Schindler, and L. V. Gool, Coupled Detection and Trajectory Estimation for Multi-Object Tracking. In *International Conference on Computer Vision (ICCV'07)*, 2007.

[19] Khanloo, B. Y. S., F. Stefanus, M. Ranjbar, Z.-N. Li, N. Saunier, T. Sayed, and G. Mori, A Large Margin Framework for Single Camera Offline Tracking with Hybrid Cues. *Computer Vision and Image Understanding*, Vol. 116, No. 6, 2012, pp. 676–689.

[20] Saunier, N. and T. Sayed, Automated Road Safety Analysis Using Video Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2019, 2007, pp. 57–64, presented at the 2007 Transportation Research Board Annual Meeting.

[21] Laureshyn, A., K. Åström, and K. Brundell-Freij, From speed profile data to analysis of behaviour. *IATSS Research*, Vol. 33, No. 2, 2009, pp. 88–98.

[22] Saunier, N., T. Sayed, and C. Lim, Probabilistic Collision Prediction for Vision-Based Automated Road Safety Analysis. In *The 10th International IEEE Conference on Intelligent Transportation Systems*, IEEE, Seattle, 2007, pp. 872–878.

[23] *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*. `http://www.cvg.rdg.ac.uk/slides/pets.html`, 2000-2013.

[24] *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*. `http://pets2013.net/`, 2013.

[25] *CAVIAR dataset (Context Aware Vision using Image-based Active Recognition)*. `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`, 2013.

[26] Oh, S., A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3153–3160.

[27] *PASCAL Visual Object Classes*. `http://pascallin.ecs.soton.ac.uk/challenges/VOC/`, 2005-2012.

[28] Gavrila, D. M., *Daimler Pedestrian Benchmark Data Sets*. `http://www.gavrila.net/Research/Pedestrian_Detection/Daimler_Pedestrian_Benchmark_D/daimler_pedestrian_benchmark_d.html`, 2001-2012.

[29] Enzweiler, M. and D. M. Gavrila, Monocular Pedestrian Detection: Survey and Experiments. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, Vol. 31, 2009, pp. 2179–2195.

[30] Saunier, N., *Wiki: Video-based transportation data collection, Free Online Datasets*. `http://wiki.polymtl.ca/transport/index.php/VideoTracking#Free_Online_Datasets`, 2013.

[31] Mariano, V., J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, H. Li, D. Doermann, and T. Drayer, Performance evaluation of object detection algorithms. In *IEEE International Conference on Pattern Recognition*, IEEE, 2002, Vol. 3, pp. 965–969.

[32] Vondrick, C., D. Patterson, and D. Ramanan, Efficiently Scaling up Crowdsourced Video Annotation. *International Journal of Computer Vision*, Vol. 101, No. 1, 2012, pp. 184–204.

[33] Bernardin, K. and R. Stiefelhagen, Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing*, Vol. 2008, No. 1, 2008, p. 246309.

[34] Jackson, S., L. Miranda-Moreno, P. St-Aubin, and N. Saunier, A Flexible, Mobile Video Camera System and Open Source Video Analysis Software for Road Safety and Behavioural Analysis. In *Transportation Research Board Annual Meeting Compendium of Papers*, 2013, 13-3229. Accepted for publication in Transportation Research Record: Journal of the Transportation Research Board.

[35] Saunier, N., *Traffic Intelligence*. `https://bitbucket.org/Nicolas/trafficintelligence`, 2011-2013, software under the open source MIT License.

[36] Bradski, G. and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, Inc., 2008.

[37] Ismail, K., T. Sayed, and N. Saunier, A Methodology for Precise Camera Calibration for Data Collection Applications in Urban Traffic Scenes. *Canadian Journal of Civil Engineering*, Vol. 40, No. 1, 2013, pp. 57–67.

[38] Tsai, R., A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, Vol. 3, No. 4, 1987, pp. 323–344.

[39] Kuhn, H. W., The Hungarian method of solving the assignment problem. *Naval Res Logistics Quart*, Vol. 2, 1955, pp. 83–97.

[40] Brosseau, M., S. Zangenehpour, N. Saunier, and L. Miranda-Moreno, The Impact of Waiting Time and Other Factors on Dangerous Pedestrian Crossings and Violations at Signalized Intersections: a Case Study in Montreal. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 21, 2013, pp. 159–172.

[41] Ardö, H., M. Nilsson, A. Laureshyn, and A. Persson, Enhancements of traffic micro simulation models using video analysis. In *The 17th European Conference on Mathematics for Industry*, 2012.