25th ITS World Congress, Copenhagen, Denmark, 17-21 September 2018

Paper ID # EU-TP1375

Cloud Based Large Scale Video Annotations to improve mapping and mobility for connected, cooperative and automated transport

Marcos Nieto, mnieto@vicomtech.org

Gorka Vélez, gvelez@vicomtech.org

François Fischer, f.fischer@mail.ertico.com

Erwin Vermassen*, e.vermassen@mail.ertico

1. Vicomtech, Spain

2. Vicomtech, Spain

3. ERTICO, Belgium

4. ERTICO, Belgium

Abstract

Annotating vast amounts of video information, containing Petabytes of data is today still a very difficult, tedious and almost impossible to achieve task. Automated driven vehicles as an example need a very detailed view of the environment they operate in. Unexpected obstacles such as potholes, work areas, lane restrictions and so on could easily put automated vehicles in a problematic situation. A fast, almost real-time annotation of video data, produced by the vehicle on-board cameras and other sensors is therefore imperative to realize the vision of automated transport in a near future. The resulting annotated video streams improve Advanced Driver Assistance Systems (ADAS) and Map creation to benefit both car and drivers. This is the goal of the European funded, H2020 Cloud LSVA project. Cloud LSVA, short for Large Scale Video Annotation in the Cloud uses sensor fusion technology and Artificial Intelligence deep learning algorithms to realize semi-automated video annotation.

Keywords:

VIDEO ANNOTATION, SLAM

Technical Challenges

Increasingly vehicles and road infrastructure is equipped with ever more sensors. The Car to Car and Car to Infrastructure research and implementation efforts effectuated for almost 20 years now, start to find their way into today’s vehicles. In the next few years cars, trucks, busses and even bicycles and pedestrian will be equipped with cameras, sensors, smartwatches, smart glasses etc. sending petabytes of data into the internet cloud. This data forms the basis for a better, safer and more performant transport and mobility experience, if at least we are able to mine this enormous mountain of information fast and adequate enough to harness the rough sensor data into useful information, useful for both man and machine alike.

In the short future one can expect vehicles to be equipped with multiple cameras and Lidar scanners just like almost every car today has integrated parking sensors. The data and video streams produced by these sensors amount into the petabytes just for a few hours of recording per vehicle. Making sense out of this data is a serious technical challenge. Especially annotating these large video streams is an almost impossible task to achieve. Due to the huge amount of human interaction needed this operation is too expensive, impractical, very slow and probably most inconsistent. Especially related to the automotive sector where we want these data streams to feed Mapping and ADAS functionality the pure human annotation is not feasible.

When looking at autonomous driving, the challenge becomes even more complex as these vehicles need to know the exact details of their surroundings. This translates into the need for a hard real-time delivery of extremely accurate information. Human manual annotation is a showstopper for automated drive.

The non-existence of the appropriate standards for Video Data Annotation (VDA) is another technical challenge to overcome. Without an agreed-on standard, used by the entire industry, it will be difficult to realize interoperable systems either in the cloud or in the vehicle. The same standards need to be used worldwide by every OEM and every software supplier.

These are the challenges which must be solved. Thanks to the advances in the data communication technology and the dawn of the automation of the transport of persons and goods, a wealth of data is available to drive the next generation of mobility technology but not without taking a major step forward in the processing of these enormous data streams. The Cloud LSVA consortium is working towards this goal and provides a first solution to the set problems.

The Proposed Solution

Annotations for automotive purposes

Before jumping into the discussion of the solution proposed by the project, we should first make clear what is meant with video annotation in the field of automotive and mobility. In this context annotations identify all kinds of road traffic objects, events and scenes which are critical for training and testing computer vision techniques. These techniques are at the heart of ADAS and Cartography (navigation) systems. In general the following annotations types are derived from the Use Case definitions:

Region-Based annotation – Polygons and closed polylines are drawn around the subject regions for each frame. The regions can represent objects such as cars and busses but also indications of the lens state, such as blurred, soiled and clear. These annotations help to identify the lens state. Other forms of region-based annotations are bounding boxes and pixel-wise annotations of elements.

Road reconstruction – These annotation types help to determine the position of objects and help to identify which objects need to be monitored by ADAS systems. Examples of such objects are: Car in front for autonomous emergency braking, traffic density to drive adaptive cruise control etc.

Functional annotation – These annotations go beyond the simple determination of pixel regions but are bound to new functionalities demanded by specific ADAS systems such as time-to-collision and time-to-brake indicators. As a consequence these measurements are variable, vary from frame to frame and are time dependent. In some cases additional geometry must be introduced to support this functional annotation as shown by the example of figure 1.

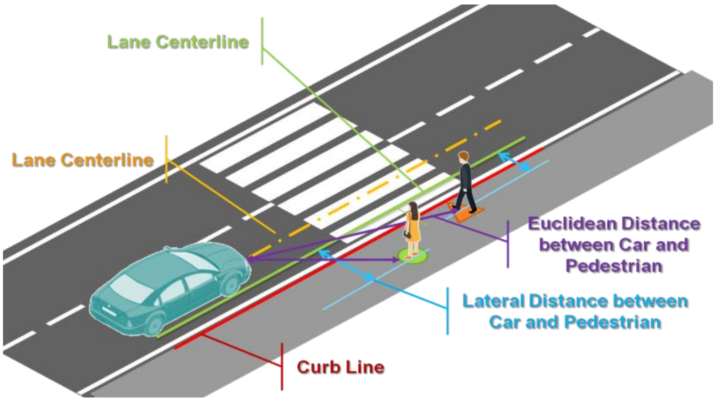

Lane Centerline
Lane Centerline
Euclidean Distance between Car and Pedestrian
Lateral Distance between Car and Pedestrian
Curb Line

Figure 1 Functional annotations for pedestrian crossing

Steps towards an automated video annotation

It becomes clear that introducing the annotations as described by the previous paragraph, would need an immense army of human interaction. Today some tools exist to support the annotations of video streams but they are by far not adequate to efficiently allow the annotations generated by the multiple cameras mounted on a reference or test vehicle. In order to automate the annotation of large scale video recordings, machine learning is introduced. Using machine learning technologies and AI introduces false positives. So in short, not all annotations will be correct. Room must be left for correcting wrong annotations and process these annotations into ground truth. This results effectively in a semi-automated video annotations workflow which should answer to a quality assurance process.

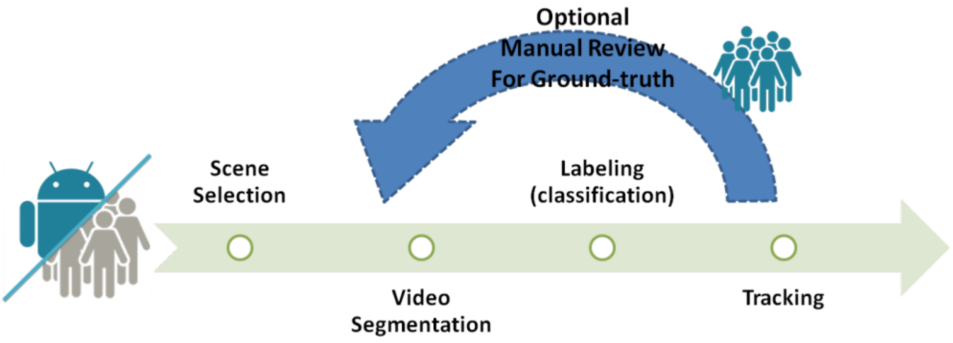


Figure 2 The semi-automated annotation process

To enhance the annotation process additional sensors are used to enrich the video content. This can be telemetry car sensors, data made available from field operation tests, digital cartography etc.

Processing of this data is done by a Cloud based software and application layer where multiple components work in concert to produce the annotated data sets which drive the in vehicle Navigation Systems, Smart Devices and ADAS functionality. Figure 3 on the next page provides an overview of the overall solution provided by the Cloud LSVA project.

Cloud Infrastructure

The infrastructure for running the software implementation is divided in three distinct layers. A first layer provides the application layer providing all the functional engines available to the end-users. This layer runs:

- The Analytics engine
- The Annotation engine
- The Dataset engine
- The Search engine
- The upload engine
- The Web engine

The modules running on this layer interact by means of open and standard communication technologies such a RESTful web services. The lower platform layer contains the Cloud based components which support the upper application layer. The components running on this layer take care of redundancy and scalability. On this layer you will find tools such as Docker, Kubernetes etc. . Finally the lower platform layer takes care of the storage, networking, computation etc.

The orchestration between the different components is based on the OASIS industry standard TOSCA, short for Topology and Orchestration Specification for Cloud Applications. OpenTOSCA was chosen for the implementation of the Cloud platform.

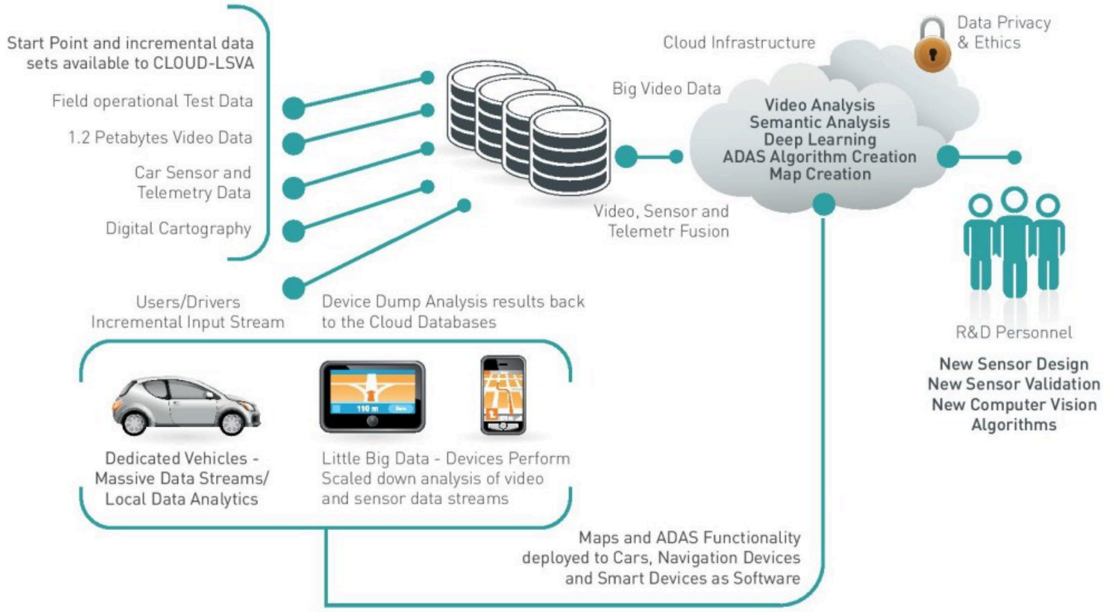

Start Point and incremental data sets available to CLOUD-LSVA
Field operational Test Data
1.2 Petabytes Video Data
Car Sensor and Telemetry Data
Digital Cartography
Users/Drivers Incremental Input Stream
Device Dump Analysis results back to the Cloud Databases
Cloud Infrastructure
Data Privacy & Ethics
Big Video Data
Video Analysis
Semantic Analysis
Deep Learning
ADAS Algorithm Creation
Map Creation
Video, Sensor and Telemetr Fusion
R&D Personnel
New Sensor Design
New Sensor Validation
New Computer Vision Algorithms
Dedicated Vehicles - Massive Data Streams/ Local Data Analytics
Little Big Data - Devices Perform Scaled down analysis of video and sensor data streams
Maps and ADAS Functionality deployed to Cars, Navigation Devices and Smart Devices as Software

Figure 3 Solution overview

Access to the tools is either enabled by means of a Web Application based Graphical User Interface and a web API. The GUI part is role based and serves in general 4 types of users:

- Annotators
- Engineers
- Scientists
- System managers

Open Data Formats and Standardization

A second major point remains the description of knowledge about objects and groups of objects (scenes) and their relation in a specific context. This is known as a conceptualization which is an abstract simplified view of the world that we wish to represent. “An ontology is an explicit specification of a conceptualization. For Video Annotation purposes it allows to derive the description of the elements related to the description of the scenes. Many languages exist to describe ontology. For the purpose of the Video Data Annotation platform the W3C Web Ontology Language (OWL) was selected as this language is the most widely used.

Today there is a lack of suitable standards for annotated video and sensor data. The work plan foresees a task to establish an open standard for video and sensor data annotation suitable for ADAS and cartography applications. This open standard should answer to the following requirements:

- Semi-automated annotation, supported by large scale video analysis in the cloud as well as local by using machine learning algorithms.
- Provide an interaction mechanism for fast manual annotation and validation.

- Automatically extraction of traffic related information related to ADAS and Cartography scenarios for automatic annotation, traffic security related event recognition and scene classification.

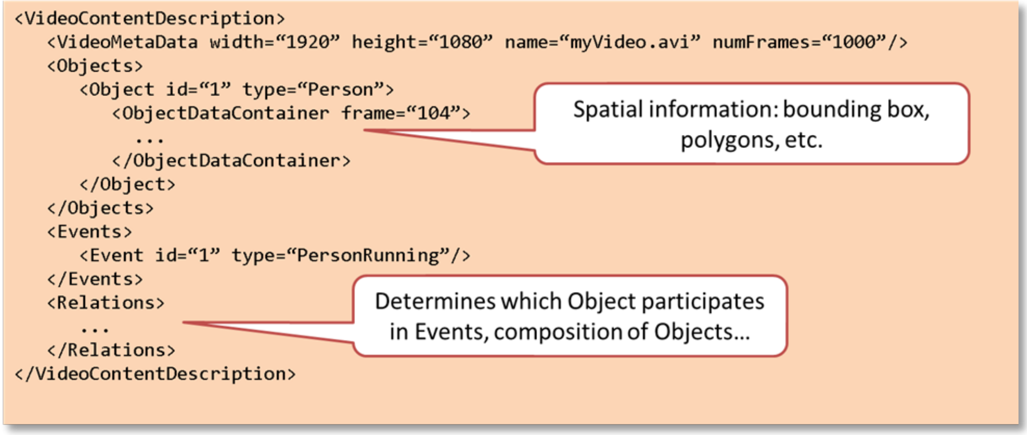
The data model should be fit to describe multi-dimensional objects, traffic related events and connect to the defined ontologies.

Since some of the information provided by the video and sensor data may contain personal information, data anonymization, safe and secure data storage, obfuscation of personal data and sharing data with 3rd party users must be taken care of. This includes the obfuscation of license plates and faces, removing GPS coordinates from annotations etc. A standardized message format must also be performant and efficient to marshal into industry standard formats such as XML and JSON.

A number of data models exist, models such as KITTI, RoadMark, and Cityscapes. However these models are not sufficient for the objectives of the project. The definition of an open standard for Video Annotation becomes therefore imperative and a *conditio sine qua non* for use in the fields of Connected and Automated transport.

To effectively annotate Video Data a number of elements are defined:

- **Objects** – Entities such as Car, Person, Pedestrian, information inside an object is organized in ObjectDataContainers which contain the ObjectData describing the object.
- **Events** – A point in time that triggers some action or appearance of an object.
- **Actions-** Represents a situation such as an action or activity of one or more objects.
- **Contexts** – Description of the contextual information of the scene.
- **Relations** – The connection between elements.



```
<VideoContentDescription>
   <VideoMetaData width=“1920” height=“1080” name=“myVideo.avi” numFrames=“1000”/>
   <Objects>
      <Object id=“1” type=“Person”>
         <ObjectDataContainer frame=“104”>
           ...
         </ObjectDataContainer>
      </Object>
   </Objects>
   <Events>
      <Event id=“1” type=“PersonRunning”/>
   </Events>
   <Relations>
      ...
   </Relations>
</VideoContentDescription>
```

Spatial information: bounding box, polygons, etc.

Determines which Object participates in Events, composition of Objects...

Figure 4 Example of a Video Content Description using XML.

The Video Description Content assembled by using these elements can easily be marshalled into XML files or JSON messages as shown by Figure 4.

For a standard it is important to be used by every stakeholder in the industry, if not system are not interoperable. So to make the work of the Cloud LSVA consortium future proof a most appropriate standardization organization must be found to continue the development and standardization of the VCDA standard.

Usage of the project results

SLAM

Cartography and ADAS are the main usages of the projects' results. SLAM, short for Simultaneous Localization and Mapping is a main technology to enhance maps with the data recorded by vehicle sensors. Visual Camera images are fused with LIDAR and odometer information to continuously improve the "vision" of an (autonomous) vehicle. SLAM data processing uses machine learning algorithms to extract environmental information from the collected data. This is done by means of the Google TensorFlow platform running in a Cloud environment. This system is able to process a huge amount of data either forwarded from the vehicle over a wireless data link from the test vehicle or manually transferred in the cloud environment.

ADAS

Vision based ADAS is a research topic for many years now. The last few years machine learning strategies such as deep learning techniques witnessed a tremendous progress. The availability of heavy lift computing power provided by advanced CPUs and Graphical Processing Units (GPUs) makes it possible to prepare extended training data sets to be used by the Deep Learning process and greatly improve the in-vehicle ADAS features. This makes it possible to detect pedestrians under different and difficult conditions and track vehicles. The code running these algorithms must be highly optimized and make use of the full CPU/GPU hardware available. The presentation will show the result details of the efforts taken by the Cloud LSVA consortium to realize these objectives.

Conclusion

This abstract summarizes the work done by the H2020 Cloud LSVA project. The presentation discusses the challenges of the annotation of Large Scale Video data, the need for an Open Vehicle Data Annotation standard and the proposed solutions to overcome the technical problems which prevent the fast and efficient annotation of petabytes of video data. This solution uses deep learning techniques to realize the semi-automated identification of objects in a scene, the events and relation between the objects and their environment. The main goal of this effort is to improve the information available to both drivers and cars in the form of a better cartography by using SLAM techniques and improve Advanced Driver Assistance Systems (ADAS). The presentation details the experience obtained by applying deep learning technologies and the technological restrictions which prevent to realize some of the goals set forward by the project.

References

1. Manuel Reis-Monteiro, Marcos Nieto, Joachim Kreikemeier (2017*). D1.1 – Requirements, specifications and reference architecture*

2. François Fischer, *(*2016), *D6.6 Initial Standardisation plan*

3. Cloud LSVA Consortium, (2016) Public Deliverables.

4. OpenTOSCA, http://www.opentosca.org/