Journal of Mechatronics, Automation and Identification Technology
Vol. 4, No. 4. pp. 1- 6

Vision Guidance for Autonomous Vehicles: Comparative Study Of 3D Video and Point Cloud

Dragorad MILOVANOVIĆ, Dragan KUKOLJ, Sandra NEMET

University of Belgrade
Faculty of Technical Sciences, University of Novi Sad
RT-RK Institute of Computer-based Systems

dragoam@gmail.com, dragan.kukolj.rt-rk.uns.ac.rs, sandra.nemet@rt-rk.com

Abstract— **Recent advances in 3D sensing and volumetric data compression are encouraging research efforts to meet the demands for vision guidance of autonomous vehicle. Nowadays, autonomous vehicle technology is successful in a high percentage of common road scenarios. However, new research efforts are required to meet the demands for higher performance. The diversity of traffic scenarios in the urban environment presents great challenges, foremost for video-based environment recognition. Autonomous driving technology and other automated assistance systems process huge amounts of data, thus efficient data compression, storage and retrieval is necessary. Key technologies for autonomous vehicles, i.e. 3D video/Point Cloud compression solutions for camera and LiDAR sensor systems are presented in this paper.**

***Keywords*—autonomous driving, 3D video, point cloud**

I. Introduction

The research in 3D sensing and the development of embedded systems strongly support the vision guidance for autonomous vehicles. Advanced sensing technology includes machine vision image recognition technology, i.e. radar (laser, millimetre wave, ultrasound) scanning of the traffic scene. Camera generates 20-60 MB/s, radar upwards of 10 kB/s, sonar 10-100 kB/s, GPS runs at 50 kB/s, and LiDAR between 10-70 MB/s. Based on these numbers, each autonomous vehicle generates approximately 4 TB of data a day. Thus, efficient compression, storage and retrieval are necessary to process huge amounts of data in real-time. Using multi-sensor fusion, a consistent environmental model is generated from the sensing data for the purposes of a scene mapping, obstacle recognition, and for identifying moving objects [1, 2].

An autonomous car is a vehicle that is capable of sensing its environment and moving with little or no human input. There are five essential steps in autonomous vehicle guidance. In the localization step, a vehicle gets the data from all the above mentioned sensors and determines its perception with the highest possible accuracy. Perception is how cars sense and understand their environment. In the prediction step, cars predict the behaviour of every object (vehicle or human) in their surroundings. How they will move, in which direction, at which speed, what trajectory they will follow. Path planning is how the car plans the route to follow, or in other words, how it generates its trajectory. Finally, the control system uses a trajectory generated in the previous step to accordingly change the steering, acceleration or deceleration of the car.

In the first part of the paper, requirements for the next-generation autonomous vehicles are pointed out. Principles of technical operation and performance of 3D sensing devices and research and development of digital video codecs are presented in the second part.

II. Requirements for the next-generation AIV

In the near future, automotive intelligence (AIV) will boost the next generation of vehicles and provide human-level intelligence. Future intelligent vehicles will go in the direction toward environmental protection, energy saving, intelligence, personalization, safety and comfort [1, 3].

Advanced driver assistance system (ADAS) makes use of various kinds of in-car sensors; collects real-time information about the environment, recognizing the static, as well as dynamic objects, and then recommends the most suitable driving actions to the driver to avoid dangerous situations. ADAS relies on inputs from multiple data sources and supports automotive imaging, RaDAR, LiDAR, image processing, computer vision, and in-car networking. Normally, ADAS includes a GPS navigation system, intelligent transportation services (ITS), automatic parking (AP), adaptive cruise control (ACC), and lane keeping assist system (LKAS).

The Society of Automotive Engineers (SAE) classification system is based on six different levels, ranging from fully manual (Level 0) to fully automated systems (Level 5). This classification system is based on the amount of driver intervention and attentiveness required. Driver assistance systems that enable autonomous driving from Level 3 need at least three types of sensor systems: video camera, RaDAR and LiDAR systems. Cameras are mainly used to detect the near objects and to obtain the image regions with possible obstacles considering its rich information and high sensitivity to the lateral displacement. RaDAR and LiDAR are real-time depth data sensors capable of detecting far or longitudinally moving objects.

Automotive intelligence is a trend in the automotive research area which requires high-precision maps with a high update rate. To reach the state of fully automated driving, a high-precision map is the foundation, but real-time information is also required. The world model aims at providing a precise representation of the world. Precision is the key parameter for measuring the performance of a map for autonomous vehicles. The concept of global navigation still dominates other approaches to autonomous navigation. The vehicle is guided by trajectories derived from planning algorithms operating in global metric maps of the environment [1].

All autonomous vehicles being tested rely on one or more of the four main technologies in the industry: video cameras, RaDAR, LiDAR, and the Global Positioning System (GPS). The specific technologies under these umbrellas are continually advancing, and the way in which each R&D utilizes them is different, but all ways lead back to these four areas. Many regular cars on the market today have rear cameras or cameras that monitor activity around the vehicle and warn drivers when they are about to change lanes unsafely or collide with an object. For autonomous vehicles, video cameras are more numerous and more advanced. Even though the cameras on autonomous vehicles are high-performance, low light and distractions like driving under rain can mislead cameras. With RaDAR, electromagnetic waves are emitted away from the source and travel until they collide with an obstacle, at which point they bounce back towards their source. Self-driving cars use both short- and long-range radars to glean information, especially on how fast the vehicles around them are traveling, although they struggle with the height of objects (like bridges). LiDAR uses pulses of near-infrared light instead of radio waves, which are emitted millions of times in just one second, and it relays that exact measurements to a computer, where a 3D map of the world around an autonomous vehicle can be created and used to guide its path [2].

Local environmental perception refers to the ability of an autonomous system to collect information and extract relevant knowledge from the traffic scene. It is one of the main challenges in the field. Environmental conditions, like lighting or colours, are permanently changing, and there are a lot of static, as well as dynamic objects in the scene to be taken into account. The best perception results are typically achieved by the strengths of different sensors. Vehicles must be able to recognize their environment and take control. The vehicle must perceive relevant objects, including other traffic participants and infrastructure information, but also assess the situation and generate appropriate actions. To perceive the 3D information, normally two steps are involved: *segmentation* and *classification*. Some may include a third step, time *integration*, to improve the accuracy and consistency. Segmentation of a point cloud is the process of clustering points into multiple homogeneous groups, while classification is meant to identify the class of the segmented clusters. After the segmentation, each cluster needs to be categorized into different objects [4].

Different sensors have different strengths and weaknesses. Sensor fusion techniques are required to make full use of the advantages of each sensor. A video camera is able to provide rich appearance data with much more object details, but its performance is not consistent across different illumination conditions. Furthermore, a camera does not implicitly provide 3D information. LiDAR is able to produce 3D measurements and it is not affected by the illumination of the environment, but it offers little information on objects' appearances. Sensor fusion is performed in most current systems, especially when complementary sensors like colour cameras (with good angular resolution, no distance information) and range measuring devices (with no colour, bad angular resolution, and precise distance information) are available. Sensor fusion may then proceed at different levels of abstraction. The fusion could be done at the feature level, tracking a road boundary based on 3D measurements from the LiDAR, and image features from vision. The last level of fusion operates on the object level, fusing the detected objects.

LiDAR refers to a light detection and ranging device, which sends millions of light pulses per second in a well-designed pattern. With its rotating axis, it is able to create a dynamic, 3D map of the environment. The outputs from the LiDAR are sparse 3D points reflected back from the objects, with each point representing an object's surface location in 3D, with respect to the LiDAR. Three main representations of the points are commonly used, including point clouds, features, and grids. *Point cloud*-based approaches directly use the raw sensor data for further processing. This approach provides a finer representation of the environment, but at the expense of increased processing time and reduced memory efficiency. To mitigate this, usually a voxel-based filtering mechanism is applied to the raw point cloud to reduce the number of points. Feature-based approaches first extract parametric features out of the point cloud and they present the environment using the extracted features. The features that are commonly used include lines and surfaces. Grid-based approaches discretise the space into small grids, each of which is filled with information from the point cloud, such that a point neighbourhood is established. An adaptive octree is proposed to guide the segmentation from coarse grids to fine ones [2].

III. 3D CODECS RESEARCH AND DEVELOPMENT

The prerequisite for any kind of autonomous driving is sensor detection of the vehicle environment. In order to meet the high requirements in terms of reliability, field of vision, and range, several different sensor systems are generally used. Using multi-sensor fusion, a consistent environmental model is generated from the measurement data for the purposes of mapping, obstacle recognition, and for identifying moving objects [5, 6].

Large-scale 3D maps of outdoor environments can be created using devices that provide localization combined with depth and colour measurements of the surrounding environment. A combination of LiDAR point cloud data and camera images generates a 3D map. These maps are further combined with road markings, such as lane information and road signs, in order to enable autonomous navigation of vehicles. Multiple map layers will be stored and exchanged across the network, including static maps that do not change very frequently and dynamic maps that include real-time information about dynamic objects in a scene, such as vehicles or pedestrians.

For environment perception, both image-based sensors like monocular and stereo cameras (monochrome and colour), and range sensing devices like RaDAR and LiDAR are used. RaDAR sensors are additionally able to determine the object's relative velocity directly. Light detection and ranging (LiDAR) sensors are commonly used in perception for autonomous vehicles because of their high accuracy, speed, and range. Distance-providing image-based sensors are mostly based on a time-of-flight principle.

A. Principles of technical operation and performance

Sensing devices are the essential components for acquiring the information of environmental conditions and the surrounding objects. The implementation of a comprehensive number of sensory inputs from different types of sensing modalities provides more reliable and complete information. The commonly used sensing devices include a video camera, radio detection and ranging (RaDAR) transducer, and light detection and ranging (LiDAR) transducer. Each of these sensing modalities retains characteristics and behaviours that might either beneficially enhance or adversely decimate the sensory performance, depending on certain conditions due to the distinct principle of technical operations [2].

The principle of operation for video camera imaging is conducted by receiving light information that is reflected from the surrounding objects and the environment from the external light sources. The imaging sensor is sensitive to light interference because the perceived quality of the acquired image depends on the presence of either inferior or excessive amount of light. The RaDAR is an active transducer which uses radio frequency waves to measure the time of flight between the transmitter and the receiver on a certain degree of the field of view. LiDAR is an active transducer which uses modulated infrared (IR) waves to measure the time of flight on a full-round of the field of view. It is obvious that both of the surveyed internal and external constraints are interfering with data quality (Table 1).

TABLE I COMPARISON OF 3D SENSING DEVICES.

Video camera	RaDAR	LiDAR
• passive sensor type • 50m sensing range • 60° field of view • dense resolution • mildly sensitive on weather conditions • highly sensitive on illumination and sun-exposure	• active sensor type • 150m sensing range • 30° field of view • highly sparse resolution • mildly sensitive on weather conditions	• active sensor type • 100m sensing range • 360° field of view • sparse resolution • highly sensitive on weather conditions • mildly sensitive on sun-exposure

The performance of non-contact sensing techniques of a video camera, RaDAR and LiDAR can be interfered by the external constraints. In adverse weather condition, the medium for probe wave propagation is occupied with unwanted particles that reduce the visibility. The internal constraints that influence the sensing performance are sensing range, field of view, and data resolution. Each of these constraints can be seen as a trade-off function. However, through the implementation of multi sensing modalities and multi-level fusion methodology, it is possible to achieve optimal performance. Video camera and RaDAR are a perspective combination that provides a dense resolution and the broadest sensing range. However, from the perspective of the field of view, LiDAR offers the widest field of view that covers 360°, or a full-round view.

B. 3D Video codecs

The spatial resolution UltraHD provides drivers with enhanced visual experience via a wide field of view (FoV), both horizontally and vertically. 4K UltraHD is four times the high-definition (HD) resolution, and thus can deliver a larger amount of visual information from the sensors. Traffic scenes with complicated lighting conditions have a high dynamic range (HDR). Dynamic range is the ratio between the values of the largest light intensity and the smallest possible light intensity in a scene. The UltraHD HDR video format defines enhanced parameters of a video camera: higher spatial resolutions (3,840x2,160 and 7,680x4,320 image samples), frame rates (up to 120Hz), sample bit depths (up to 12 bits for HDR high dynamic range support) and a wider colour gamut (ITU-R BT.2020). The format requires increased storage capacity and bandwidths, so video compression is an essential and important first step.

The latest video compression MPEG standard, high efficiency video coding (3D-HEVC) and new versatile video coding (VVC) support 3D vision formats [5, 6]. VVC efficiency coded three types of camera video: standard dynamic range (SDR), high dynamic range (HDR), and 360° field of view (omnidirectional view). Various technologies are being proposed and evaluated in the MPEG JVET's exploration process (Table 2). The target compression performance has a 30–50% bit-rate reduction compared to HEVC for the same subjective video quality. MPEG-I Part 3 VVC is scheduled to reach the FDIS stage (the final draft International Standard) in October 2020.

TABLE II MPEG JVET DEVELOPMENT PROCESS.

Meeting	Working process
Oct. 2015	Joint Video Exploration Team (JVET) developed the JEM (Joint Exploration Model) reference software
June 2016	FVC requirements for the functionalities and performance
April 2017	Call for Evidence (CfE)
Oct. 2017	Call for Proposal (CfP) - 26 responses from 21 organizations
April 2018	ISO/IEC 23090 MPEG-I Part 3 - Versatile Video Coding (118 working documents)
July 2018	VVC Working Draft 2, Test Model VTM 2 (559 documents)
Oct. 2018	VVC Working Draft 3, Test Model VTM 3 (690 documents)
Jan.2019	VVC Working Draft 4, Test Model VTM 4 (897 documents)

March 2019	VVC Working Draft 5, Test Model VTM 5 (858 documents)
Oct. 2020	VVC FDIS

Some of the principal tools adopted in VVC so far are listed in Table 3. Many proposals are intensively evaluated and selected at each meeting.

TABLE III Principal coding tools adopted in VVC.

Category	VVC Working Draft 4
Partition structure	Maximum size 28x128, Quad-/Ternary-/Binary-tree, CST (chroma separate tree)
Intra prediction	DC+Planar+65 directional prediction+28 wide angle+3 CCLM (cross-component linear model), PDPC (position dependent prediction combination), MLIP (multi-line intra prediction), CPR (current picture referencing)
Inter prediction	AFF (affine motion compensation), CIIP (combined intra/inter prediction), triangular, BWA (bi-directional weighted average), decoder-side motion refinement (BDOF bi-directional optical flow), DMVR (decoder-side motion vector refinement), MV prediction (ATMVP alternative temporal motion prediction), HMVP (history-based motion vector prediction), AMVR (adaptive motion vector difference), PMC (pairwise merge candidate), MMVD (merge with motion vector difference)
Transform	Square and rectangular transforms (up to 64x64 size), shape adaptive transform, MTS (multiple transform set DCT2, DST7, DCT8)
In-loop filter	ALF (adaptive loop filter), large-block adaptive DF, LADF (luma-adaptive DF), SAO
Entropy coding	DQ (dependent quantization), template context

The most recent reference software for VVC, VTM5.0, achieves a 33.14% bit-rate reduction with an encoding runtime of 6.71 times, and a decoding runtime of 1.03 times compared to HEVC reference software HM16.19, under the random access coding structure. More details are given in Table 4. Faster coding tools with more coding gain are demanded.

TABLE IV Coding performance: VVC codec VTM5.0 vs. HEVC HM16.19

Coding structure	Luminance-Y BD-rate	Encoding runtime	Decoding runtime
All Intra	-23.14%	22.46	1.04
Random Access	-33.14%	6.71	1.03
Low delay B	-24.69%	4.82	0.89
Low delay P	-28.15%	4.39	0.92

C. *Compression of dynamically acquired point clouds*

Point clouds (PC) have recently emerged as representations of the real world, enabling more immersive formats to better understand and navigate it. MPEG 3D content categories (static objects and scenes, dynamic acquisition, dynamic object) are typically captured using various setups of multiple cameras, depth sensors, and LiDAR scanners. The PC codec standard targets efficient geometry, an attributes compression, scalable/progressive coding, and coding of sequences of point clouds captured over time. In addition, the compressed data format should support random access to subsets of the point cloud in the reconstruction of object/scene as a composition of points [7, 8].

Key requirements for PCC automotive application are:

- high precision is needed to support localization needs,
- low complexity and/or support for real-time encoding/decoding is needed,
- low delay is needed for real-time communication of dynamic parts of the map,
- region selectivity is important to maintain and access the map data,
- colour attributes coding is needed for realistic rendering and visualization,
- additional attributes coding for reflectance and other scene properties.

Various technologies are being proposed and evaluated in the MPEG exploration process. The target compression performances are presented in Table 5. Video-based PCC (V-PCC) of MPEG-I Part 5 is scheduled to reach the FDIS stage in January 2020, and Geometry-based PCC (G-PCC) of MPEG-I Part 9 is scheduled to reach the FDIS stage in April 2020.

TABLE V Compression requirements for mapping in automated driving.

Compression format (lossy, good quality)	Compression ratio [bpp]
2D Flat UHD (Intra/Inter)	0.25-0.5 / 0.025-0.1
3D (2D + smooth depth)	2D Flat + 25% (for parallax)
3D (Multi-View)	2D * No. views * 75%
Point Cloud Coding	Geometry 1-3, Texture 0.5-2

In 2014, MPEG began an exploration activity on PCC. Observing the fast-growing interest of point cloud-based applications from industry, MPEG made a call for proposals (CfP) for PCC in 2017 and evaluated a set of 13 technical proposals in October 2017. As an outcome, three different technologies were chosen as test models for three disparate content categories (static objects and scenes, dynamic acquisition, and dynamic objects). As an input format to PCC, MPEG currently uses the Polygon File Format to represent point clouds. In this format, every point has a 3D position and its associated attributes. Generally, this includes colours and eventually other properties, such as reflectance attributes. Each point is supposed to have the same number of attributes as a legitimate input to PCC. Typical MPEG test point clouds have the following characteristics (Table 6):

- millions to billions of 3D points with up to 1cm precision
- colour attributes with 8-12 bits per colour component
- normals and/or reflectance properties as additional attributes.

A point cloud is defined as a set of (x,y,z) coordinates, where x, y, z have a finite precision and dynamic range.

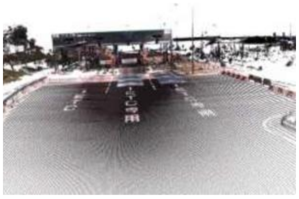
Each (x,y,z) coordinate can have multiple attributes associated to it (a_1, a_2, a_3...). Typically, each point in a cloud has the same number of attributes attached to it.

$$\text{Point}_v = ((x,y,z), [c],[a_0 \ldots a_A]):\ x,y,z \in \text{R} \qquad (1)$$
$$c \in (r,g,b) \mid r,g,b \in N,\ a_i \in [0,1]$$

The point cloud is then simply a set of K points without a strict ordering:

$$\text{Point Cloud}_A = \{(v_i): i=0,\ldots,K\text{-}1\} \qquad (2)$$

TABLE VI DYNAMICALLY ACQUIRED (FUSED POINT CLOUDS) TEST MATERIAL.

Data set (sequence number)	**Points number, geometry precision**	**Attributes**
CityTunnel (28) 	21.163.706 32 bit	RGB, I
OverPass (29) 	5.326.157 32 bit	RGB, I
ToolBooth (30) 	7.148.520 32 bit	RGB, I

For dynamically acquired point clouds data, the G-PCC compressed geometry is typically represented as an octree from the root all the way down to a leaf level of individual voxels (octree geometry codec). There are 3 attribute coding methods: Region Adaptive Hierarchical Transform (RAHT) coding, interpolation-based hierarchical nearest-neighbour prediction (*Predicting Transform*), and interpolation-based hierarchical nearest-neighbour prediction with an update/lifting step (*Lifting Transform*).

Let A and B denote the original and the compressed point cloud, respectively. Consider evaluating the compression errors, as denoted in the point cloud relative to the reference point cloud. In the case of a lossy compression, the number of points in the set and/or the positions x,y,z are not identical to the original. For objective evaluation metrics, the geometric PSNR is defined as the peak signal over the symmetric distortion:

$$\text{PSNR}=10 \log_{10} (3p\text{^}2/\text{MSE}) \qquad (3)$$

where p is the peak constant value defined for each reference point, and MSE is the mean squared point-to-point (D1) or point-to-plane (D2) error.

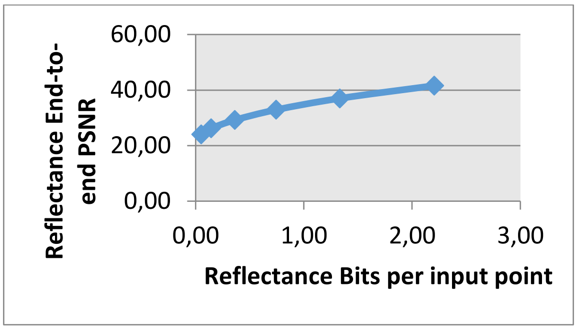


Fig. 1 Performance of G-PCC Octree reference codec TMC3 v6: PSNR vs. Reflectance Bits per input point for data set *CityTunnel*.

For lossy attribute coding, the attribute PSNR value is computed as MSE for each of the three color components (Fig. 1). A conversion from RGB space to YUV space is conducted using the ITU-R BT.709, since the YUV space correlates better with human perception. A symmetric computation of the distortion is utilized in the same way as it is done for geometric distortions. In the case of lossless compression, the decoder returns exactly the same set of (x,y,z) coordinates, with exactly the same attributes.

IV. CONCLUSIONS

In order to safely operate with autonomous vehicles, a multitude of problems in perception, navigation, and control have to be solved. High-accuracy and high-efficiency 3D sensing and associated data processing techniques are urgently needed. While progress has been made with applying 3D sensory data to those applications, many essential questions remain regarding the processing and understanding of such massive 3D data. Dynamic point clouds can be used for some newly emerging applications, such as autonomous navigation based on 3D largescale dynamic maps.

Our main intention with this article was to provide an overview of the standardization trends in the MPEG PCC and to acquaint readers with the most recent work of the G-PCC codec with dynamically acquired data. From the requirements of PCC to future technical specification development, we address the basic framework of PCC and briefly discuss what lies ahead until MPEG PCC becomes the international standard in 2020. Researchers will continue to explore ways to improve and enhance autonomous vehicle technology. As advancements are made in self-driving vehicle technology, these features will become more prevalent, and likely less expensive.

ACKNOWLEDGEMENTS

This work was partially supported by the Ministry of Education, Science and Technology Development of the Republic of Serbia under the Grant TR32034.

REFERENCES

[1] J. Li et al., “Survey on Artificial Intelligence for vehicles”, Automotive Innovation, 1:2-14, 2018.

[2] D. Milovanović, D. Kukolj, S. Nemet, “Recent advances in 3D video modalities for autonomous vehicle guidance”,

Journal of Mechatronics, Automation and Identification Technology, Vol. 3, No. 4, 2018, pp. 1-5.

[3] T.Luettel, M.Himmelsbach, H-J.Wuensche, "Autonomous ground vehicles - Concepts and a path to the future", Proceedings of the IEEE, Vol. 100, May 2012

[4] D.E. Dickmanns, "Vision for ground vehicles: History and prospects", International Journal of Vehicle autonomous systems, Vol. 1, No.1, 2002, pp.1–44,

[5] D. Milovanović, D. Kukolj, Z. Bojković, "Recent advances on 3D video coding technology: HEVC standardization framework", Connected media in the future Internet era (Eds. A.Kondoz, T.Dagiuklas), Springer 2016, pp.77-106

[6] D. Milovanović, D. Kukolj et al., "Recent developments in video compression with capabilities beyond HEVC. 3D", Visual content creation, coding and delivery (Eds. P.Assuncao, A.Gotchev), Springer 2018, pp.41-96

[7] P.A. Chou, M. Koroteev, M. Krivokuća, "A volumetric approach to Point Cloud compression", Part I: Attribute Compression, IEEE Trans. Image Processing, March 2019, DOI: 10.1109/TIP.2019.2908095

[8] M. Wien et al., "Standardization status of immersive video coding," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, Vol. 9, No. 1, 2019, pp. 5-17.