INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 2
MPEG TECHNICAL REQUIREMENTS

ISO/IEC JTC 1/SC 29/WG 2 M56105
Online – January 2021

Title: [VCM] Requirements for stereoscopic and multiview video in VCM environment
Source: WG 2 MPEG Technical requirements

Author(s):
Marek Domański, Tomasz Grajek, Sławomir Maćkowiak,
Sławomir Różek, Olgierd Stankiewicz, Jakub Stankowski,
Poznań University of Technology, Poznań, Poland

Status: Approved

Abstract

This contribution, we consider stereoscopic and multiview video coding in the use case for video coding for machines. Here, the machine is a GPU that estimates depth, i.e. estimates distances. The case is related to autonomus driving and depth estimation from remotely acquired video.

1. Introduction

In the document [1], stereoscopic and multiview video coding was recognized as one of the sub-tasks for Video Coding for Machines project. Stereoscopic and multiview video coding is huge topic, hundreds of papers are provided, and even some standards like multiview and 3D profiles of AVC and HEVC exist. In the context of MPEG VCM activity, stereoscopic and multiview video coding was already considered in [2]. Here, we consider the important potential applications related to stereoscopic and multiview video coding. We focus on the important topics related to intelligent transportation where distance estimation is a task of paramount importance. Such measurements are provided using dedicated depth cameras, lidars or radars, or the distances are estimated using video analysis. The depth cameras, lidars or radars need to illuminate the scene by some radiation and their measurements relay on the reflected radiation. Therefore, they are prone to interference between the devices as well to the environmental factors. Unfortunately, lidars have the disadvantage of high cost, relatively short perception range, and sparse information [3]. Similar problems are reported for the depth cameras. With the rapid development of GPUs, distance estimation provided with the use of video analysis is growing in importance.

2. Application scenario

Connected vehicles can share video data and features to improve navigation performance and driving safety. Among many tasks, distance estimation is crucial for safety of autonomous and semi-autonomous driving. Therefore, cars will probably exchange data in order to obtain more reliable and more comprehensive measurements. On the other side, certain legal issues may discourage the car makers to rely on the video analysis done by the products provided by other vendors. Therefore, one of the potential application scenarios is to share the row video data, whereas the analysis is done individually in each car according to the needs of the navigation and security systems of a given car (Fig. 1).

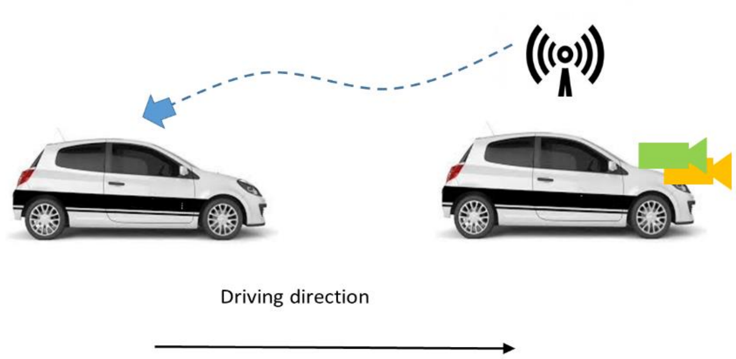

Fig. 1. The leading car provides video streams to the following car(s) that processes the video onboard.

In Fig. 1, we see two cars, the cameras on the first car (on the right) acquire video that depicts longer distance ahead while the cameras onboard the second car (on the left) catch the video of short distance between cars. For the second car, it would be advantageous to analyze also the view from the first car. Such analysis could warn the system of the second car in the case of sudden obstacle in front of the first car. Thus, the driving system of the second could start to react earlier, when the obstacle is not visible in the cameras of the second car.

The scenario from Fig. 1 is typical for Video Coding for Machines: the video from the leading car has to compressed, transmitted and decoded before it consumed by machines in the following car. In this case, the machine vision task is depth estimation.

The fidelity and accuracy of depth estimation depends on all the preceding stages:

1. Video acquisition (camera resolution, frame rate, shutter type, accuracy of camera parameter estimation, camera set configuration – their distances, number of camera);
2. Video compression (video quality after decoding);
3. Transmission reliability (e.g. bit error rate, error concealment);
4. Depth estimation (efficiency and reliability of the software and hardware).

Here, we are going to consider the abovementioned factors except of Item 3.

3. Video acquisition

The problem of the limitations of the accuracy of depth has been studied with respect to several aspects. The depth values are needed to estimate distances to the objects and the dimensions of the objects.

In general, accuracy of the estimated dimensions of the objects are limited by several factors [6]:

- Accuracy of the stereo pair calibration.
  - Accuracy of the intrinsic and the extrinsic camera parameters estimation.

    Focal length and principle point localization have always limited accuracy. Those values are commonly estimated using some camera calibration procedure, while theirs true values remain unknown. The use of distorted value introduces errors in back-projected position of the 3D points of the object, resulting in dimension estimation inaccuracy.
- Accuracy of the system.
  - Digital nature of the image.

    The accuracy of positions of points in digital images is always limited. The position of the given point is usually represented with up to one sampling period. However, some algorithms can estimate point's position with higher, but still finite accuracy. The same is for disparity, which can be estimated with up to 2 sampling periods (accuracy of one sampling period for both the first and the second image pixel position).

First, let us consider [4] the depth estimation (e.g. [7]) for only one camera pair. The focal length of both cameras is f, the base distance is b. The depth of a point object is z and the disparity of the object images is d. Assuming $f \ll z$ we get [36]:

$$z = \frac{f \cdot b}{d} \ . \tag{1}$$

Let us assume two objects with the depths $z_1$and z_2, respectively. Their positions may be distinguished if the respective disparity difference $|d_1 - d_2|$ exceeds a minimum value Δd:

$$|d_1 - d_2| \geq \Delta d \ . \tag{2}$$

Δd is the disparity accuracy, i.e. 2 to 3 distances between the centers of the pixels in the sensors. From (2) we get $d_1 = \frac{f \cdot b}{z_1}$, $d_2 = \frac{fb}{z_2}$, and we can denote average depth as $z = \sqrt{z_1 \cdot z_2}$. Therefore, depth values z_1 and z_2 may be distinguished when:

$$|z_1 - z_2| \geq \frac{z^2}{f \cdot b} \Delta d \ . \tag{3}$$

The abovementioned reasoning explains the well-known fact that the depth map can be estimated with a high accuracy for a long base of a camera pair. Therefore, for the sake of the spatial accuracy, the depth estimation should be performed from a camera pair with the longest base. For multiple cameras, the above considerations imply that the depth estimation should be performed with the use of the longest available base, which is between two furthest cameras in the system. In complex scenes, individual points of a scene are acquired by different sets of cameras. Each camera set exhibits its longest base that corresponds to the two outer cameras of this set. Nevertheless, the view pairs with large bases suffer from increased occlusions that deteriorate measurement of dimensions, e.g. of round objects.

The limitations of the accuracy of the measurements of object size were already studied , e.g. in the context of the vehicle (car or truck) dimension estimation [5,6,8-12]. The theoretical limitations of the vehicle dimension estimation using stereoscopic video analysis due to sensor resolutions and geometrical limitations are depicted in Figs. 2 and 3.

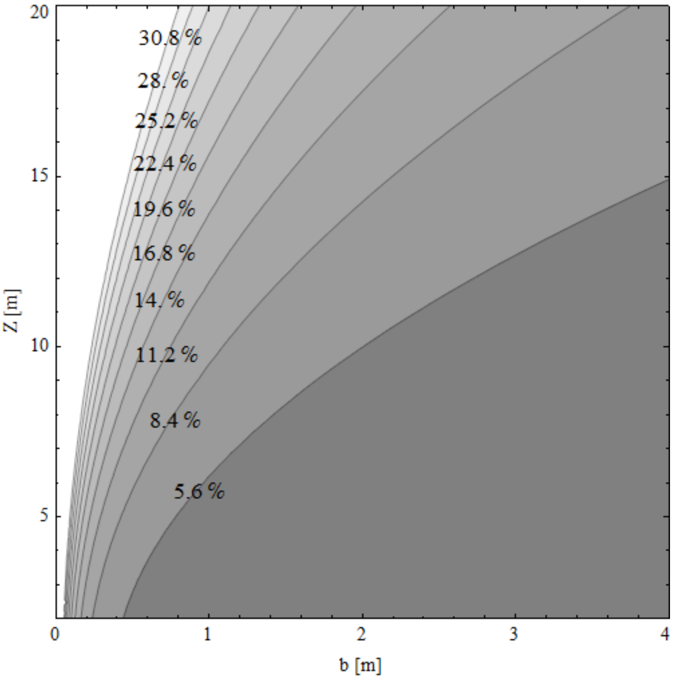

Fig. 1. Estimated accuracy with respect to distance of vehicle from the camera system (Z) and the camera baseline (b), assuming constant vehicle length L = 4.2 m and angle $\alpha = 0°$ (vehicle moving towards the camera), f = 9.6 mm [6].

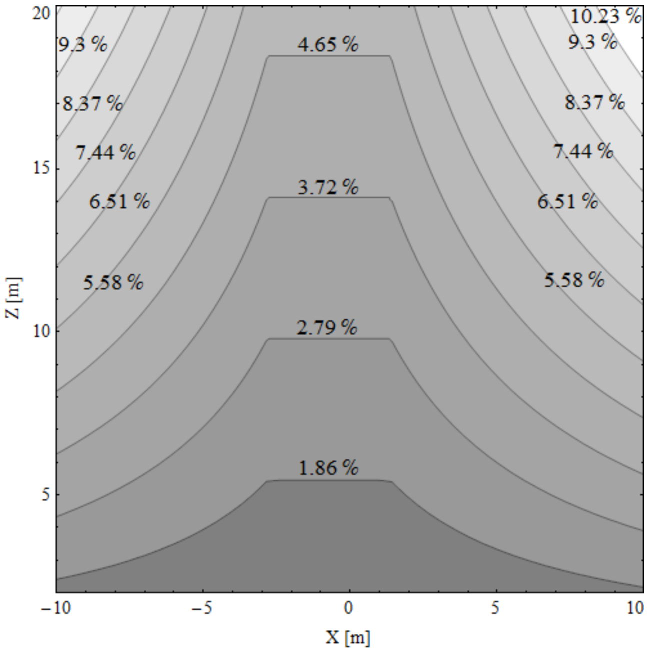

Fig. 2. Measurement accuracy with respect to distance of vehicle from the camera system (Z) and the position of the vehicle (X), assuming constant baseline b = 0.175 m, vehicle length L = 4.2 m and angle $\alpha = 90°$(vehicle moving perpendicularly the camera axix) , f = 9.6 mm [6].

The experimental data obtained for a pair of wide angle cameras with the base 1.445 meter is provided in Table 1. The data are provided by Poznań University of Technology.

TABLE I. RESULTS OF ESTIMATION OF THE VEHICLE LENGTH [5]

Vehicle	***b*** **[m]**	**Distance from the camera [m]**	**Vehicle length [m]**		**Accuracy [%]**	
			True	***Estimated***	***Measured***	***Estimated***
Alfa Romeo 147	1.445	8.865	4.223	**3.994**	5%	**4.3%**
Fiat 126p		8.540	3.054	**2.955**	3%	**5.1%**
Daewoo Tico		7.460	3.340	**3.393**	2%	**4.1%**
Opel Corsa		8.580	3.990	**3.863**	3%	**4.3%**
VW Caddy		8.560	4.405	**4.070**	8%	**4.0%**
Honda Concerto		8.260	4.415	**4.258**	4%	**3.9%**
VW Polo		8.310	3.916	**3.543**	10%	**4.2%**

Table 1 demonstrates that the aforementioned calculations of the theoretical limitations of the measurement accuracy are close to the real values.

Conclusion:

For the further work on video coding for machines, in the considered scenario, new test sequences will be needed. The acquisition should be obtained in cars/trucks in motion. For cars, the cameras could be mounted at the upper corners of the windscreen with the base of about 1.5 meter. The height over ground would be 1.2 – 1.6 meter. For trucks, the base could be even 1.8 – 2.0 meters. The focal length should be chosen (Formula 3) according to the requirement for high accuracy at distances of 3-200 meters. Such a range is quite wide. Therefore probably video should be acquired with 2-3 pairs of cameras with different focal lengths.

4. Depth estimation from compressed video

The stereoscopic or multiview video shall be compressed for transmission to another vehicle. Obviously, coding degradations yield some errors in the depth maps estimated from decoded.

Multiview video compression can be performed using different approaches. One of them is to use dedicated multi-layer encoder, such as MV-HEVC [13]. This method provides the best compression efficiency, however multi-layer encoders are not popular due to their complexity and limited number of applications. Another solution is simulcast encoding, which means all views that compose multiview video are encoded separately with single-layer encoder, such as HEVC [14,15]. This technique is much simpler, has many implementations, but it does not exploit similarities between the views. For the more efficient and never VVC codecs [16], no multiview video coding tool is included into the standard. Therefore, the application of Screen Content Coding appears as an interesting option for multiview video coding as proposed in [17,18]. It was demonstrated that the coding performance of Screen Content Coding for multiview video is virtually the same as that MV-HEVC, so the same approach may be used for VVC as well [2].

The quality of the depth map is often measured by the quality of the virtual views synthesized using these depth maps. This approach is currently under consideration by Immersive Video group that is also working on immersive video coding profile where depth is estimated in the decoder [19]. Previous research has demonstrated that for AVC, for QP < 21, the loss of quality for depth estimated from compressed video is negligible [20], i.e. mostly below 0.3 dB. Similar results are reported for HEVC.

More exact study is needed to define the optimum configuration of HEVC and VVC codecs for the decoder-side depth estimation. In particular, we suggest to test the Screen Content Coding in this applications [2].

5. Conclusions

For the application scenario of autonomous cars, for the usage of stereoscopic and multiview video, we propose to:
Gather the appropriate test video clips as required in Section 3;
Test and optimize HEVC and VVC codecs, in particular, with Screen Content Coding.

6. Conclusions

The outcomes from experiment are similar to the anchors reported in [7, 8], but are not the same. For further experiments, to achieve consistency and uniqueness of the results, it is necessary to exact define the neural network weights and parameters set to use. It is also recommended to provide more precise experiment descriptions in the contributions, with more technical details, to simplify eventual test repetition by 3rd party.

7. Acknowledgement

This work was supported by Project 0314/050.

8. References

[1] Doc. ISO/IEC JTC1/SC29/WG2 N18, Use cases and requirements for Video Coding for Machines, Online meeting, October 2020.

[2] M. Domański, J. Samelak, S. Różek, T. Grajek, S. Maćkowiak, O. Stankiewicz, [VCM] Stereoscopic and multiview video coding for machines, Doc. ISO/IEC JTC1/SC29/WG11 M54407, Online meeting, June 2020.

[3] I. Ashraf, et al., "An investigation of interpolation techniques to generate 2D intensity image from LIDAR data", *IEEE Access*, vol.5, Apr. 2017, pp.8250-8260.

[4] O. Stankiewicz, M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch, J. Samelak, "A free-viewpoint television system for horizontal virtual navigation", *IEEE Transactions on Multimedia*, Vol. 20, No. 8, IEEE, 5th January 2018, pp. 2182-2195

[5] R. Ratajczak, M. Domański, K. Wegner, "Vehicle size estimation from stereoscopic video", *19th International Conference on Systems, Signals and Image Processing (IWSSIP 2012)*, Vienna, April 2012, pp: 405-408.

[6] T. Grajek, R. Ratajczak, K. Wegner, M. Domański, "Limitations of Vehicle Length Estimation Using Stereoscopic Video Analysis", *20th International Conference on Systems, Signals and Image Processing (IWSSIP 2013)*, Bucharest, July 2013, pp: 27-30.

[7] D. Mieloch, O. Stankiewicz and M. Domański, "Depth Map Estimation for Free-Viewpoint Television and Virtual Navigation," in *IEEE Access*, vol. 8, pp. 5760-5776, 2020.

[8] M. Ito, Y. Takada, T. Hamamoto, „Distance and Relative Speed Estimation of Binocular Camera Images Based on Defocus and Disparity Information", *28th Picture Coding Symposium*, Nagoya, Japan, p. 278 - 281, December 2010.

[9] R. Ratajczak, T. Grajek, K. Wegner, K. Klimaszewski, M. Kurc and M. Domański, "Vehicle dimensions estimation scheme using AAM on stereoscopic video," *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Kraków, 2013.

[10] J. Elfring, R. Appeldoorn, M. Kwakkernaat, "Multisensor simultaneous vehicle tracking and shape estimation", *Intelligent Vehicles Symposium (IV) 2016 IEEE*, pp. 630-635, 2016.

[11] Chi Hak Lee, J. Pil Ahn, Y. Mo Kim, "Vehicle's Model Classification Using a Vertical Stereo Camera", *Applied Mechanics and Materials*, vol. 744-746, pp. 1960, 2015.

[12] Y. Liu, M. Reynolds, Du Huynh, G. Hassan, "Study of Accurate and Fast Estimation Method of Vehicle Length Based on YOLOs", *Artificial Intelligence and Information Systems (ICAIIS) 2020 IEEE International Conference on*, pp. 118-121, 2020.

[13] G. Tech, Y. Chen, K. Müller, J. R. Ohm, A. Vetro and Y. K. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding", in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 35-49, Jan. 2016.

[14] ISO/IEC Int. Standard 23008-2: 2015 "High efficiency coding and media delivery in heterogeneous environment – Part 2: High efficiency video coding" and ITU-T Rec. H.265 (V3) (2015), „High efficiency video coding".

[15] G. J. Sullivan, J. Ohm, W. J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard", in *IEEE Transactions on Circuits Systems for Video Technology*, vol. 22, no. 12, pp. 1649-1668, Dec. 2012.

[16] ISO/IEC DIS 23090-3 (2020) / ITU-T Recommendation H.266 (08/2020) "Versatile video coding".

[17] J. Samelak, J. Stankowski, M. Domański, "Efficient frame-compatible stereoscopic video coding using HEVC Screen Content Coding", *IEEE International Conference on Systems, Signals and Image Processing IWSSIP 2017*, Poznań, Poland, May 2017.

[18] J. Samelak, M. Domański, Unified Screen Content and Multiview Video Coding - Experimental results, JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11 12th Meeting: Doc: JCTVC- M0765, Marrakech, MA, 9–18 Jan. 2019, also ISO/IEC JTC1/SC29/ WG11 MPEG Doc. M46332.

[19] B. Salahieh, J. Boyce, F. Julien, Ch. Bertrand, D. Renaud, "MIV View with Decoder-Derived Depth", Doc. ISO/IEC JTC1/SC29/WG11 M54492, Online meeting, June 2020.

[20] K. klimaszewski, "Algorithms for multiview video compression", PhD Dissertation, Poznań Univ. Technology, 2012.