INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11
MPEG2017/M40019
January 2017, Geneva, Switzerland

Source **Poznań University of Technology**
Status **Input**
Title **Depth map formats used within MPEG 3D frameworks**
Author **Krzysztof Wegner, Olgierd Stankiewicz, Tomasz Grajek, Marek Domański**

1. Introduction

According to the plain definition using in computer vision the depth of a point M is the distance between the optical centre of the camera lens and the plane containing point M and being perpendicular to the camera optical axis. The set of all depth values related to the individual points of an image constitutes a depth map. Obviously depth is closely related to disparity measured on the image planes of a stereoscopic pair of cameras. Nevertheless, in the course of research, development, design and implementations of the practical 3D video systems and computer vision systems, various definitions of depth are used.

During recent years MPEG has been working on 3D video compression and related technologies that employ depth maps. Depending on the particular application and framework, various depth format have been used [1][2]. Since FTV and Lightfield groups are commonalities and differences between various 3D scene representation formats, this document tries summarize and describe the depth formats used within MPEG. Often, any map of depth or disparity is named as a depth map. Therefore, in order to avoid ambiguity, the depth defined as above is called the z-distance. Even disparity is used in different formats, therefore this document tries summarize and describe the depth formats used within MPEG.

2. Camera projection principles

Let us first consider point M positioned in 3D space at coordinates $M = [X \quad Y \quad Z]^T$. This point is captured by a camera positioned at position $T = [T_x \quad T_y \quad T_z]^T$ looking at direction described by rotation matrix R. The projection of point M onto image plane of the considered camera can be mathematically described as

$$z \cdot m = K \cdot [R \quad -R \cdot T] \cdot \begin{bmatrix} M \\ 1 \end{bmatrix}$$

where $m = [u \quad v \quad 1]^T$ is a position of point M on image plane of the considered camera. z is distance of the projected point M from the camera in direction perpendicular to the image plane.

3. Depth map formats

3.1. z-distance format

Each point m of the captured image is associated with some value z which is a distance of that point from the camera (and of course is associated with position M in 3D space).
Of course this information is lost in a process of image acquisition, so z value must be attained otherwise – e.g. measured by z-camera, estimated algorithmically, or taken from 3D model in case of computer graphics.
Depth map in a z-distance format is an array of directly stored z values for each image point.

3.2. Disparity format

The term of disparity is related to the properties of the human visual system. The retinal disparity is an important binocular cue of depth. In general the definition of disparity d is the following. Assuming u_1 and u_2 are the horizontal coordinates of the images of A, in the right and the left view, respectively, we have

$$d = u_1 - u_2 = f \cdot \frac{b}{Z}$$

where

f is the focal length,
b is the base of the camera pair,
Z is the z-distance, i.e. the depth according the strict definition.

The second part of the abovementioned formula holds under the assumption that Z is much larger than f.

In general, we can consider a stereoscopic acquisition system. Let's assume that point M is observed not by one camera but by two cameras. Those two cameras are positioned side by side, on one line in a way that optical axis of both camera are parallel to each other. For simplicity we can assume that both camera looks in z-axis direction i.e.:

$$R_1 = R_2 = I$$

Moreover let's assume that both cameras are identical (their intrinsic parameters are the same)

$$K_1 = K_2 = \begin{bmatrix} f & 0 & o_u \\ 0 & f & o_v \\ 0 & 0 & 1 \end{bmatrix}$$

Moreover, let's assume that one of the cameras is placed at the origin of the 3D coordinate system and second is shifted by value b, along x-axis i.e.:

$$T_1 = [0 \quad 0 \quad 0]^T \quad T_2 = [b \quad 0 \quad 0]^T$$

b is called baseline of the camera system.
In such conditions cameras observing point M positioned in 3D space at $M = [X \quad Y \quad Z]^T$ see it as point $m_1 = [u_1 \quad v_1 \quad 1]^T$ and $m_2 = [u_2 \quad v_2 \quad 1]^T$ onto image plane of first and second camera respectively.

After putting these into projection equation we get:

$$z_1 \cdot m_1 = \begin{bmatrix} f \cdot X - o_u \cdot Z \\ f \cdot Y - o_v \cdot Z \\ Z \end{bmatrix}$$

$$z_2 \cdot m_2 = \begin{bmatrix} f \cdot (X - b) - o_u \cdot Z \\ f \cdot Y - o_v \cdot Z \\ Z \end{bmatrix}$$

which directly leads to

$$z_1 = z_2 = Z$$

$$u_1 = f \cdot \frac{X}{Z} - o_u$$

$$u_2 = f \cdot \frac{X - b}{Z} - o_u$$

$$v_1 = v_2 = f \cdot \frac{Y}{Z} - o_v$$

So both cameras see image of point M at exactly the same row $v_1 = v_2$ and is different columns (horizontal positions) - shifted in horizontal direction for distance d , which is called disparity

$$d = u_1 - u_2 = f \cdot \frac{b}{Z}$$

So position of image of point M in both images is directly related to distance Z of a point M to the cameras. (in direction perpendicular to the image plane of the cameras).

Therefore, disparity d can be understood as shift in horizontal direction between an image of a point M positioned Z units away from the camera taking the image and some other, second identical camera shifted by b units in x-axis. This second camera may be even non-existing (be a virtual one) but its baseline distance b from the first camera is crucial for the definition of disparity format.

So knowledge about disparity, along with baseline distance b and focal length f is enough to fully define 3D position of observed objects/points.

3.3. Normalized disparity format

Depth in disparity format is very convenient way of storing depth information, but it also has several disadvantages. First of all, depth in disparity format describe depth information as seen from one perspective (camera) but is inseparably connected to some other second camera (with which it creates a stereo pair) shifted by b units in x-axis. Secondly, disparity stored directly do not exploits efficiently whole range of possible values of depth maps in binary format. For example, in 8-bit image format, representable values are form 0 to 255.

In order to overcome these disadvantages, we can use disparity normalization. Instead of storing disparity as binary values, we can scale the dynamic range, so that it matches the representable binary values:

$$v = \frac{d - d_{min}}{d_{max} - d_{min}} \cdot v_{max}$$

where v is a normalized disparity value, v_{max} is maximal binary value, i.e. 255 in case of 8-bit depth maps, d_{min} and d_{max} are minimal and maximal disparity value in a given scene. If we substitute disparity terms with their definition based on z-values from point 3.2, we get:

$$v = \frac{f\frac{b}{Z} - f\frac{b}{Z_{far}}}{f\frac{b}{Z_{near}} - f\frac{b}{Z_{far}}} \cdot v_{max}$$

Where b is baseline distance, f is focal length, and d_{min} is a disparity of a farthest object and d_{max} is a disparity of a nearest considered object in the scene.

$$d_{min} = f\frac{b}{Z_{far}} \quad d_{max} = f\frac{b}{Z_{near}}$$

After simplification we get:

$$v = \frac{\frac{1}{Z} - \frac{1}{Z_{far}}}{\frac{1}{Z_{near}} - \frac{1}{Z_{far}}} \cdot v_{max}$$

When normalized disparity is used, we not only efficiently use whole dynamic range of depth map to represent disparity but also we eliminate its dependency on baseline distance b and focal length f. Instead of d_{min} and d_{max} we introduce two more intuitive new constant Z_{near}, Z_{far}, which represent z-distance to the nearest and farthest considered object within the scene.

4. Summary

Three depth formats used within MPEG frameworks has been described:

- z-distance,
- disparity,
- normalized disparity.

If some new depth map formats will be introduced within future works on FTV or Lightfield, it is proposed that MPEG should issue an document defining and clarifying them, so that other MPEG documents could use it as a reference in order to avoid misunderstandings.

Acknowledgement

The work has been supported by the public funds under the DS project from Poznań University of Technology.

References

[1] ISO/IEC IS 23002-3:2007: Representation of auxiliary video and supplemental information, 2007.

[2] T. Senoh, K. Yamamoto, R. Oi, T. Mishina, M. Okui: Consideration of depth format. ISO/IEC JTC1/SC29/WG11 MPEG dok. m15047, Antalya, Turkey, Jan. 2007.