GLOBAL DEPTH ESTIMATION FOR MULTI-VIEW VIDEO CODING USING CAMERA PARAMETERS

Xiaoyun Zhang, Weile Zhu
University of Electronic Science and Technology of China, Chengdu, Sichuan, China

George Yang
Panovasic Technology Co. Ltd., China

Keywords: Multi-view video coding, multi-view video plus depth, global depth, depth-based view synthesis.

Abstract: Multi-view video plus depth (MVD) data format for Multi-view Video Coding (MVC) can support rendering a wide range continuum of views at the decoder for advanced 3DV and FVV systems. Thus, it is important to study global depth to reduce the rate for depth side information and to improve depth search efficiency. In this paper, we propose a global depth estimation algorithm from multi-view images using camera parameters. First, an initial depth is obtained from the convergent point of the camera system by solving a set of linear equations. Then, the global depth is searched around the initial depth to minimize the absolute difference between the synthesized view and the practical view. Because the initial depth can provide appropriate depth search range and step size, the global depth can be estimated efficiently and quickly with less computation. Experimental results verify the algorithm performance.

1 INTRODUCTION

It has been recognized that multi-view video coding (MVC) is a key technology that serves a wide variety of future interactive multimedia applications, including FVV (free-viewpoint video systems), 3DTV (3D television), immersive tele-conference and surveillance. Because of the huge amount of data, efficient compression is more important than ever in the storage and transmission of multi-view video. In MVC related techniques, one crucial point is to improve the coding efficiency by utilizing the inter-view correlations besides the temporal and spatial correlations within a single view video.

In April 2007, Smolic *et al.* propose multi-view video plus depth (MVD) data format for advanced 3D video systems in the JVT (Joint Video Team of ISO/IEC MPEG & ITU-T VCEG) proposal (Smolic *et al.* 2007). The proposal points out that MVD can meet the central requirement of 3DV and FVV applications, i.e., a format that allows rendering a wide range continuum of views at the decoder, and JVT is considering to work align with MVD format. Therefore, block-based or pixel level depth map estimation from multi-view has been an important issue for MVC.

Although there have been a lot of related work (Okutomi, et al. 1993) (Kauff, *et al.* 2007) (Zitnick, *et al.* 2004) on depth estimation from images, most of them are only suitable for image pairs captured by parallel cameras or calibrated image pairs. In these work, horizontal disparity is first computed using stereo correspondence algorithms, and then the depth is computed according to the relationship that depth is inverse proportional to disparity. Readers may refer to the work of Scharstein (Scharstein, *et al.* 2002) for a good review of stereo matching.

However, in many multi-view video applications, cameras are arranged in arc or along a curved line with rotated angles in order to capture the same scene from different viewpoints, where projection distortion exists between views and the translational block based disparity estimation is not suitable. Thus, depth search method (Yea, *et al.* 2006) is developed to estimate the depth directly and efficiently for view synthesis prediction for rotated camera systems.

In the disparity compensated multi-view video coding, the JVT proposal (Ho, *et al.* 2006) propose

Zhang X., Zhu W., Zhu W., Zhu W., Yang G., Yang G. and Yang G. (2008).
GLOBAL DEPTH ESTIMATION FOR MULTI-VIEW VIDEO CODING USING CAMERA PARAMETERS.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 631-634
DOI: 10.5220/0001071706310634

global disparity compensation, where the global disparity between two views is calculated as the displacement vector minimizing the mean absolute difference between the one translated view and the other view. After global disparity compensation, the disparity vectors mainly distribute around zero value, and thus the search range for disparity can be reduced to improve coding efficiency.

Similarly, in MVC using view synthesis prediction and MVD data format, global depth is studied to reduce rate for depth side information and to improve depth search efficiency (Vetro, 2007). In addition, appropriate depth search range and step size play important roles in the depth estimation (Yea, *et al.* 2007), and estimating global depth can provide useful information for determining range and step size. Therefore, global depth estimation also becomes a key problem for MVC and depth search.

In this paper, we propose a method for global depth estimation from multi-view images using camera parameters. First, an initial depth value is obtained from the convergent point of the camera system by solving a set of linear equations. Then, the global depth is searched to minimize the absolute difference between the synthesized view and the practical views.

2 INITIAL DEPTH

Suppose the camera coordinate system of camera c_i $(i = 1, \cdots, m)$ is $o_i - x_i y_i z_i$ and m is the number of cameras. The orientation and position of the camera coordinate system $o_i - x_i y_i z_i$ relative to the world coordinate system $o - xyz$ is described by a 3 by 3 rotation matrix R_i and a translation vector t_i. If a scene point is described as $p = [x, y, z]$ in $o - xyz$ and $p_i = [x_i, y_i, z_i]$ in $o_i - x_i y_i z_i$, then they satisfy the following equation

$$p = R_i p_i + t_i \tag{1}$$

And, according to perspective projection, the camera coordinate p_i is mapped into the image plane with pixel coordinate $[u_i, v_i]$ by

$$z_i P_i = A_i p_i \tag{2}$$

where $P_i = [u_i, v_i, 1]$ is the homogeneous coordinate of the pixel $[u_i, v_i]$, and A_i is the intrinsic matrix of camera c_i which contains information about focal length, aspect ratio and principal point.

In order to capture the same scene from different viewpoints, in many multi-view video applications cameras are arranged in arc or along a curved line with rotated angles with optical axes convergent to one common point. In practice, although the optical axes of all cameras may not converge to one point, it is always possible to determine one 3D point which is closest to all optical axes.

The 3D convergent point is generally the center of the observed scene, and can be considered as the representative point of the scene. Therefore, the depth of the convergent point provides important information about the scene depth, and can be set as the initial value and used to determine appropriate search range and step size for the global depth search. In the following, the convergent point is computed by solving a set of linear equations.

Suppose the convergent point is represented by M_c and M_i in the world and camera coordinate system respectively. Because the point is on the optical axes, M_i has non-zero value only in z component and can be expressed as follows

$$M_i = [0, 0, z_i] \tag{3}$$

According to the formula (1), we have

$$M_c = R_i M_i + t_i \tag{4}$$

From (3)(4), we can get $3m$ linear equations about M_c and the depth values $z_1, z_2, \cdots, z_m$ for a capture system with m cameras, which can be easily solved by linear least square method.

For parallel camera systems, since there is no convergent point, we can first compute the global disparity using the method in the JVT proposal (Ho, *et al.* 2006), and then get depth information according to the inverse proportion relation between depth and disparity.

3 GLOBAL DEPTH SEARCH

To reduce the rate for depth side information and to improve code efficiency, in this section the global depth is searched to minimize the absolute difference between the synthesized view and the practical view. And the search range and step size is determined according to the initial depth value.

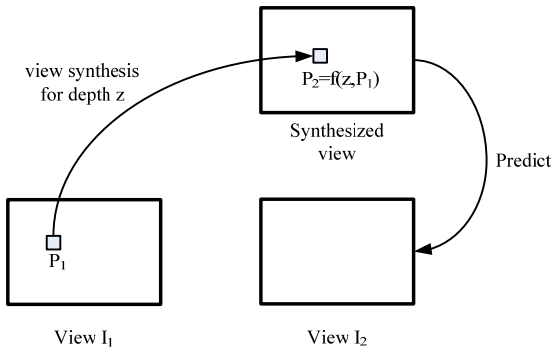


Figure 1: Depth based view synthesis and prediction.

Suppose a pixel P_1 in view I_1 has depth value z_1 in its correspondent camera coordinate system, and its correspondent pixel P_2 in view I_2 has depth value z_2 in the correspondent camera coordinate system. Then, according to formula (1) (2), we have

$$z_1 R_1 A_1^{-1} P_1 + t_1 = z_2 R_2 A_2^{-1} P_2 + t_2 \tag{5}$$

From (5), we can further derive that

$$z_1 B P_1 + Ct = z_2 P_2 \tag{6}$$

where $C = A_2 R_2^{-1}$, $B = A_2 R_2^{-1} R_1 A_1^{-1}$ and $t = t_1 - t_2$. Because P_1 and P_2 are homogeneous coordinates with the third component being one, we can get the correspondent pixel P_2 in view I_2 for pixel P_1 in view I_1 by eliminating depth z_2

$$P_2 = \frac{z_2 P_2}{z_2} = \frac{z_1 B P_1 + Ct}{z_1 b_3^T P_1 + c_3^T t} \hat{=} f(z_1, P_1) \tag{7}$$

where b_3 and c_3 are the third rows of B and C respectively. Formula (7) shows that the correspondent pixel P_2 in view I_2 for pixel P_1 in view I_1 is the function of the image coordinate P_1 and its depth z_1 when given with known camera parameters.

For view synthesis prediction using depth information, the pixel P_1 in view I_1 with the given depth z is projected to the correspondent pixel P_2 in the synthesized view using the camera parameters. According to the common assumption that the same scene point has the same intensity value in different views, thus the image value at pixel P_2 for the synthesized view $Synthesized\_I_2$ are set the same as P_1 in view I_1, i.e.,

$$Synthesized\_I_2(P_2) = Synthesized\_I_2(f(z, P_1)) = I_1(P_1) \tag{8}$$

The synthesized view is then used to predict view I_2, and the prediction error between synthesized and practical view is encoded, as shown in Fig. 1. In order to improve code efficiency, it is reasonable to minimize the absolute difference between the synthesized view and the practical view. Define the global sum of absolute difference between the synthesized view and the practical views as

$$\begin{aligned} GSAD(z) &= \sum_{k=2}^{m} \sum_{P \in I_1} \left\| Synthesised\_I_k(f(z,P) - I_k(f(z,P)) \right\| \\ &= \sum_{k=2}^{m} \sum_{P \in I_1} \left\| I_1(P) - I_k(f(z,P)) \right\| \end{aligned} \tag{9}$$

where the absolute difference is summed over the whole image and the other m -1 views. Then, the global depth is obtained by solving the following problem

$$\min_{z \in \{z_{\min}, z_{\min}+z_{step}, \cdots, z_{\max}\}} GSAD(z) \tag{10}$$

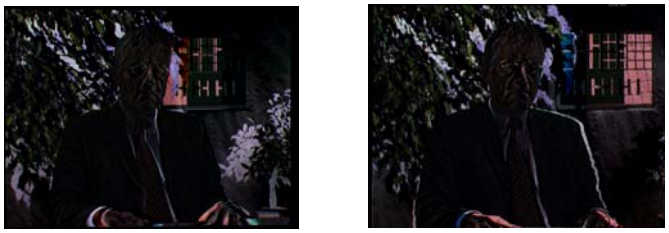
where $[z_{\min}, z_{\max}]$ is the depth search range and z_{step} is the step size.

It is asserted that the search range and step size have a substantial effect on coding gains. Yea *et al.* (Yea, *et al.* 2007) present a method to determine appropriate range and step using matching points, which is only effective when the feature track algorithm succeeds in finding the right match points robustly. However, in our work, the initial depth value provides important reference information for determining the search range and step size. Generally, we can set a local search range centered at the initial depth, and set the step size as a scale of the initial depth.

4 EXPERIMENTS

Experiments is tested on the standard test data "Uli" for MVC, which is provided by Heinrich-Hertz-Institute (HHI, German) and available at https://www.3dtv-research.org/3dav_CfP_FhG_HHI. "Uli" data is captured by 8 convergent cameras.

(a) view 6 (b) view 7 (c) view 8



(d) (e)

Figure 2: "Uli" images and difference images: (a)~(c): images from view 6, 7 and 8 respectively; (d)~(e): the difference images between practical views and synthesized views using estimated global depth, (d) for view 6 and (e) for view 8.

For simplicity, we only use view 6, 7 and 8 in the experiment, as shown in Fig.2. With the given camera parameters, the initial depth value is obtained by computing the convergent point through solving linear equations. The computed initial depth for view 7 is 3030.7mm, and the ground truth value of the bright reflection on the glasses of left eye is 3076.2mm, which is computed from the provided scene point M_W=[35.07, 433.93, -1189.78] in world coordinates. Since the depth change of the "Uli" scene is not large, the estimated initial depth is reasonable.

After get the initial depth, we can set appropriate search range and step size which makes the global depth estimated efficiently and quickly with less computation. In the experiment, the search range is set ±20% of the initial depth, and the step size is set 1% of the initial depth. When the projected pixel in the synthesized view does not fall on the integer-grid point, it is rounded to the nearest integer pixel. If the projected pixel is out of image, the image value for the pixel is set as that of the nearest image edge point. And the difference between the synthesized view and practical view is summed over the whole image using all the RGB components. The final estimated global depth for view 7 is 3151.9mm.

To intuitively show the estimation performance of the global depth, we give out the difference image between the practical view and the synthesized view based on the estimated global depth. From Fig. 2 (d) (e), we see that the differences in most areas are around zero (black areas), which show the estimated global depth is reasonable and right.

REFERENCES

Smolic, A. and Mueller, K. et al., 2007. Multi-View Video plus Depth (MVD) Format for Advanced 3D Video Systems. *ISO/IEC JTC1/SC29/WG11,Doc.JVT-W100.* San Jose, USA.

Scharstein, D. and Szeliski, R., 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Problem. *International Journal of Computer Vision* 47(1/2/3): 7-42.

Okutomi, M. and Kanade, K., 1993. A Multiple-Baseline Stereo. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 15 (4): 353– 363.

Kauff, P. and Atzpadin, N. et al., 2007. Depth Map Creation and Image-Based Rendering for Advanced 3DTV Services Providing Interoperability and Scalability. *Signal Processing: Image Communication*, Volume 22, Issue 2, Special Issue on 3D Video and TV, pp. 217-234.

Zitnick, C.L., Kang, S.B., Uyttendaele, M. Winder, S. and Szeliski, R., 2004. High-quality Video View Interpolation Using a Layered Representation. In *Proceedings of the ACM SIGGRAPH*, Los Angeles, CA, USA, pp. 600–608.

Ho, Y. S., Oh, K. J., et al., 2006. Global Disparity Compensation for Multi-view Video Coding. *ISO/IEC JTC1/SC29/WG11,Doc.JVT-T136*, Klagenfurt, Austria.

Vetro, A., 2007. Summary of BoG Discussion on View Interpolation Prediction. *ISO/IEC JTC1/SC29/ WG11, Doc. JVT-W133 ,* San Jose, USA.

Yea, S. and Vetro, A., 2007. Report of CE6 on View Synthesis Prediction. *ISO/IEC JTC1/SC29/WG11 , Doc. JVT-W059*, San Jose, USA.