Rate-Distortion Optimal Motion Estimation for Depth Map Coding

Kwan-Jung Oh*, Byung Tae Oh†, Kyungae Moon*, and Jinwoong Kim*

*Electronics and Telecommunications Research Institute, Daejeon, Korea
†Korea Aerospace University, Goyang, Korea

Abstract—This paper presents a rate-distortion (RD) optimal motion estimation for depth coding. The view synthesis distortion is widely used in 3D video coding since it is a good approach to reduce the depth bit-rate without degradation of rendering quality. However, view synthesis distortion method is only applied to a mode decision process until now. The proposed method applies the view synthesis distortion to motion estimation process and achieves 0.41% bit-rate saving for AVC-compatible 3D video coding.

I. INTRODUCTION

This document shows guidelines for preparing a final camera-ready manuscript in the proceedings of APSIPA ASC 2013. The format here described allows for a graceful transition to the style required for that publication. The depth map is an image that contains distance information of the surfaces of scene objects from a camera or certain viewpoint. It is represented as a gray scale image, where darker denotes greater depth value, and conversely, brighter means less depth value. The depth map is used in many fields such as robotics applications, medical imaging, three-dimensional (3D) applications such as 3D television, 3D model recovery, and digital holography. In general, the depth map can be obtained from two approaches: active approach and passive approach [1], [2].

In active approach, the special signal such as infrared (IR) signal, laser, structured light, and audio signal is projected into the objects and then the returned signal is received by detector. The depth map is then calculated based on the difference between the projected signal and detected signal. Time-of-flight (TOF), structured light pattern, light detection and ranging (LiDAR), and ultrasonic ranging are representative active sensors and Kinect is also included in active sensor. In passive approach, the depth map is indirectly estimated from the captured texture images without any active sensing techniques. The passive approach includes stereo matching, depth from focus, depth from defocus techniques. Generally, the passive depth acquisition system is cheaper than active method, but accuracy of depth information is relatively lower than active method. On the other hand, the active method can create more accurate depth but is more expensive than passive approach and have many limitations in acquisition environment.

The depth map compression has been studied in video coding field based on conventional video coding algorithms. The depth map coding algorithms can be classified into following three approaches. The first approach focuses on depth own characteristics. The depth image consists of multiple large planes and distinct edges. Thus, efficient edge preserving depth coding methods based on intra prediction [3] and in-loop filter [4] are proposed. The second approach utilizes the correlation with corresponding texture data. Inside view motion prediction (IVMP) [5] and motion parameter inheritance (MPI) [6] are good example tools to exploit motion correlation between texture and depth data. The last approach focuses on depth functionality itself. In many depth based applications, depth accuracy has a trade-off relation between rendering quality and bit-rate. The rendering quality means not depth quality itself, but the quality of the rendered results using decoded depth map. In the 3D video coding, the quality of synthesized views from the decoded texture and depth denotes rendering quality. There are several approaches for considering above concept using view synthesis distortion. In the previous method, the view synthesis distortion (VSD) [7] is calculated by direct or indirect method and applied to the rate-distortion (R-D) mode decision process only. The VSD is also can be combined with existing other distortion metrics such as sum of absolute difference (SAD) and sum of squared error (SSE).

In this paper, we propose a view synthesis distortion based motion estimation for depth coding. The motion estimation is a key technique in inter coding and it is a prior process of a mode decision. Although, VSD is adopted as distortion metric in mode decision process its efficiency is not maximized since currently motion estimation process adopts the SAD and SSE instead of VSD. That is, the best motion vector is only tested with VSD metric for certain inter mode. The proposed depth motion estimation metric is designed based on both view synthesis distortion and conventional distortion metrics. The proposed algorithm is tested with the 3D-AVC compatible codec which used in joint collaborative team for 3D video (JCT-3V) group.
II. RD OPTIMIZATION IN 3D VIDEO CODING

Rate-distortion (RD) optimization in video compression is a problem, how to maximize the video quality for a given bit rate or to minimize the bit rate necessary for representing a certain level of video quality. In 3D video coding, two types of RD optimization are used. For texture coding, conventional RD optimization used in 2D video coding is used. The depth coding also employs similar RD optimization but view synthesis distortion is used as distortion metric. More details are described in following sub-sections.

A. 3D Video Coding

The 3D video has been studied to satisfy viewer’s demands on more realistic 3D scene and multiview plus depth (MVD) system in Fig. 1 is one of the promising solutions of 3D video system. The MVD data of a single view consists of a 4:2:0 texture image and a 4:0:0 depth image. In general, the number of views for depth is either the same as or less than the number of views for texture [8].

![Fig. 1 Overview of the 3D video system.](image)

To support the MVD format, MPEG issued a call for proposals (CfP) for 3D video coding in March 2011. At the 98th MPEG meeting held in Genova, various H.264/AVC compatible and HEVC (high efficiency video coding) compatible proposals were evaluated and now both AVC-based and HEVC-based 3D video coding have been studied in JCT-3V group. The 3DV-AVC codec supports the mixed resolution 3D video consisting of full-size texture video and reduced-size depth video. It also supports flexible coding order for the texture and depth components. In this paper, our works are done based on 3DV-AVC.

B. Lagrangian RD Optimization

Generally, Lagrangian RD optimization [9] is used in image and video coding fields. Lagrangian optimization is model as

\[ J = D + \lambda R \]  \hspace{1cm} (1),

where \( J \) denotes the Lagrangian cost and a smaller \( J \) value means the optimal status. \( D \) represents distortion, \( R \) denotes bit rate, and \( \lambda \) is a Lagrangian multiplier and a smaller \( \lambda \) give more emphasis to minimizing distortion allowing a higher bit rate, whereas a larger \( \lambda \) tend to minimize bit rate at the expense of a higher distortion. That is, \( \lambda \) represents the relation between bit rate and quality and modeled relating to quantization parameter (QP) in general.

C. RD Optimization in Texture Coding

The Lagrangian method is used to find the optimal motion vector for inter coded block and the optimal coding mode for any block. In the H.264/AVC motion estimation, the motion information is determined by minimizing

\[ J(mv, \lambda_{motion}) = D(mv) + \lambda_{motion} R(mv - pmv) \]  \hspace{1cm} (2),

where \( mv = (m_x, m_y) \) represents the candidate motion vector, \( pmv \) denotes the predicted motion vector, and \( \lambda_{motion} \) is the Lagrangian multiplier. The rate term \( R(mv - pmv) \) denotes the estimated bit rate to encode motion information. \( D(mv) \) denotes the SAD and it is calculated as

\[ D(m_x, m_y) = \sum_{(x,y) \in B} |I_o(x, y) - I_{c}(x + m_x, y + m_y)| \]  \hspace{1cm} (3),

where \( I_o \) and \( I_c \) represent original video signal and coded video signal, respectively, and \( (x, y) \) denotes a pixel position in block \( B \). All possible motion vectors within a motion search range are investigated based on above criterion and the one with minimum cost is selected as the best motion vector.

Afterward, the best mode decision is performed to select the RD optimal coding mode by minimizing

\[ J(o, r, mode, \lambda_{mode} | QP) = D(o, r, mode | QP) + \lambda_{mode} R(o, r, mode | QP) \]  \hspace{1cm} (4),

where \( o \) means original block and \( r \) denotes reconstructed block, \( \lambda_{mode} \) is the Lagrangian multiplier, and \( QP \) denotes quantization parameter. The rate term \( R(o, r, mode | QP) \) is the real bit rate calculated after entropy coding. \( D(o, r, mode | QP) \) denotes the distortion for certain mode with QP and sum of squared error (SSE) is used in mode decision. \( D(o, r) \) is calculated as

\[ D(o, r) = \sum_{(x, y) \in B} (I_o(x, y) - I_r(x, y))^2 \]  \hspace{1cm} (5),

where \( I_o \) denotes the original video signal and \( I_r \) represents the reconstructed video signal, and \( (x, y) \) denotes a pixel position in block \( B \).

In (2) and (4), the Lagrangian multiplier \( \lambda_{motion} \) and \( \lambda_{mode} \) have the following relation with QP:

\[ \lambda_{motion} = \sqrt{\lambda_{mode}} \]  \hspace{1cm} (6)
D. RD Optimization in Depth Coding

In 3D video coding, most texture coding tools can be directly applied to depth coding. However, RD model should be modified proper to depth coding. In depth coding, rendering quality is more important than depth quality itself. Thus, the distortion metric in RD model is newly designed in terms of rendering quality. The view synthesis distortion is the representative rendering distortion metric used in depth coding and it can be written as

\[
VSD(T, S) = \sum_{(x,y)} \left| S_{x,y} - \tilde{S}_{x,y} \right|^2
\]  

where \( T \) and \( D \) denote texture and corresponding depth views, \( S \) represents synthesized view, and \((x', y')\) is warped coordinate from the \((x, y)\) by warping function \( f_w \). However, the warping is too complex calculation to be done for every pixel and not proper to block based processing since sometimes there would be no matched pixel in the synthesized view by occlusions. Thus, we introduced a simple view synthesis distortion metric in our previous work [7], where the VSD is modeled as

\[
VSD(T, S) = \sum_{(x,y)} \left| S_{x,y} - \tilde{S}_{x,y} \right|^2 \approx \sum_{(x,y)} \left| \tilde{r}_{x,y} - \tilde{r}_{x,y - \Delta p(x,y)} \right|^2
\]  

where \( \Delta p \) denotes translational rendering position error and it is proportional to depth error as

\[
\Delta p(x, y) = \alpha \times (D_{x,y} - \tilde{D}_{x,y})
\]

where \( \alpha \) is proportional coefficient determined by

\[
\alpha = \frac{f \cdot B}{255} \left( \frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right)
\]

where \( f \) denotes a focal length, \( B \) represents a baseline distance between current view and synthesized view, and \( Z_{near} \) and \( Z_{far} \) are the nearest and farthest depth values, respectively. \( \alpha \) determines the ratio between depth distortion and its view position shift in the synthesized view domain.

III. PROPOSED RD OPTIMAL DEPTH MOTION ESTIMATION

Motion estimation (ME) process is a key part in video coding since it takes most of the encoding time and total coding performance of video compression highly depends on it. To reduce the complexity of motion estimation process, many fast ME algorithms have been studied. Enhance predictive zonal search (EPZS) and unsymmetrical-cross multi-hexagon grid (UMHexagon) search are representative fast ME methods used in H.264/AVC. However, more important thing is a metric used in ME process.

As mentioned in previous section, we developed VSD in our previous work and applied it mode decision in depth coding. However, depth motion estimation is still done by \( J_{motion} \) in (2). It means that depth motion estimation process is not optimized in terms of VSD. Thus, we propose RD optimal motion estimation based on conventional VSD. Fig. 2 shows the overall concept of the proposed RD optimal motion estimation for depth coding.

In our previous work, we also proposed the joint cost function based on both VSD and conventional \( J_{motion} \). The new cost function is defined as

\[
J_{joint} = p \cdot J_{motion} + q \cdot J_{VSD}
\]

\[
= \frac{p \cdot [SAD + \lambda_{motion}] + q \cdot [VSD + \lambda_{VSD}]}{p \cdot SAD + q \cdot VSD} + \frac{p \cdot \lambda_{motion} + q \cdot \lambda_{VSD}}{p + q}
\]

where \( p \) and \( q \) denote weighting factor for SAD and VSD, and \( p+q=1 \). In depth coding, the best coding performance is achieved when \( p \) is larger than \( q \) in motion estimation and \( p \) is smaller than \( q \) in mode decision process. The optimal values can be obtained experimentally.

IV. EXPERIMENTAL RESULTS

The proposed RD optimal depth motion estimation was implemented in 3DV-ATM version 0.9 and tested under common test condition (CTC) of the JCT-3DV. All seven test sequences were run on a 3-view configuration and all adopted tools were enabled. The coding results compared using the BDBR (Bjøntegaard delta bit rate) metric [10] in terms of the total bit rate and the PSNR of decoded texture views, as well as the total rate and PSNR of rendered virtual views. The virtual views were synthesized using VSRS 1-D fast software [11].

We used \( J_{joint} \) in (11) instead of \( J_{motion} \) in (2) for depth motion estimation and several combination of \( p \) and \( q \) values are tested. Table I and Table II show the coding results on proposed RD optimal depth motion estimation.
As shown in above results, the best coding gain is achieved when \( p:q = 2:1 \). For reference, the best combination of \( p \) and \( q \) values for mode decision is known as \( p:q = 1:4 \). It means that to keep the depth quality is more important in motion estimation process than in mode decision process. Encoding time increases about 10%.

V. CONCLUSIONS

The depth coding performance can be improved by RD optimization considering rendering quality. In our previous work, we proposed VSD metric which is simple and efficient cost function designed considering rendering distortion. In this paper, we proposed RD optimal motion estimation for depth coding. In experimental results, 2:1 ratio between depth quality and rendering quality shows the best 0.41% BD gains in terms of total bit-rate and synthesized PSNR. In the future, we will study on reduction of complexity caused by RD optimal motion estimation.

ACKNOWLEDGMENT

This research was supported in part by GigaKOREA project (GK14D0100, Development of Telecommunications Terminal with Digital Holographic Table-top Display) and Basic Science Research Program through the National Research Foundation of Korea (NRF funded by the Ministry of Science, ICT & Future Planning (NRF-2013R1A1A1057779).

REFERENCES


