# FAST DEPTH CODING IN 3D-HEVC USING DEEP LEARNING

A DISSERTATION
SUBMITTED TO THE
DEPARTMENT OF ELECTRONIC AND INFORMATION ENGINEERING
OF THE HONG KONG POLYTECHNIC UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Zhen-xiang WANG
October 2017

CERTIFICATE OF ORIGINALITY

I hereby declare that this dissertation is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written nor material which has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

_____(Date)

# Abstract

The 3D Extension of the High Efficiency Video Coding standard (3D-HEVC), which has been finalized by the Joint Collaborative Team on Video Coding (JCT-VC) in February 2015, is the new industry standard for 3D applications. The 3D-HEVC provides plenty of advanced coding tools specifically for addressing the coding of auto-stereoscopic videos which have the format of multiple texture views along with the depth maps which are responsible for synthesising intermediate views with sufficient quality for auto-stereoscopic display. The provided tools take advantage of the statistical redundancies amongst texture views and depth maps in the video sequences, as well as the unique characteristics of depth maps to significantly shrink the bit-rate while preserving the objective visual quality of the 3D videos. However, those tools with high capability in terms of compression come with the high complexity of computation which has made the encoding time of the 3D video sequences much longer than ever by traversing a lot more candidates, calculating time-consuming RD Cost for each of them, especially in the wedgelet searching process for depth maps. While this full-search style method can promise to find the best candidate in depth intra mode decision, the time cost is expensive.

In this dissertation we address the time cost by presenting a new intra mode decision method for depth maps, leveraging the deep convolutional neural networks to predict the wedgelet angles for the depth blocks. The predictions from the learned models are capable of reducing the number of wedgelet candidates by half as well as the angular modes in depth map coding. The size of the neural network has been carefully designed to balance the trade-off between the time cost of model prediction and the model prediction accuracy. Confusion matrix is used to monitor the training process. Top-K criteria is employed for the prediction. We have integrated the learned models into the reference software of 3D-HEVC for the experiments. The compiled executable binaries are able to harness the power of the simultaneous computation of CPU, as well as the parrallel computation of GPU to accelerate the predictions. The simulation results show that the proposed algorithm provides 64.6% time reduction in average while the BD performance has a tiny decrease comparing with the state-of-the-art 3D-HEVC standard.

# Acknowledgments

Allow me first to give sincere thanks to my supervisor, Dr.Yui-Lam Chan, for his extremely generous support, most insightful advices and innumerable yet constructive feedback. I learned from him to first identify a problem, by reading a vast amount of articles to know what people have achieved and what bottlenecks they have encountered. I learned how to read papers, how to organize them to become the inner comprehension. He guided me to use the machine learning approach to solve the problem that has been found in the first stage. Without his guidance I will not have the idea to learn the deep learning technology and apply it to optimize the video coding. His encyclopedic knowledge and charming personalities made him my mentor in both research and life. I wish to thank Dr.Sik-Ho Tsang, for our in-depth discussions from which I can always find useful clues to proceed to next step. His great expertise in video coding significantly benefits me during my intensive period of learning. Also I would like to thank my friends Alex and Jacky, for our extensive discussions about artificial intelligence and their applications. Finally thank you my parents, for the great love and constant encouragement which give me confidence to face and handle all the challenges at every moment.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Video is the medium to record, copy, playback, broadcast and display the motion images in an electronic style [1]. Watching videos is becoming an important way for our entertainment as well as education. The high definition (HD) and ultra high definition (UHD) video are increasingly demanding nowadays. People prefer videos with higher definitions than those with lower resolutions because the former one provides much better viewing experience. However, challenges emerged for delivering videos with high definition. HD videos typically contain much more information in every picture frame than the standard definition videos. More data needs to be squeezed into the same capacity for transmission. For example, the uncompressed video with the dimension 720 x 480 at 30 frames per second requires 0.03 gigabytes per second, while the uncompressed video with the dimension 2880 x 2048 at 120 frames per second requires 2.12 gigabytes per second. Since bit rate is proportional to system bandwidth for transmission [2], and expanding the bandwidth in a large scale is too expensive, the significantly increased bit rate for transmitting the video data is becoming one of the major obstacles for HD video services.

To cope with the growing need for higher compression of moving pictures [3], Joint Collaborative Team on Video Coding (JCT-VC) [4] has developed the High Efficiency Video Coding standard which is the newest international video coding standard for substantially ameliorate the compression performance against the previous standards. Comparing with the H.264 Advanced Video Compression Standard [5], the H.265 High Efficiency Video Coding Standard provides fifty percent bit rate reduction while maintaining the objective video quality at the same level.

While Two-dimensional video is the most common video type, Three-dimensional (3D) video has been brought to market via lots of ways, including Blu-Ray disc, cable and satellite transmission, terrestrial broadcast, and streaming or downloading from the Internet [6]. 3D video provides the perception of depth information which augments the vividness of the video contents. Currently most

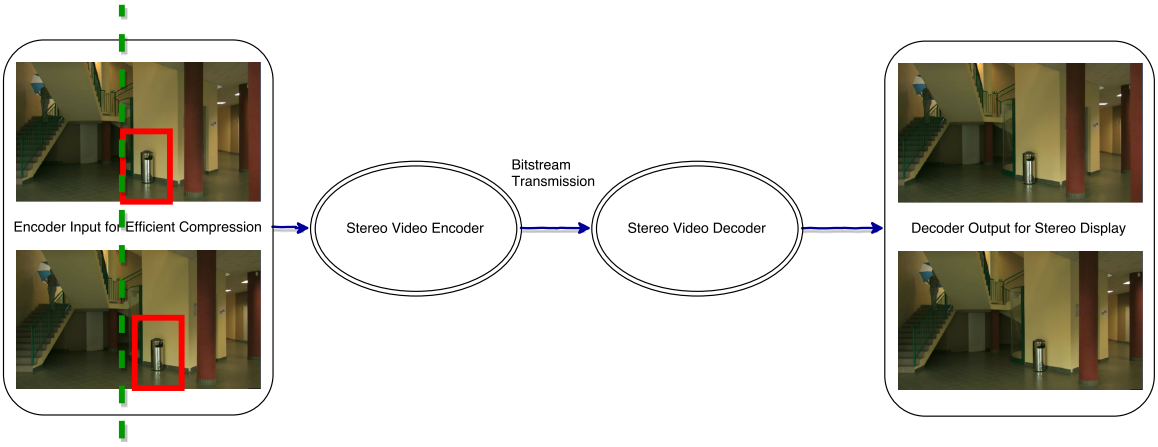


Figure 1.1: System Structure for transmitting videos targeting stereo display.

3D videos in the market are using stereo display technology. Two similar views, one for left eye, the other for right eye, are presented at the same time with the multiplexing techniques enabling the adjustments of video geometry information [7] to provide the 3D effect. Figure 1.1 illustrates the typical system structure for transmitting videos targeting stereo display. It can be observed that there exists a displacement between the two views. The green vertical left margins of the red rectangles in the two views at encoder side are different. Such a displacement is the visual disparity for 3D perception. Stereoscopic videos [8] have achieved great profitability for movie theatres in recent years. For example, IMAX 3D has became the most popular one that offering the immersing multimedia experiences around the world. Special 3D glasses are needed for watching the IMAX 3D movies. The current 3D film industry is very successful in terms of attracting customers, however, it is not the end of the story. Myopic people do not like to wear one more pair of glasses when watching 3D movies. Some people will experience discomfort after wearing the 3D glasses for a period of two hours. To get rid of the undesired 3D glasses, autostereoscopic multi-view technology [8] is coming to our rescue. The two major different characteristics between stereo display and autostereoscopic display are listed in Table 1.1 [9]. The impact of different view numbers for autostereoscopic display is shown in Table 1.2 [9]. Comparative ease can be brought to the 3D video audience since they do not need to wear 3D glasses for watching autostereoscopic videos. At each different view position, scenes with minor differences are available from multiple stereo pairs which

Table 1.1: Characteristics Comparison of Stereoscopic Display and Autostereoscopic Display

| Characteristic | Stereo Display | Autostereoscopic Display |
|---|---|---|
| Glass-Free | No | Yes |
| Multiple Stereo Pairs | No | Yes |

Table 1.2: Impact of Available View Amount for Autostereoscopic Display

| Characteristic | Small Number of Views | Large Number of Views |
|---|---|---|
| Seamless View Transition | No | Yes |
| High Quality of Scene Depth | No | Yes |

are provided by autostereoscopic display [9]. As a result, when audience make a move for various view positions, scenes not viewable from the previous locations are revealed during the movement. The autostereoscopic multi-view display demands more than two views. With a sufficient amount of views present in autostereoscopic display, the disparities between every two adjacent views can be small enough to offer seamless transitions from scene to scene, such that when multiple views meet eyes sequentially, the scenes as a whole can be gorgeous. The visual quality of the autostereoscopic display is highly proportional to the number of available views. Due to limited available bandwidth, transmitting arbitrary number of views is not practical. Researchers have proposed a new format which only requires limited number of view and their associated depth maps for the capability of generating arbitrary amount of views theoretically. The typical system structure for using this new format to compress and supply 3D video resources is shown in Figure 1.2. An enormous amount of views in the medium positions which are able to guarantee the high quality of the 3D video can be synthesized from the decoded texture frames in combination with decoded depth maps.

To employ multi-view plus depth format for 3D video, efficient compressing methods are desired, which has led to the 3D Video Coding Extension of the High Efficiency Video Coding Standard

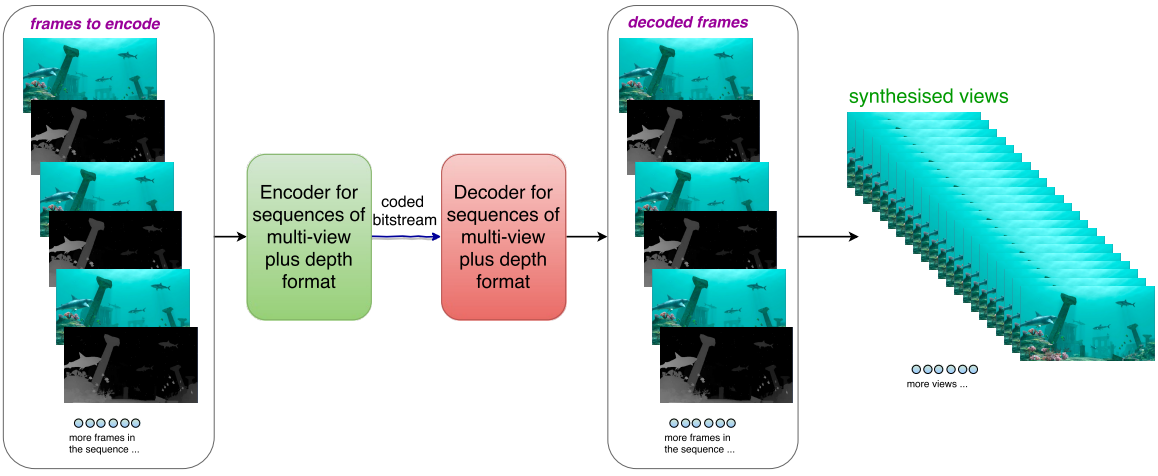


Figure 1.2: System Structure for transmitting videos of Multi-view Plus Depth format.

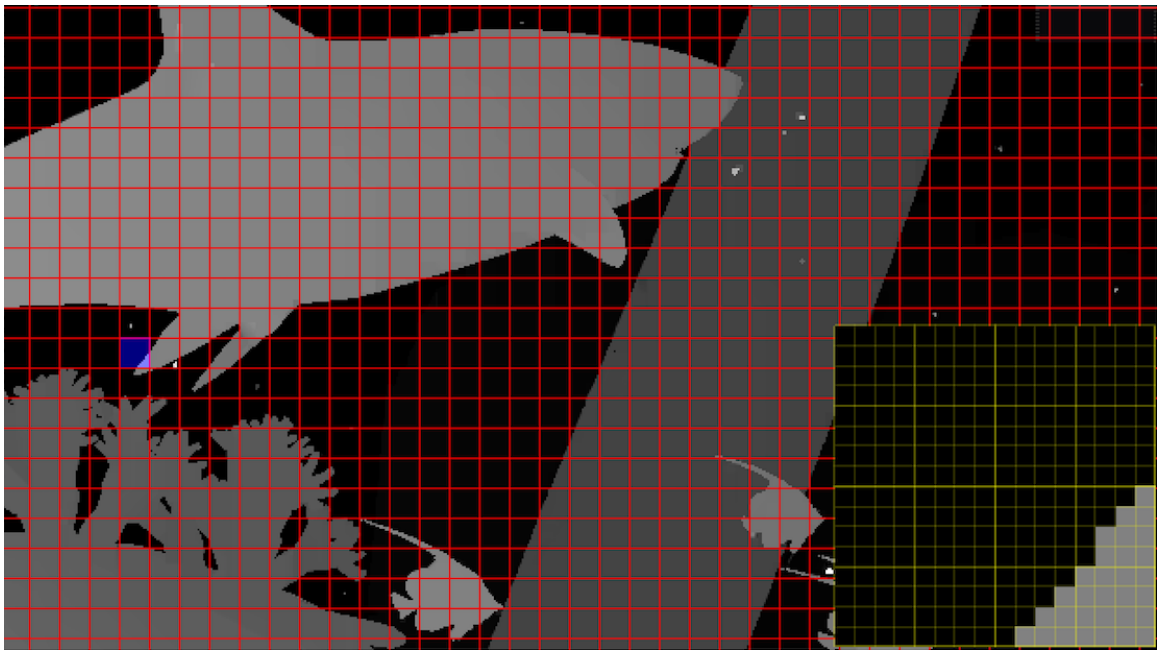

Figure 1.3: Example of wedgelet partition in a block of size 16 by 16 in depth map from Shark video sequence.

(3D-HEVC) by the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) [10]. The 3D Extension of the HEVC standard gives extra coding efficiency for encoding a few texture views along with the corresponding depth maps by using new tools which exploit the redundancies amongst texture and depth views, and pay attention to the unique characteristics of the depth maps, such as large homogeneous regions separated by sharp boundaries [11].

Depth information measures of the distance between the object in the far position and the object in the near position from a static viewpoint, which is expressed in the format of depth map. Instead of presenting depth maps directly to the viewer, views in the medium positions are generated by Depth-Image-Based Rendering (DIBR) technique. The qualities of the depth maps are vital to the DIBR process. Corona artifacts (a.k.a. ringing artifacts) can be discovered in synthesized views if the edge sharpness in depth maps can not be well preserved. Therefore, retaining the edge sharpness in depth map is the key to avoid the artifacts in the synthesized views. In 3D-HEVC, new intra-picture prediction tools and residual coding methods have been applied to preserve the special properties of depth maps. Depth Modelling Mode (DMM) which is one of the new intra-picture prediction tools, is designed to provide much more granularity for encoding the depth maps than the normal angular intra prediction modes. DMM is more capable of approximating the depth maps to be encoded due to the fact that it provides a vast amount of non-rectangle partitions. Figure 1.3

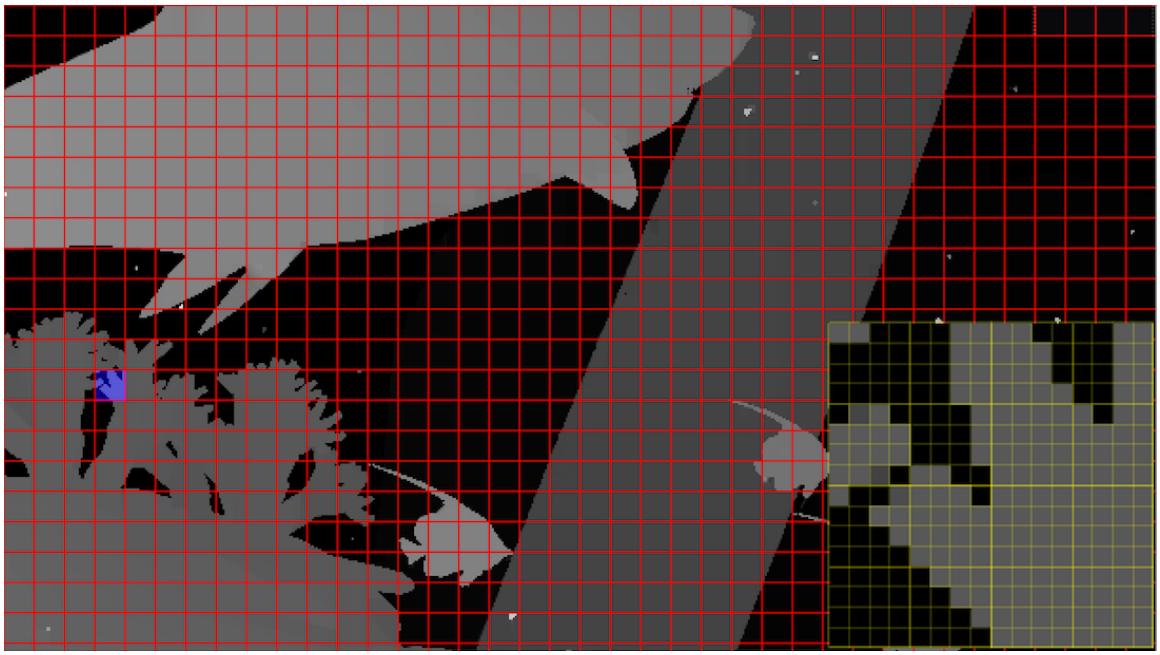

Figure 1.4: Example of contour partition in a block of size 16 by 16 in depth map from Shark video sequence.

presents an example of the wedgelet partition from the depth map in Shark video sequence. The small block highlighted by blue color amongst the blocks separated by the red grid is magnified at the right-bottom position in Figure 1.3. A straight line is used for the partition in wedgelet mode. Figure 1.4 shows a sample of the contour partition from the same depth map as Figure 1.3. The partition pattern comprises contour lines instead of one single straight line. Wedgelet partition and contour partition for depth maps are enabled by DMM1 and DMM4 separately.

## 1.1 Motivation

The idea of this work originates from the discovery of the computational complexity of the wedgelet searching process in depth modelling modes. The immense complexity for searching the best wedgelet candidate lead to the strongly marked increase of encoding time. The time consumed for compressing a single depth map in 3D-HEVC encoder is roughly a sixfold increase relevant to the encoding time of a single texture frame wherein the all intra configuration in HTM-16.2. Thus we designed a deep neural network architecture which is trained subsequently for predicting the most probable wedgelet candidates. The learned model achieves 92.2% to 97.3% top-16 accuracy for various block sizes. The inference engine is integrated into the reference software (HTM-16.2) of 3D-HEVC. The learned models reduce roughly half of the wedgelet searching candidates. It provides 64.6% time

reduction in average while the BD performance has a negligible decrease comparing with the unmodified 3D-HEVC encoder.

**Motivation for Wedgelet Candidates Reduction:** The encoding time consumed in HTM-16.2 encoder for each view (including both texture and depth) by default can be observed from the terminal (Linux/macOS) outputs or command line (Windows) outputs. An example of the encoding time for each view from the GhostTownFly sequence is shown in Table 1.2.

## 1.2 Contribution and Dissertation Outline

# Chapter 2

# Background

With the rising popularity of the high definition videos, the new standard termed High Efficiency Video Coding (HEVC) for compressing videos in a more efficient way comparing with previous standards, such as H.264/AVC, has emerged under the efforts from the Joint Collaborative Team on Video Coding (JCT-VC). In the meanwhile, five extensions of the HEVC standard, comprising Format Range Extension (RExt), Scalability Extension (SHVC), Multi-view Extension (MV-HEVC), 3D Extension (3D-HEVC), Screen Content Coding Extension (SCC), have been finalized from 2014 to 2016 to support fulfill extra requirements in various scenarios.

## 2.1 Video Coding

... ...

# Bibliography


[1] *Video*, Web Page, 2017. [Online]. Available: http://hidefnj.com/video.html.

[2] C. E. Shannon, *The mathematical theory of communication*. Urbana: Urbana : University of Illinois Press, 1949.

[3] "Itu-t recommendation database," 2017. [Online]. Available: http://www.itu.int/ITU-T/recommendations/rec.aspx?rec=12905&lang=en.

[4] "Jct-vc - joint collaborative team on video coding," 2017. [Online]. Available: http://www.itu.int/en/ITU-T/studygroups/2013-2016/16/Pages/video/jctvc.aspx.

[5] I. E. Richardson, *The H.264 Advanced Video Compression Standard*, 2nd ed.., ser. H.264 Advanced Video Compression Standard 2e. Hoboken: Hoboken : Wiley, 2010.

[6] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the h.264/mpeg-4 avc standard," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642, 2011, ISSN: 0018-9219. DOI: 10.1109/JPROC.2010.2098830.

[7] J. Konrad and M. Halle, "3-d displays and signal processing," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 97–111, 2007, ISSN: 1053-5888. DOI: 10.1109/msp.2007.905706.

[8] I. Sexton and P. Surman, "Stereoscopic and autostereoscopic display systems," *Signal Processing Magazine, IEEE*, vol. 16, no. 3, pp. 85–99, 1999, ISSN: 1053-5888. DOI: 10.1109/79.768575.

[9] Mu, amp, X, K. Ller, P. Merkle, and T. Wiegand, "3-d video representation using depth maps," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 643–656, 2011, ISSN: 0018-9219. DOI: 10.1109/JPROC.2010.2091090.

[10] "Jct-3v - joint collaborative team on 3d video coding extension development," 2017. [Online]. Available: http://www.itu.int/en/ITU-T/studygroups/2013-2016/16/Pages/video/jct3v.aspx.

[11] G. Tech, K. Ying Chen, J.-R. Muller, A. Ohm, A. Vetro, and A. Ye-Kui Wang, "Overview of the multiview and 3d extensions of high efficiency video coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 26, no. 1, pp. 35–49, 2016, ISSN: 1051-8215. DOI: 10.1109/TCSVT.2015.2477935.