

Reviewer's opinion
on Ph.D. dissertation authored by
Dominika Klóska
entitled:

Immersive Video Coding Using Decoder-Side Depth Estimation

1. Problem and its impact

What is, in your opinion, the most important problem discussed in the dissertation?

This doctoral dissertation focuses on immersive video coding, particularly Decoder-Side Depth Estimation (DSDE) techniques. The work's main objective is to improve the efficiency and visual quality of 6DoF (Degrees of Freedom) video systems by contributing to and improving the MPEG Immersive Video (MIV) standard.

The author introduces novel approaches to selecting views that consider occlusions in 3D scenes, and proposes more effective depth estimation algorithms to reduce artefacts caused by lossy video compression. The dissertation also proposes a hybrid coding approach that uses additional depth information to accelerate processing while maintaining moderate bandwidth requirements.

The solutions developed have contributed to the international MIV standard and resulted in a granted patent. This research improves the efficiency and visual quality of 3D video systems, enabling users to explore freely virtual 3D environments while preserving high-quality visual presentation.

Is it a scientific one?

Yes, the undertaken research problem is a scientific one. The author undertakes the theoretical mathematical aspects of multi-view video compression by formulating three theses:

1. *The selection of views transmitted for Decoder-Side Depth Estimation should depend on scene characteristics, specifically the proportion of occlusions.*
2. *It is possible to modify the depth estimation method to improve the quality of depth maps estimated from encoded views.*
3. *It is possible to enhance the quality of virtual view synthesis while maintaining the same bitrate as the state-of-the-art Decoder-Side Depth Estimation methods by leveraging additional input depth information at the decoder side.*

Ad. Thesis 1: Spatial sampling

Classical computer vision approaches to 3D scene reconstruction estimate depth from multiple viewpoints (at least two), forming a stereoscopic camera arrangement. Image blocks are then matched in the respective cameras using stereoscopic matching. Dense depth maps can be computed in this way, provided that there are no occlusions obscuring the view of a scene fragment in one of the cameras. A lack of correspondence between views is particularly severe when the baseline between the stereoscopic cameras increases. This is because the regions of the scene that are occluded become larger.

The author has proposed a solution to the above limitation by defining the spatial uniformity coefficient, $U = b/d$, where b is the baseline of the stereo pair (i.e. the distance between the two cameras in each pair), and d is the distance between the two stereo pairs (see Fig. 2.3.1.1 in the thesis). By monitoring the value of U in the camera setup, the author can model the projective geometry constraint and analyse the effect of camera arrangement on the Decoder-Side Depth Estimation process. This enables the encoder to dynamically select which camera configurations to transmit.

Significant quality gains can be achieved if occlusions exceed a 10% threshold (for smaller occlusion ratios, a uniform camera layout is maintained). This adaptive view selection scheme enables visual quality improvements of up to 1.44 dB according to the BD-IVPSNR (*Bjontegaard Delta Immersive Video Peak Signal-to-Noise Ratio*) metric for highly occluded sequences, such as Martial Arts, while keeping the encoding computation demand below 0.5%

Ad. Thesis 2: Robust Optimization Under Lossy Codec Distortion

In non-compressed stereovision systems, the matching procedure is based on undistorted images, and the corresponding image blocks are compared using correspondence algorithms. However, for lossy video encoders, image distortions can complicate this task because image features (particularly high-frequency components) are altered in a nonlinear manner.

To mitigate the issue described above, the author proposes the Point-to-Block Matching method, which is defined by Eq. (3.1.1). The underlying idea of this method is to modify image comparisons by evaluating pixel similarity over a localised spatial window (3x3 or 5x5 pixels) and selecting the minimum matching error. A drawback of this method is that the computational burden increases ninefold, however, the process can be efficiently parallelised.

Additionally, to address errors in foreground boundary propagation caused by lossy compression during graph-cut optimisation, the author modifies the hierarchical alpha-expansion method to dynamically use depth labels from neighbouring superpixel regions.

Evaluations conducted on five multiview sequences (with varying search space parameters) validated Thesis 2. These evaluations showed that the proposed optimisation method achieves an average 6.6% reduction in IV-PSNR-based BD-rate and accelerates decoding and rendering computation time by 62.2%.

Ad. Thesis 3: Cooperative Hybrid Coding

To prove thesis 3 the author puts forward and evaluates two novel hybrid methodologies, i.e. Input Depth Map Assistance (IDMA) and Extended IDMA (eIDMA). In these methods the

encoder selects a highly optimized and sparse subset of key views and the depth patches representing occluded elements to transmit. At the decoder side, these compressed depth maps are used to constrain and focus the search range of the decoder's global graph-cut optimization, allowing the estimator to correct compression errors. Additionally, a special content-adaptive mechanism is proposed that automatically switches between IDMA and eIDMA depending on whether the scene is natural or computer-generated. This hybrid cooperative framework achieved a BD-rate reduction exceeding 50% in comparison to unassisted state-of-the-art DSDE, thus successfully validating Thesis 3.

The methods used to prove the theses are based on scientific methodology utilizing a systematic framework evaluated under standardized MPEG Common Test Conditions across various natural and computer-generated test sequences. Furthermore, the proposed methods have been quantitatively verified using signal metrics (IV-PSNR, WS-PSNR, and BD-rate) and subjectively validated by human perception trials with 45 viewers conducted under standardized ITU methodologies (cited in references [40-41]).

Does it have a practical meaning?

The research problem addressed in the dissertation has significant practical potential. Multiview video technology is used in a wide range of advanced multimedia applications, including 3D television (3DTV), free-viewpoint television (FTV), virtual reality (VR), augmented reality (AR), and immersive video. For these techniques to be developed and used widely, effective video compression techniques must be implemented to make the deployment of these multimedia techniques affordable for users while offering high-quality video signals.

The specific practical applications of the proposed methods include:

- hybrid compression methods, specifically Input Depth Map Assistance (IDMA) and Extended IDMA (eIDMA), were formally ratified and adopted into the new Decoder-Side Depth Estimation (DSDE) profile of the second edition of the MPEG Immersive Video (MIV) standard,
- the developed algorithms have been officially integrated into standard reference software packages, including the Immersive Video Depth Estimation (IVDE) and Test Model for MPEG Immersive Video (TMIV) software,
- the proposed coding framework pipeline has been filed under USPTO patent no. US12530806.

4. Contribution

What is the main, original contribution of the dissertation?

The candidate's primary achievement is co-inventing and development of a patented, standardized hybrid immersive video coding methods termed Depth Map Assistance and Extended Input Depth Map Assistance.

The above scheme is based on three original processing procedures:

1. The developed occlusion-aware view selection method which allows the encoder to dynamically group transmitted cameras into stereopairs to minimize non-estimable depth regions in highly occluded scenes.

2. The Point-to-Block Matching algorithm that replaces conventional point-to-point correspondence with a localized spatial search window. This approach makes the estimator resistant to compression noise, edge-shifting and ringing artifacts.
3. An accelerated hierarchical graph-cut optimizer that incorporates second-cycle label offset and adjacent segment depth label queries. This mathematical scheme significantly reduces the computational demand for the optimization while preserving sharp physical depth boundaries and correcting spatial distortions caused by video codecs.

5. Correctness

The dissertation consists of five chapters (counting 74 pages), list of references, followed by copies of the publications of the author. The theses of the dissertation formulated in Chapter 1 have been addresses and proved in consecutive Chapters 2, 3 and 4 correspondingly.

The Author has applied a suitable theory, appropriate materials, experimental methods, and verification tools to prove the theses. In particular:

- projective geometry, camera models (intrinsic and extrinsic parameters), and 3D point reprojection rules were used to mathematically define and detect non-estimable regions,
- depth estimation is correctly formulated as a global pixel-labeling cost-function minimization problem,
- to solve NP-hard multi-label labelling problem, the hierarchical alpha-expansion-based graph-cut optimization was used,
- the experiments were conducted using a comprehensive dataset (both Natural Content (NC) and Computer-Generated (CG) content) including perspective, linear, arc, and spherical/semi-spherical multi-camera sequences (e.g., BBB Flowers, Carpark, Frog, Painter, Street, CBABasketball, MartialArts, Chess, ChessPieces, ClassroomVideo, Museum, Hijack),
- direct integration of the developed algorithms into the MPEG standardized reference software stack consisting of TMIV, IVDE, VVenC/VVdeC (VVC), x265 (HEVC) and VVS/VWS has been conducted,
- all objective and subjective evaluations were executed according to the MIV Common Test Conditions (CTC) established by the ISO/IEC MPEG Video Coding group, also a widely accepted metrics for verification the results such as: Bjøntegaard Delta (BD-rate & BD-PSNR/BD-IVPSNR), IV-PSNR (Immersive Video PSNR), and WS-PSNR (Weighted-to-Spherically-Uniform PSNR) were used.

A very helpful for the reader, but missing part of the thesis is the list of shortcuts and abbreviations.

6. Knowledge of the candidate

The candidate possesses expert knowledge in Decoder-Side Depth Estimation and immersive video coding. She co-designed the innovative IDMA and eIDMA coding methods, which have been granted USPTO patent titled: *Method for decoding immersive video and method for*

encoding immersive video accepted in 2023 and officially integrated into the international MPEG Immersive Video (MIV) standard.

Her peer-reviewed publication record includes four international journals, including a high-impact *IEEE Transactions on Broadcasting* (IF: 4.8) journal and three publications in national journals, showing high research expertise with contribution shares reaching 80% of the candidate. She is the co-author of three conference publications and the co-author of ten ISO/IEC MPEG documents. Her mathematical and algorithmic skills encompass spatial view selection, point-to-block matching, and graph-cut optimization. She has extensive experience of using software tools, including the Immersive Video Depth Estimation (IVDE) and the Test Model for MPEG Immersive Video (TMIV).

The wide bibliography included in candidate's dissertation counting more than 100 items shows her broad, multidisciplinary knowledge. It demonstrates that her expertise extends beyond conventional video compression and cover academic literature from a wide range of complementary fields.

7. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 13 of the *Act of 14 March 2003 of the Polish Parliament on the Academic Degrees and the Academic Title* (with amendments)¹, my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with X)

☒	☐	☐	☐	☐
Definitely YES	Rather yes	Hard to say	Rather no	Definitely NO

B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**, and particularly the area of?

☒	☐	☐	☐	☐
Definitely YES	Rather yes	Hard to say	Rather no	Definitely NO

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

☒	☐	☐	☐	☐
Definitely YES	Rather yes	Hard to say	Rather no	Definitely NO

Moreover, considering the following Author's research achievements, i.e.:

- the Patented IDMA/eIDMA Hybrid Framework that enables to reduce the transmission bitrate by up to 53% while accelerating decoding speed up to three times,
- Occlusion-Aware and Inhomogeneous View Selection that implements 3D point reprojection for dynamic camera grouping,

¹ http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf

- Noise-Resilient Matching and Accelerated Graph-Cuts which is a point-to-block matching similarity search to preserve sharp depth boundaries while accelerating optimization times,
- publications in IF journals and presentations at international conferences,

I **recommend distinguishing** the dissertation for its quality².

I hereby submit a request to the Scientific Council of Information and Communication Technology at Gdańsk University of Technology to admit the doctoral thesis of Dominika Klóska, MSc (Eng), to a public defence.



Signature

² Obviously, this sentence is optional.