

Research Center for Industrial Electronics and  
Multimodal Systems (CEIMM)

Universidad Politécnica de Madrid (UPM)  
eduardo.juarez@upm.es

**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?

Is it a scientific one?

Does it have a practical meaning?

The dissertation concerns the compression and transmission of immersive video, i.e, video which lets the viewer freely change position and viewing direction within a recorded scene. Such systems must deliver not only texture but also the geometry of the 3D scene. The usual solution, sending depth maps compressed with conventional video codecs, is inefficient because those codecs were not designed for depth data. This dissertation focuses on an alternative, the so-called Decoder-Side Depth Estimation (DSDE), in which depth is not transmitted but estimated at the decoder from the decoded views.

From my perspective, the most important problem discussed is how to make DSDE-based coding efficient in both rate-distortion performance and decoder complexity when the decoder has only a small number of lossy-compressed views to work with. The research breaks the problem down into three questions: which views should be transmitted, how the depth estimator should be adapted to the compressed input, and if a limited amount of encoder-side depth can improve the result at the same bitrate.

Definitely, the problem is a scientific one. It requires the understanding of how view arrangement, compression artefacts and the optimisation strategy affect depth estimation and the quality of synthesised views. To get this understanding an experimental approach is followed.

The problem also has a clear practical meaning. Bitrate, pixel rate and decoding time are the main obstacles to practical immersive video services, and the proposed methods address all three. The practical relevance is confirmed by the adoption of the proposals, co-authored by Dominika Klóska, into the second edition of the MIV standard and its reference software.

## **2. Contribution**

What is the main, original contribution of the dissertation?

If appropriate, you can make a distinction between what the Ph.D. candidate claims to be the main contribution and what you consider as the main contribution. If this is the case, indicate the reason for which you do not agree (e.g. it could be that somebody else has already proposed a given idea or it can be original but not correct due to some flaws described in Sec. 3 of the reviewer's opinion). You can also comment on practicality of the proposed solutions (it could be that the problem is highly practical, but the proposed solution is not). If applicable, you can refer to other quality indicators you know about (e.g. quality of publications by the candidate, patents authored by the candidate, citations, existing applications of the proposed solutions etc.).

The candidate presents four contributions:

*(a) View selection (Chapter 2; [P4], [P2]).* An analysis showing that evenly spaced views are preferable in a general transmission scenario, a modified selection rule for omnidirectional content that reduces the significance of vertical distances, and an occlusion-aware method for DSDE that chooses between evenly spaced views and stereo pairs depending on the proportion of occluded regions.

*(b) Point-to-block matching (Section 3.1; [P5]).* A modification of the data term of the cost-function in depth estimation problems that compares a pixel with a small block around the corresponding point, intended to tolerate decoded artifacts.

*(c) Graph-cut optimisation (Section 3.2; [P3]).* Two modifications of hierarchical alpha-expansion, a label offset in the second cycle of graph-cut optimisation and the examination of labels from neighbouring segments, which shorten the estimation and improve its average quality.

*(d) Input depth map assistance (Chapter 4; [P1], [P6]).* Two coding schemes, Input Depth Map Assistance (IDMA) and Extended Input Depth Map Assistance (eIDMA), in which a subset of depth maps and depth patches for eIDMA is transmitted and used by a modified estimator to refine and complete depth at the decoder. Additionally, an automatic mechanism to select between the two is introduced.

I agree with Dominika that (d) is the main contribution of the dissertation. It changes DSDE from a purely texture-based scheme into a hybrid one, yielding the largest reported gains. Additionally, it is the part that was adopted into the DSDE profile of the second edition of MIV.

The proposed solutions are compliant with the MIV bitstream, they were implemented in the reference software (TMIV, IVDE), and their computational cost is reported: the view selection adds 0.44% to encoding time and requires no decoder changes, and the modified graph cut reduces decoding and rendering time to about 62% of the baseline.

### **3. Correctness**

Can we trust what is claimed in the dissertation? Are the arguments correct? Indicate the flaws you have noticed, if any. Also point out those aspects concerning correctness that you value most (elegance of proofs, design of experiments, analysis of empirical data, quality of prototype software/hardware etc.).

I have some critical remarks that, from my perspective, do not invalidate the main conclusions of the dissertation. They only concern the strength of some of the evidences.

In Section 2.2, the objective results in the DSDE configuration show clear losses (for Museum, a BD-rate increase of up to 57.2%; BD-PSNR losses of up to 13.2% for the other sequences). The explanation is illustrated for one sequence at one rate point only (Table 2.2.2.3). From my perspective, an evaluation excluding the polar views, or weighting views by viewing probability for all sequences would make the argument more conclusive.

In Section 2.3, the conclusion is based on six sequences, four transmitted views in a single atlas, and one-dimensional camera arrangements. The data, from my point of view, could be also consistent with other threshold values, or with an influence of scene properties other than the occlusion ratio.

In Section 3.1, the gain is also present for uncompressed input (38.83 to 39.35 dB) and is of similar size to the gain for low compression, while for medium compression it is only 0.06 dB. The improvement seems not to be specific to compression artefacts, and the proposed mechanism is not isolated by the experiment.

In Chapter 4, atlas encoding time for eIDMA rises from 1.5 to 66 seconds per frame. This result could be acceptable offline, but it is relevant for the live scenarios.

The four contributions are evaluated separately, but the research does not show whether the gains add up when occlusion-aware view selection, point-to-block matching, the modified graph cut and input depth map assistance are used together. A joint experiment, even on a few sequences, would be valuable.

From the methodology point of view, no remark. The experiments follow the MPEG common test conditions and use publicly available reference software (TMIV, IVDE, VVenC), which makes the results reproducible and comparable with those of other groups. The dataset of Section 3.2 (21 sequences) and Chapter 4 (14 sequences) are broad and include both computer-generated and natural content. Computational cost is reported alongside quality in Sections 2.3 and 3.2 and in Chapter 4. Results are explained analysing how the bitrate is divided between texture, depth and metadata.

## 4. Knowledge of the candidate

What are the chapters of the dissertation (or sections in chapters) that resemble a tutorial and thus confirm a general knowledge of the candidate in the discipline of **Information and Communication Technology**. What areas of that discipline are covered by those chapters/sections? What do you think about quality of those chapters/sections? What is your opinion on the list of references? What is the degree of its completeness? Provide any other arguments in favour or against the claim that the candidate has general knowledge and understanding of the **Information and Communication Technology** discipline.

Section 1.1 presents the development from conventional and stereoscopic video to immersive video with six degrees of freedom, the Multiview Video plus Depth representation, depth maps and their acquisition, virtual view synthesis, simulcast coding, and the MIV standard with its Main and Geometry Absent profiles (view labelling, pruning, packing, atlases). The introductions to Chapters 2, 3 and 4 review, respectively, earlier work on view selection, the sources of inter-view matching errors and

similarity measures, and existing methods of assisting DSDE. Sections 3.1 and 3.2 explain depth estimation as energy minimisation and its solution by graph cuts (alpha-expansion and alpha-beta swap). The experimental sections show familiarity with objective quality metrics (WS-PSNR, IV-PSNR), and subjective testing according to ITU recommendations.

These parts cover several areas of the discipline: video compression and standardisation, three-dimensional and multiview video systems, depth estimation and stereo matching, discrete optimisation, and video quality assessment. They are clearly written and well-illustrated from my point of view.

As a small concern, neural scene representations appear only as future work and alternative immersive formats, such as point clouds, are not discussed.

The list of references contains 108 items. It is complete with respect to MIV standardisation and includes the classical works on graph cuts and superpixels. There are many ISO/IEC MPEG documents cited. Although it is understandable in the context of the research, for readers outside the MPEG community could be a barrier to get deeper into the details given the difficulties that can be faced to access the documents.

From my perspective, the dissertation shows a working knowledge of the whole immersive video chain, from acquisition and calibration to coding and quality evaluation.

## 5. Other remarks<sup>1</sup>

## 6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 187 of *the Act on Higher Education and Science of the Polish Parliament* (Dz. U. 2018 poz. 1668 with amendments)<sup>2</sup>, 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 **Immersive Video Chain**?

☒

*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*

---

<sup>1</sup> Optional

<sup>2</sup> [http://www.nauka.gov.pl/g2/oryginal/2013\\_05/b26ba540a5785d48bee41aec63403b2c.pdf](http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf)

Moreover, taking into account the excellent research work presented in the dissertation and the amplitude of the achieved results, I **recommend to distinguish** the dissertation for its quality<sup>3</sup>.

Firmado por JUAREZ MARTINEZ EDUARDO -  
\*\*\*3525\*\* el día 06/10/2026 con un certificado  
emitido por AC FNMT Usuarios

---

*Signature*

---

<sup>3</sup> Obviously, this sentence is optional.