

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 most important problem. In my opinion, it is not any one of the three theses taken separately, but the question that Decoder-Side Depth Estimation opens and that the dissertation answers as a whole: given a coding framework in which geometry is derived at the receiver, which textures and which depth information have to be coded and transmitted, and which can deliberately be kept out of the bitstream and recovered by estimation. This is a genuinely novel formulation, because it reverses the conventional split between encoder and decoder — information is no longer classified by what it represents, but by whether the decoder is able to reconstruct it. The dissertation attacks this question consistently at three levels: which views of texture have to be transmitted so that the geometry is estimable at decoder side (Chapter 2); how the estimator itself has to be adapted when it is fed with decoded rather than pristine textures, and how it has to be accelerated in order to fit the complexity budget of a decoder (Chapter 3); and which depth information should be reintroduced into the bitstream, for which views and in what form (Chapter 4, IDMA and eIDMA). That these three chapters are three answers to one and the same question is, in my view, the main intellectual achievement of the work, and it is a merit of the dissertation that all three are validated inside the same standardised framework.

Is it a scientific problem? In substance, yes — in its treatment, only partly. The underlying questions are genuinely scientific: how accurately geometry can be identified from a given camera configuration and baseline under occlusion is an identifiability question; replacing point-to-point correspondence by a minimum over a neighbourhood is a concrete claim about the robustness of a matching cost under displacement; how to select and encode part of depth maps to improve the geometry estimation at decoder side is also a valuable scientific problem. The dissertation, however, treats all of this empirically, and that is my main reservation.

Does it have a practical meaning? Yes — and this is where the work is strongest. MPEG introduced the Geometry Absent profile precisely because depth maps are expensive to transmit; anything that makes geometry reconstruction work at the decoder is directly a bandwidth result, and bandwidth is the binding constraint in immersive and volumetric delivery. The contributions have been ratified at the 140th MPEG meeting in October 2022 and integrated into the TMIV and IVDE reference software to support the DSDE profile of the second edition of *MPEG Immersive video* standard, and the results of Chapter 4 are covered by a granted US patent (US 12,530,806 B2, granted 20 January 2026). The

measured effects are of practical size: for eIDMA the decoder-and-rendering time falls to about 32 % of that of the unassisted DSDE anchor, the adaptive configuration reports a BD-rate improvement exceeding 50 % with respect to the same anchor, and the texture/depth bitrate split is explicitly tabulated (about 81.6 % texture and 18.4 % depth at the highest rate point).

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.).

What the candidate claims. The dissertation is built on three theses: (1) the selection of views transmitted for DSDE should depend on scene characteristics, specifically the proportion of occlusions; (2) the depth estimation method can be modified so as to improve the quality of depth maps estimated from encoded views; (3) the quality of virtual view synthesis can be improved while maintaining the bitrate of state-of-the-art DSDE methods, by leveraging additional encoder-side depth information at the decoder. I accept all three as formulated, and I consider each of them validated within the experimental paradigm adopted by the author.

As supplementary, I would like to indicate that the first of all contributions of the dissertation is, within the standardised framework of MIV with decoder-side depth estimation, raising the question of what must be transmitted and what may safely be left to the decoder as the variation of input contents: which textures are needed for the geometry to be estimable, which depth is worth its bits, and how the estimator has to be adapted once it works on decoded rather than pristine textures. Posed in this generality the question is new — the MIV specification leaves decoder-side depth estimation explicitly non-normative — and it is the right question for a coding scheme whose geometry is inferred instead of transmitted. I therefore regard the formulation itself as a key contribution, independently of how far each individual answer goes.

Practicality. It is high, and well evidenced. All proposals are codec-agnostic and bitstream-compliant with MIV. All experiments follow the ISO/IEC MPEG Common Test Conditions for immersive video, use the reference encoder (TMIV) and reference estimator (IVDE), and report the metrics the community itself uses (IV-PSNR, BD-rate, decoder-side runtime). The proposals are not simulations of an idea but working code inside the reference software, which MPEG then adopted. This is a level of validation that few dissertations reach.

Other quality indicators. The dissertation is composed of five publications: one in IEEE Transactions on Broadcasting ($M = 140$, $IF = 4.8$), two in Applied Sciences ($M = 100$ each) and two in the WSCG conference series ($M = 70$ each), plus a granted US patent listed as a supporting publication. The candidate is the first author of the two publications underlying Chapter 2 and a co-author of those underlying Chapters 3 and 4, and she declares her own contribution in each case; the declared shares are consistent with the author order. She additionally co-authored eleven ISO/IEC MPEG contributions in the relevant area and took part in research projects on immersive video coding. All of this is in the mainstream of the field and directly related to the dissertation.

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.).

General assessment. Yes — with the qualifications set out below, the claims of the dissertation can be trusted. The reported numbers are internally consistent, the experimental chain is the standard one, the metrics are appropriate, negative results are reported rather than hidden, and in almost all cases the conclusions drawn are no stronger than the evidence supports. My reservations are not about the correctness of what was measured; they are about the extent to which the measured effects can be attributed to the mechanisms proposed and carried over to other tools.

What I value most. Three things. First, the *design of experiments*: the use of the MPEG Common Test Conditions throughout, the choice of IV-PSNR and BD-rate for rate–quality curves that would otherwise not overlap, the per-GOP fixed view selection in Chapter 2, the explicit bitrate-split tables in Chapter 4, and the one-factor-at-a-time variation in Chapter 3 (compression strength, number of views, number of depth layers, superpixel count). Second, the *honesty of the empirical analysis*: results that contradict the hypothesis are reported rather than omitted (Street, -0.03 dB in Table 2.3.2.2; the configurations in which Point-to-block Matching is worse than point-to-point matching), and the author explains candidly why the decision threshold had to move from the 25 % quoted from the literature to approximately 10 % with modern tools. Third, the *quality of the prototype software*: the proposals were not evaluated in a private test bed but implemented in C++ inside the MPEG reference software (TMIV, IVDE), submitted to MPEG, adopted into the standard and, in the case of Chapter 4, protected by a granted patent. Very few dissertations can claim that their algorithms are inside an ISO/IEC standard.

Conclusions I consider established. I regard the following as sound and adequately supported. (i) Chapter 2: the proportion of occluded pixels is a useful and cheaply computable predictor of whether a camera arrangement will be favourable for decoder-side depth estimation, and clustering the transmitted cameras measurably helps for sequences with high occlusion (about $+1.2$ dB for MartialArts) while being neutral or harmful for sequences with low occlusion. (ii) Chapter 3: Point-to-block Matching with a 3×3 window improves the synthesised views on average over point-to-point matching in the tested configurations, and the combination of the second-cycle label offset with neighbouring-segment labels improves quality and speed at the same time relative to the hierarchical baseline. (iii) Chapter 4: transmitting depth selectively, for a subset of views and a subset of pixels, improves rate–quality and reduces decoder-side runtime relative to both unassisted decoder-side depth estimation and SEI-based geometry assistance. I found no error of measurement and no conclusion stronger than its evidence.

Chapter 2 — flaws. (a) The decision rule is incomplete. The occlusion ratio is defined precisely and measured cheaply (0.44 % of encoding time, once per GOP), but the paper never says which cameras are to form the stereopairs, how many pairs are to be used, or how to select the uniformity coefficient of acceptable pairing of cameras based a measured occlusion ratio instead of an empirical range (0.2–0.7). (b) The scope is narrow: one-dimensional line or arc rigs, four views, and effectively only the placement of the two inner cameras is optimised, while the outer cameras and the number of transmitted views are fixed.

Chapter 2 — recommendation. The encoder-side view selection should be formulated as a general, content-adaptive decision rule rather than as a heuristic over the middle two views of a four-view, one-

dimensional rig. Concretely, I would ask the author to (i) state the selection as a function of measurable content properties, e.g. occlusion ratio, (ii) evaluate it for a variable number of transmitted views and for two-dimensional rigs.

Chapter 3 — flaws. (a) The causal claim is not sufficiently supported by the ablation: there is no sufficient tests or analysis on the test results to demonstrate that the technologies proposed in 3.1 and 3.2 are designed for improving the accuracy of depth estimation with input of decoded texture. (b) The per-sequence dispersion is very large (for example -89.1% for Guitarist and $+25.4\%$ for Hijack in Table 3.2.3.3) and is averaged away, and BD-rate itself is unstable when the rate–quality curves are non-monotonic — a difficulty the author acknowledges in Chapter 2.

Chapter 3 — recommendation. First, a *dedicated ablation*: the same estimator, the same sequences, the same metrics and preferably the same rate points should be run on pristine textures and on decoded textures, and the gain obtained in the two conditions reported side by side. In the same spirit, estimation time should be reported for, and the per-sequence results should be shown in addition to the means. Second, an analysis and explanation *on the variation* performance across sequences should be provided; and a condition of usage of the proposed graph cut optimization can be suggested then after.

Chapter 4 — flaws. (a) The comparison is incomplete. IDMA and eIDMA are evaluated against two references only: the unassisted MIV DSDE anchor and DSDE assisted by encoder-derived features (Geometry Assistance SEI). What is missing is the other extreme of the design space, namely *MIV Main*, in which depth is transmitted for all basic *and* additional views. This comparison is essential, and not merely for completeness: MIV Main has by far the lowest decoder complexity, since no depth estimation is needed at all, so the claim that the proposed methods reduce decoder-side runtime must be measured against it, and not only against an anchor that is itself expensive. The dissertation mentions MIV Main once, in Table 4.1.1, and never tests it. I recommend a full three-way comparison — nothing transmitted (MIV GA/DSDE), everything transmitted (MIV Main), and the proposed selective transmission — including rate–quality, decoder runtime and encoder runtime. (b) The switch between IDMA and eIDMA is automatic but not content-adaptive in a defended sense: no validation of the threshold 0.1% for the portion of inconsistent pixels has been provided. (c) The partition into essential and non-essential views is inherited from the standard MIV view-labelling algorithm (a second pass over the basic views); no rule is given for how many views should be essential for a given content, and the influence of this number on the rate–quality–complexity trade-off is not studied.

Chapter 4 — recommendation. I would ask the author to (i) add the MIV Main anchor to the comparison, so that the proposed methods are positioned between no geometry transmitted solution and all geometry transmitted solution on both the rate–quality and the complexity axes; (ii) validate the fixed 0.1% threshold of the portion of inconsistent pixels as a content-dependent rule for choosing between IDMA and eIDMA, and report how often the rule chooses the better variant; and (iii) derive, rather than inherit, the number of essential views, and show how the rate–quality–complexity trade-off depends on it. These three additions would turn Chapter 4 from a very good engineering result into a general method, which is what the thesis statements promise.

These comments are intended as directions for the final version or the author’s further work; they do not change my overall assessment. The reservations I have raised concern the extent to which the results can be explained and generalised, not the correctness of what was measured, nor the practical value of what was achieved.

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.*

The parts of the dissertation that have a tutorial character are Chapter 1 (Sections 1.1–1.3), the introductory parts of each of Chapters 2, 3 and 4, and Chapter 5. Chapter 1 surveys the MPEG Immersive Video standard, the Geometry Absent profile, the evolution from DERS and VSRS to IVDE, and the metrics used in the field; the introductory parts of Chapters 2–4 explain, respectively, the view-labelling and atlas-packing machinery of TMIV, the graph-cut (alpha-expansion) formulation of multi-view depth estimation and its hierarchical variant, and the modifications of the MIV encoder and decoder needed to carry depth information. Together these parts cover: video compression (HEVC/VVC, atlas-based representation, rate–quality methodology), multiview and immersive video representation and rendering, camera geometry and calibration, depth estimation as an energy-minimisation problem, objective quality assessment for synthesised views (IV-PSNR, BD-rate), and the practice of standardisation in ISO/IEC MPEG. Their quality is good: the exposition is concise, technically correct, up to date with respect to the second edition of MIV, and consistently referenced to primary sources (standards and MPEG documents) rather than to secondary literature.

The list of references contains 108 items and is, in my opinion, adequate and up to date (18 % of the items are from 2023–2024). It is properly balanced between journal and conference publications, ISO/IEC standards and MPEG working documents, and reference-software manuals, which is appropriate for work of this kind. Its one limitation is that it is almost entirely confined to classical, non-learning-based depth estimation and to the MPEG ecosystem: learning-based estimators are mentioned only once, in passing. For a dissertation whose aim is to improve a standardised reference estimator this is defensible, but a short discussion of how the proposed selection principles would relate to learned depth estimators would have strengthened the contextualisation and would have made the limitations of the empirical conclusions clearer.

On the basis of the above, and of the publication record discussed in Section “Contribution”, I have no doubt that the candidate possesses general theoretical knowledge and understanding in the discipline of Information and Communication Technology, and in particular in the area of immersive video coding, multiview geometry and depth estimation. The knowledge is not merely book knowledge: it is demonstrated by working implementations inside reference software and by contributions accepted by an international standardisation body.

5. Other remarks¹

The dissertation is a publication-based thesis of 87 pages composed of five papers and one patent. This format is accepted and common, but it shows here **one improvement** that I consider worth introduce in the final version: test on the combination of the three major contributions to demonstrate the

¹ Optional

improvement on coding efficiency of immersive contents with DSDE, as the three chapters discuss separately about the three topics, selection of base views transmitted for DSDE, modification of depth estimator to decoded views, and allocation of essential base views and non-essential base views for transmission of depth map to assist DSDE, in the current version of dissertation.

Editorial remarks: Fig. 6, Fig. 8, Fig. 2.1, Fig. 2.3.1, and the other similar figures: the Virtual view synthesis does not belong to MIV Main decoder nor MIV GA decoder, while the Bitstream parsing belongs to MIV Main decoder or MIV GA decoder. Bitstream formation does belong to MIV Main encoder or MIV GA encoder.

A bibliographic detail: the patent is listed inconsistently inside the dissertation — on p. 7 it appears as granted US 12,530,806 B2 of 20 January 2026, while the list of the author's publications on p. 85 still gives the pre-grant application publication (US 2023/0230285 of 20 July 2023) with the status "accepted". I verified the record at the USPTO: the patent was in fact granted on 20 January 2026 and is currently active, so the statement on p. 7 is correct and the entry on p. 85 should be updated. These remarks are intended as directions for the final version and for the author's further work; they do not change my overall assessment. The reservations I have raised concern the extent to which the results can be explained and generalised, not the correctness of what was measured, nor the practical value of what was achieved.

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)², 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 coding, multiview geometry and depth estimation?

☒	☐	☐	☐	☐
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


Signature

² http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf