Inside the Frame: A Plan for Audio-Visual Feature Analysis of Video Recommendations for Children

Erasmo Purificato[1,†]

[1]European Commission, Joint Research Centre (JRC), Ispra, Italy

Abstract

Algorithmic recommendation systems are increasingly shaping children's digital consumption, but there is limited understanding of how audio-visual features impact the visibility and popularity of videos aimed at young audiences. Regulatory frameworks, such as the EU Digital Services Act, demand greater transparency and accountability, particularly regarding content targeted at minors. However, current systems often overlook the influence of content design on engagement. In this position paper, we propose a research agenda to systematically analyze interpretable visual and audio features, such as color vividness, motion intensity, vocal dynamics, and musicality. By linking these elements to engagement outcomes, we aim to discover consistent patterns that can inform the design of child-sensitive recommendation systems, algorithmic audits, and compliance with policies, as well as establish a foundation for more accountable algorithmic media environments for children.

Keywords

Video recommendation, Audio-visual feature extraction, Algorithmic transparency, Children engagement

1. Motivation and Literature Gaps

Recommendation systems play a crucial role in shaping the digital media consumption of millions of users, particularly children, who are increasingly exposed to *algorithmically curated content* on platforms such as YouTube [1]. While traditional approaches to recommendation rely heavily on *behavioral signals* (e.g., click-through rate, watch time) [2, 3, 4], recent attention has turned toward incorporating *content-level features* to enhance both recommendation quality [5] and fairness [6]. In parallel, concerns from policymakers and researchers have grown regarding the *excessive engagement of children* with digital platforms, with emerging regulations, such as the EU Digital Services Act (DSA) [7] and the related recently-published guidelines on the protection of minors[1], placing new obligations on platforms to *limit manipulative design and addictive recommendation patterns*, especially when targeting minors. Despite this, little is known about the role of **low-level audio-visual features** (e.g., color saturation, motion patterns, pitch dynamics) in *driving engagement and influencing recommendation outcomes*. In practice, such features are often disregarded in the design or audit of recommendation pipelines, despite their potentially significant influence on viewer retention and algorithmic amplification.

To address this gap, we propose a new research perspective that *systematically analyzes* the role of intrinsic audio-visual features in shaping the popularity and virality of children's videos. We argue that understanding how low-level multimedia signals *correlate with engagement* is significant to ensure transparent, fair, and child-appropriate recommendation systems. Our position is that such content features, despite their potential psychological and perceptual impact, are *critically understudied* in both academic research and regulatory scrutiny. Our perspective requires the development of a multimodal analysis pipeline that can extract interpretable audio-visual descriptors from full-length video content. By investigating how these features vary across high-visibility and low-visibility content aimed at

DaQuaMRec 2025: 1st International Workshop on Data Quality-Aware Multimodal Recommendation, co-located with the 19th ACM Conference on Recommender Systems (RecSys 2025), September 22–26, 2025, Prague, Czech Republic

† **Disclaimer**: The view expressed in this paper is purely that of the author and may not, under any circumstances, be regarded as an official position of the European Commission.

erasmo.purificato@ec.europa.eu (E. Purificato)
https://erasmopurif.com (E. Purificato)
0000-0002-5506-3020 (E. Purificato)

[1]https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-protection-minors

children, future research can uncover methodical engagement patterns and evaluate their alignment with *child protection principles* and *legal obligations* under frameworks such as the DSA.

Despite interest in video content analysis, especially for child safety and moderation, gaps remain in understanding how intrinsic audio-visual features influence the visibility and popularity of children's videos on platforms like YouTube. Most existing research focuses on **high-level content classification** or **metadata-driven moderation**. Deep learning (DL) models are commonly used to detect *inappropriate content* [8, 9, 10], while popularity is often predicted using metadata such as views, likes, comments, emotional valence, or linguistic style [8, 11, 12, 13]. However, these approaches overlook how low-level video attributes, such as colour saturation, motion patterns, or pitch dynamics, contribute to engagement or recommendation outcomes. Although general video mining includes broad measures, such as *visual variation* or *content richness* [14], a systematic analysis of these features in the context of children's media is still missing. Studies on children's preferences tend to focus on **static elements** (e.g., book cover characteristics [15, 16] or music genres [17]) for personalising recommender systems, without addressing how dynamic video features affect algorithmic amplification. Evaluation frameworks for children's YouTube content often emphasize educational or design quality [18, 19, 20], but not the role of intrinsic audio-visual traits in affecting appeal. The "*Elsagate*" phenomenon [21, 22, 23] exposed the limitations of metadata filtering, demonstrating the need for deeper analysis of the multimedia attributes that drive engagement, irrespective of content appropriateness. Moreover, many high-performing classification models offer **limited interpretability**. Tools like *class activation maps* highlight relevant image regions but do not explain how specific audio-visual features influence viewer behavior [9]. Research on video engagement in educational settings (e.g., comparing lecture capture to infographic videos) reveals the impact of visual dynamics [24]. However, these findings do not generalize to children's entertainment content. Crowdsourcing is used for content moderation, allowing human judgment to identify and segment inappropriate content [25]; yet, this relies on *human perception of appropriateness*, lacking interpretability into which multimedia elements propel video virality.

2. Planned Methodology

We propose to systematically extract low-level audio-visual features from children's videos to **uncover correlations** between content characteristics and their popularity or recommendation exposure. We argue that formalizing the influence of content-level attributes can provide valuable insights for more fair and transparent recommenders. In practice, in the initial iteration, we plan to analyze the top-25 most-viewed YouTube videos for children compared with the overall top-25 most-viewed videos.

On the **visual side**, we plan to compute *global color statistics* (saturation, brightness, contrast) to capture the overall **vividness** and **visual salience** of each video, factors previously shown to affect children's attention [26]. *Texture descriptors*, such as *local binary patterns*, will be used to distinguish **visually detailed scenes** [27]. *Motion characteristics* will be quantified using the *dense optical flow* metric [28], which captures the proportion of highly dynamic frames, an indicator of **editing pace** and **visual stimulation**. We also intend to include features that reflect **social and narrative structure**, such as *face presence* and *wave-ratio*, to approximate the visibility of on-screen presenters or characters [29]. To identify **differences in animation style**, we compute depth statistics and assess consistency across frames to infer whether the video adopts a *flat 2-D layout* or *3-D CGI production* [30]. To complement these handcrafted features, we plan to extract vector embeddings from pre-trained *deep video models* such as VideoMAE [31] and SlowFast [32], which can summarize spatial and temporal patterns to describe the video's overall **visual style** and **movement patterns**.

On the **audio side**, we plan to extract *rhythmic and harmonic descriptors*, such as spectral contrast, melody, and tempo, to distinguish between **speech, music, and song-based content** [33, 34]. Additional features, including short-term energy, zero-crossing rate, spectral centroid and roll-off, harmonic ratio, pitch, and silence ratio, will serve as proxies for **vocal intensity** and **excitement level**, which have been linked to children's emotional arousal [35].

The extracted features will be used in comparative analyses between content explicitly targeted at

children and general-audience videos to measure whether children's videos exhibit **unique design signatures**. In conclusion, this planned methodology seeks to inform both the auditing of content-driven engagement mechanisms and the development of recommender systems that are more accountable and appropriate for child audiences.

Declaration on Generative AI

The author have not employed any Generative AI tools.

References

[1] J. Radesky, E. Bridgewater, S. Black, A. O'Neil, Y. Sun, A. Schaller, H. M. Weeks, S. W. Campbell, Algorithmic Content Recommendations on a Video-Sharing Platform Used by Children, JAMA Network Open 7 (2024) e2413855. doi:10.1001/jamanetworkopen.2024.13855.

[2] P. Wang, Y. Jiang, C. Xu, X. Xie, Overview of Content-Based Click-Through Rate Prediction Challenge for Video Recommendation, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 2593–2596. doi:10.1145/3343031.3356085.

[3] Y. Zheng, C. Gao, J. Ding, L. Yi, D. Jin, Y. Li, M. Wang, DVR: Micro-Video Recommendation Optimizing Watch-Time-Gain under Duration Bias, in: Proceedings of the 30th ACM International Conference on Multimedia, MM '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 334–345. doi:10.1145/3503161.3548428.

[4] H. Zhao, L. Zhang, J. Xu, G. Cai, Z. Dong, J.-R. Wen, Uncovering User Interest from Biased and Noised Watch Time in Video Recommendation, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 528–539. doi:10.1145/3604915.3608797.

[5] Y. Deldjoo, M. G. Constantin, H. Eghbal-Zadeh, B. Ionescu, M. Schedl, P. Cremonesi, Audio-visual encoding of multimedia content for enhancing movie recommendations, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 455–459. doi:10.1145/3240323.3240407.

[6] W. Chen, L. Chen, Y. Ni, Y. Zhao, Causality-Inspired Fair Representation Learning for Multimodal Recommendation, ACM Transactions on Information Systems (2025). doi:10.1145/3744240.

[7] European Union (EU), Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), 2022. URL: https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng.

[8] K. Yousaf, T. Nawaz, A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos, IEEE Access 10 (2022) 16283–16298. doi:10.1109/ACCESS.2022.3147519.

[9] R. Tahir, F. Ahmed, H. Saeed, S. Ali, F. Zaffar, C. Wilson, Bringing the kid back into YouTube kids: Detecting inappropriate content on video streaming platforms, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '19, Association for Computing Machinery, New York, NY, USA, 2020, pp. 464–469. doi:10.1145/3341161.3342913.

[10] A. El Bakri, A. Yehia, M. Ali, R. Osama, Z. Adel, O. Gamal, S. N. Saleh, The Eye: An AI-Powered Video Streaming Platform to Protect Children from Inappropriate Content, in: 2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES), 2024, pp. 529–532. doi:10.1109/NILES63360.2024.10753227.

[11] S. Wu, M.-A. Rizoiu, L. Xie, Beyond Views: Measuring and Predicting Engagement in Online Videos, Proceedings of the International AAAI Conference on Web and Social Media 12 (2018). doi:10.1609/icwsm.v12i1.15031.

[12] L. Stappen, A. Baird, M. Lienhart, A. Bätz, B. Schuller, An Estimation of Online Video User

Engagement From Features of Time- and Value-Continuous, Dimensional Emotions, Frontiers in Computer Science 4 (2022).
[13] A. C. Munaro, R. H. Barcelos, E. C. F. Maffezzolli, J. P. S. Rodrigues, E. C. Paraiso, To engage or not engage? The features of video content on YouTube affecting digital consumer engagement, Journal of Consumer Behaviour 20 (2021) 1336–1352. doi:10.1002/cb.1939.
[14] X. Li, M. Shi, X. S. Wang, Video mining: Measuring visual information using automatic methods, International Journal of Research in Marketing 36 (2019) 216–231. doi:10.1016/j.ijresmar.2019.02.004.
[15] Y. Beyhan, M. S. Pera, Covering Covers: Characterization Of Visual Elements Regarding Sleeves, in: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, UMAP '23 Adjunct, Association for Computing Machinery, New York, NY, USA, 2023, pp. 28–33. doi:10.1145/3563359.3597404.
[16] A. Milton, L. Batista, G. Allen, S. Gao, Y.-K. D. Ng, M. S. Pera, "Don't Judge a Book by its Cover": Exploring Book Traits Children Favor, in: Proceedings of the 14th ACM Conference on Recommender Systems, RecSys '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 669–674. doi:10.1145/3383313.3418490.
[17] L. Spear, A. Milton, G. Allen, A. Raj, M. Green, M. D. Ekstrand, M. S. Pera, Baby Shark to Barracuda: Analyzing Children's Music Listening Behavior, in: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 639–644. doi:10.1145/3460231.3478856.
[18] M. M. Neumann, C. Herodotou, Evaluating YouTube videos for young children, Education and Information Technologies 25 (2020) 4459–4475. doi:10.1007/s10639-020-10183-7.
[19] D. Poveda, M. Matsumoto, E. Sundin, H. Sandberg, C. Aliagas, J. Gillen, Space and practices: Engagement of children under 3 with tablets and televisions in homes in Spain, Sweden and England, Journal of Early Childhood Literacy 20 (2020) 500–523. doi:10.1177/1468798420923715.
[20] J. Zhang, Y. Huang, M. Gao, Video Features, Engagement, and Patterns of Collective Attention Allocation: An Open Flow Network Perspective, Journal of Learning Analytics 9 (2022) 32–52.
[21] J. Balanzategui, Examining the "Elsagate" Phenomenon: Disturbing Children's YouTube Content and New Frontiers in Children's Culture, AoIR Selected Papers of Internet Research (2019). doi:10.5210/spir.v2019i0.10921.
[22] W. Han, M. Ansingkar, Discovery of Elsagate: Detection of Sparse Inappropriate Content from Kids Videos, in: 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), 2020, pp. 46–47. doi:10.1109/ZINC50678.2020.9161808.
[23] P. Soustas, M. Edwards, The Elsagate Corpus: Characterising Commentary on Alarming Video Content, in: R. Mitkov, S. Ezzini, T. Ranasinghe, I. Ezeani, N. Khallaf, C. Acarturk, M. Bradbury, M. El-Haj, P. Rayson (Eds.), Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, Lancaster, UK, 2024, pp. 147–152.
[24] S. Lackmann, P.-M. Léger, P. Charland, C. Aubé, J. Talbot, The Influence of Video Format on Engagement and Performance in Online Learning, Brain Sciences 11 (2021) 128. doi:10.3390/brainsci11020128.
[25] S. K. Mridha, B. Sarkar, S. Chatterjee, M. Bhattacharyya, ViSSa: Recognizing the appropriateness of videos on social media with on-demand crowdsourcing, Information Processing & Management 57 (2020) 102189. doi:10.1016/j.ipm.2019.102189.
[26] D. R. Anderson, H. L. Kirkorian, Attention and Television, in: Psychology of Entertainment, Routledge, 2006.
[27] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 971–987. doi:10.1109/TPAMI.2002.1017623.
[28] G. Farnebäck, Two-Frame Motion Estimation Based on Polynomial Expansion, in: J. Bigun, T. Gustavsson (Eds.), Image Analysis, Springer, Berlin, Heidelberg, 2003, pp. 363–370. doi:10.

1007/3-540-45103-X_50.
[29] R. F. Kizilcec, J. N. Bailenson, C. J. Gomez, The instructor's face in video instruction: Evidence from two large-scale field studies, Journal of Educational Psychology 107 (2015) 724–739. doi:10.1037/edu0000013.
[30] R. Ranftl, A. Bochkovskiy, V. Koltun, Vision Transformers for Dense Prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12179–12188.
[31] Z. Tong, Y. Song, J. Wang, L. Wang, VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, Advances in Neural Information Processing Systems 35 (2022) 10078–10093.
[32] C. Feichtenhofer, H. Fan, J. Malik, K. He, SlowFast Networks for Video Recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6202–6211.
[33] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Transactions on Speech and Audio Processing 10 (2002) 293–302. doi:10.1109/TSA.2002.800560.
[34] J. Salamon, E. Gomez, Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics, IEEE Transactions on Audio, Speech, and Language Processing 20 (2012) 1759–1770. doi:10.1109/TASL.2012.2188515.
[35] H. E. Kragness, M. J. Eitel, A. M. Baksh, L. J. Trainor, Evidence for early arousal-based differentiation of emotions in children's musical production, Developmental Science 24 (2021) e12982. doi:10.1111/desc.12982.