The Institution of Engineering and Technology

IBC2024

The Best of IET and IBC 2024

The Institution of Engineering and Technology

OPEN ACCESS

Be seen.
Be cited.

Engineering and technology research helps solve society's challenges.

This is why IET journals are openly accessible to all. Submit your articles to further your career and shape a better future.

Talk to us and learn more about:

- our range of open access journals across engineering and technology
- the IET Journal Transfer Network to increase your chance of getting published
- your eligibility for funding.

ietresearchhub.org

Contents

Advance your learning in

Media
Imaging
Vision
Aerospace

Come and learn from experts in industry. We have over 20 technical topics covered, ranging from Aerospace to Vision and Imaging, our technical webinars are sure to have a topic to match your discipline.

theiet.org/technical

#ForThoseWhoDoMore #IETCommunities

The Institution of Engineering and Technology

Sign up to our monthly digest of Technical Network online events so you don't miss a thing.

Introduction

Welcome to The Best of IBC2024. This is an annual joint publication between the IET and IBC. It features the ten papers which have been judged by IBC's Technical Papers Committee to be the best of those presented at this year's conference in Amsterdam. The twenty Committee members were not only looking for contributions which are highly novel, but which are also: topical, analytical, entertaining, educational, well-written and which have the potential to make a significant impact upon the media industry. Often the best papers will also have demonstrated their success through simulation, prototype development or full practical trials.

It is worth taking a moment to set these top papers in the context of the competitive process by which IBC selects its conference papers: In January of every year, media technology researchers from all over the world submit brief synopses of their prospective papers, hoping that theirs will be chosen for inclusion in the event. This year IBC received 327 such offers spanning the entire technical spectrum from the streaming of cricket matches across the Indian sub-continent, to the finer mathematical algorithms of AI-based metadata generation.

In the initial peer review, every synopsis was assessed by at least five specialists and the result was a ranked list of about 40 synopses which showed the greatest potential. Authors of these synopses were then asked to write and submit their full papers. Finally, with involvement of the whole Committee, 27 full papers were chosen for publication. These contributed to the nine themed technical conference sessions.

Of course, the Committee members have a privileged view of all the synopses and this gives them a unique perspective on the technological direction of the entire industry and especially those new ideas and concepts which are in the ascendancy. It is this perspective which informs the planning of the conference sessions, to ensure that all who attend are able to gain a genuine insight into some remarkable new developments, whether they are emerging from multi-national media companies or small university research groups.

It is interesting to remember that in the Best of IBC 2019 we wrote that about 50% of all the synopses submitted that year involved AI in some way. Now five years later, we no longer bother to record this statistic. AI has become almost ubiquitous in media technology research and is now an established and powerful enabling tool which not only provides performance enhancements to many systems, but has fundamentally changed the way in which we think about problem-solving in engineering. This is a revolution, the only parallel to which was the move from analogue to digital.

We are delighted that in IBC's best ten papers presented here, the diversity of ascendant technologies is well represented including the latest ideas in: volumetric video, sustainable low-energy technologies, live audio and video streaming, video compression, adaptive sensors for panoramic images, AI-based text processing and media provenance.

As we begin to summarise these, we must first refer to IBC's Best Technical Paper Award 2024 winner, 'Advancements in Radiance Field Techniques for Volumetric Video Generation: A Technical Overview,' by Joshua Maraval, Nicolas Ramin and Lu Zhang. This paper seeks to find a solution to the complexity of capturing and rendering volumetric video, a complexity which, it argues, is holding back the mainstream creation of VR content.

The paper begins with a very clear review of the various approaches taken to generate 3D video of real-world scenes, as they would appear from any given point in space, by employing only a collection of conventional 2D video camera views (a process known as Novel View Synthesis). The team then describes its enthusiasm for a 2020 ground-breaking paper in a new field called Neural Radiance Fields (NeRFs). Not only do they coherently explain the operation of these, but they build upon these ideas and perform a comparison with several of the best methods from current research literature. It was from this comparison that they demonstrated the significant advantages of a new NeRF-based method called Space-Time Gaussian Feature Splatting which provides excellent quality rendering and proves rugged in the presence of difficult scene geometries. The prospects for their further work in this area are significant. For those wishing to follow this work, the paper also gives an extensive list of 70 references to key published papers.

As in previous Best of IBCs, we have interviewed the winning authors, so you can discover what motivates them and what they like to do when they are not in the lab. While on the subject of NeRFs, do see the videoconferencing paper from Shanghai Jiao Tong University. Their approach has been to use pre-trained NeRFs for the real-time representation and rendering of 3D human heads (including facial expressions) at a mere 5kbit/s.

Another rapidly advancing area of research and standardisation is media provenance. Developing mechanisms through which we are able to trust distributed media, especially news, is one of the most fundamental issues in our industry and in democratic society, more widely. We present two papers on this topic, one by the BBC which looks at standards for content credentials and which reports on a trial which assesses subsequent gain in audience trust. The second by Verance Corp in the USA, provides a more technical approach where watermarking and cryptography are discussed in some depth. Both papers relate to standards emerging under the Coalition for Content Provenance and Authenticity (C2PA).

Energy efficiency is a vital consideration as organisations strive towards net-zero targets. One such study by Arqiva and the BBC explores the reduction of the energy required to transmit terrestrial DAB. This reveals potential savings of between 12% and 17% on many DAB transmitters at the expense of minor signal degradation and an optimisation of their power supplies.

How does a streaming provider manage system resources and data strategies when concurrently delivering live services up to 59 million customers and where traffic can surge by over 1 million viewers per minute? Find out in the paper by AWS, India.

Panoramic and 360° image sensors very often have to contend with light-levels which vary significantly across the width of the device. A team from NHK in Japan describe how they have developed a novel sensor which can optimally adapt resolution, frame-rate and exposure time across the device.

How well do mobile phone networks cope with live streaming? This is the question posed by the BBC's Project Timbre and their paper answers it with the help of their own staff who employ a novel experimental approach to determine a real-world Quality of Experience metric.

Video compression technology has been advancing for decades and the work of IMAX-SCT in Canada, shows that it will continue to do so. This impressive work introduces a Large Multi-modal Model to optimally choose video encoding configurations across an entire TV programme by understanding the semantics of the content. While results are significant, rate-distortion theory reveals that there is still potential for further compression gains. In order for news content to be profitably re-used it must be tagged with metadata word-labels to allow the material to be recognised in diverse contexts. This is a perfect task for an AI application, however NHK has discovered that where only a limited set of labels is desired, probabilistic assignment introduces a bias favouring poorly matching words. In their paper they analyse this problem and propose a novel solution.

We are extremely proud that so many media professionals continue to choose IBC for the publication of their technical work and as a forum for discussion with their fellow technologists and market strategists. This journal is a tribute to all those individuals who submitted synopses this year, whether successful or not. If you are inspired by the papers and stories presented here and would like to tell us about your own research or innovation, then please look out for our call for papers in January. And if your synopsis was not successful this year, then please try again - we work hard to accommodate as many papers as we possibly can.

We hope that you enjoy reading this collection of the best papers as much as we and IBC's committee of specialist peer reviewers did. We would like to convey our thanks to everyone involved in the creation of this year's volume, both at the IET and at IBC.

Dr Nicolas Lodge
Executive Producer, Technology

Dr Paul Entwistle
Chair, IBC Technical Papers Committee

Who we are

IBC2024

IBC is where the future of the global media and entertainment industry is redefined. Energising the market, enabling content everywhere and inspiring new conversations, IBC brings the creative, technology and business communities together to collaborate, learn and unlock new opportunities. Exhibitors and speakers from around the globe come to IBC to showcase game-changing innovations and tackle the media sector's most pressing trends and issues – changing perceptions and meeting the needs of the world-leading broadcasters, content owners, rightsholders, service providers and others attending the highly-respected peer-reviewed Conference and comprehensive four-day trade show. With a focus on inclusivity, IBC propels change – driving thought leadership, sparking discussion, shifting expectations, accelerating creativity, and enabling real business outcomes. IBC's mission is to empower our 250,000-strong global community to explore new opportunities, build knowledge, and play an active role in the technological transformation and broader change sweeping the industry worldwide.

IET The Institution of Engineering and Technology

The IET is one of the world's leading professional societies for the engineering and technology community, with more than 156,000 members in 148 countries, and offices in Europe, North America and the Asia-Pacific region. It is also a publisher whose portfolio includes a suite of 48 internationally renowned peer-reviewed journals covering the entire spectrum of electronic and electrical engineering and technology. Many of the innovative products that find their way into the exhibition halls of IBC will have originated from research published in IET titles, with more than a third of the IET's journals covering topics relevant to the IBC community (e.g. IET: *Image Processing*; *Computer Vision*; *Communications*; *Information Security*; *Microwave Antennas & Propagation*; *Optoelectronics*; *Circuits & Systems*; *Signal Processing* and *Electronics Letters*).

Winners of IBC's Best Paper Award 2024

Congratulations to the winners of the IBC Best Paper Award 2024 - Joshua Maraval, Nicolas Ramin and Lu Zhang for 'Advancements in Radiance Field Techniques for Volumetric Video Generation: A Technical Overview'.

Interview with Joshua Maraval

Tell us a little about yourselves and how you each contributed to your fascinating paper.

Joshua: I am a third-year PhD student working under the supervision of Lu and Nicolas. The foundation of this state-of-the-art paper is based on my vision of volumetric video, which has been shaped over my years of research in radiance fields. This vision has been continuously refined through the valuable guidance and mentorship I've received from both Lu and Nicolas.

How long have you researched in the field of novel viewpoint synthesis and do you have backgrounds relevant to this?

Joshua: I have been working on novel view synthesis for the past three years as part of my PhD. My research began just after the rise of radiance field methods, which makes my work especially relevant to this rapidly evolving field. Before this, I worked on view synthesis during my master's degree, where I used classical descriptor methods. However, with the paradigm shift towards neural rendering, my background in deep learning-based computer vision became even more pertinent.

How would you describe the significance of the 2020 landmark paper on NeRFs by Mildenhall et al, to the field of volumetric rendering?

Joshua: The 2020 NeRF paper by Mildenhall et al. is a foundational work that has had a profound influence on the field of novel view synthesis. It brought unprecedented attention to the area and set a new standard for volumetric rendering. The paper expertly combines established theories with modern computational capabilities, laying the groundwork for later innovations. It is already regarded as a seminal piece in the field, and rightfully so.

Your valuable comparison of radiance field methods must have taken considerable effort, what part of this proved to be the most challenging?

Joshua: Comparing radiance field methods was facilitated by the open-source culture in the research community, with many authors sharing their code. However, the biggest challenge was the lack of standardization across the field. We had to adapt datasets to align with the conventions of various methods and even proposed our own multiview video dataset for more comprehensive evaluations.

Joshua Maraval

Nicolas Ramin

Lu Zhang

We took special care to ensure a diverse range of scene content to allow for fair comparisons, as different methods tend to excel in different types of content due to their unique limitations.

How fundamental is the problem of image flicker due to ambiguous geometry and temporal consistency? Will it ever be completely solved?

Joshua: Despite the impressive progress in recent years, volumetric video is still in its infancy when it comes to industrial applications. Achieving temporal consistency is a critical milestone in delivering a fully immersive experience. Today, we've grown accustomed to flawless, high-quality 2D video and naturally expect the same from volumetric video. For instance, novel view synthesis was prominently featured during the recent Olympic Games but was limited to static shots due to these challenges. Just as 2D video eventually overcame its early limitations, I'm confident that temporal flickering in volumetric video will be fully resolved as the field matures.

When do you foresee that the advances to which you have contributed here, will allow us to efficiently capture time-varying scenes using NeRFs?

Joshua: Radiance fields are still relatively new, and their industrial adoption is limited by the lack of experience in developing complete end-to-end products. However, with growing interest from industry, I believe we'll see rapid progress in the near future. I'm optimistic that by the next Olympic Games, we'll witness some truly remarkable volumetric video experiences.

Will your future work focus on STG methods? What are the next steps in your research?

Joshua: Since NeRF's release, the field has been evolving rapidly, and to remain at the forefront, we've had to explore multiple approaches. Currently, 3D Gaussian Splatting (3DGS) is one of the most promising developments and has stayed on the mainstage for longer than any other recent method. As the community stabilizes around these advancements, we are starting to tackle more complex tasks such as dynamic model reconstruction. STG, as well as other dynamic implementations of 3DGS, are incredibly promising. We are focusing on STG and are developing a headset demonstrator for immersive training using volumetric video. b<>com will be sharing more details soon, so stay tuned!

Finally, what part do visual media play in your life outside of work? And do you use immersive devices now and do you look forward to a world where they are commonplace?

Joshua: The consumption of visual media continues to rise every year, and interactive devices like mobile phones are leading the charge. Immersive devices are also becoming more accessible, and while they provide the most engaging experiences, they still cater to a niche audience. I believe that volumetric video's widespread impact will be felt on 2D displays, and we're already seeing glimpses of this in sports broadcasts. I think the real potential lies in mobile phones, which now have highly accurate motion tracking capabilities and could unlock exciting possibilities for volumetric video. However, as a tech enthusiast, I personally look forward to enjoying volumetric video at home on virtual headsets or holographic displays as the technology continues to advance.

“Despite the impressive progress in recent years, volumetric video is still in its infancy when it comes to industrial applications. Achieving temporal consistency is a critical milestone in delivering a fully immersive experience.”

Advancements in Radiance Field Techniques for Volumetric Video Generation: A Technical Overview

Joshua Maraval[1,2], Nicolas Ramin[1] and Lu Zhang[1,2]
[1]IRT b<>com, France and [2]CNRS, IETR-UMR 6164, France

Abstract
Over the past decades, video consumption and video devices have become widespread globally. In 2014, mainstream virtual reality headsets marked a pivotal moment for 360° video accessibility. Advanced immersive devices, like the Apple Vision Pro as well as smartphones and tablets with advanced spatial capabilities can now provide users with real-time 6 Degrees of Freedom (6DoF) navigation experiences.

However, the lack of engaging content is hindering potential applications in areas such as training and entertainment. Volumetric video is a promising solution. However, its production poses challenges, such as the need for natural 3D+t reconstruction, coding, and rendering, which still require intensive computational resources.

In 2020, the ground-breaking Neural Radiance Field (NeRF) paper introduced a new way to generate natural free-viewpoint renderings of real scenes from sparsely captured views. Follow-up research has led to faster and more flexible methods, such as the widely used 3D Gaussian Splatting. However, these approaches require independent models for each frame, posing a challenge for volumetric video representation. To address temporal limitations, extensions of radiance field techniques use temporal redundancy to create a compact, temporally consistent, and editable volumetric video representation.

This paper offers a comprehensive overview of state-of-the-art volumetric video methods based on neural radiance fields, including their respective advantages and drawbacks. Using a diverse multi-view video dataset of diverse real-world scenarios, we present an objective evaluation of these methods for video volumetric content generation in entertainment and training.

Introduction

Novel view synthesis (NVS) is a long-standing challenge of 3D computer vision: the rendering of unseen views of a scene from a set of captured views. NVS has a growing impact on a wide array of video applications including media consumption [1], sports retransmission [2], immersive training [3] and telepresence [4]. The applications fall into one of two categories: visual effects or immersive experiences. One common visual effect with NVS is the virtual rendering of non-captured camera movement. An illustrative is the Intel True View technology [5], which proposes frozen time 360 degree replays of sports stadiums. In contrast, immersive experiences rely on real-time NVS to display position-dependant views to a user, allowing them to navigate freely within a virtual scene as if they were in the real location.

Early NVS methods interpolated viewpoints from depth information [6]. These methods were capable of rendering realistic novel views in ideal conditions, but they had limited light effect rendering capacity and were restricted to rendering views that were close to the reference views. Concurrently, novel devices, including the Apple Vision Pro, were making real-time 6 Degrees of Freedom (6DoF) tracking increasingly accessible. Free navigation of real scenes requires reconstructing a complete representation of the scene from recorded videos. In recent years, significant progress has been made towards volumetric-based approaches. One such approach is Neural Radiance Fields (NeRF) [7] which was first published in 2020.

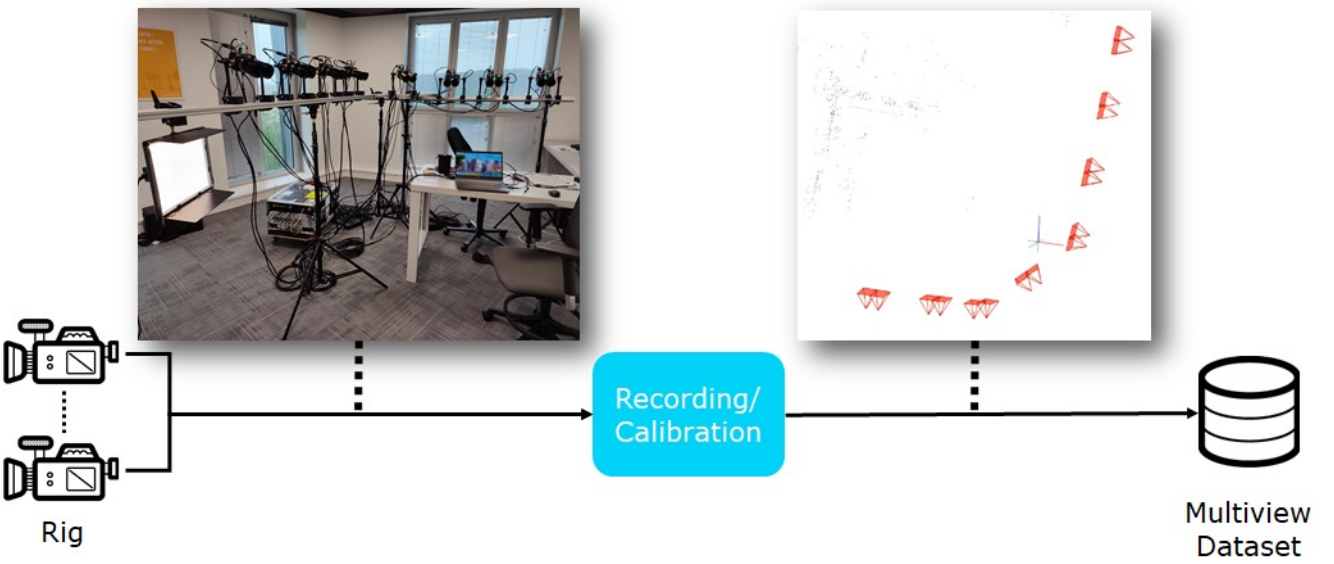


Figure 1: The input images for NVS undergo a pre-processing calibration step. The camera parameters are retrieved from the input images with the Structure-from-motion software COLMAP. A Multiview Dataset is constituted of the input images and meta-parameters of the scene, including camera parameters.

This method enables the generation of high-quality views by modelling the scene's geometry and radiance. NeRF demonstrated the capacity of radiance field methods to represent complex real scenes with accurate light effects. Following the publication of NeRF, radiance fields have rapidly become the most regarded approach for NVS of natural content.

Early radiance field methods, including NeRF necessitated slow reconstructions for every scene to be reconstructed, and could not render novel views in real time. New approaches have been implemented from NeRF for faster processing [8], [9], higher quality rendering [10], [11] and more stable reconstructions [12]. The recent radiance field method 3D Gaussian Splatting techniques [13] has gained considerable popularity due to its demonstration of state-of-the-art rendering quality with in-real-time rendering capacity.

Building up upon the latest advances in radiance field methods, the volumetric video field has undergone a rapid evolution in recent years. New techniques have extended the applicability of radiance field to a range of classical 2D video tasks, including semantic segmentation [14], streaming [15] and edition [16]. The development of real-time dynamic representations of radiance fields has opened the door to 6DoF+t navigable content in real-time.

However, each approach has its own advantages and drawbacks. We propose an overview of NVS methods, focusing on radiance field approaches. We review the latest advances towards to real-time 6DoF+t navigation and evaluate the state-of-the-art methods on a dataset of scenes showcasing complex human interactions in diverse environments.

Novel View Synthesis Methods

From multiple images capturing a single scene, NVS algorithms render novel viewpoints that have not been previously observed. In order to interpret the image's information, it is necessary to understand the position and orientation of the camera relative to the other images. There exist methods that generate novel views only from input images without requiring prior knowledge of camera parameters, simultaneously based on SLAM [17], [18] or state-of-the-art radiance fields [19], [20], [21]. These camera-parameters free methods are considered out of the scope of this review. This review focuses solely on methods that use known camera parameters of input images. In order to obtain camera parameters from multi-view images of a scene, a pre-processing step of structure-from-motion is typically used. In this paper, this calibration step is achieved using the structure-from-motion software COLMAP [18], as illustrated in Figure 1.

Interpolation-based View Synthesis

Interpolation-based NVS methods generate novel views by interpolating pixel information between input views (11). Some approaches, called Depth Image Based Rendering (DIBR) leverage information from the associated depth maps of the input images for more accurate translation of the input pixels to the novel view [6], [22]. DIBR can achieve high-quality rendering of intermediate views, but often results in errors at object edges and occluded areas and lacks light effect rendering. Light field approaches [23], [24] interpolate

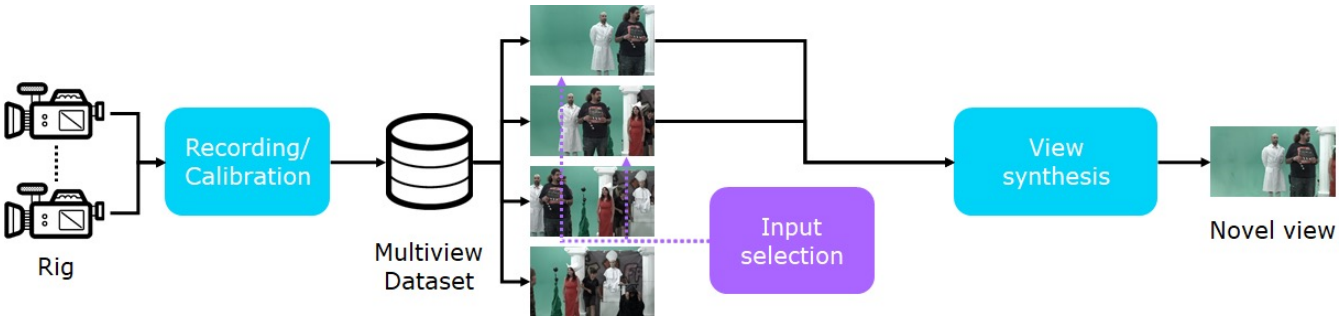


Figure 2: nterpolation-based view synthesis. The pixels from the input views are interpolated to render the novel view.

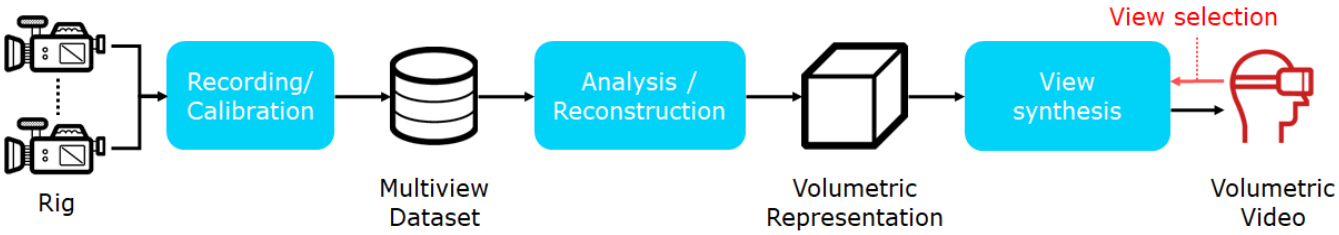


Figure 3: 3D Model learning-based View synthesis from volumetric representation. A volumetric representation of the scene is reconstructed from the input images. After complete training, novel views are rendered by inference of the volumetric representation.

all pixels in an intermediate 3D space, which is then inferred to render novel views. Light fields can render views with complex occlusions and light effects but this necessitates a dense array of input views.

Following the democratization of Convolutional Neural Networks (CNNs) based deep learning by Krizhevsky et al. in 2012 [25], CNNs gained popularity for view synthesis methods. In recent works, CNNs have been trained as a post-filter to improve the rendered images of DIBR-based methods, effectively removing some artefacts [26]. Other methods train a CNN-based architecture in place of the interpolation function for DIBR [27]. Generative Adversarial Networks (GANs) have demonstrated that a few images of a scene can be used to predict other views [28], [29] or enhance NVS renders [11]. While methods based on deep CNNs for NVS may generate high-quality renders, they are inherently slow to infer which results in low rendering frame rates.

Learning-based volumetric representations

The rendering of views from a volumetric representation is a well-studied subject of computer vision. For instance, photorealistic models can be rendered from synthetic scenes modelled as meshes using the latest rendering technologies. However, models used to design synthetic data are limited representations of the real world and differ significantly from the light physics behind the human vision. Some works have proposed the use of CNN architectures for higher quality mesh rendering [30], [31], [32]. However, these representations are partially differentiable, which makes them difficult to optimize without a high density of input images. Seminal works proposed the use of a plenoptic function as a volumetric representation of scene, capable of rendering light coherent novel views [23], [24]. The plenoptic function is a 5D function describing the light flow at any 3D position of any 2D orientation. With recent advances in machine learning, a learning-based approach has become a viable option.

Most modern approaches to learning-based volumetric representations feature a complete or partial representation of the plenoptic function. A generic workflow is presented in Figure 3. Prior to rendering novel views, the volumetric representation must be reconstructed from the input images. Input views are rendered from the representation, and then an error loss is computed based on the difference between the rendered image and the reference images. The loss is then propagated backwards to adjust the volumetric model parameters into a new model that more closely renders the reference images, until complete convergence is achieved.

At the condition of having a fully differentiable rendering pipeline, a variety of volumetric representation can be trained to render novel views of a scene. Multi-plane images (MPI) approaches divide the plenoptic function into successive planes that store colour and transparency information [33], [34], [35]. While these approaches are fast and capable of photorealistic renderings, MPI are limited to rendering views facing a single direction, as a perpendicular view to the planes cannot be rendered. Broxton et al. [36] demonstrated that a structure of spherical planes can be used for efficient 360° scene rendering, with the limitation that rendered views must be close to the circle's centre.

Neural Radiance Fields

NeRF

The Neural Radiance Fields (NeRF) model, published in 2020 by Mildenhall et al. [7], marked a significant shift in the field of volumetric rendering. The paper attracted considerable and growing attention, as evidenced by the citations graph in Figure 4. NeRF introduces a fully connected deep network that outputs volume density and view-dependent radiance at any point in space. A ray-casting strategy is proposed to retrieve the colours of any view. The pixel ray is projected into the three-dimensional space, sampled into three-dimensional points, and the density and radiance of each point are inferred by a multilayer perceptron (MLP). The pixel colour is obtained with a classical ray-casting rendering. The MLP is trained from scratch for any new scene reconstruction. NeRF is a powerful representation capable of generating photorealistic renders of complex scenes and a flexible and simple volumetric representation.

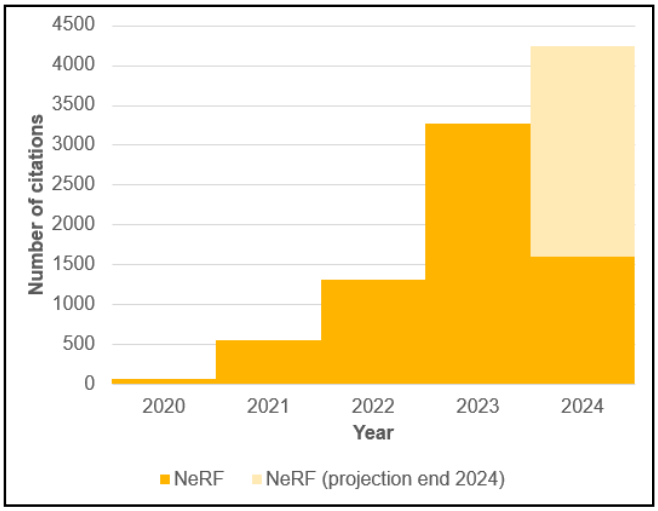


Figure 4: NeRF paper citations over the years.

Following the introduction of NeRF, numerous radiance field architectures have been proposed. IBRNet [37] blends classical interpolation-based view synthesis with a non-scene-specific radiance field MLP. Optimized radiance field architectures been demonstrated to have real-time rendering capacities [38], [39], [40].
More efficient sampling strategies have been studied for faster rendering [41], anti-aliased and generalizable NeRF for boundless scenes [10], [42]. Many extensions of radiance fields have been proposed to extend the applications to other research fields. Large-scale NeRF extend the capacities of radiance fields to city-scale models [43], [44], [45]. Other contributions on scene understanding integrate NeRF into a scene graph [46], [47], for an editable volumetric representation. A large amount of NeRF studies are focused on more specific tasks such as avatar or face reconstruction [48], [49], [50].

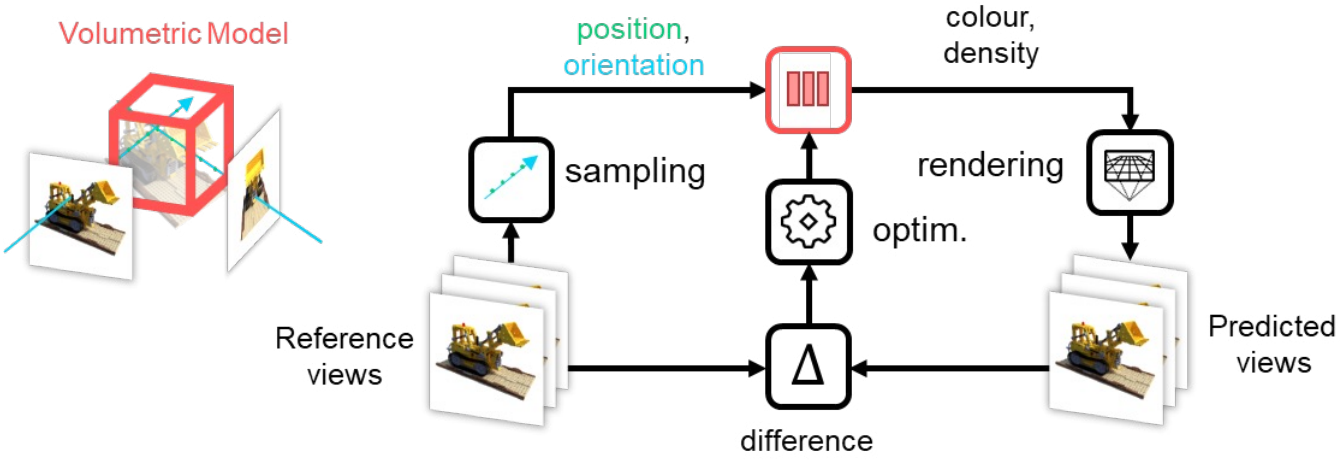


Figure 5: Radiance Fields reconstruction pipeline.

	Regularization type				
	Homogeneity		Sparsity		Appearance
Method	sample	ray	sample	ray	image
Neural Volumes	TV		Beta		
NeRF					
NSVF			Beta		
Baking-NeRF				Cauchy	
PlenOctree				Cauchy	
MIP-NeRF 360	Dist				
DirectVOXGO	RGB			Entropy	
Plenoxels	TV		Beta	Cauchy	
RegNeRF		Depth			Colour
InfoNeRF		Gain		Entropy	
Dense Depth Priors		Depth			

Table 1: Overview of regularisations for radiance fields

Reconstruction of radiance fields from a limited set of views is possible, but failure may occur if the captured views are too sparse. The reconstruction process involves recovering a 5D function from 2D images, which is an under-resolved problem. Significant advances have been made in improving radiance field stability with regularizations, which are rules to constrain the radiance field function to converge towards a coherent model. Three main regularizations have greatly improved radiance field stability. The homogeneity regularization, proposed as Total Variation (TA) by Lombardi et al. [51], encourages the model to have homogeneous zones. This means that the model must feature compact objects with diffuse colour. The regularization can be applied on 3D points [8], [52], [53] or encouraged on adjacent pixels of the rendered views [12], [54], [55]. Sparsity regularization encourages the emptiness of the model, thereby reducing the occurrence of unstructured artefacts. Beta-loss [8], [51], [51], Cauchy-loss [8], [39], [56] and entropy loss [53], [54] have been demonstrated to be efficient losses for sparsity regularization. Finally, appearance regularization encourages renders to appear correct, with the use of a trained CNN [12] or GAN [11].

Spatially Encoded Radiance Fields

NeRF represents a significant advance in the field of computer vision, but it comes at a cost. The MLP, a key component of the NeRF model, is a relatively slow network to infer, requiring multiple inferences for a single pixel. One of the major advances in radiance field research has been the simplification of the implicit radiance function. To reduce the complexity load on the MLP, it can be spatially decomposed into smaller functions, as demonstrated in [57]. Yu et al. demonstrated that the radiance function can be reduced to a simple MLP-free representation, encoding density directly and orientation-dependent colour in a simple parametrization [8], [56].

Other research has investigated the use of a voxel grid to store feature vectors in the three-dimensional space. The feature vectors are trained alongside the MLP [53] and inputted to the MLP depending on the sampling location. The volumetric model information is divided into smaller batches that are more focused on local features. Consequently, equal or higher rendering quality to NeRF can be achieved with a smaller MLP architecture, resulting in faster rendering. As demonstrated in [39] and [58], feature vectors can be stored in sparse voxel grids. Müller et al. developed

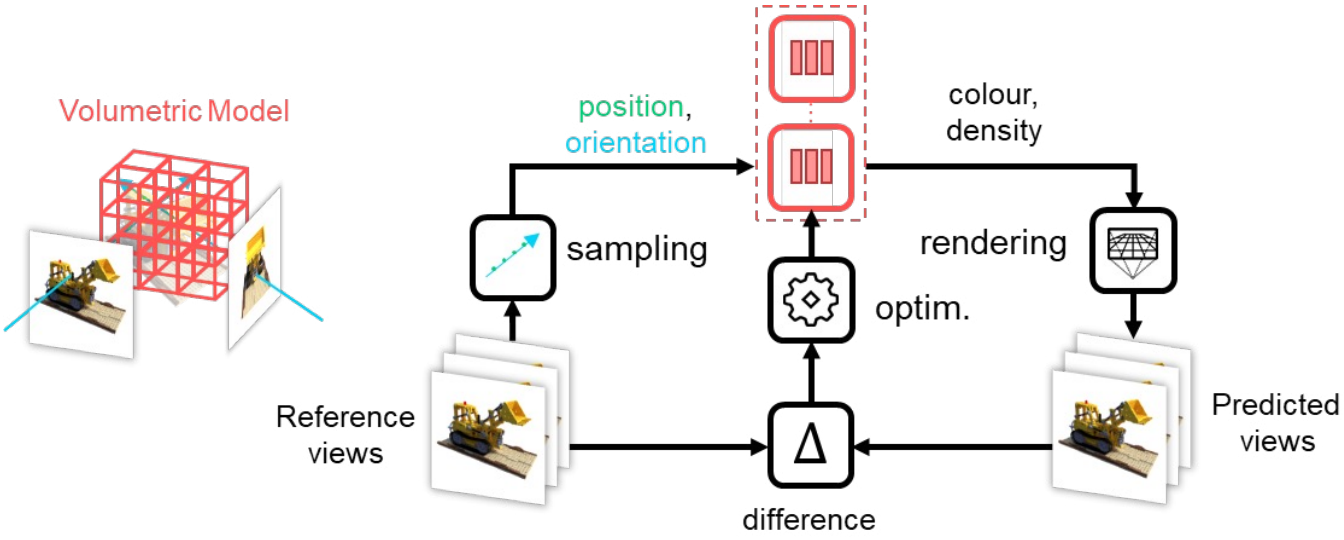


Figure 6: Example representation of spatially encoded radiance field.

Figure 7: Illustration of the temporal flickering in Gaussian Splatting renderings of adjacent frames on the Carpark scene [60]

Instant-NeRF [59] a multi-resolution feature voxel grid-based radiance field, which enables the rendering of higher-quality radiance with faster rendering speeds.

3D Gaussian Splatting (3DGS) is a method published in 2023 [13] that has rapidly gained recognition in volumetric reconstruction research. It is commonly associated with radiance field methods, and is a differentiable point cloud-based approach for learning-based volumetric rendering. The 3DGS model is composed of 3D Gaussians with geometry, orientation–dependant colour and density trainable parameters. Novel views are rendered through rasterization of the Gaussians onto the new view image plane, which is a faster process than ray casting, while maintaining most properties of radiance field rendering. In contrast, previous learning-based point cloud rendering approaches [32], 3DGS generates and prunes points during reconstruction, and is not dependent on a dense point cloud initialization.

Radiance field methods are flexible yet powerful trainable representations of 3D scenes. Intuitively, training radiance field models using successive video frames as input results in a 4D representation of a dynamic scene. While this is true, the lack of temporal constraints associated with the underlying under-resolution of the radiance field reconstruction results in temporal artefacts during rendering of dynamic scenes. Figure 7 illustrates the flickering that can occur with sparse input views. The phenomenon of flickering is particularly evident in reconstructed zones with a higher degree of shape ambiguity, where numerous potential reconstructions could satisfy the reference images rendering. Consequently, windows reflections and occluded areas exhibit more pronounced flickering artefacts.

Method	*PSNR ↑*	*SSIM ↑*	*LPIPS ↓*	*Recon time ↓*	*Memory Usage ↓*
Plenoxels	22.963	0.810	0.368	45 min	2,85 Go
Nerfacto	24,662	0,816	0,261	**8 min**	167 Mo
3DGS	27.803	0.866	0.220	**8 min**	140 Mo
STG	**29.523**	**0.916**	**0.177**	10 min	**5 Mo**

Table 2: Methods comparison results. Average reconstruction and rendering metrics over 14 Multiview sequences.

Dynamic Radiance Fields

The integration of the temporal dimension in a radiance field-based volumetric representation offers two benefits. Firstly, it increases temporal homogeneity, reinforcing spatial information with temporal redundancy. Secondly, it addresses temporary occluded areas. Deformation-based approaches achieve this by dividing the dynamic radiance field into a spatial radiance field and a dynamic deformation field [49], [61], [62]. All temporal instants respect morphologically consistent changes. Similar proposals have been published for dynamic 3DGS [63], [64]. Other dynamic radiance field papers demonstrate excellent performance by inputting the temporal dimension to a first MLP [15] or the feature grid [65].

In contrast to NeRF-based methods, 3DGS features a fully explicit volumetric model that can be more directly extended to the temporal dimension. Yang et al. [66] extend the 3D Gaussians with a temporal dimension and constrain them to coherent movement. In SpaceTime Gaussian Feature Splatting (STG) [67], polynomials are trained to represent the 3D Gaussians movement, forcing smooth movement.

Methods Comparison

The comparison of Radiance Field methods is a challenging task, given it is an evolving field. Evaluation of methods is often conducted on older datasets that may not fully reflect the capabilities of state-of-the-art methods. Many methods have metadata prerequisites, such as the scene bounding box or depth maps. Moreover, a significant number of radiance field implementations are designed for specific content, such as full frontal views, inside captures and concentric views. We evaluate the methods on scenes of the MUVOD Dataset [68], a compilation of multiview video sequences from various sources. The sequences feature varied content, environments, and capture rigs, and include scenes focusing on human interaction.

The evaluation dataset comprises 14 sequences made of between 8 and 30 views. The view calibration is conducted using COLMAP [18]. One middle view is excluded from the training and retained for evaluation purposes for each sequence. Following training, the evaluation view is rendered and compared with the reference. The evaluation metrics employed are PSNR, SSIM and LPIPS. The higher the PSNR and SSIM the better, the lower the LPIPS the better. The reconstruction time is the training time of the volumetric representation. The rendering time is the average rendering time for a frame. The memory usage is the average size of the volumetric model for a sequence.

The following methods were evaluated: Plenoxels [8], Nerfacto [9], 3DGS [13] and STG [67]. Plenoxels and Nerfacto were selected to represent spatially encoded radiance fields. Nerfacto is a community-driven implementation of Instant-NeRF [59]. Plenoxels is trained on a high-resolution grid of 10243 voxels. Both Nerfacto and Plenoxels utilize the COLMAP calibration sparse point cloud, as described in [69], to initialize their bounding boxes. 3DGS is the trained for 7000 iterations. The STG method is the only dynamic method of the tested methods and is a dynamic extension of 3DGS. It is evaluated on five frames, and the results are averaged to be equivalent to a single frame, as with the other methods. Other state-of-the-art methods such as Mip-NeRF [10] or GANeRF [11] are not evaluated despite their high rendering quality due to their lengthy training times. The methods are trained with default parameters on a Nvidia A100 GPU, and the results are shown in Table 2.

Compared to all three other methods, Plenoxels has lower performance for each metric. Plenoxels' regular voxel grid stores feature vectors at the same resolution throughout the scene. This results in suboptimal information resolution for reconstructing foreground content, which has a higher resolution in reference images.

In Figure 8, the foreground content is visibly blurred for the Plenoxels renders. This data structure also results in long training times and higher memory consumption compared to other methods.

Nerfacto performs significantly worse than 3DGS and STG on all rendering quality metrics, but has the lowest reconstruction time tied with Nerfacto and a comparable memory usage to 3DGS. Renders of two scenes are shown in Figure 8. The rendering quality of Nerfacto is very high for simple content like the car, but the reconstruction is particularly unstable for more complex scenes. For example, the people in the background in the lower image are poorly reconstructed due to occlusions from the people in the foreground.

3DGS and STG have the best rendering quality compared to the other two methods. The 3D Gaussians are a powerful and flexible representation. Even for difficult scenes, the training converges to a coherent geometry due to the point cloud initialization with COLMAP. 3DGS and related work are a promising solution for reliable high quality volumetric video with real-time rendering.

STG renders have better quality for every metric compared to all three other methods. This demonstrates the benefit of temporal information that can resolve occlusion ambiguities. If part of the scene is visible in the other training frames, this information constrains the occluded area to coherent content. The average reconstruction time of STG for a single frame is comparable to 3DGS and Nerfacto, and could be reduced by training with more frames. STG has significantly lower memory consumption than the other methods evaluated. A single model is used for all five frames, resulting in an optimized, memory efficient model that could be further optimized by training on a larger number of frames.

The image metrics used evaluate images independently, and the video rendering of STG is not compared to the reference video. However, there is a strong gain in temporal stability of the renders compared to training independent frames with 3DGS. Figure 7 illustrates the rendering of independent frames and Figure 9 illustrates the simultaneous training of multiple frames with STG. Temporal flickering is visible for 3DGS in areas of ambiguous geometry, while renders are coherent for STGFS in static areas. While this can have an important impact on subjective quality, it is a missing piece of information for PSNR, SSIM, and LPIPS, which are the metrics classically used for volumetric video quality evaluation.

Conclusions

In this paper, we provided an overview of volumetric video, focusing on recent advances in radiance fields for real-time free navigation of natural content. Advances in radiance field training and rendering optimization, reconstruction stability, and temporal expansion were detailed. The performance of state-of-the-art radiance field methods was evaluated in terms of objective metrics, training complexity, and memory usage on a multiview video dataset of complex scenes.

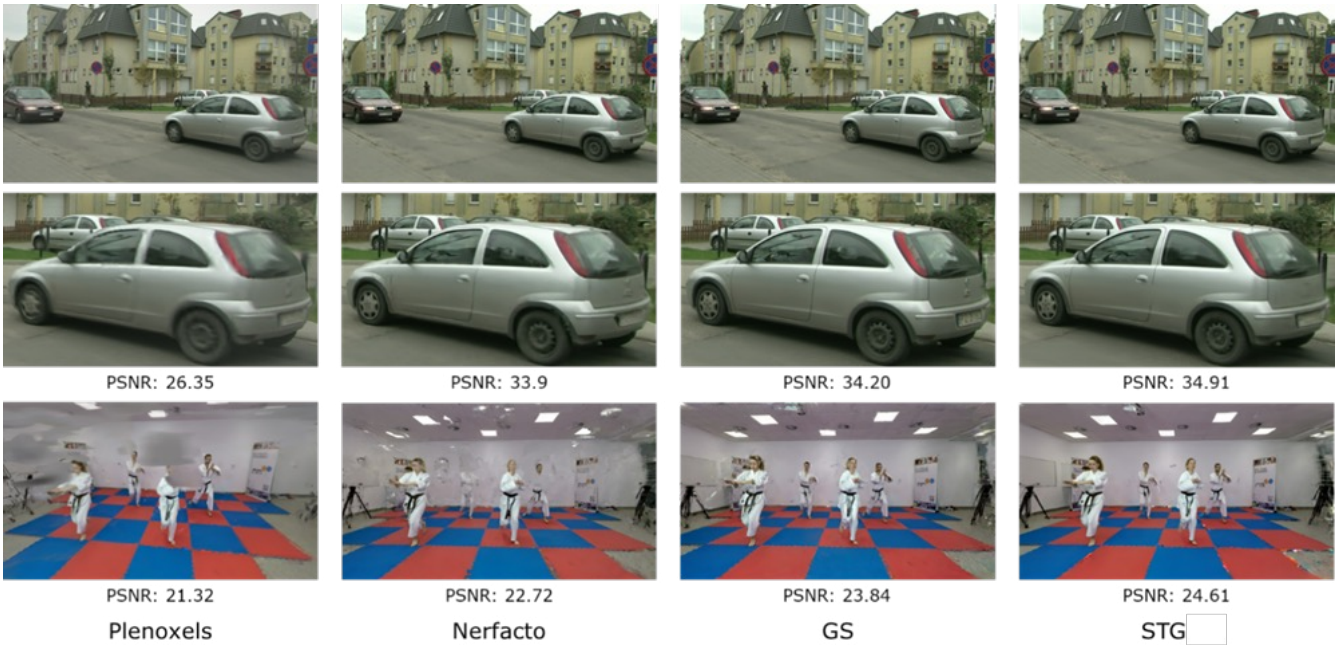


Figure 8: Renders from sequences PoznanStreet [60] and MartialArts [70]

Figure 9: Illustration of the temporal stability in Spacetime Gaussian Feature Splatting renderings of adjacent frames models on the Carpark scene [60]

Since the introduction of NeRF in 2020, volumetric video has steadily evolved towards reliable application in real-world use cases. Radiance field techniques keep improving and are close to maturity for widespread use. Volumetric video could be the long-awaited answer to the lack of engaging content on immersive displays, helping content providers create immersive experiences with minimal production costs.

References

[1] A. Smolic, « 3D video and free viewpoint video – From capture to display », Pattern Recognition, vol. 44, no 9, p. 1958-1968, 2011, doi: https://doi.org/10.1016/j.patcog.2010.09.005.

[2] G. Thomas, R. Gade, T. B. Moeslund, P. Carr, et A. Hilton, « Computer vision for sports: Current applications and research topics », Computer Vision and Image Understanding, vol. 159, p. 3-18, 2017, doi: https://doi.org/10.1016/j.cviu.2017.04.011.

[3] M. Hackett, B. Makled, E. Mizroch, S. Venshtain, et M. Mccoy-Thompson, « Volumetric Video and Mixed Reality for Healthcare Training », mai 2022.

[4] S. Orts-Escolano et al., « Holoportation: Virtual 3d teleportation in real-time », in Proceedings of the 29th annual symposium on user interface software and technology, 2016, p. 741-754.

[5] « Intel True View - Intel in Sports ». [En ligne]. Disponible sur: https://www.intel.com/content/www/us/en/sports/technology/true-view.html

[6] C. Fehn, « Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV », Proc SPIE, vol. 5291, mai 2004.

[7] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, et R. Ng, « Nerf: Representing scenes as neural radiance fields for view synthesis », in European Conference on Computer Vision, 2020, p. 405-421.

[8] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, et A. Kanazawa, « Plenoxels: Radiance Fields Without Neural Networks », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, p. 5501-5510.

[9] M. Tancik et al., « Nerfstudio: A Modular Framework for Neural Radiance Field Development », in ACM SIGGRAPH 2023 Conference Proceedings, in SIGGRAPH '23. 2023.

[10] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, et P. P. Srinivasan, « Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields », in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, p. 5855-5864.

[11] B. Roessle, N. Müller, L. Porzi, S. R. Bulò, P. Kontschieder, et M. Nießner, « GANeRF: Leveraging Discriminators to Optimize Neural Radiance Fields », ACM Trans. Graph., vol. 42, no 6, nov. 2023, doi: 10.1145/3618402.

[12] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. M. Sajjadi, A. Geiger, et N. Radwan, « RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs », CoRR, vol. abs/2112.00724, 2021, [En ligne]. Disponible sur: https://arxiv.org/abs/2112.00724

[13] B. Kerbl, G. Kopanas, T. Leimkühler, et G. Drettakis, « 3D Gaussian Splatting for Real-Time Radiance Field Rendering », ACM Transactions on Graphics, vol. 42, no 4, juill. 2023, [En ligne]. Disponible sur: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[14] M. Ye, M. Danelljan, F. Yu, et L. Ke, « Gaussian grouping: Segment and edit anything in 3d scenes », arXiv preprint arXiv:2312.00732, 2023.

[15] L. Song et al., « Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields », IEEE Transactions on Visualization and Computer Graphics, vol. 29, no 5, p. 2732-2742, 2023.

[16] J. Zhang et al., « Editable Free-viewpoint Video Using a Layered Neural Representation », arXiv e-prints, p. arXiv-2104, 2021.

[17] A. M. Barros, M. Michel, Y. Moline, G. Corre, et F. Carrel, « A Comprehensive Survey of Visual SLAM Algorithms », Robotics, vol. 11, no 1, 2022, doi: 10.3390/robotics11010024.

[18] J. L. Sch"onberger et J.-M. Frahm, « Structure-from-Motion Revisited », in Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[19] Z. Wang, S. Wu, W. Xie, M. Chen, et V. A. Prisacariu, « NeRF–: Neural Radiance Fields Without Known Camera Parameters », arXiv e-prints, p. arXiv:2102.07064, févr. 2021.

[20] Y. Jeong, S. Ahn, C. Choy, A. Anandkumar, M. Cho, et J. Park, « Self-calibrating neural radiance fields », in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, p. 5846-5854.

[21] S.-F. Chng, S. Ramasinghe, J. Sherrah, et S. Lucey, « Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation », in European Conference on Computer Vision, 2022, p. 264-280.

[22] Z. Liu, P. An, S. Liu, et Z. Zhang, « Arbitrary view generation based on DIBR », in 2007 International Symposium on Intelligent Signal Processing and Communication Systems, 2007, p. 168-171.

[23] M. Levoy et P. Hanrahan, « Light field rendering », in Seminal Graphics Papers: Pushing the Boundaries, Volume 2, 2023, p. 441-452.

[24] S. J. Gortler, R. Grzeszczuk, R. Szeliski, et M. F. Cohen, « The lumigraph », in Seminal Graphics Papers: Pushing the Boundaries, Volume 2, 2023, p. 453-464.

[25] A. Krizhevsky, I. Sutskever, et G. E. Hinton, « Imagenet classification with deep convolutional neural networks », Advances in neural information processing systems, vol. 25, 2012.

[26] J. F. N. R. L. Z. N. H. J. Maraval, « MUSE: A Multi-view Synthesis Enhancer », To be published in EUSIPCO, 2023.

[27] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, et G. Brostow, « Deep blending for free-viewpoint image-based rendering », ACM Transactions on Graphics (TOG), vol. 37, no 6, p. 1-15, 2018.

[28] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, et S. Gao, « Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis », in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), oct. 2019.

[29] R. Huang, S. Zhang, T. Li, et R. He, « Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis », in Proceedings of the IEEE International Conference on Computer Vision (ICCV), oct. 2017.

[30] G. Riegler et V. Koltun, « Free view synthesis », in European Conference on Computer Vision, 2020, p. 623-640.

[31] G. Riegler et V. Koltun, « Stable view synthesis », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, p. 12216-12225.

[32] D. Rückert, L. Franke, et M. Stamminger, « Adop: Approximate differentiable one-pixel point rendering », ACM Transactions on Graphics (TOG), vol. 41, no 4, p. 1-14, 2022.

[33] B. Mildenhall et al., « Local light field fusion: Practical view synthesis with prescriptive sampling guidelines », ACM Transactions on Graphics (TOG), vol. 38, no 4, p. 1-14, 2019.

[34] J. Flynn et al., « Deepview: View synthesis with learned gradient descent », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, p. 2367-2376.

[35] S. Wizadwongsa, P. Phongthawee, J. Yenphraphai, et S. Suwajanakorn, « Nex: Real-time view synthesis with neural basis expansion », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, p. 8534-8543.

[36] M. Broxton et al., « Immersive light field video with a layered mesh representation », ACM Transactions on Graphics (TOG), vol. 39, no 4, p. 81-86, 2020.

[37] Q. Wang et al., « Ibrnet: Learning multi-view image-based rendering », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, p. 4690-4699.

[38] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, et J. Valentin, « Fastnerf: High-fidelity neural rendering at 200fps », in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, p. 14346-14355.

[39] P. Hedman, P. P. Srinivasan, B. Mildenhall, J. T. Barron, et P. Debevec, « Baking neural radiance fields for real-time view synthesis », in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, p. 5875-5884.

[40] B. Deng, J. T. Barron, et P. P. Srinivasan, « JaxNeRF: an efficient JAX implementation of NeRF, 2020 ».

[41] T. Neff et al., « DONeRF: Towards Real-Time Rendering of Neural Radiance Fields using Depth Oracle Networks ». 2021.

[42] K. Zhang, G. Riegler, N. Snavely, et V. Koltun, « NeRF++: Analyzing and Improving Neural Radiance Fields ». 2020.

[43] M. Tancik et al., « Block-NeRF: Scalable Large Scene Neural View Synthesis ». 2022.

[44] H. Turki, D. Ramanan, et M. Satyanarayanan, « Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, p. 12922-12931.

[45] M. Zhenxing et D. Xu, « Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields », in The Eleventh International Conference on Learning Representations, 2022.

[46] J. Ost, F. Mannan, N. Thuerey, J. Knodt, et F. Heide, « Neural scene graphs for dynamic scenes », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, p. 2856-2865.

[47] M. Niemeyer et A. Geiger, « Giraffe: Representing scenes as compositional generative neural feature fields », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, p. 11453-11464.

[48] M. Mihajlovic, A. Bansal, M. Zollhoefer, S. Tang, et S. Saito, « KeypointNeRF: Generalizing Image-based Volumetric Avatars using Relative Spatial Encoding of Keypoints ». arXiv, 2022. doi: 10.48550/ARXIV.2205.04992.

[49] K. Park et al., « HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields ». 2021.

[50] A. Grigorev et al., « StylePeople: A Generative Model of Fullbody Human Avatars ». 2021.

[51] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, et Y. Sheikh, « Neural volumes: Learning dynamic renderable volumes from images », arXiv preprint arXiv:1906.07751, 2019.

[52] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, et P. Hedman, « Mip-nerf 360: Unbounded anti-aliased neural radiance fields », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, p. 5470-5479.

[53] C. Sun, M. Sun, et H.-T. Chen, « Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, p. 5459-5469.

[54] M. Kim, S. Seo, et B. Han, « InfoNeRF: Ray Entropy Minimization for Few-Shot Neural Volume Rendering », déc. 2021.

[55] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, et M. Nießner, « Dense depth priors for neural radiance fields from sparse input views », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, p. 12892-12901.

[56] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, et A. Kanazawa, « Plenoctrees for real-time rendering of neural radiance fields », in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, p. 5752-5761.

[57] D. Rebain, W. Jiang, S. Yazdani, K. Li, K. M. Yi, et A. Tagliasacchi, « DeRF: Decomposed Radiance Fields », arXiv preprint arXiv:2011.12490, 2020.

[58] L. Liu, J. Gu, K. Z. Lin, T.-S. Chua, et C. Theobalt, « Neural sparse voxel fields », Advances in Neural Information Processing Systems, vol. 33, p. 15651-15663, 2020.

[59] T. Müller, A. Evans, C. Schied, et A. Keller, « Instant Neural Graphics Primitives with a Multiresolution Hash Encoding », ACM Trans. Graph., vol. 41, no 4, p. 102:1-102:15, juill. 2022, doi: 10.1145/3528223.3530127.

[60] D. Mieloch, A. Dziembowski, et M. Domański, « [MPEG-I Visual] Natural Outdoor Test Sequences », Natural Outdoor Test Sequences, Brussels, Belgium, 2020.

[61] A. Pumarola, E. Corona, G. Pons-Moll, et F. Moreno-Noguer, « D-nerf: Neural radiance fields for dynamic scenes », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, p. 10318-10327.

[62] J.-W. Liu et al., « Devrf: Fast deformable voxel radiance fields for dynamic scenes », Advances in Neural Information Processing Systems, vol. 35, p. 36762-36775, 2022.

[63] G. Wu et al., « 4d gaussian splatting for real-time dynamic scene rendering », arXiv preprint arXiv:2310.08528, 2023.

[64] Y. Liang et al., « GauFRe: Gaussian Deformation Fields for Real-time Dynamic Novel View Synthesis », arXiv preprint arXiv:2312.11458, 2023.

[65] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, et A. Kanazawa, « K-planes: Explicit radiance fields in space, time, and appearance », in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, p. 12479-12488.

[66] Z. Yang, H. Yang, Z. Pan, X. Zhu, et L. Zhang, « Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting », arXiv preprint arXiv:2310.10642, 2023.

[67] Z. Li, Z. Chen, Z. Li, et Y. Xu, « Spacetime gaussian feature splatting for real-time dynamic view synthesis », arXiv preprint arXiv:2312.16812, 2023.

[68] B. Wei, J. Maraval, M. Outtas, K. Kpalma, N. Ramin, et L. Zhang, « MUVOD: Multi-view Video Object Segmentation Dataset ». Submission in progress, 2023. [En ligne]. Disponible sur: https://volumetric-repository.labs.b-com.com/#/muvod

[69] L. Z. M. J. N. RAMIN, « K3BO: Keypoint-based bounding box optimization for radiance field reconstruction from multi-view images », ICME workshop on Immersive Media Compression, 2023.

[70] D. Mieloch et al., « [MIV] New natural content - MartialArts ». janvier 2023.

NeRF Based 3D Generative Video Conferencing System

Jianglong Li, Jun Xu, Yuelin Hu, Zhiyu Zhang, Li Song
Institute of Image Communication and Network Engineering,
Shanghai Jiao Tong University, China

Abstract
Video conferencing, is the most demanding form of video communication in terms of real-time requirements and it remains a challenge to maintain quality in weak network conditions. Traditional block-based encoding video conferencing systems can experience freezing and significant degradation when bandwidth is extremely low or network conditions deteriorate suddenly. Recent advancements in 3D facial representation offer novel promising solutions for video conferencing under weak network conditions. This paper introduces a generative 3D video conferencing system using pre-trained Neural Radiance Field (NeRF) models for high-fidelity 3D head reconstruction and real-time rendering. Clients extract and encode facial parameters for transmission, while simultaneously receiving and decoding parameters from peers to generate visuals. Our system maintains good video quality at bit-rates under 5kbps, with objective and subjective quality comparable to HEVC encoders at 18kbps and 50kbps, respectively. By integrating real-time face tracking of facial parameters, Real-Time Communication (RTC), and real-time volumetric video rendering, our system enhances the potential for 3D video conferencing collaboration. A live demonstration showcases the significant innovation of the system, promising to forge a new paradigm for video conferencing in the context of future spatial computing.

Introduction

According to the latest report from Cisco, video traffic accounted for 82% of internet traffic in 2022 [32]. As a crucial form of video communication, video conferencing demands high real-time performance, necessitating a lightweight overall system and strong network adaptability. Meeting these requirements continues to pose a challenge to both the academic and industrial communities.

Currently, mainstream video conferencing systems such as Zoom, Microsoft Teams and Tencent Meeting typically rely on traditional video codec frameworks. Traditional video codecs like High Efficiency Video Coding (HEVC/H.265) [1], Versatile Video Coding (VVC/H.266) [2], and AV1 [3] have several advantages: (1) they aim for pixel-level fidelity, staying true to the original images; (2) they are universal and stable, providing good encoding performance across a wide range of scenarios; (3) they have low hardware performance requirements, facilitating large-scale deployment. However, traditional video encoding also has significant drawbacks, particularly its inability to handle extremely low-bandwidth conditions. This often results in freezing and a dramatic decrease in quality, leading to a poor user experience. These challenges have prompted the exploration of new encoding solutions suitable for video conferencing systems.

In the academic community, numerous studies have examined the characteristics of video within video conferencing, leading to the development of various solutions. Specifically, it has been noted that in video conferencing scenarios, the background image behind a participant is typically static, focusing attendees' attention primarily on the facial region. This implies that only the facial video needs to be encoded and transmitted. Facial images usually have similar structures and semantic meanings (such as eyes, mouths, etc.), which suggests that we can learn *a priori* on a facial dataset and then use minimal semantic information to reconstruct facial images. Recent deep learning methods have demonstrated potential in generating facial imagery from limited

information, making them promising for facial semantic communication. Feng et al [4] proposed a generative video compression framework based on FSGAN [9], achieving a low bit-rate of around 1 kB/s, but it struggles with significant facial movements. FOMM [10] uses keypoints and Jacobians to represent sparse motion, which is then used to animate a talking face. Building on FOMM, Konuko et al [8] utilize one raw frame as a reference frame and add generated frames to the reference frame pool, which may lead to error accumulation. Xu et al [11], building on [6], proposed a hybrid encoding framework based on facial keypoints. However, it needs keyframes on both sides from which to reconstruct intermediate frames. This introduces additional latency, which is less acceptable in real-time communication (RTC) scenarios. All these methods can only reconstruct 2D faces.

The recent surge in Artificial Intelligence Generated Content (AIGC) has fostered innovative 3D representation methods, such as Neural Radiance Fields (NeRF) [12] and 3D Gaussian Splatting (3D GS) [13], which promise to shift video communication towards generative approaches. These generative video communication technologies not only aim to address the issues associated with low-bandwidth networks as mentioned earlier, but also offer users novel experiences. NeRF is a technique for creating realistic 3D models from 2D images, utilizing neural networks to predict the color and density of 3D points in space, based on their coordinates and camera viewpoints, and achieving the rendering through volumetric rendering. 3D GS uses 3D Gaussians to model three-dimensional scenes explicitly and optimizes parameters using the capabilities of neural networks. There have already been applications using NeRF or 3D GS to represent parameterized 3D human heads, capable of real-time rendering [14][19]. Building on these advances within the 3D community, this paper proposes and implements an ultra-low bit-rate generative 3D video conferencing system. Our approach requires the receiver to have a personalized NeRF model [14] of the other participants.

Compared to conventional system such as WebRTC, only facial parameters are extracted and transmitted to the other end, instead of image data. The receiver end then reconstructs the 3D head based on the decoded expression parameters and the personalized NeRF model of the sender's head. In particular, the receiver can use custom pose parameters for rendering from any chosen viewpoint. This system not only addresses issues related to weak network environments but also provides an enhanced 3D viewing experience.

The main contributions of this paper are:

1) We innovatively propose a 3D video conferencing system, which is based on open-source components, with all modules being real-time and practical. The overall end-to-end latency of the system is below 90ms. To the best of our knowledge, this is the first practical real-time system that integrates a 3D representation model.

2) The parameter encoding module and pose control module allow the system to achieve ultra-low bit-rates while supporting free-viewpoint watching.

3) Experiments demonstrate that our approach outperforms traditional encoding methods. At bit-rates under 5kbps, our video quality achieves levels comparable to those of the x265 encoder at 18kbps and 50kbps for objective and subjective metrics, respectively.

Related Work

Video Conference System

As industries accelerate their digital transformation, the demand for Real-Time Communication (RTC) applications has surged dramatically. Video conferencing has become a cornerstone of professional collaboration, remote education, and personal connections. This evolution has driven widespread adoption of video conferencing across various platforms, including mobile devices, personal computers, and dedicated conference systems.

With the expansion of application scopes and the increasing demand for advanced features, many systems have transitioned from traditional on-premises setups to more scalable cloud-based solutions to enhance computational efficiency and manageability. As spatial computing has advanced, facial representation has also evolved from traditional 2D generation to 3D reconstruction. For example, Google Project Starline [34] and Apple Vision Pro [35] have introduced new possibilities for immersive experience video conferencing. Video conferencing systems have stringent requirements for latency and bandwidth efficiency to maintain optimal Quality of Experience (QoE) across diverse network conditions.
This necessitates the support of efficient transport protocols, among which the QUIC protocol is considered more promising than traditional UDP-based transmission protocols due to its enhanced efficiency and reliability.

Facial Presentation

In traditional video conferencing systems, a block-based hybrid encoding architecture, such as HEVC and AV1, is used. This approach is suboptimal when encoding video specifically for video conferencing scenarios. With the development of deep learning technologies, some approaches have adopted generative methods for 2D facial generation. These methods, such as those outlined in [5] [6] [11], primarily rely on the extraction of 2D keypoints. At the sending end, facial images are divided into keyframes and non-keyframes, with keypoints extracted from non-keyframes. At the receiving end, deep generative models use these keypoints and keyframes to reconstruct the non-keyframes. Due to their dependency between frames, these methods lack robustness in handling large movements.

As spatial computing continues to advance, research into 3D facial representation has expanded significantly, exploring a variety of innovative approaches. These methods enable the utilization of minimal semantic parameters to drive models, facilitating the accurate reconstruction of 3D faces. This technological progress has made 3D video conferencing a practical reality. Some studies focus on 3D keypoints; for example, in [7] the authors extract 3D keypoints from facial images to simulate facial movements. Although this technique allows for rendering from arbitrary viewpoints at the receiver end, it lacks robustness for large movements. Other research has adopted the parameterized NeRF model for heads. For example, NeRFace [16] inputs facial parameters to achieve dynamic 3D modeling, but the training of this model is slow and fidelity is relatively low. Research presented in [14] introduces a method for reconstructing a facial semantic NeRF model from monocular videos, facilitating rapid training and high-fidelity reconstruction. With a pre-trained NeRF model, this approach can generate 3D faces using a set of expression parameters and allows for specifying the angle of 2D rendering using pose parameters.

Additionally, some research has utilized explicit expressions of 3D GS to represent 3D human heads. In [18], the explicit representation of 3D GS is blended with a set of learnable latent features, enabling the driving of a parametrized head model with low-dimensional linear parameters. [19] employs a 3D Gaussian field to represent a parametrized facial model, capturing facial details using geometric priors, and achieves high-fidelity rendering at 300 fps on consumer-grade GPUs.

Transport Protocol

The choice of transport protocol is crucial to video conferencing systems. The most commonly used transport layer protocols are Transmission Control Protocol (TCP) and User Datagram Protocol (UDP).

While TCP provides ordered and reliable data delivery, its susceptibility to high latency due to head-of-line blocking and inefficient retransmission tactics renders it less ideal for RTC applications. Consequently, most advanced video conferencing systems leverage UDP-based protocols that prioritize timeliness over reliability. These systems commonly utilize the Real-Time Transport Protocol (RTP) [20] and its secure variant, the Secure Real-time Transport Protocol (SRTP) [21], with industry leaders like Zoom enhancing these protocols with custom extensions [22].

The Quick UDP Internet Connections (QUIC) [23] protocol has recently attracted considerable attention for its impressive performance and flexibility. QUIC dramatically reduces the time required to establish a reliable and secure connection to just one Round-Trip Time (RTT), a significant improvement over TCP's cumbersome three-way handshake. It also facilitates stream-multiplexing and connection migration, which enhance performance under fluctuating network conditions and simplify congestion control upgrades through its modular approach.

This growing interest in QUIC indicates its potential to revolutionize real-time video streaming. Ongoing research is focused on integrating QUIC with traditional video streaming protocols, such as adapting RTP to operate over QUIC [24] and extending it to support unreliable transmissions [25]. These developments position QUIC as a transformative element in the video conferencing domain.

System Architecture

Overview

The proposed 3D generative video conferencing architecture is depicted in Figure 1, consisting of a sender, network simulation, and receiver. The workflow of our video conferencing system is as follows: At the sender side, video is first captured from the camera. Then facial expression parameters and pose parameters are extracted from each frame. These parameters are encoded into bitstreams for network transmission. At the receiver side, bitstreams are received from the network, decoded to retrieve the expression parameters and pose parameters, and used to drive a NeRF-based 3D head representation model to generate and display facial images. In our system, different modules communicate through FIFO (a First-In First-Out buffer), which enable asynchronous data transfer

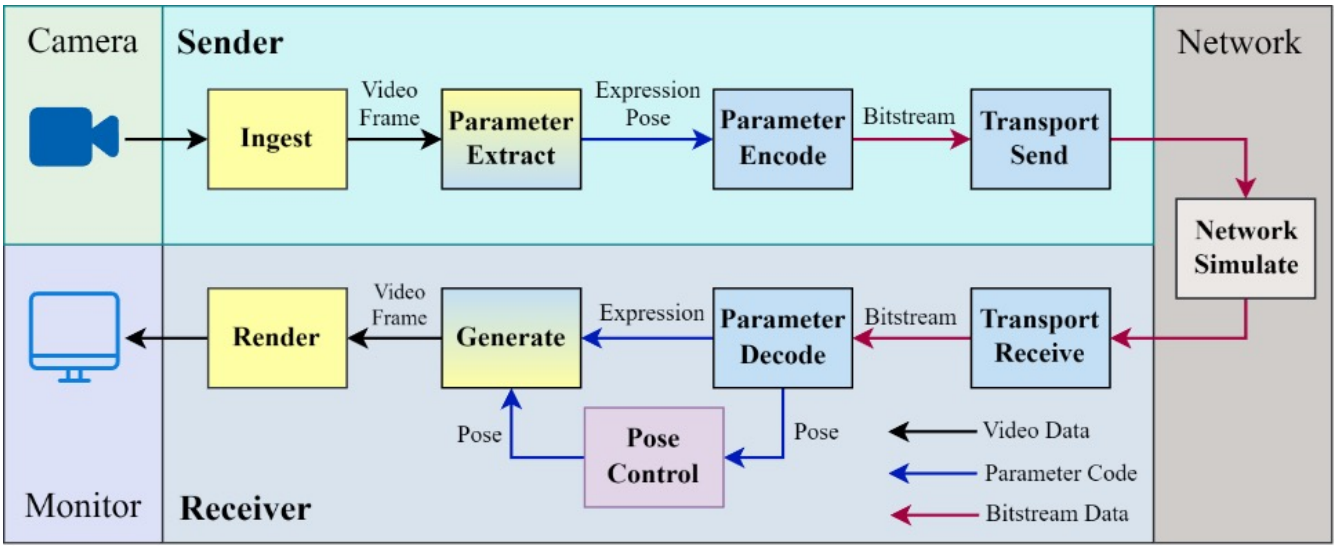


Figure 1: The proposed framework for the 3D generative video conferencing system. The data format in the yellow modules is images, while in the blue modules, the data format consists of expression and pose parameters. Black arrows represent video data, blue arrows indicate parameter codes, and red arrows represent bitstream data.

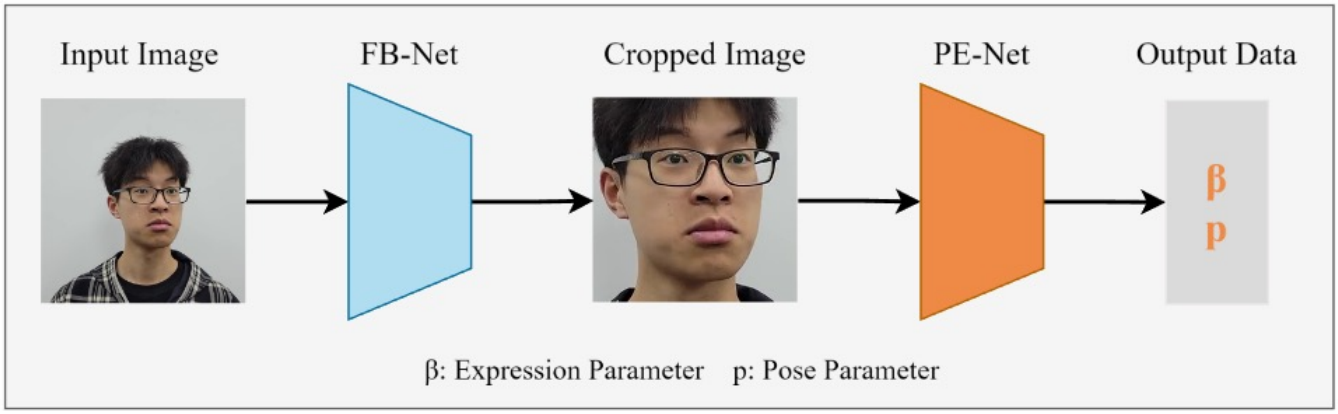


Figure 2: Parameter extraction module workflow. An input frame image is processed using a Face Bounding Network (FB-Net) to obtain a cropped image, which is then input into a Parameter Estimation Network (PE-Net) to extract expression and pose parameters.

and buffering between modules, thus facilitating real-time operation.

Our 3D video conferencing system not only supports viewing from a realistic perspective at the receiver side but also from any arbitrary angle. Moreover, since only the expression parameters and pose parameters are transmitted for each frame, the bit-rate of our system is significantly lower, while still providing video quality comparable to traditional video conferencing systems at a much higher bit-rate.

In the following subsections, we provide a detailed description of the functionalities and implementations of each module within the 3D generative video conferencing system.

Ingestion and Rendering

At the sender side, the Ingest module captures video data from a camera. This video data is subsequently segmented by frames and fed into a FIFO queue, which reliably forwards it to subsequent processing modules.

On the receiver side, the Render module is responsible for displaying the received video frames. To ensures that the video is rendered smoothly. Specifically, we have established a playback buffer for K frames and set the playback time for the first frame as the base time. Each subsequent frame is assigned a target playback time according to the frame rate.

Parameter Extract Module

This module extracts facial expression parameters and head pose parameters from video frames, employing the open-source CPEM model [15] equipped with pre-trained weights.

CPEM utilizes a linear 3D Morphable Model (3DMM) [26] as its 3D facial model, which comprises both shape and texture components. The shape component is subdivided into a facial base and an expression base. The expression parameters extracted by CPEM represent the coefficients of the expression base, guided by the FaceWarehouse [27] database, which distinctly annotates each expression base with specific semantics (e.g., Eye Close Left, Eye Squint Left). The pose parameters generated by CPEM include both rotational and translational elements. In our system, to ensure a consistent communication experience, the face is fixed at the centre, therefore only the rotational element is utilized.

The operational workflow of the Extract module is depicted in Figure 2. It starts with the Face Bounding Network, which segments out the facial region from the image. Subsequently, the segmented facial image is input into the Parameter Estimation Network, based on a ResNet50 architecture, to estimate the expression coefficients and pose parameters. The extracted parameter information is sent to a FIFO and then passed to the encoding module for encoding.

Encoding and Decoding

This process encodes the extracted expression and pose parameters (floating point numbers), including quantization, prediction, and zero-order exponential Golomb coding.

In the quantization step, each floating-point number is converted to an 8-bit integer. To improve the quantization accuracy, we record the maximum and minimum values of each dimension from our model training dataset. These bounds serve as the upper and lower limits for the expression parameters and pose parameters.

The prediction module utilizes inter-frame prediction to compute differences between successive frames, encoding only the resultant residuals via zero-order exponential Golomb coding, thus markedly reducing the volume of data to encode.

Zero-order exponential Golomb coding, an effective lossless compression technique, is particularly suited for data sequences predominantly comprising small values. This method, which builds on Golomb coding, utilizes straightforward rules for encoding non-negative integers, thereby efficiently compressing the data.

The decoding module performs inverse operations to recover the expression parameters and pose parameters.

Generate Module

The generation module at the receiver end utilizes expression and pose parameters to drive a NeRF model for generating facial images from specific viewpoints. During generation, the expression parameters dictate the facial expressions, while the pose parameters determine the viewing angle of the generated face. The receiver can either use the pose parameters transmitted from the sender for true-to-perspective facial generation or generate from arbitrary viewpoints through a pose control module. Our system supports free-viewpoint generation along the x, y, and z axes, within a range of ±45°.

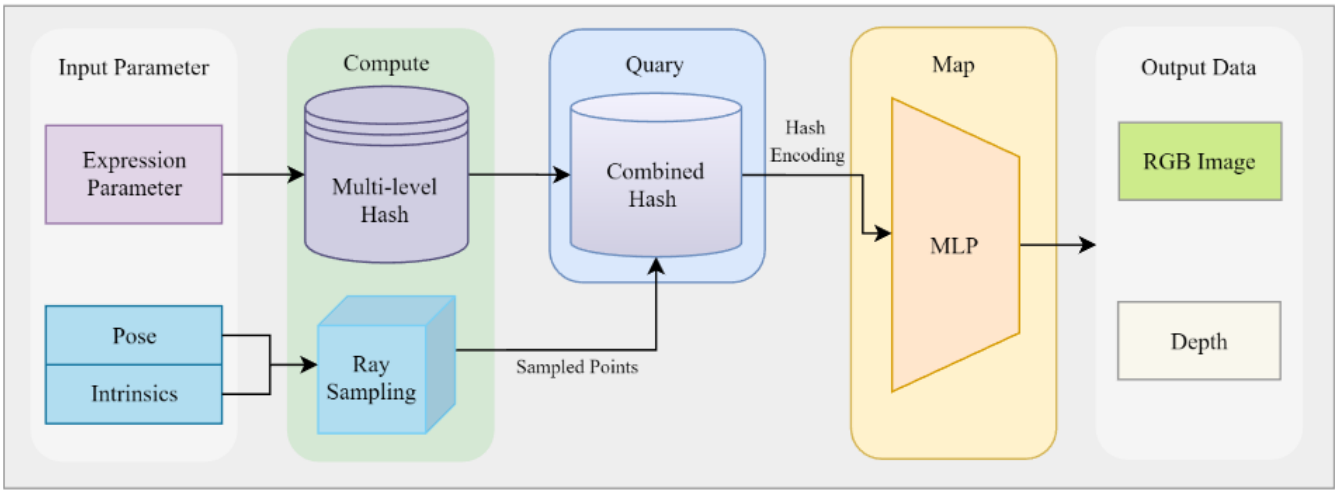


Figure 3: Generate module workflow

This functionality is implemented using the open-source NerfBlenderShape [14][17]. NerfBlenderShape is a semantic facial model based on NeRF that can be trained in 10-20 minutes using short monocular RGB video inputs and can render facial images in tens of milliseconds based on expression and pose parameters. It represents facial semantics as an MLP-based implicit function and links expression bases with multi-level hash tables. Each hash table corresponds to specific facial semantics, and expression coefficients can combine encodings from multiple hash tables. Camera parameters used during rendering include intrinsic parameters and extrinsic parameters represented by the head pose.

The detailed rendering process of NerfBlenderShape is illustrated in Figure 3. For each image to be rendered, the input parameters include expression parameters and camera parameters. The expression parameters are used to linearly combine multi-level hash tables in linear space to form a combined hash table. Camera parameters are fed into the Ray Sampling module to cast rays and obtain sample points. These sample points are then queried in the combined hash table to obtain hash encodings. Finally, these hash encodings are input into a lightweight MLP to produce the RGB image and depth information.

Building on the open-source methodology provided by NerfBlenderShape, we retrained the NeRF model using our proprietary dataset to derive the weights employed in the Generate module. Detailed descriptions of the training process are presented in Implement section.

Sending and Receiving

The sending and receiving modules use QUIC as the underlying transport protocol to enhance the efficiency and reliability of data transmission. The sending module first encapsulates the encoded facial expression and pose parameters into QUIC packets. These packets are then sent over the network, where they quickly and reliably reach the peer via the QUIC protocol. Leveraging QUIC's sequential characteristics and reliable transmission, the system ensures the continuity and integrity of expression and pose data. Additionally, QUIC's one Round-Trip Time (1RTT) handshake and robust congestion control mechanisms effectively reduce transmission latency.

We also provide a network simulation module that utilizes the TC tool to simulate network traces, which facilitates the verification of our proposed system's adaptability to weak network conditions. To better simulate various and extremely weak network conditions, we have opted to deploy our system on a single host. Network transmission occurs over the local loopback, controlled using the TC tool.

Implementation

We implement our 3D generative facial video conferencing system on an Ubuntu 20.04 64-bit operating system using Python. The entire system operates on two Intel® Xeon® Gold 6240 CPUs at 2.60GHz with 256 GB of RAM, and utilizes two GeForce RTX 4090 GPUs for neural network computations. To ensure real-time performance, the extraction module on the sender side and the generation module on the receiver side are each run on separate GPUs.

In our system, the dimension of the expression parameters is set to 46, and the dimension of the pose parameters (rotational elements only) is set to 3. This section details the specific implementation of key system modules.

Nerf Model Training

Following the methodology provided by RAD-NeRF [28], we constructed our own dataset. The dataset construction involves the following steps:

1. Video Recording: A 100-second video capturing a variety of facial expressions and head poses was filmed.
2. Semantic Segmentation: Each frame was semantically segmented to identify the background, head, neck, and torso.
3. Background Extraction: A stable background image was intelligently extracted by analysing the foreground and background within the image sequence.
4. Torso Image Extraction: Using the semantic segmentation images from step 2 and the original images, torso images were extracted, and the backgrounds in the original images were replaced with the stable background from step 3 to create ground truth images.
5. Facial Tracking: Pose parameters for each image were obtained using facial tracking, including 3-dimensional Euler angles and 3-dimension translations, which were then converted into standard 4x4 position matrices.
6. Expression Parameter Extraction: 46-dimensional expression parameters were extracted from each image using the CPEM model.
7. Training File Preparation: Expression and pose parameters were compiled into a training file.

Using the training methodology provided by NerfBlenderShape, we retrained our NeRF head model to suit our conference system. For optimization, we employed the Adam optimizer with an initial learning rate of 0.001, and momentum betas configured to (0.9, 0.99). To facilitate effective learning rate management throughout the training, we incorporated a MultiStepLR

scheduler that dynamically adjusted the learning rate at predetermined epochs. The entire training process was designed to run 200 epochs.

Network Transmission Based on QUIC

Quiche [29] is an implementation of the QUIC transport protocol and HTTP/3 as specified by the IETF. It provides a QUIC kernel implemented in Rust and offers C/C++ APIs. Utilizing the Quiche library, we designed send and receive modules using C++. Communication between the Python-based system and the C++-based network transmission modules is achieved through a FIFO named pipe.

Specifically, at the sender side, a 2-byte delimiter is appended to the encoded data of each frame to distinguish between data of different frames. This bitstream, including the delimiter, is then written into the sender's FIFO. The C++-based send module reads the bitstream from the FIFO and transmits it using the QUIC protocol. On the receiving end, the receive module reads data from the QUIC stream and writes it into the receiver's FIFO. Subsequently, the system reads the received bitstream from the FIFO and separates it into individual frames based on the delimiter.

Accelerating Model Inference Speed

In order to facilitate real-time functionality in our 3D video conferencing system, we have adopted various methods to accelerate inference. Specifically, at the sender side, the facial segmentation and CPEM models are exported and inferred in the ONNX format, rather than utilizing direct PyTorch implementations. On the receiving end, the rendering of facial images is conducted using fp16 precision. Furthermore, both modules undergo a pre-warming process prior to system activation to improve the efficiency of model inference.

Experiment

In this section, we present the experimental results of the proposed system. Bit-rate and latency performance are the most critical indicators for video conferencing systems, directly impacting the user experience. We conducted experiments on these two metrics separately.

Bit-rate

The test videos include ten video sequences, each with a length of 1000 frames. Each frame is cropped to 512x512 and encoded with an 8-bit quantization depth.

We conduct a comparative analysis with the traditional HEVC encoder, specifically x265, using quantitative metrics such as PSNR, SSIM, and LPIPS [33]. Given that our system generates images containing only the head, while the x265 encodes image blocks, we adopt the following experimental setup to ensure comparability and fairness. We use semantic segmentation to replace the background of the test video with a plain white background, preserving only the head portion as our reference source video. We then encode the source video into a specific bit-rate H.265 stream using the FFmpeg tool. An example command is: *ffmpeg -i input -c:v libx265 -x265-params bframes=0 -b:v bitrate -r 20 output*. The background used at the reception side for generating facial images is also plain white. Figure 4 compares the performance of our NeRF-based approach and x265, and illustrates the impact of quantization on generation quality during the codec process.

As shown in Figure 4, with encoding process at 20fps and 512x512 resolution, our system achieved an average bit-rate of only 4.94kbps (red star), compared to 25.6kbps without encoding process (blue star). Despite a fourfold difference in bit-rate, the video quality achieved by both methods is nearly identical.

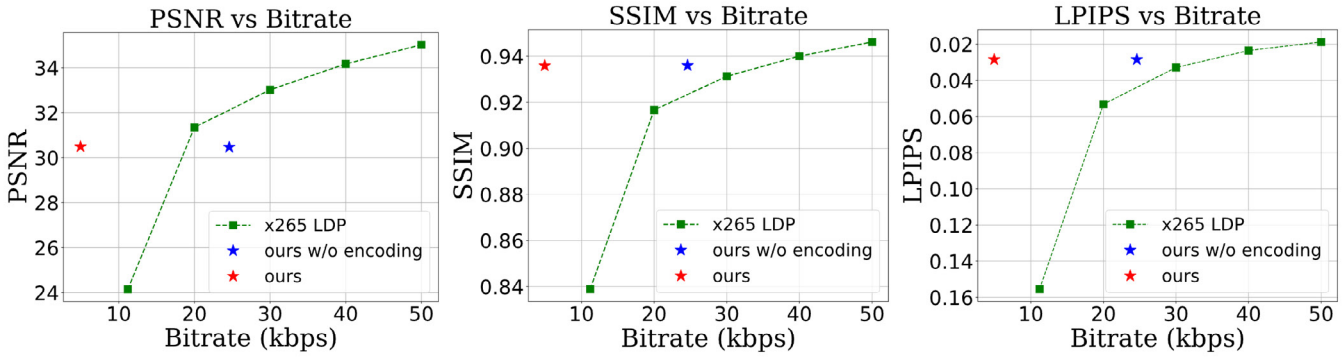


Figure 4: RD performance comparison.

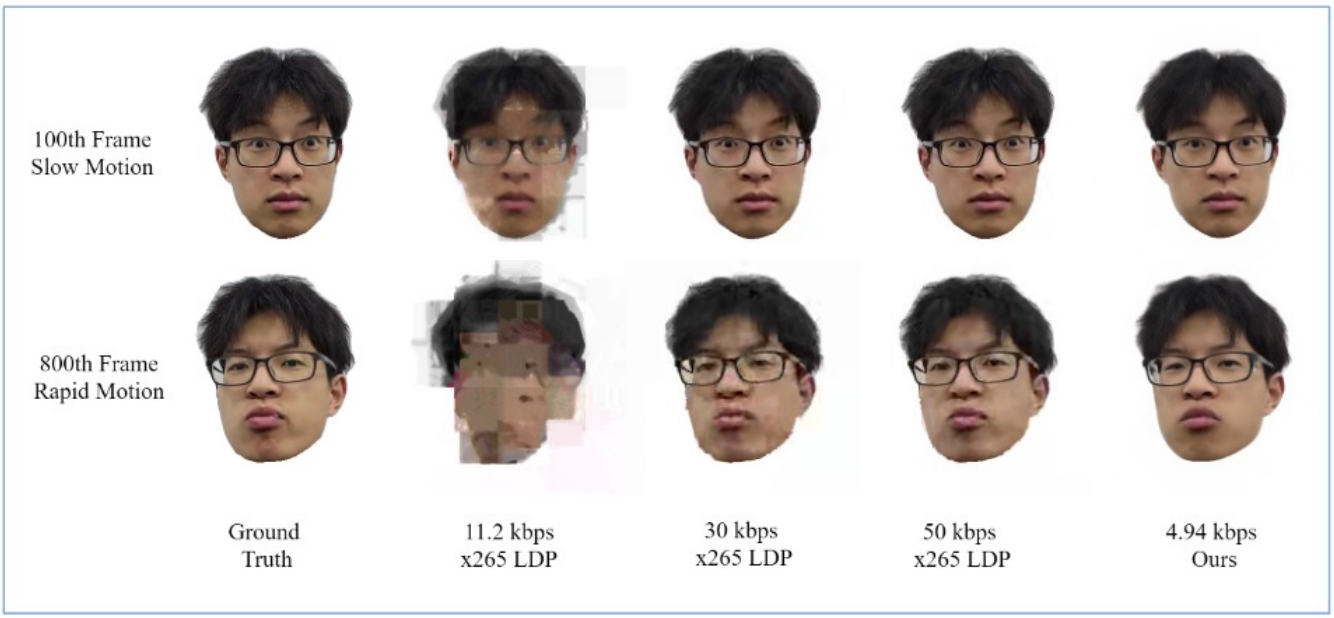


Figure 5: Subjective performance comparison

This is attributed to our facial transmission being semantics-based, which is robust against minor semantic parameter errors, having negligible impact on pixel-based objective metrics. Therefore, the information loss due to quantization does not affect our generation quality, allowing us to save a significant amount of bit-rate.

The curves for PSNR, SSIM, and LPIPS for x265 are shown with green line. Due to bit-rate compression constraints, x265 can only encode test video at a minimum bit-rate of 11.2kbps, at which the PSNR is 24.15dB. It is noteworthy that our encoding scheme at 4.94kbps achieves a PSNR of 30.49 dB, equivalent to x265 at 18kbps. Our SSIM and LPIPS metrics are 0.9360 and 0.0285, respectively, comparable to x265 at 35kbps. Our video transmission is generation-based and does not aim for pixel-perfect recovery, yet we still outperform x265 significantly in objective assessment metrics.

Figure 5 displays a human perception comparison of the test images. It is evident that our advantages are more pronounced during fast head movements. For rapidly moving images, our system achieves higher subjective quality using only one-tenth the bit-rate of x265. This is because x265 relies on prediction between adjacent frames, whereas our semantics-based method independently processes expression parameters and pose parameters for each frame.

Latency

We set the Round-Trip Time (RTT) of the link to 20ms through our network simulation module. We have measured the computational latencies of individual modules when operating independently and as part of the integrated system, as shown in Table 1.

During full system operation, the latency of each module was less than 50ms, indicating that our system can operate in real-time at a frame rate of 20 fps. Moreover, the system's end-to-end latency of 89.5ms supports a smooth real-time communication (RTC) experience. The predominant contributor to the overall system latency is the generate module at the receiver end, which imposes a computational delay of 49.2ms, thus limiting our frame rate to 20 fps. Currently, our generate module operates using the PyTorch framework. Transitioning to ONNX for inference acceleration could yield higher frame-rates and further reduce the end-to-end latency.

Discussion

This section discusses the limitations of our system and outlines future work:

- Due to current limitations in 3D reconstruction technology, our system only generates images of the head. In the future, we plan to include the upper body to enhance realism.

Delay test	Extract	Encode	Transport	Decode	Generate	End to End
Individual(ms)	14.7	0.5	12.46	0.15	27.6	---
Overall(ms)	25.6	0.8	13.69	0.21	49.2	89.5

Table 1: Latency of each system module and end-to-end latency

- Our system transmits only facial information, and the generated images lack backgrounds. Future efforts will focus on integrating 2D virtual backgrounds or placing participants' 3D head avatars within the same 3D virtual environment.

- We employ residual-based predictive coding for encoding and decoding. Network packet loss can impact the decoding of adjacent frames. To combat weak network conditions, we will introduce a group of pictures (GOP) strategy to limit decoding dependencies within a GOP. In the event of packet loss, we will retransmit parameters from I-frames, using them as references for subsequent frames to ensure reliable transmission.

- Our conferencing system requires participants to have pre-trained NeRF models of each other, and these NeRF-based head models are substantial in size (exceeding 500MB). For future practical deployments, we plan to incorporate online training. For new participants without models, we will initially conduct traditional video conferences and train individual NeRF head models based on received facial images within 20 minutes. When network conditions deteriorate, these pre-trained models can be utilized for conferencing. Additionally, there has been some work on compressing these models to less than 20MB [30][31], which we aim to integrate into our system.

Conclusions

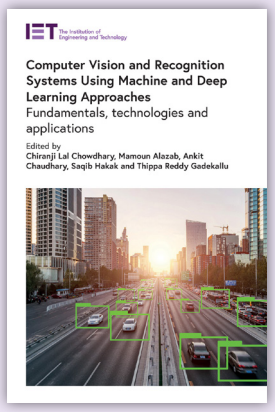
In this paper, we propose and implement an ultra-low bit-rate generative 3D video conferencing system. Utilizing the latest advancements in 3D facial reconstruction technology, we achieved the goal of conducting 3D video conferences at an ultra-low bit-rate while providing acceptable video quality. Our system supports viewing from any angle, offering participants an immersive experience. In our future work, we will further refine our system, providing a new paradigm for video conferencing systems.

References

1. G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
2. B. Bross et al., "Overview of the Versatile Video Coding (VVC) Standard and its Applications," IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 10, pp. 3736– 3764, 2021.
3. J. Han et al., "A Technical Overview of AV1," Proc. IEEE, pp. 1–28, 2021.
4. D. Feng, Y. Huang, Y. Zhang, J. Ling, A. Tang, and L. Song, "A Generative Compression Framework For Low Bandwidth Video Conference," in 2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW), Jul. 2021, pp. 1–6.
5. M. Oquab et al., "Low Bandwidth Video-Chat Compression using Deep Generative Models," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, Jun. 2021, pp. 2388–2397.
6. Tang, Anni, et al. "Generative compression for face video: A hybrid scheme." 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022.
7. T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, Jun. 2021, pp. 10034–10044.
8. Konuko, Goluck, Giuseppe Valenzise, and Stéphane Lathuilière. "Ultra-low bitrate video conferencing using deep image animation." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.
9. Nirkin, Yuval, Yosi Keller, and Tal Hassner. "Fsgan: Subject agnostic face swapping and reenactment." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
10. Siarohin, Aliaksandr, et al. "First order motion model for image animation." Advances in neural information processing systems 32 (2019).
11. J. Xu et al., "An ultra-low bitrate video conferencing system with flexible virtual access patterns" in IBC 2022.

12. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis", Communications of the ACM, vol. 65, no. 1, pp. 99-106, 2021.
13. Kerbl, Bernhard, et al. "3d gaussian splatting for real-time radiance field rendering." ACM Transactions on Graphics 42.4 (2023): 1-14.
14. Gao, Xuan, et al. "Reconstructing personalized semantic facial nerf models from monocular video." ACM Transactions on Graphics (TOG) 41.6 (2022): 1-12.
15. Mo, Langyuan, et al. "Towards accurate facial motion retargeting with identity-consistent and expression-exclusive constraints." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 2. 2022.
16. G. Gafni, J. Thies, M. Zollhöfer and M. Nießner, "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8649-8658, June 2021.
17. Xuan, USTC-3DV/NeRFBlendShape-code. 2024. Accessed: May. 12, 2024. [Online]. Available: https://github.com/USTC3DV/NeRFBlendShape-code
18. Dhamo, Helisa, et al. "Headgas: Real-time animatable head avatars via 3d gaussian splatting." arXiv preprint arXiv:2312.02902 (2023).
19. Xiang, Jun, et al. "FlashAvatar: High-Fidelity Digital Avatar Rendering at 300FPS." arXiv preprint arXiv:2312.02214 (2023).
20. H. Schulzrinne, S. L. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," Internet Engineering Task Force, Request for Comments RFC 3550, 2003. doi: 10.17487/RFC3550.
21. K. Norrman, D. McGrew, M. Naslund, E. Carrara, and M. Baugher, "The Secure Realtime Transport Protocol (SRTP)," Internet Engineering Task Force, Request for Comments RFC 3711, 2004.
22. B. Marczak and J. Scott-Railton, "Move Fast and Roll Your Own Crypto: A Quick Look at the Confidentiality of Zoom Meetings," University of Toronto, Citizen Lab Research Report No. 126, Apr. 2020.
23. A. Langley et al., "The QUIC Transport Protocol: Design and Internet-Scale Deployment," in Proceedings of the Conference of the ACM Special Interest Group on Data Communication, Los Angeles CA USA, Aug. 2017, pp. 183–196.
24. C. Perkins and J. Ott, "Real-time Audio-Visual Media Transport over QUIC," in Proceedings of the Workshop on the Evolution, Performance, and Interoperability of QUIC, Heraklion Greece, Dec. 2018, pp. 36–42.
25. M. Palmer, T. Krüger, B. Chandrasekaran, and A. Feldmann, "The QUIC Fix for Optimal Video Streaming," in Proceedings of the Workshop on the Evolution, Performance, and Interoperability of QUIC, Heraklion Greece, Dec. 2018, pp. 43–49.
26. Blanz, Volker, and Thomas Vetter. "A morphable model for the synthesis of 3D faces." Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023. 157-164.
27. Cao, Chen, et al. "Facewarehouse: A 3d facial expression database for visual computing." IEEE Transactions on Visualization and Computer Graphics 20.3 (2013): 413-425.
28. Tang, Jiaxiang, et al. "Real-time neural radiance talking portrait synthesis via audio-spatial decomposition." arXiv preprint arXiv:2211.12368 (2022).
29. Cloudflare, cloudflare/quiche. 2024. Accessed: May.12, 2024. [Online]. Available: https://github.com/cloudflare/quiche
30. Z. Zhang, A. Tang, C. Zhu, G. Lu, R. Xie and L. Song, "High-Fidelity Free-View Talking Head Synthesis for Low-Bandwidth Video Conference," 2023 IEEE International Conference on Visual Communications and Image Processing (VCIP), Jeju, Korea, Republic of, 2023, pp. 1-5
31. Zhang, Zhiyu, et al. "Efficient Dynamic-NeRF Based Volumetric Video Coding with Rate Distortion Optimization." arXiv preprint arXiv:2402.01380 (2024).
32. Cisco, "Cisco Visual Networking Index: Forecast and Trends, 2017–2022",2018
33. Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
34. Lawrence, Jason, et al. "Project starline: A high-fidelity telepresence system." ACM Transactions on Graphics (TOG) 40.6 (2021): 1-16.
35. Apple Vision Pro. Apple. Accessed May 12, 2024. [Online]. Available: https://www.apple.com/apple-vision-pro/

The Institution of Engineering and Technology



Computer Vision and Recognition Systems Using Machine and Deep Learning Approaches Fundamentals, technologies and applications

Edited by: Chiranji Lal Chowdhary, Mamoun Alazab, Ankit Chaudhary, Saqib Hakak, Thippa Reddy Gadekallu

Written by a team of International experts, this edited book covers state-of-the-art research in the fields of computer vision and recognition systems from fundamental concepts to methodologies and technologies and real-world applications. The book will be useful for industry and academic researchers, scientists and engineers.

IET Members save 25%

Order online at: shop.theiet.org

Virtual Reality and Light Field Immersive Video Technologies for Real-World Applications

Authors: Gauthier Lafruit, Mehrdad Teratani

Inspired by MPEG-I and JPEG-PLENO standardization activities, this book is for readers who want to understand 3D representations and multi-camera video processing for novel immersive media applications. The authors address new challenges that arise beyond compression-only, such as depth acquisition and 3D rendering.

IET Members save 25%

Order online at: shop.theiet.org

Media Provenance – Signing your Content in Practice

D. J. Bevan[1], N. C. Earnshaw[1], L. C. Ellis1, C. H. M. Halford[1], M. Fjellhaug[2], M. H. Iversen[2], M. F. Marcus[1], J. S. Parnall[1], L. Strappelli[1], H. O. Svela[2]
[1]BBC, UK, [2]Media Cluster Norway, Norway

Abstract
Disinformation is a threat to the healthy functioning of democratic society. The advent of Generative AI technology means it is even easier for anyone to create false or misleading information and distribute it at scale making it almost impossible for people to trust anything they see online. This is a problem for most news organisations whose audiences trust their output less and less each year. How can organisations rebuild a trusted relationship with their audiences and consumers empowered to make their own decisions about the accuracy and veracity of what they see online?

Content credentials may be one solution. This technical standard exposes the origin and provenance of media, showing users exactly who published a piece of content and how it was made. This paper will summarise findings from public trials that assess the impact of content credentials on audience trust in news media. It will also review the adoption efforts at the BBC and in the Norwegian media industry.

Introduction

Trust in news has been declining for some time 'Newman (1)'. There are many intersecting reasons for this, but recent advances in generative AI, and the fact that almost anyone can make synthetic (fake) images quickly and easily, are making it a lot more difficult for people to trust what they see online.

Studies have shown that it is currently very difficult for organisations and individuals to consistently detect what has been fabricated or manipulated to the very high accuracy level required for broadcast or publication 'Chuangchuang et al (2)'. We therefore argue that it is better to look at a hybrid strategy that includes methods to positively assert who created the content and how it was made, and whether an AI tool was used. If you make this information available to the public, they can make an informed decision themselves whether to trust it. This strategy can include a combination of media provenance, watermarking and finger printing technologies.

As well as assertions from the original creator, content provenance is useful in cases where complex stories include contributions from multiple sources by making a 'chain of provenance' visible to the audience. This also has the benefit of increasing the transparency of the editorial process.

To ensure an interoperable and core set of signed media provenance features are available for all, the Coalition for Content Provenance and Authenticity (C2PA) has published an open standard describing a way to cryptographically sign content and associated standardised metadata 'C2PA 2024 (3)'. After signature verification, the metadata can be decoded by a "validator", such as a web a browser extension, and displayed to viewers at a level of detail that they wish to receive. This capability is being rolled out under the user facing brand of Content Credentials, with adoption by a number of AI tools and early implementations by broadcasters.

There are many organisations actively investigating the value that C2PA provenance data can bring to their audiences and journalists and trialling how this new dimension can fit within their workflows. This paper covers trials for the practical use of content credentials from the BBC and Project Reynir – a collaboration led by Media Cluster Norway.

Standards

There has been an awareness of the misrepresentation of published media for some time as the tools for manipulating media objects have become commonplace, just as the danger of images being separated from their original context has grown. In response, Project Origin was formed in 2019 by BBC Research & Development, CBC / Radio-Canada, Microsoft and the New York Times to consider a technological solution for application to the news media industry 'Aythora et al (4)'. In 2020 this group came together with the members of the Content Authenticity Initiative (CAI), led by Adobe, and together with other partners formed the Coalition for Content Provenance and Authenticity (C2PA) to develop an open standard for content provenance.

C2PA swiftly put together a plan for an initial "version 1" of the specification 'Earnshaw et al' (6), focussing on a few key areas:

- The assertions (or metadata) that the signers of content might want to assert about content (e.g. where a photo was taken, or its original caption)
- The provenance chain model (called "ingredients" in C2PA), allowing multiple stages of editing to be linked into a complete history
- The trust model, linking the assertions being made to the identity of the entity signing them.

Whilst the application of the Content Credentials can span a number of media-related verticals and industrial applications, Project Origin, now expanded in membership, remains focused on their application to news media and how to sustain robust and effective news distribution in the internet age using this technology.

How it works

The C2PA specification is fundamentally a way to do a few things:

1. Make (or "assert") some statements about the content, such as capture date, or description
2. Add links back to previous versions of the media, or its components, to optionally show the provenance of the media all the way back to the capture device
3. Add a cryptographic hash to the media, so that the provenance is tied to the media in a tamper-evident way
4. Add a tamper-evident digital signature (by the "signer") over all the previous data, showing who or what made the provenance statements

This data is recorded in a structure called a Manifest, and multiple Manifests are included and linked, via 2, to show a full provenance history if required, in a "Manifest Store". A Manifest Store can be embedded in a piece of media's extension / metadata section (multiple "embeddings" are specified for different file types), or it can be stored in an internet-accessible location and a link to that location included in the media.

Finally, when some media is being consumed, all this data is accessed and read, and a "validation algorithm" is run to ensure the certificate that signed the media is valid, and all the links and the hashes are valid too. A "valid" Manifest can then be shown to the consumer, allowing them to see a secure description of what a "signer" says about the provenance of the content.

User Experience Research and Design

The BBC have been researching the user experience aspects of provenance since 2021. Our studies have helped us establish a greater understanding of:

- What image provenance information is important to show
- How people want to see that provenance information
- The impact on trust in content when provenance is shown

Some of this work has fed into the BBC's live trial (below) where provenance information was disclosed in an expandable User Interface (UI) beneath image and video content on various BBC Verify articles, which was reported on the BBC Blog 'Monday et al (5) '
The following sections explain our research and subsequently our design decisions.

What information is important and relevant

We conducted a survey of 200 people as well as one-to-one interviews with 15 people to uncover qualitative and quantitative findings. We found that there are certain types of information that people are more interested in seeing when evaluating images, including: description, time, date, location, if the image has been verified, the verification checks conducted, who published the image, who the image is owned by, the creator/photographer, any image edits, and whether AI was used.

76% of respondents preferred a medium amount of information; 52% wanted additional information about the image; 89% agree/agree strongly showing the information is useful. It found that under 35's were more concerned with contextual info e.g. creator/photographer, who published it, and who it was owned by, whilst over 35's were more concerned with factual information: date time etc.

Another important aspect we found was the use of language when surfacing this information; some words have certain associations or can be perceived more negatively or positively. According to our qualitative research findings, there were a few key areas that require more exploration around the use of language: edits, verification and the use of AI.

- **Edits:** Participants from further interviews primarily wanted to know if an image had or had not been edited. Some wanted to know more about the kind of edits or be told if significant edits had been done, with the primary concern being misrepresentation or change of context, meaning or perception.

- **Verification:** People primarily wanted to know if the image had been verified or not, and they were also interested in the kind of checks that were done to verify it. Ideally, they wanted these checks to be conducted by an independent body and wanted the ability to learn more about the verification process if desired.

- **Generative AI:** They primarily wanted to know whether or not generative AI was used in the creation of a particular image. People's level of digital literacy and use of language was a key factor in users' perceptions of this. Usage of generative AI was not viewed as acceptable in a news context, other than reporting on the use of AI. They also wanted this information to be flagged clearly to them.

How to show provenance information

We have been researching what design patterns people find most usable, helpful and engaging. C2PA recommend four levels of progressive disclosure, each with more detail:

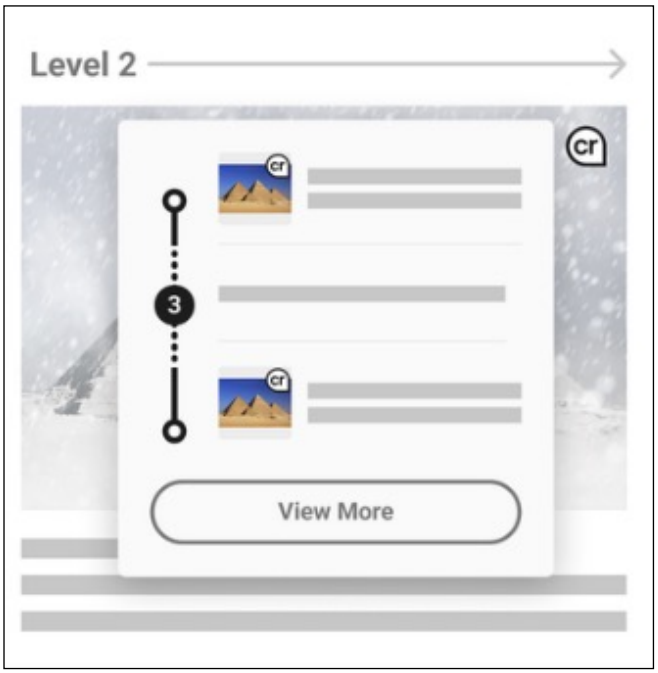


Figure 1: level 2 display

- the first indicating the information is available
- the second with a summary of the key information users want
- and additional levels with considerably more detail.

A key challenge with disclosing provenance information is that there are large amounts of data which if shown all at once would be overwhelming and time consuming to have to read. People also have diverse needs; some are less interested in the details or unfamiliar with content production processes, while others are interested in interrogating the content further.

Medium amount of information – show all that is essential but more can be available.	Stay on the same page	Should be consistent across device/platform
See the image and text at the same time	Unobstructed image	Clear visual indication if the image used AI / was edited
Preference for drop down but expectation of pop-up	Link to further info.	Wanted bullet points

Table 1: Primary considerations for Visual Preferences

Our research has focused primarily on the first and second levels, where key information is summarised. We have explored how to show information in a way that is objective but concise, while responding to the various user needs.

We also explored potential layouts for presenting users with image provenance information. People wanted the information to be consistently displayed across devices and platforms and to be able to easily scan it to find key points of interest. However, they did acknowledge that this may not be possible given how different organisations may implement the technology on their platforms. From the research, we were able to pull out key visual principles that are important for people when seeing provenance information for images.

It's important to consider how presentation can impact people's perceptions and ensure that it is done in a way that does not influence them but allows them to make their own decisions. For example, people may want a system to flag content that may be misleading, using a traffic light system, however we know this can lead to misunderstanding and misrepresentation of the facts.

In the BBC trial below provenance information relating to BBC Verify content and therefore we had to consider technical restrictions around what was possible to show to people, such as the available provenance data and web/app platform constraints. Our goal was to ensure the overall article experience was not negatively affected while making the provenance information clearly discoverable to people. We explored a number of possibilities within the constraints we had, and through testing these solutions we identified an approach using dropdown lists as the most effective and desired format.

How provenance information impacts trust

We found that surfacing this information through the methods and examples outlined above had a primarily neutral impact on trust, with some indication of instances of increased trust. This is true of users that do and do not use the BBC. Adding provenance information was also shown to increase trust in content published on BBC News for non-BBC users.

We conducted our tests based on three different types of images: editorial, stock image, and user generated content (UGC). Adding provenance information served as an equaliser of trust across the image types, with them all eliciting the same levels of trust. Where no provenance information is shown, UGC had the highest levels, followed by stock image and then editorial images.

There are still knowledge gaps around this space that we are looking to research further. These include education, the impact of the lack of consistency across sources, accessibility considerations, how to flag the use of AI and the impact on behaviour. Drawing on these learnings, we have built a lightweight way for journalists to sign and share the content credentials of key imagery we publish, giving audiences an insight into not just what we know, but how we know it. We have launched a live trial on the BBC News website to get a better and more representative understanding of whether this helps audiences increase their trust in news.

BBC Trial

Content credentials can help audiences discover how a piece of content was made and who made it. They help audiences assess whether a piece of content was made by the organisation who claims that they published it, and if they trust that organisation they can trust the content. As more organisations use them, a larger proportion of authentic media will have content credentials included, and internet users will increasingly be able to judge the authenticity of what they are seeing or hearing for themselves.

It will take time for credentials to be widespread and visible across third party platforms, so we now describe a trial carried out in collaboration with BBC News. For the trial we looked at where we could use content credentials on our own platforms to add value to our content, by providing reassurance, transparency and more information about how we know what we know.

At the BBC, we strive to report the facts accurately, holding ourselves to the highest journalistic standards. BBC journalists do rigorous manual verification of the media we publish to ensure it is an authentic depiction of events. We do so by checking the content against other sources, examining the metadata, comparing locations, weather, and searching for other instances of the material online. We do everything we can to ensure we're not furthering the spread of disinformation. And where we do find fakes, we call them out. All this takes time, but we would rather be right than be the first to a story. We now have a dedicated team within the newsroom, BBC Verify, using a range of forensic investigative skills and open source intelligence (OSINT) capabilities to validate incoming material.

"At BBC News we know that trust is earned. When our audiences know not just what we know, but how we know it, they feel they can trust our journalism even more." **– Deborah Turness, CEO of BBC News**

We also recognise that it is not enough to ask audiences to trust us at face value and that we need to show audiences how we sort fact from fiction and do so securely, binding the context to the content so any tampering will be evident. This is why we have combined the cryptographic signing mechanism of C2PA with the details of manual verification from our journalists for images that come into the newsroom from sources lacking C2PA-enabled devices.

This not only means that audiences can make up their own minds about whether a piece of content is trustworthy, based on how much they trust the BBC, but also means that the provenance can be traced back to the BBC if the material is shared elsewhere.

"In a world of deep fakes, disinformation and distortion, this transparency is more important than ever."
– Deborah Turness, CEO of BBC News

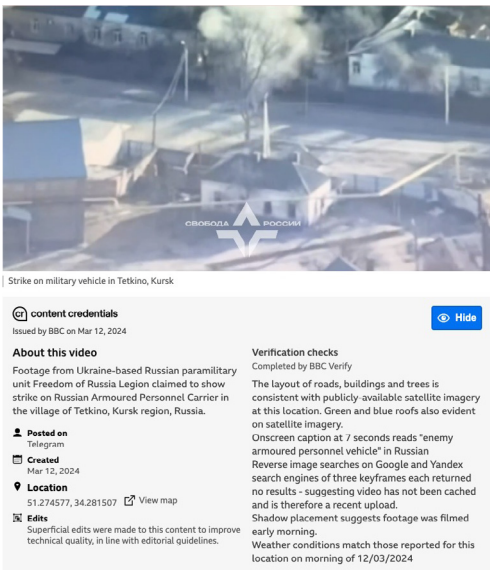
Approach and Methodology

For the trial we worked with BBC Verify to integrate their work, carefully compiling their in-depth verification of the authenticity of an image or video into a claim that could be associated with the published media.

We focused on user generated content (UGC) because, from the research above commissioned by the BBC, it is the media people seem to trust the most, ahead of professionally captured media. It is also the most likely to arrive from an unverified source, and so requires manual verification before we can publish it as a piece of news. This process creates a proxy for the rich provenance data that could be provided directly by cameras and other devices in the future, and so we sought to use it to demonstrate the value of transparently surfacing context through provenance data.

To deliver content credential for BBC published material we built a C2PA-signing plugin that extends the BBC's internal UGC management tool that the BBC Verify team use to collect, collate, prepare, review, and approve provenance and authenticity metadata for UGC assets. The trial system was designed to add almost no additional work for journalists, integrating with existing systems and providing interfaces to give them editorial control.

Our design drew on the research done with BBC audiences to create an element that can be embedded into an article, like any other piece of media, showing the provenance data. For the trial we gathered feedback from users who interact with it about whether it affects the trust they have in it.



Strike on military vehicle in Tetkino, Kursk

Figure 2: Article with Content Credentials on bbc.co.uk/news

The plugin maps the manual verification data to C2PA assertions. As we are focusing on provenance claims, rather than inventing our own metadata schema, we adopted an existing review schema from schema.org - https://schema.org/ClaimReview. While we do not use every field in that data type, we do map as much of our data as fits that definition as possible. C2PA is a flexible standard which permits this form of flexible extendibility as a core feature.

The certificate we use to identify ourselves is issued by a Certificate Authority that is trusted by both the Content Credentials Verify and the Origin Verify tool – https://truepic.com/certificate-authority/ (see Certificate section). Should we elect to sign media that already contains existing C2PA credentials – for example, inserted and signed by a Leica or Sony camera body – the signing process can append our assertions to the manifest to achieve the "chaining of trust" as was earlier discussed.

To surface this metadata to audiences, we built a Content Credentials component that journalists insert into articles on the BBC News website.

We surface both the media asset, and its attendant provenance and verification metadata. The component features a hyperlink that opens the asset in the Content Credentials website - a page that features a C2PA "decoder". Once decoded, the embedded C2PA metadata found in the file is validated. This site also checks that the metadata and image shown have not been tampered with or altered, as the C2PA bundle is signed against a hash of the media at the point of

publishing. Any changes to metadata or content will immediately invalidate the integrity of the file, both on the Content Credentials website and on our own platform, and this is visible to the audience.

Project Reynir: The Sandbox for Implementing C2PA

Project Reynir aims to secure an 80% implementation of C2PA in Norwegian newsrooms by the end of 2026. This will mean that Norway could be the first country with wide-scale implementation of the technology.

A pressing and practical issue for any industry seeking to implement C2PA is the extent to which companies spend a lot of time and money reproducing a set of processes across each individual organisation. In mitigating this challenge, Project Reynir is taking a holistic and collaborative approach, aiming to unite the efforts of the media industry in Norway. The main objective for the project is to establish an ecosystem of collaboration and the sharing of knowledge and experiences across an industry with strong individual actors.

Value Creation

Project Reynir is creating value through:

- **Accessibility:** Making the technology accessible as a common good that accelerates innovation and provides an industrial boost for the media industry.
- **Democratisation:** Granting small actors like the 200 local newspapers in Norway access to the same technology and industry standards as the world's largest newsrooms
- **Expertise:** Securing expert knowledge and upskilling in the industry through the sharing of knowledge, technology and information.
- **Collaboration:** Contributing to the creation and development of value chains from academia, research and the private sector to end-users.
- **Innovation testing:** Providing a large-scale sandbox for proof-of-concept across a larger media ecosystem.

Collaboration and Partnerships

These objectives are being addressed through collaborative efforts that include developing shared standards, sharing best practices and creating a common platform for collaboration.

Key partners include editorial companies such as Schibsted, TV 2 Norway and NRK; technology firms like Vizrt and Factiverse; and academic collaborators such as the University of Agder and Teklab at the University of Bergen. Other partners include the Norwegian news agency NTB and the fact-checking organization Faktisk, to mention just a few.

The first media tech companies in Project Reynir who are in the pipeline to support C2PA are Vizrt, CuttingRoom, Mimir and Wolftech. Vizrt is a world-leading provider of tools for visual storytelling in live broadcasts. They are seeking to implement C2PA for the purposes of live video, as well as for their MAM-system. CuttingRoom provides a cloud-based editing suite as well as a recording tool for journalists. They will implement C2PA for use in video production and editing. Mimir is a cloud-based media asset management system that are going to implement C2PA as a part of their pipeline in order to serve the media companies that use the service. Wolftech provides a workflow tool for journalists and will provide a seamless C2PA pipeline for the users.

There is a risk that the different companies will encounter a varying degree of complexity and face different challenges when implementing the technology. Some will support the viewing of content credentials for end users. Some will sign the content themselves, some will sign on behalf of their media company costumers. Other companies deliver products that are involved in the production and processing of media assets. Others still are looking into new types of media assets and how the technology could be integrated here. By being part of the Project Reynir ecosystem, companies are able to build on each other's work, sharing knowledge obtained from working with the technology and assisting each other in the process.

Since the pre-study of Project Reynir kicked-off in August 2023, we have successfully managed to create substantial interest in the media industry, as well as built an ecosystem consisting of a variety of stakeholders including vendors, news media and academic institutions. We have aided in facilitating the technological implementation through connecting technology partners with the current CA, Truepic, to enable them to be first movers on implementing the technology. Our academic partners have already started conducting research projects into C2PA-relevant questions, as well as the application of the technology in a Norwegian context in collaboration with the newsrooms. Furthermore, after gathering all stakeholders and interested parties to the Project Reynir summit, we successfully managed to formalize the collaboration workflow and established a project structure for the road ahead.

In considering the enablement of wide adoption of C2PA in national market, we think that it has been an advantage that a neutral party, namely Media Cluster Norway, has facilitated the collaboration. Through Project Reynir, the entire industry has come together to solve a huge challenge and agree on a common way forward to implement Content Credentials through the

chain. This approach could very well be reproduced and employed in other areas, be they countries, regions or particular industries.

Managing Certificates

In order to adopt Content Credentials, organisations need to obtain a certificate to sign their content. The International Press Telecommunications Council, in conjunction with Project Origin has established a working group to create and manage a C2PA-compatible list of verified news publishers. These publishers will be recognised by the system through a corresponding identification program.

The group has created the "Origin Verified Publisher" logo to convey the fact that content has been signed by a certificate granted to a publisher that has been verified according to the Project Origin process.

Origin Verified Publisher Certificates will ensure that the identity of established news organisations are protected from imposters. The certificates confirm organisational identity and do not make any judgement on editorial position. Liaison agreements with other groups in the media ecosystem will be used to accelerate the distribution of certificates.

The initial trust list uses Truepic as a certificate authority, with the BBC and CBC/Radio-Canada as trial participants. In the early phases of trust list roll-out, the IPTC Media Provenance Committee will be designing the policies required to issue a certificate, as well as slowly expanding the trial to include more organisations. These are likely to be drawn from IPTC members in the early trial phase, with the final intention being that it is possible for any organisation (regardless of IPTC membership status) to request a certificate.

Conclusion

Media Provenance is being trialled in BBC and Norway and the adoption is increasing within both news organisations, vendors and platforms, including creators work using of generative AI. Initial work focuses on the point of publication (signing media and metadata from a publisher), as there is not yet enough adoption for realistic "glass to glass" (e.g. signing media from the point of capture through to credential display on a user device) workflows. This is expected to change rapidly as the ecosystem evolves.

Early trials are showing success in the development of trust relationships and helping users understand which media to trust, and these will be explored in more detail as trials continue. However, challenges remain in helping users understand the signals from C2PA provenance and there will need to be a concerted effort to explain to the concepts to general users.

Figure 3: Origin Verified Publisher Logo

The Trust list, a critical part of the ecosystem, is at an early stage but is getting ready to expand to C2PA adopters in news. Media Provenance has the potential to make significant impact on the challenge increasing trust in news and the security of other media online.

References

1. Newman, N. 2023. Digital News Report. Reuters Institute. 2023
2. Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, Yunchao Wei, 2023. Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection. ArXiv. December 2023
3. C2PA Specifications, 2024. C2PA website, https://c2pa.org/specifications/specifications/2.0/index.html 2024
4. Aythora J., Burke R., Chamayou A., Clebsch S., Costa M., Earnshaw N., Ellis L., England P., Fournet C., Gaylor M., Halford C., Horvitz E., Jenks A., Kane K., Lavallee M., Lowenstein S., MacCormack B., Malvar H., O'Brien S., Parnall J., Shamis A., Sharma I., Stokes J., Wenker S., Zaman A.. 2020. Multi-stakeholder Media Provenance Management to Counter Synthetic Media Risks in News Publishing. International Broadcasting Convention. September 2020.
5. Monday L, Strappelli L, Does provenance build trust?, https://www.bbc.co.uk/rdnewslabs/news/does-provenance-build-trust June 2024
6. Fighting Misinformation with Authenticated C2PA Provenance Metadata, 2023. Earnshaw N., Dupras J., MacCormack B., NAB Broadcast Engineering and InformationTechnology Conference. April 2023

Acknowledgements

The authors would like to thank their colleagues for their contributions to the work supporting this paper. We would like to thank BBC Verify and BBC Product for enabling us to implement the trial on the BBC News website. We would also like to thank Truepic for the Certificate Authority, Microsoft for all their support in development, CBC / Radio-Canada, for joining us in launching the Origin Trust List and the Content Authenticity Initiative, for their open-source C2PA library.

They would also like to thank the International Broadcasting Convention for permission to publish this paper.

Interoperable Provenance Authentication of Broadcast Media using Open Standards-Based Metadata, Watermarking and Cryptography

John C. Simmons, Joseph M. Winograd
Verance Corporation, USA

Abstract
The spread of false and misleading information is receiving significant attention from legislative and regulatory bodies. Consumers place trust in specific sources of information, so a scalable, interoperable method for determining the provenance and authenticity of information is needed. In this paper we analyse the posting of broadcast news content to a social media platform, the role of open standards, the interplay of cryptographic metadata and watermarks when validating provenance, and likely success and failure scenarios. We conclude that the open standards for cryptographically authenticated metadata developed by the Coalition for Provenance and Authenticity (C2PA) and for audio and video watermarking developed by the Advanced Television Systems Committee (ATSC) are well suited to address broadcast provenance. We suggest methods for using these standards for optimal success.

Introduction

In our interconnected world, information flows ceaselessly, shaping opinions, policies, and societies. Within this digital torrent, false and misleading information often obscures the truth.

False information may take the form of misinformation, spread when well-intentioned individuals share what they found online, neglecting to verify what they found. Or it may be disinformation, false or misleading information intentionally created and spread to deceive.

Both forms of false information are harmful, and both thrive in the global digital ecosystem. Social media platforms amplify their reach, turning falsehoods into viral storms. A rumor, a manipulated video, a fabricated statistic—these can cascade across screens, eroding public discourse.

Provenance and Authenticity

Any attempt to address false information on the web must proceed from an understanding of how people come to place trust in information.

The prevalence of information 'bubbles' demonstrates that people primarily place trust in specific sources of information. If information appears unaltered and from a trusted source, we often consider that information to be factual.

In other words, most of us judge what is factual based on the provenance and authenticity of the information, where provenance refers to the origin, history, and chain of custody of a piece of audio-video content, and authenticity refers to whether the content has been manipulated or altered in a way out of the control of the trusted source of the information.

The Role of Standards

There are two general methods for conveying provenance and authenticity metadata in association with audio-video content. Metadata can be cryptographically bound to the audio-video content, perhaps stored at the audio-video container level. Metadata can also be embedded as a watermark in the audio-video elementary stream.

For practical reasons described in this paper, these two metadata approaches are interdependent. Both cryptographic and watermarking provenance and authenticity methods will be required to provide a reasonable degree of provenance assurance.

A critical issue to address is the impact of adopting proprietary solutions on interoperability and scalability, an issue often encountered. For example, fifteen years ago, Digital Rights Management (DRM) on the web had not yet been standardized. Prior to the ISO/IEC Common Encryption standard [14] playback devices would have to support every major variety of digital rights management software, and there would be as many versions of the audio-video content as there were DRM systems. Had this continued it would have resulted in a combinatorial explosion, an effective barrier to large scale growth of commercial web media. It is no wonder that Netflix was one of the first companies to recognize the value of the common encryption standard.

It is reasonable to expect that the same will hold true for provenance and authenticity. For scalability and interoperability, the cryptographic metadata bound to the audio-video container and the watermark metadata embedded in the audio-video elementary streams must include open standard options.

Scope and Goals of this Paper

A solution for provenance and authenticity for broadcast content distributed on social media platforms is outlined, utilizing metadata, watermarking and cryptographic standards. Our goal is to show how this can be used with broadcast news content while pointing out several important implementation considerations.

Provenance and Authenticity Success Scenarios

A provenance and authenticity use case can encompass multiple scenarios, including success scenarios, where everything goes roughly as intended and various exception scenarios, which lead to undesirable outcomes. All these scenarios should be describable as discrete programmatic steps to uncover the functional requirements for addressing provenance and authenticity in practice.

In preparing this paper, we examined the details of one specific, provenance and authenticity use case – the posting of what appears to be broadcast news content to a social media platform. We were particularly interested in the interplay between provenance validation using tamper-evident cryptographic bindings and metadata retrieval using elementary stream watermarks, with a focus on the constituent 'success' and 'exception' scenarios.

A broadcaster produces content for linear distribution by an affiliate/network/platform operator. This content consists of a series of audio-video programs comprising a single linear broadcast TV channel.

There are a variety of scenarios where some of the broadcaster content finds its way into internet distribution and is uploaded to a social media platform. At a minimum it will then be transcoded into a platform's preferred framerates, resolutions, bitrates, codecs, and container formats. It may also be truncated to meet the platform's maximum size limits.

Verifying Authenticity

Before being posted to a social media platform, broadcast news content may be altered such that there are observable, meaningful differences between what was depicted in the original broadcast and the posted video. This manipulation could be done for artistic, creative, or deceptive reasons, depending on the intention of the editor.

One way to characterize these differences is to ask whether the posted video is an authentic representation of the original, whether it is true to the original, without any judgment as to whether the original itself depicted what transpired in front of the camera lens and microphone.

In this definition, an authentic representation of the broadcaster content may not be bit-wise identical to the original, it may be an unaltered clip from the original, or it may be a transcoding of the original, but it may nonetheless accurately represent what was depicted by the original.

The C2PA standard provides a mechanism for editing the original – e.g., transcoding or clipping – and a means to cryptographically verify the authenticity of the result, but this requires that the tool used to alter the broadcaster content supports the use of this standard. We believe it is highly unlikely that in the near-term social media platforms will reject content that was edited using a tool that does not implement cryptographic metadata standards.

Verifying Provenance

When consuming video in a linear TV receiver, consumers quite reasonably believe that the network/platform operator is accurately identifying the channel and content creator.

Video content on the internet might misrepresent the identity of the original content creator, the identity of who subsequently transcoded the video and what authorized or unauthorized changes were made.

One way to characterize this history is to use the term 'provenance,' meaning, the identifiable source of the content and an accurate history of the content's transformation from that source.

There are cryptographic methods for verifying the provenance of video posted to a social media platform using the C2PA standards, but again we believe it is unlikely in the short-term that social media platforms will reject videos that do not enable the use of these methods to identify the source of the original content.

Canonical Representation of a Media Object

A tamper-evident cryptographic binding to an audio-video media object which contains provenance information can be used to validate the provenance and authenticity of that object. It is surely a successful outcome if the provenance is validated, but what should happen if the provenance and authenticity fail to be verified? What constitutes success in this scenario?

There are multiple scenarios where the content may have been innocently modified by the user when preparing to post to a social media platform, since even a single bit change to the content will invalidate a cryptographic binding. Generating numerous alerts for innocent alterations to the media could lead to 'security alert fatigue,' diminishing user trust in the alert's salience. Doing nothing is also an unattractive option because it leaves users blind to the 'trust signal' conveyed by the presence of a provenance assertion.

Should the cryptographic verification of the provenance and authenticity of a media object fail, a successful outcome is for the social media platform to use an embedded watermark to retrieve the authoritative version – the canonical representation of the media object. This can be done by first using the watermark to retrieve the cryptographic metadata associated with the original media object as distributed and then use that trusted metadata to retrieve the media object's canonical representation.

An Approach to Authentication using Metadata and Watermarking

Architecture

The provenance and authenticity approach in this paper builds on the relationship between the registered content distributors, media objects, their embedded watermarks, associated cryptographic metadata, and the canonical representations of the media object itself. This relationship is shown in Figure 1.

Security Model

If the tamper evident cryptographic metadata associated with a media object is stored as a component of the media object container, it is relatively easy to remove. A durable embedded watermark can enable cryptographic metadata to be brought back into association with the media object.

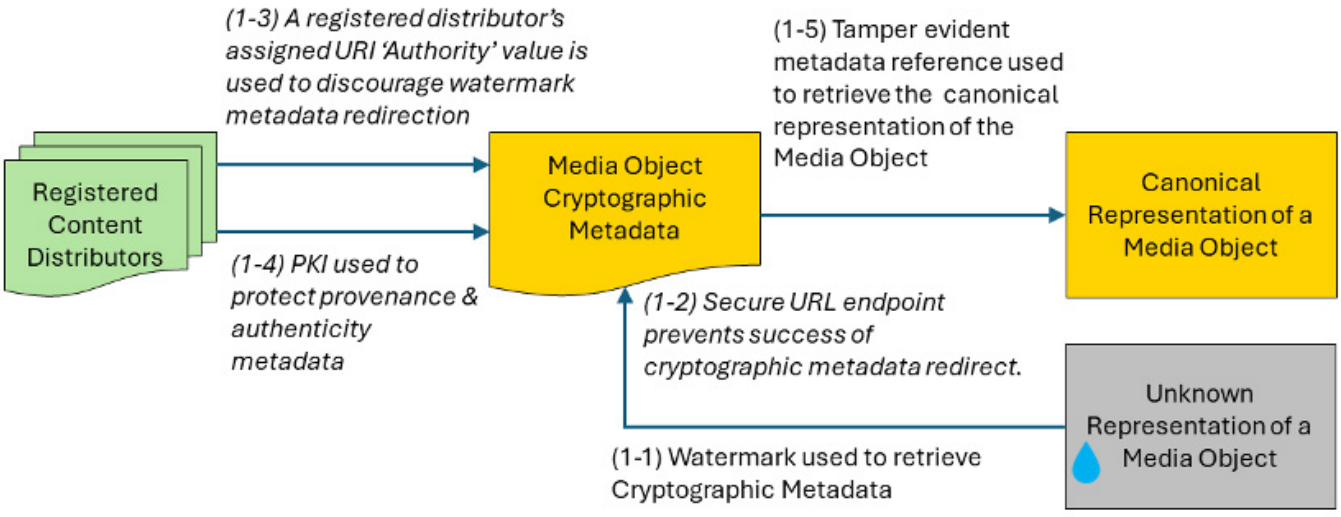


Figure 1: Metadata and Watermarking Architecture

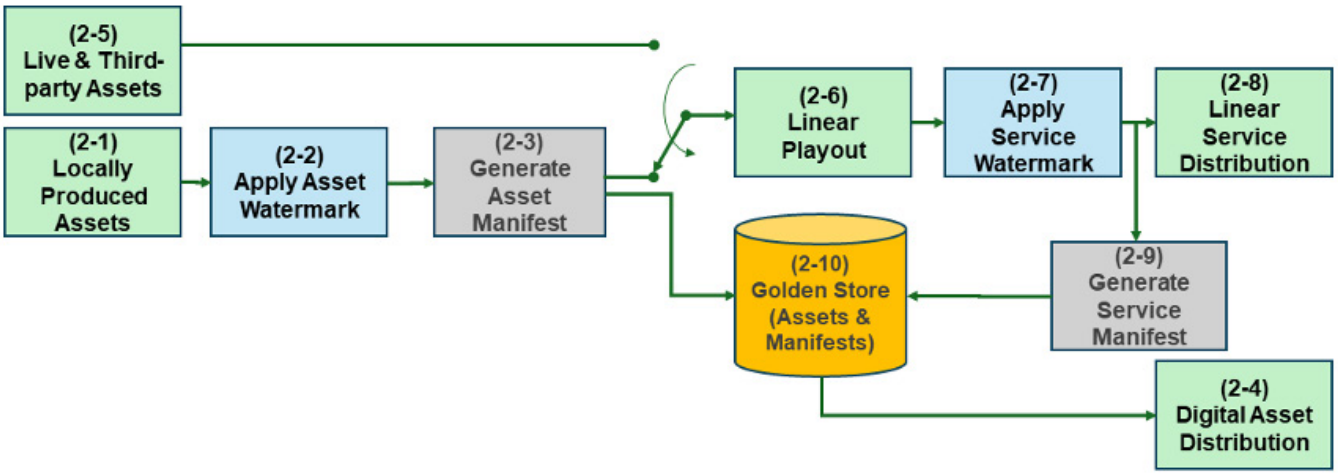


Figure 2: Service and Asset Watermarks

Watermark security is typically maintained by making the watermark difficult to remove or alter by keeping the watermark technology secret. This approach works against availability and interoperability by demanding hardened implementations and strict access controls. It can also provide only weak security assurances because its secrecy impedes comprehensive security assessment. Because recorded broadcast content has a long lifespan on the internet, the security of 'closed' watermarks requires successful long-term protection of the associated secrets. And furthermore, recent advances in attacks on watermarking have demonstrated that advances in artificial intelligence render even robust, secret watermarks automatically removable [10], further diminishing their potential advantages.

This motivates a security approach that does not treat the watermark as a root of trust. Instead, we assume that they are durable, i.e. that they survive content processing that causes traditional metadata formats to be lost, but that they are otherwise as mutable as traditional metadata and can be modified or removed by any intermediary. Like traditional metadata, data conveyed via watermarking is treated as untrusted and must be validated using cryptographic methods.

This same approach was advocated by England et al. [12] in their foundational work on media provenance authentication. That work, however, assumed the presence of a signature in the watermark payload. We view that signature as unnecessary and assume that the watermark carries only a URL and media timeline. The root of trust is a manifest that has been retrieved using the watermark and cryptographically validated using an appropriate trust list.

Watermarking Audio-Video Content

Our success scenario demands a path to validated content regardless of distribution source, which for broadcasters must encompass both linear and on-demand delivery. To achieve this, it must be possible to apply watermarking in asset-based digital publishing as well as within the live production chain.

Figure 2 illustrates an exemplary production flow in which watermarks are applied to enable provenance across multiple distribution paths, with asset watermarks applied to pre-recorded assets and service watermarking continuously applied to a linear playout stream.

The broadcaster may produce content to be published on their website (2-1). They would apply an asset watermark (2-2), generate cryptographic metadata or a 'manifest' for that content (2-3), store the asset (2-10) and distribute the asset to their website (2-4).

The broadcaster may also want to take live and third-party assets (2-5) and prepare them for linear playout (2-6). They would apply a service watermark with a time-varying component (2-7), distribute the content (2-8) and periodically generate cryptographic metadata or 'manifest' information for that broadcast (2-9).

A watermark can be used to retrieve the associated, static cryptographic metadata and canonical content. And the time-varying service watermarks can be used to retrieve the associated, time-varying cryptographic metadata and canonical content (2-10).

Validating Content Authenticity using Watermarking

Figure 3 and Figure 4 summarize how watermarks can be integrated into the content validation process.

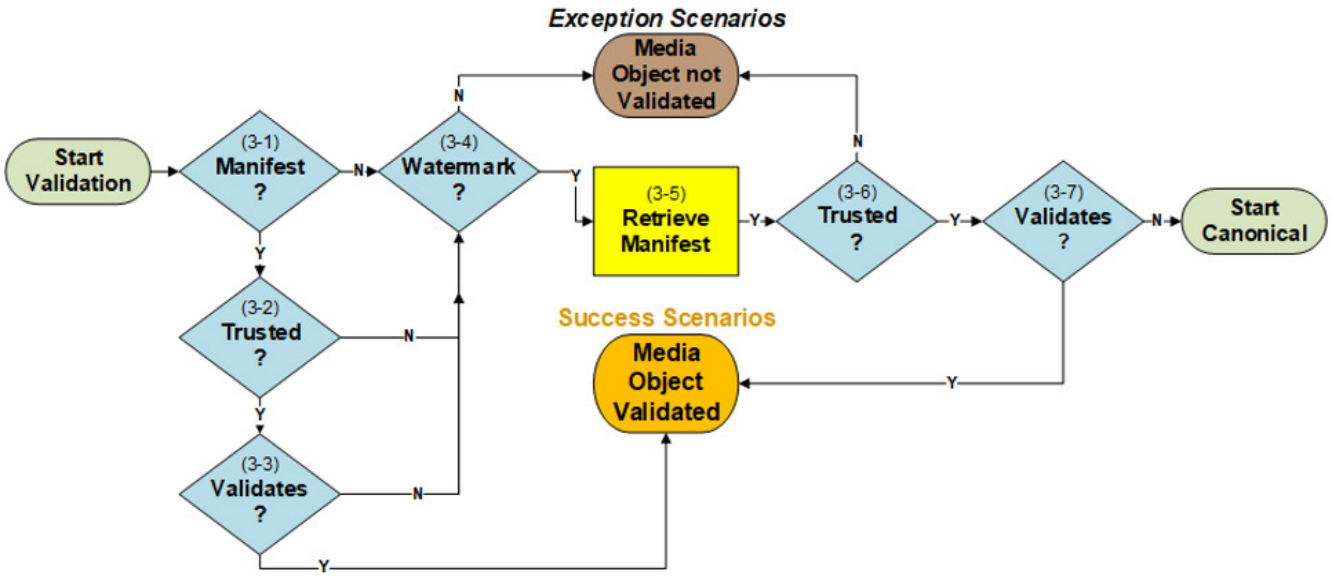


Figure 3: Media Object Validation Scenarios

Media validation uses the cryptographic metadata which may be stored in the media object, distributed with the media object and/or retrievable from the cloud. This object is referred to as a 'manifest' in the C2PA standard [4].

Media object validation scenarios

If the content can be validated by a contained or retrieved manifest, then a successful outcome does not require utilizing canonical content.

If the media contains a manifest (3-1), the manifest corresponds to a registered distributor (3-2), and the manifest validates the media object (3-3), validation is achieved without reference to a watermark. We would view this as a success scenario.

Otherwise, if the media object does not contain a watermark (3-4), the media remains unvalidated, this is an exception scenario.

If the media does contain a watermark (3-4) then the manifest is retrieved from the manifest cloud store (3-5). If this retrieval fails, for example if the URI Authority Field provided in the watermark does not correspond to a registered broadcaster, the media object is not validated; an exception scenario.

If the retrieved manifest's digital signature does not correspond to an approved broadcaster (3-6), then the media object cannot be validated. Another exception scenario.

Otherwise, if the manifest's digital signature is trustworthy (3-6) and the manifest validates the content (3-7), validation is achieved by using the watermark. A success scenario.

Media object canonical representation scenarios

If a retrieved manifest (3-5) is trustworthy (3-6) but it does not validate the content (3-7), then it is the view of this paper that the only success scenarios involve canonical processing.

The decision to perform canonical processing (4-8) can be made by the user posting the content or by the platform supporting the validation logic, depending on the policy being adhered to by the social media platform.

If the decision is to perform a canonical process (4-9), the validator retrieves the canonical content (4-10).

The previously retrieved manifest (3-5) should always validate the canonical content (4-11). If it does not, it is an error and an exception scenario.

We view validation of the retrieved canonical content as optional because its retrieval location has been established as trusted through validation of the manifest that contains it (3-6) (3-7).

Media object canonical processing

The availability of a canonical version of the media object presents the social media platform with additional success scenario opportunities, including one or more of the following:

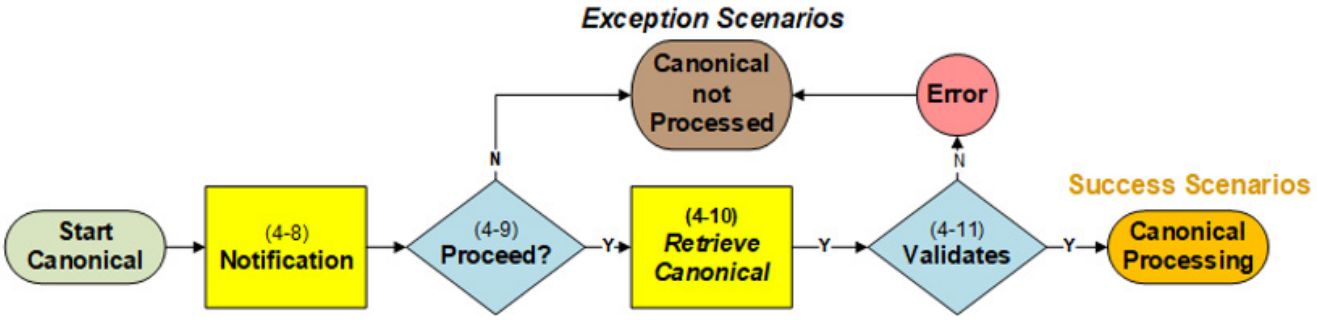


Figure 4: Media Object Canonical Processing Scenarios

- Posting the uploaded content together with the retrieved asset, or a link to it;
- Providing the uploader with a choice between which version of the content should be posted and posting that version with an appropriate label;
- Performing an automated comparison of the uploaded and reference asset to determine the nature and amount of difference between the two;
- Automatically replacing the uploaded content with the valid asset content; or
- Forwarding the uploaded content and the retrieved asset to an internal content moderation process.

This Approach Applied to Live Broadcast

Low Latency Considerations

Real-time broadcast and live streaming, often referred to as 'glass to glass' is a process where content is captured through a camera lens and transmitted to a viewer's screen with minimal delay. Although it is a real-time transmission, it always involves some degree of delay or latency, incidental and/or intentional.

Live scenarios may be categorized by the degree of latency required. This depends on the content's nature and the desired viewer experience. Real-time, low latency live is essential for live sports and breaking news, where timely viewing is important. Higher latency can be introduced for any number of reasons. For example, content is often recorded, edited, or processed before broadcast.

Using digital signatures to protect provenance metadata for 'glass to glass' real-time live streaming scenarios is technically challenging. The primary difficulty is that performing a digital signing operation on a Content Delivery Network (CDN) edge server is not adequately secure and performing that operation in a Hardware Security Module (HSM) is unlikely to achieve the low latency desired.

However, any live scenario where the content is captured downstream, edited, and subsequently posted will introduce an inherent latency sufficient to allow the use of an HSM for provenance metadata protection.

Live Broadcast News Content Posted to a Social Media Platform

Suppose that a 30-minute evening news program is broadcast. The live broadcast is captured and recorded on a device downstream of an HDMI port. A 20-second clip of the news broadcast is created as an MP4 file and posted to a social media platform.

No manifest can be present with the content since only the elementary stream will make its way beyond the HDMI port. The social media platform can examine the posted video for a watermark, but what would be the success scenario?

The fragmented MP4 broadcast replica

We believe that the following approach can provide a reasonable degree of provenance and authenticity assurance for broadcast news content posted to social media platforms.

The live news program is broadcast with a watermark consisting of a constant service/asset identifier component and a time-varying index code (Figure 5). This watermark approach is in use today for delivering metadata for interactive television services [1][2][3] and can readily support the retrieval of provenance metadata without the need to modify the watermark itself.

The broadcaster or the network/platform operator on behalf of the broadcaster produces a secure transcoding of specified portions of the live linear broadcast into a fragmented MP4 format – an 'fMP4 Replica' of the portion of the linear broadcast for which provenance and authenticity is to be established.

Watermark Constant	Service/Asset Identifier									
	Data Hash Segment #M					Data Hash Segment #N				
	Cryptographic Metadata, C2PA Manifest #M					Cryptographic Metadata, C2PA Manifest #N				
	fMP4 Replica, range #M					fMP4 Replica, range #N				
	begin time (M)	+1 Δt	+2 Δt	...	end time (M)	begin time (N)	+1 Δt	+2 Δt	...	end time (M)
Watermark Time code	BINX(M)	+1	+2	...	EINX(M)	BINX(N)	+1	+2	...	EINX(N)

Figure 5: Data Hash Segments, Cryptographic Metadata and Watermarks

Periodically a C2PA manifest is produced for this fMP4 Replica. In this design the portion of the Replica that each manifest corresponds to is defined as the Data Hash Segment (DHS). The real-time duration of a DHS defines a minimum lag time behind the linear live edge for the availability of DHS Replica Manifests.

The Replica itself consists of a sequence of fragmented MP4 segments or chunks for each track, adequate to cover the length of the Data Hash Segment. Each segment or chunk includes auxiliary 'c2pa' boxes [4] which can be used by a C2PA validator to validate any portion of the DHS Replica, as described below.

Apart from the addition of c2pa-specific ISOBMFF boxes, the Replica format is identical to the format in common use for adaptive bitrate streaming, the Common Media Application Format or CMAF [15].

The fragmented MP4 replica C2PA manifest

The fMP4 Replica Manifest is constructed in the exact same way as a C2PA Manifest for audio-video streaming (section 9.2.3 of [4]).

Before the manifest is generated, a DHS initialization segment is produced for the content stored in the DHS Replica. The cryptographic metadata stored in this initialization segment is identical to that specified by C2PA for adaptive bitrate delivery (section 9.2.3 of [4]).

The c2pa-specific box in each track's initialization segment will contain the C2PA manifest, which, as is the case for adaptive delivery, must be identical across tracks. The Manifest's c2pa.bmff.hash assertion will contain CBOR with an array of Merkle rows, one per track.

In the C2PA specification for adaptive delivery provenance validation, the Merkle tree associated with the entire video stream enables piecewise validation of individual fragment components of the stream without access to the entire stream [4]. The same mechanism enables piecewise validation of arbitrary portions of the DHS using a single DHS manifest.

Producing a canonical live recording

If the watermark in the live recording is time-varying, it can be used create a canonical live recording, as shown in Figure 6.

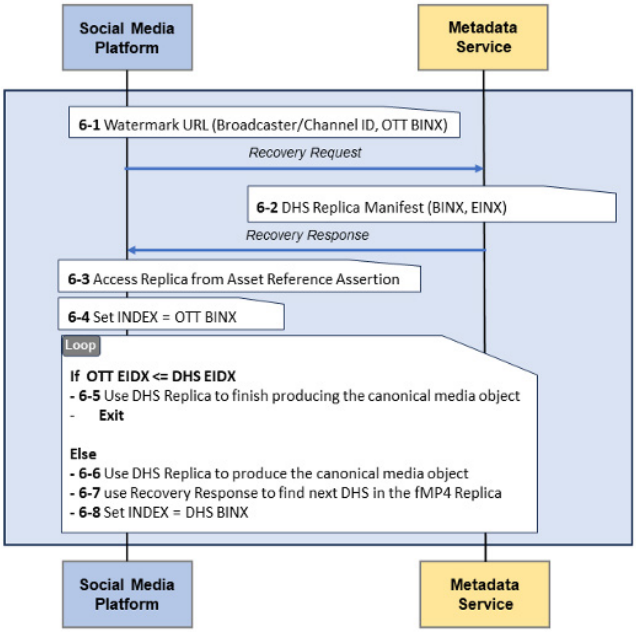


Figure 6: Producing a Canonical Media Object

The time-varying watermark is used to derive the manifest recovery URL. OTT BINX and OTT EINX are the time indices corresponding to the start and end of the posted video, respectively.

A recovery request (6-1) is sent. The DHS Manifest is provided in a recovery response (6-2).

This recovery request response is identical to the method used today for interactive television. The only change is in the payload of the response from the provenance-authenticity server.

The DHS corresponding to the manifest is accessed from the Asset Reference Assertion in the DHS Manifest (6-3).

The fMP4 Replica is used as an authenticated mezzanine format, to produce a canonical MP4 representation of the posted live content. Any portion of the Data Hash Segment can be validated with the DHS Manifest.

Beginning with the OTT BINX (6-4), the algorithm walks the Data Hash Segments provided in the Recovery Response (6-7) until the EIDX of the posted live recording is reached (6-5).

This Approach Applied to Web Published Content

Differences without a Distinction

During validation of content posted to a social media platform, even the slightest alteration to the content can cause it to be flagged as inauthentic. There are multiple scenarios where the content may have been innocently modified by a user, making their edits from a provenance perspective a 'difference without a distinction'.

As discussed, we believe the successful outcome for a validation failure to be for the social media platform to recover the original content and use it in one of the ways we outlined. There are cases, however, where this too will result in an unsuccessful outcome.

Clipped Web Published News Content Posted to Social Media

One of the most likely such cases is what we are calling 'the clipped news segment' scenario. Consider the following example.

The broadcaster publishes a 30-minute evening news program to their website. The published video file includes cryptographic metadata, and it is watermarked. A user wishes to share a 20-second clip from that 30-minute program. They download the broadcaster published video file, edit it to produce a 20-second clip, and attempt to post it to a social media platform.

If the editing tool used by the user, removed the manifest, the social media platform can recover the metadata using the watermark, as described above. Regardless, the file will be flagged as inauthentic. And recovering the canonical version of the content will result in a 30-minute post.

Producing the canonical news clip

If the watermark in the clipped news content is time-varying, it is used to derive the manifest recovery URL, a recovery request (1) is sent where BINX is the time index corresponding to the start of the clip. The retrieved DHS Manifest in a recovery response (2) can be used to produce a canonical version of the news clip, as shown in Figure 6, following the same steps as producing a canonical live recording.

Since social media platforms transcode posted video into a multitude of targeted formats, it is likely that they would treat the DHS as a canonical mezzanine format to produce a wide variety of device-targeted formats. Using fragmented MP4 as a mezzanine format is commonly done.[1]

Standards-Based Solutions

A fundamental element of the provenance challenge is that the global internet facilitates content following an unconstrained path from its place of creation to that of presentation. This makes interoperability between the systems that generate and validate authenticity signals essential.

The use of incompatible technologies creates problems at both ends of the system. If different content producers choose to label their content using incompatible authentication technologies, a platform that could receive uploads of content originating from any of these producers will need to support all the various technologies that they use. Conversely, if different platforms choose to validate content using incompatible technologies, a content producer whose media could be uploaded to any of these platforms would need to be able to generate authentication signals using all of the platforms' technologies in their content.

The establishment of open technical standards for provenance authentication presents a promising path towards addressing these problems. In the interest of advancing progress in this area, we have considered two candidate standards applicable to broadcast content – one cryptographic, the other watermarking. For each, we assess their suitability in the context of the broadcast use cases described above and, where necessary, identify improvements.

[1] In addition, the MPEG DASH specification [Annex C.4, 25] provides support to access segments of presentations at a specified media time through the use of an MPD Anchor, using a query parameter 't=' that a client can append to an MPD URL with either an NPT or UTC time. This could be used to access portions of the fMP4 Replica.

The Coalition for Content Provenance and Authenticity (C2PA) Standards

At the January 2019 World Economic Forum in Davos the director of a major news organization asked Eric Horvitz, Chief Scientific Officer at Microsoft, what could be done about deepfake videos. Within a year of that conversation Microsoft had initiated the Authentication of Media via Provenance (AMP) project, the New York Times had launched the News Provenance Project (NPP), and the BBC and CBC together had created the BBC/CBC Provenance Project.

The convening of these parties took place in London, May 2019, resulted in a combined activity - the Origin Project. Also in 2019 Adobe, the New York Times and Twitter founded the Content Authenticity Initiative (CAI), with the stated goal to build a system to provide provenance and history for digital media.

Realizing there was a need for a single, scalable interoperable standard for provenance and authenticity, on February 22, 2021, Microsoft and BBC teamed up with Adobe, Arm, Intel and Truepic to create the Coalition for Content Provenance and Authenticity (C2PA), combining the specification efforts of Adobe, which had focused primarily on still images, and Microsoft, who had focused almost entirely on video assets.

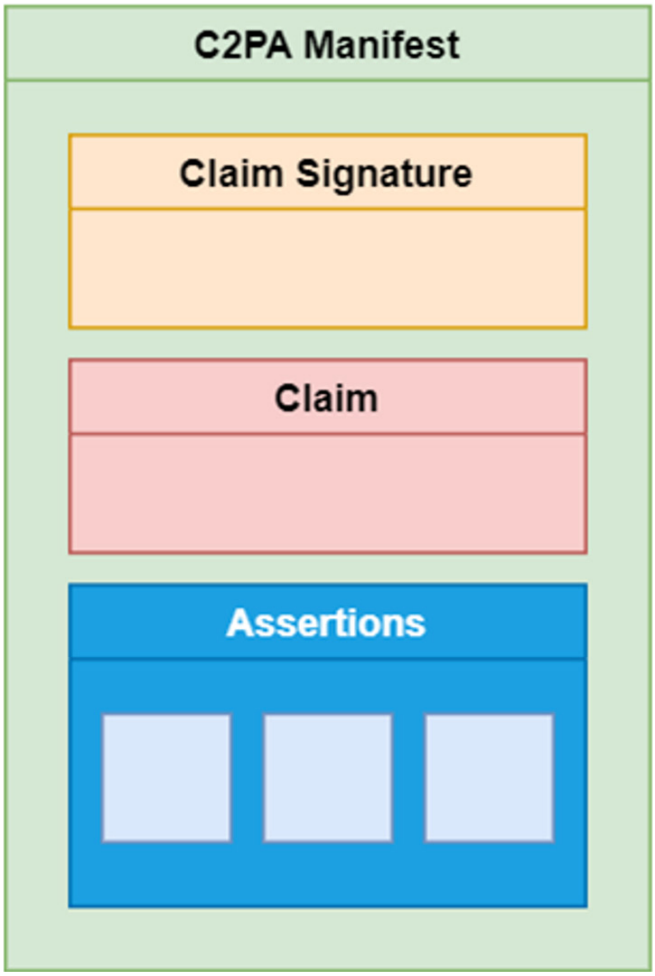


Figure 7: C2PA Manifest [4]

Today C2PA includes 120 companies and continues to grow. The C2PA Steering committee consists of Adobe, BBC, Google, Intel, Microsoft, OpenAI, Sony and Truepic. Members include Amazon Web Services, ARM, Canon, Leica, New York Times, Nikon, and NHK.

C2PA Standards Overview

Because Adobe and Microsoft had independently made considerable progress developing specifications and prototypes before 2021, the C2PA, version 1.0 specification was released to the public in less than one year - January 26, 2022.

At the time of writing, the most recent version is 2.0, released January 2024 [4].

Normative and informative C2PA documentation

There are presently two C2PA technical specifications, 1) 'Content Credentials: C2PA Technical Specification' and 2) 'Attestations in the C2PA Framework' [4].

Released along with these normative specifications are a collection of informative documents, including an explainer, guidance for implementers, user experience guidance, security considerations, harms modelling and guidance for artificial intelligence and machine learning [16].

The following overview introduces a few of the central concepts in the C2PA specifications, skipping over topics not central to this paper. Details can be found in the specifications themselves.

Content credentials overview

C2PA metadata for an asset conveys assertions such as asset metadata, actions performed, thumbnails, and cryptographic bindings to the content. These assertions convey the provenance of the asset. Assertions are combined with additional information to create a claim. The set of assertions referenced by a claim are collected into a logical construct referred to as the assertion store. The claim is digitally signed, creating the claim signature.

Assertions, Claims, and Claim Signatures and some additional information are combined to form the C2PA Manifest. For some formats, the C2PA Manifest may be embedded in the content. (see Figure 7).

For each manifest there is a single assertion store. However, multiple manifests can be associated with an asset, each one representing a specific series of assertions.

Hard and soft bindings

There are two types of bindings supported by C2PA – hard bindings and soft bindings. These are described as hard binding assertions or soft binding assertions in the assertion store.

The hard binding uses a cryptographic hashing algorithm over some or all of the bytes of an asset. It can be used to detect tampering.

The hard binding of an ISO Base Media File Format (ISO BMFF) formatted asset is described in section 9.2.3 of the Content Credentials specification [4]. There is one binding for a monolithic MP4 file where the mdat box is validated as a unit, and a different binding for when the asset is a monolithic MP4 file where the mdat box is validated piecemeal, or when the asset is a fragmented MP4 file. In this paper we reference both the monolithic MP4 and the fragmented MP4 bindings.

Soft bindings may be a perceptual hash computed from the digital content, or a watermark embedded in the digital content. In this paper we describe a soft binding using an open standard watermark.

Manifest store

C2PA data is serialized into a JUMBF-compatible box structure [22]. The outermost box is the C2PA Manifest Store, also known as the Content Credentials. How the Content Credentials, the constituent Manifests and assertion stores utilize the JUMBF-box structure is described in section 11.1 of [4].

Update manifests

Most C2PA manifests are standard manifests, containing exactly one hard binding to the associated asset. To accommodate provenance workflows where additional assertions are provided but the digital content is not changed is done using an Update Manifest (see 11.2.3 of [4]).

Embedding manifests

C2PA specifies how to embed manifests to a wide variety of formats, including JPEG, PNG, SVG, FLAC, MP3, GIF, DNG, TIFF, WAV, BMF, AVI, WebP, RIFF, BMFF as well as fonts (see 11.3 of [4]).

Asset reference assertion

This assertion indicates one or more locations where a copy of the asset may be obtained. The location is expressed as a URI. This paper makes extensive use of this assertion to enable canonical processing of watermarked content which fails provenance and authenticity validation (see 18.15 [4]).

C2PA Standards Suitability

Open standard for provenance and authenticity

C2PA provides an open standard to associate a cryptographic binding to monolithic MP4 files with a single or multiple mdat boxes as well as to fragmented MP4 content used with adaptive streaming. It supports a tamper-evident manifest store so that the transcoding history of a media object can be provenance-ensured. And it provides a mechanism through soft bindings to recover the cryptographic metadata associated with a media object.

Gap analysis

A central premise of this paper is that canonical content processing should occur when the provenance and authenticity of an asset cannot be validated, and a watermark is present. C2PA supports an asset reference assertion, but at the time of writing does not provide normative language on how this reference should be used should validation fail.

The C2PA specification deals principally with an asset which either contains a manifest or, at one time, contained a manifest. This paper suggests that for the case of broadcast content, it is important to define a success scenario for when a watermark is present, but the asset was never published to the web, associated with an asset watermark.

These are implementation issues and would not impact interoperability.

ATSC Watermarking

In 2016, the US-based digital television broadcast standards organization ATSC, published standards for use of watermarking technology [1][2][3] in connection with their development of the ATSC 3.0 ('NextGen TV') system.

These standards provide open specifications for the use of watermarking technology and associated network protocols to deliver arbitrary timed metadata associated with media content to network-connected clients through distribution paths that include media processing (e.g. transcoding) and metadata removal (e.g. HDMI, analog reconversion).

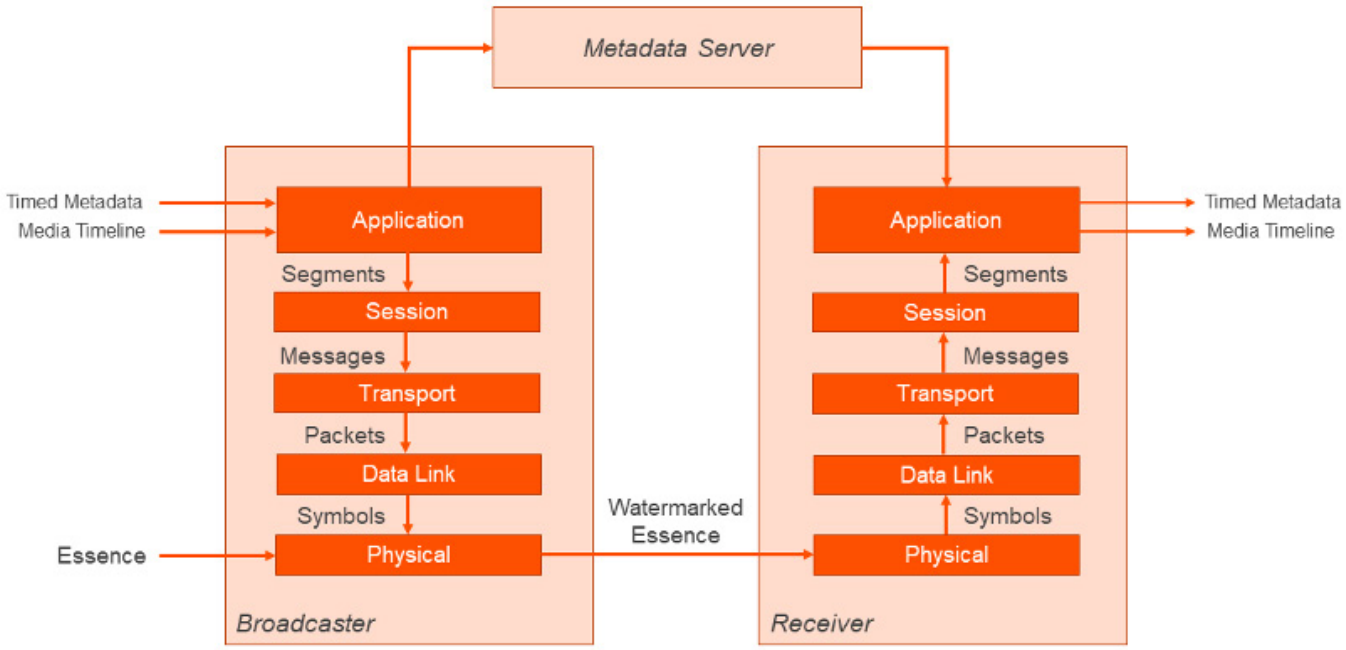


Figure 8: OSI Abstraction Model for ATSC Watermarking

ATSC's primary motivating use case for watermarking is enabling access to NextGen TV interactive (two-way) services for viewers who have purchased compatible TVs but who continue to receive broadcast services from other distribution paths, such as STBs, streaming media players, or ATSC 1.0 transmissions. These standards have been commercially deployed in the United States by multiple broadcasters and television equipment manufacturers.

The technology has also been found to be suitable for use with other broadcast systems. Since 2020, the HbbTV and DVB have published a series of standards that employ ATSC watermarking to enable interactivity and targeted advertising in their platforms. These standards are currently being readied for commercial deployment in Germany.

Description

The ATSC watermark system is specified in the publicly available standards ATSC A/334 [1], A/335 [2], and A/336 [3]. Its function is to deliver arbitrary timed metadata using audio and/or video watermarks embedded into media essence. It supports methods for conveying metadata directly in watermark messages or indirectly, by reference, via carriage of a time-tagged URL that identifies a network resource containing the metadata. The architecture of the ATSC watermark system can be understood using the OSI abstraction model shown in Figure 8.

Taken from bottom to top, essence is the baseband audio or video signal components. The physical layer consists of a stream of raw binary symbols conveyed as audio and video watermarks in the essence. Audio watermarks are conveyed using autocorrelation modulation in the 2.5k-5kHz band. Video watermarks are conveyed using luma modulation in the top two lines of active video. While watermark insertion and detection are performed on baseband (decoded) essence, the system is compatible with a wide range of media processing algorithms applied to watermarked essence, such as low bit-rate coding, that are typically found in digital media distribution. Both audio and video watermark physical layers also permit watermark energy to be adapted to the content to preserve perceptual quality. In formal testing conducted by ATSC technical committees, the audio watermark was demonstrated to be capable of surviving HE-AACv2 encoding at 32kbps stereo without performance loss while preserving perceptual transparency. The video watermark was demonstrated to be capable of surviving AVC encoding at 2.5Mbps 1080p/30.[2]

The data link layer differs for audio and video watermarks, with the audio watermark carrying a sequence of data cells of 1.5 seconds duration, each carrying a 50-bit data packet along with a synchronization header and BCH error protection. The video watermark data link layer conveys a 168-bit data packet in each video frame along with a synchronization header, message framing, and CRC error protection.

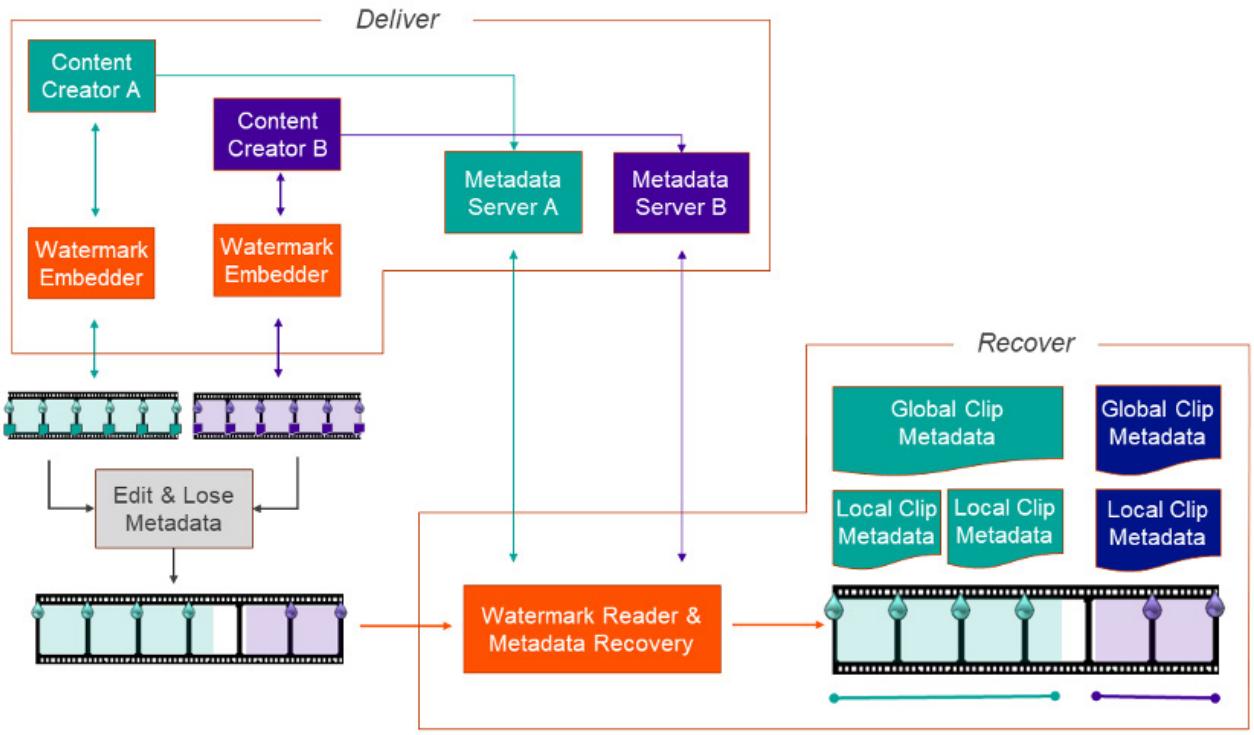


Figure 9: Typical Open ATSC Metadata Retrieval Architecture

The transport layer also differs for audio and video watermarks. For the audio watermark, the transport layer conveys a single packet format, the VP1 payload, that conveys a 'tiny URL' encoded into two fields – a server code that identifies a network server and an interval code that identifies a metadata resource on that server associated with the location in the media content where the watermark is embedded.
For the video watermark, the transport layer can convey metadata by reference using the VP1 payload or directly, using a variety of messages associated with known broadcast metadata types such as stream events, presentation timestamps, and content identifiers. Server codes are assigned values for which ATSC maintains registry authority.

The session layer specifies constraints on the arrangement of watermark messages within assets that enable receivers to perform reliable decisioning regarding the arrangement of watermarked content, including when a particular watermarked asset starts and ends and where on the media timeline a given media sample lies. This is particularly important in contexts where the content arriving at the receiver has been composed from multiple different sources, such as a 'mash-up' of multiple sources or edited version of content.

The VP1 payload is relied on in the session layer to provide the context boundary for a media asset. A watermark media asset (which can be either an individual program item or a continuous program stream) carries audio or video watermark segments comprised of contiguous, watermarked 1.5-second content intervals with a constant server code value and incrementing interval code values.

At the application layer, directly conveyed metadata becomes valid at the location in the content where it is placed. For metadata delivery over a network, a RESTful application layer protocol between the receiver and a metadata server is specified wherein receivers retrieve arbitrary timed metadata using VP1 payload data.

This protocol was adapted by ATSC from an existing 3GPP MBMS protocol [6]. In it, receivers request a metadata resource using a URL constructed from the first VP1 payload that they encounter in a watermark segment. The authority portion of the URL is an internet hostname determined using DNS resolution of the server code within a second-level domain specified by the standard. The path portion of the URL includes the interval code within a predefined template. The response is a multipart/related MIME object [23] containing some protocol-specific metadata objects and some number of additional metadata objects.

[2] Formal perceptual quality evaluation was not performed for the video watermark technology.

The protocol-specific metadata includes a mapping of the VP1 interval code value onto a media timeline, boundaries on the media timeline for which each of the additional metadata objects is valid, and guidance to receivers on where and when updates to the metadata objects should be requested. Receivers request subsequent metadata updates only as necessary, in correspondence with the validity periods of metadata objects that they have received and the time periods of watermarked segments of the media timeline of content that they process.

Any IANA-registered media type [24] can be provided as an additional metadata object. Receivers are expected to route these objects to application-specific handlers and ignore media types that they do not support.

Suitability

The ATSC watermark system provides a number of technical capabilities important to the provenance authentication use case.

Open Architecture

The ATSC watermark system shares the same underlying architecture as the modern internet. The system stack is based on publicly available specifications that enable independent development of interoperable implementations of all components. It uses federated DNS namespace management governed by ATSC, a not-for-profit, internationally recognized standards development organization. And it does not rely on any siloed or proprietary services, freeing broadcasters to host metadata services on servers of their choosing with the ability to transition to new hosts at will. A typical use case illustrating an open metadata retrieval architecture is provided in Figure 9.

Durability

Formal testing overseen by ATSC during the standards development process and subsequent commercial deployments have demonstrated that the audio and video watermarks are reliably recoverable through typical content uses. For audio, this includes common transcoding, downmixing, dynamic range compression, equalization, and analog reconversion processing. For video this includes common transcoding, frame-rate conversion, resolution conversion, image enhancement, and analog conversion processing.

Transparency

Formal testing overseen by ATSC during the standards development process demonstrated perceptual transparency of the audio watermark.[3] While formal evaluation of the video watermark has not been performed, the technology has been found acceptable in commercial broadcast settings and representative content is available for public review [7].

Data Capacity

The approach of carrying time-based references to metadata objects on servers avoids limitation of the size of associated metadata to what can be conveyed by the watermark.

One benefit of this approach is that it allows multiple metadata objects to be associated with content, enabling provenance authentication metadata (e.g. C2PA manifests) to be retrieved using the established protocols for delivering broadcast television signaling (e.g. interactive application data).

Another benefit is that it enables late-binding of metadata to media, which is valuable for live broadcast scenarios, and the ability to revise the metadata associated with content already in distribution simply by posting updated resources to metadata servers, for example to update signatures following a certificate or key revocation.

Precision

The latency, on the content media timeline, to recovery of the first watermark packet with data sufficient to enable recovery of metadata from a network server, is 1.5 seconds (minimum) and 2.25 seconds (average) for the audio watermark, and 1 video frame (minimum) and 83 milliseconds (average) for the video watermark.

After initial metadata recovery, session layer logic enables the media timeline to be tracked within the segment with accuracy of ±2 milliseconds from the audio watermark. Media timeline accuracy for the video watermark depends on whether and how message multiplexing is being used but, in any case, it will be between frame accurate and ±83 milliseconds. Segment (i.e. edit) boundary decisioning can be determined ±1 second.

The ability to precisely recover the media timeline and edit points of a segment is an essential capability for provenance authentication of broadcast content. When broadcast content is redistributed, e.g. on social media platforms, it will of necessity be just a clip of the linear program stream.

[3] Independent testing using ITU-R BS.1116-3 [8] methodology found critical samples of audio watermarked content statistically indistinguishable from original unwatermarked content to trained listeners in reference listening conditions.

The ability of the watermark technology to determine the clip boundaries with precision enables the client to access metadata associated specifically with that region of the media timeline, align manifest data for validation, resolve asset references to corresponding time segments of valid assets, and support any of the use cases described above (e.g. side-by-side review, automated analysis, content substitution, etc.).

Mutability

While ATSC audio and video watermarks are durable, in that they survive content processing that causes traditional metadata formats to be lost, they otherwise have the same mutability properties as traditional metadata, which is that they can be modified or removed as needed.[4]

The mutability characteristic provides a benefit in scenarios where watermarked content is repurposed, such as when an editorial segment includes a news clip that includes syndicated footage, and it is desired that only the provenance information associated with the combined work be referenced by watermarking. In this case, each successive actor can update the watermark to reference a manifest that includes comprehensive provenance information for their output. This approach aligns with the C2PA approach to manifest metadata updates, wherein revisions are consolidated within a unitary store [4] rather than simply appended.

Conclusions

The trust placed by the public in broadcasters as an authoritative source of information provides disinformation agents with a strong incentive to falsely adopt the broadcaster's imprimatur. Given that powerful generative AI models have been open-sourced, reliance on safety mechanisms being implemented by the generative AI systems against these risks cannot be relied on as a meaningful countermeasure.

There is a critical need to identify open standards that enable media creators, distributors, and consumers to be able to authenticate the provenance of content securely and reliably.

In this paper we described a broadcaster authentication approach based on the application of open, interoperable, standards-based technologies. Watermarking standards like that defined by ATSC [1][2][3] already in use by broadcasters enable automatic identification of content, media timeline, and provenance metadata by platforms during content ingest or presentation. And provenance metadata standards like that defined by C2PA [4] provide cryptographically-assured validation of authenticity, as well as access to authoritative versions of content.

In our analysis we examined one particular use case in detail, the posting to a social media platform of what appears to be broadcast content. We found that an architecture based on these open standards provides a workable framework for broadcast authentication.

We also identified methods to minimize the impact of authentication failures by using these standards to access the authoritative version of the content, even when that version was never itself published on the internet. While the challenges addressed in the paper are motivated by study of the broadcast use case, we believe our analysis is applicable to many other classes of audio-video use cases.

This approach holds promise for achieving interoperability and access at scale, but practical progress depends on production and presentation deployment that are mutually dependent. At the time of writing, governmental review of available technical paths is being undertaken in some geographies with some urgency, providing those who take initiative, with an opportunity to exert substantial influence on the path taken.

References

1. ATSC A/334:2024. Audio Watermark Emission. Advanced Television Systems Committee. https://www.atsc.org/atsc-documents/a3342016-audio-watermark-emission/.

2. ATSC A/335:2024. Video Watermark Emission. Advanced Television Systems Committee. https://www.atsc.org/atsc-documents/a3352016-video-watermark-emission/.

3. ATSC A/336:2024. Content Recovery in Redistribution Scenarios. Advanced Television Systems Committee. https://www.atsc.org/atsc-documents/a3362017-content-recovery-redistribution-scenarios/.

4. C2PA Specifications v2.0. 2024. Coalition for Content Provenance and Authenticity. https://c2pa.org/specifications/specifications/2.0/index.html.

5. Parsons, A. 2024. Durable Content Credentials. Content Authenticity Initiative Blog. https://contentauthenticity.org/blog/durable-content-credentials.

[4] See, for example, ATSC A/339: Audio Watermark Modification and Erasure [9].

6. ETSI TS 126 346 v13.3.0 (2016-01). Universal Mobile Telecommunications Systems (UMTS); LTE; Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs (3GPP TS 26.346 version 13.3.0 Release 13). European Telecommunications Standards Institute. http://www.etsi.org/deliver/etsi_ts/126300_126399/126346/13.03.00_60/ts_126346v130300p.pdf.

7. Targeted Advertising Verification and Validation Test Streams – Watermarking. 2023. DVB. https ://dvb.org/specifications/verification-validation/targeted-advertising-watermarking/.

8. Recommendation ITU-R BS.1116-3: Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multi-channel Sound Systems. 2014. ITU Radio Communication Assembly.

9. ATSC A/339:2024. Audio Watermark Modification and Erasure. Advanced Television Systems Committee. https://www.atsc.org/atsc-documents/3392017-atsc-recommended-practice-audio-watermark-modification-erasure/.

10. Zhao, X., et al. Invisible Image Watermarks Are Provably Removable Using Generative AI. Preprint. https://doi.org/10.48550/arXiv.2306.01953.

11. Kirwan, C., et al. Repetition of Computer Security Warnings Results in Differential Repetition Suppression Effects as Revealed With Functional MRI. 2020. Frontiers in Psychology. Vol 11, 2020. https://doi.org/10.3389/fpsyg.2020.528079.

12. England, P. et al. 2021. AMP: Authentication of Media via Provenance. 12th ACM Multimedia Systems Conference. July 2021. https://doi.org/10.48550/arXiv.2001.07886.

13. SMPTE 2064-1:2015. Fingerprint Generation. Society of Motion Picture and Television Engineers. https://ieeexplore.ieee.org/document/7395520.

14. ISO/IEC 23001-7:2006, 'Information technology – MPEG systems technologies – Part 7: Common Encryption in ISO base media file format files', first edition.

15. ISO/IEC 23000-19:2020, 'Information technology – Multimedia application format (MPEG-A) – Part 19: Common media application format (CMAF) for segmented media'.

16. C2PA Explainer, release 1.4, 2023, https://c2pa.org/specifications/specifications/1.3/explainer/Explainer.html

17. C2PA Implementation Guidance, release 1.4, 2023, https://c2pa.org/specifications/specifications/1.3/guidance/Guidance.html

18. C2PA User Experience Guidance for Implementers, release 1.0, 2023, https://c2pa.org/specifications/specifications/1.4/ux/UX_Recommendations.html

19. C2PA Security Considerations, release 1.0, 2023, https://c2pa.org/specifications/specifications/1.0/security/Security_Considerations.html

20. C2PA Harms Modeling, release 1.0, 2023, https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html

21. C2PA Guidance for Artificial Intelligence and Machine Learning, release 1.3, 2023, https://c2pa.org/specifications/specifications/1.3/ai-ml/ai_ml.html

22. ISO/IEC 19566-5:2023, Information technologies, JPEG systems, Part 5: JPEG universal metadata box format (JUMBF)

23. RFC 2387, 1998, The MIME Multipart/Related Content-type, Internet Engineering Task Force. https://datatracker.ietf.org/doc/html/rfc2387.

24. RFC 6838, 2013, Media Type Specifications and Registration Procedures, Internet Engineering Task Force. https://datatracker.ietf.org/doc/html/rfc6838.

25. ISO/IEC ISO 23009-1:2022, Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats

Reducing the Energy Consumption of Terrestrial DAB Transmission

P. D. Kesby[1], R. H. M. Poole[2], A. Wolmarans[2], P. G. Brown[1]
[1]Arqiva, UK, [2]BBC, UK

Abstract
The requirement to develop sustainable broadcast platforms that support global net-zero targets increases year on year. This paper discusses initiatives developed by the BBC and Arqiva to reduce power consumption in their terrestrial DAB broadcast network. In some cases, it is possible to improve efficiency without complete replacement of the transmitters.

Improvements in transmitter technology now allow optimisations to be considered that were not practical previously. Modern transmitter design often takes advantage of Doherty techniques and power supply optimisation. However further reductions may be possible, going beyond previous assumptions, with negligible effect on consumer reception.

This work evaluates the opportunity for a trade-off between modulation quality and energy consumption. The impact of modulation error ratio (MER) on DAB reception has been investigated through theoretical modelling and validated using laboratory testing with channel simulation. The results from these studies show that the impact on reception of reduced MER can be less than previously assumed. Field trials in the operating network, using a transmitter with reduced MER, showed a small to negligible impact on reception. This points to power savings that can be made across DAB networks.

The combination of MER change and related power supply optimisation is expected to reduce the total energy consumption of the existing in-service Doherty transmitters by between 12 and 17%. This approach will be rolled out to 165 present Doherty transmitters during 2024. This technique could be considered for use in future transmitter equipment designs.

Introduction

Many broadcasters have recently developed carbon reduction and net-zero strategies. Reducing energy consumption will help meet net-zero targets with significant financial and carbon reduction savings. The BBC has recently published its carbon reduction strategy and plans to reduce its energy consumption on UK terrestrial platforms by 28% by 2030 [1]. The BBC terrestrial transmission networks in the UK are owned and operated by its transmission partner Arqiva who together with the BBC are investigating a range of energy saving initiatives. However, measures taken must be proportionate to the benefits whilst maintaining minimal impact on the audience. Improving the efficiency of the terrestrial networks can be achieved by deploying new, more efficient transmitters but it can be preferable to optimise existing, already operational equipment. Improving existing equipment avoids the carbon impact and cost of replacement. Finally, efficiency measures must maintain compliance with regulatory requirements and must not compromise the long-term reliability of equipment.

The BBC DAB network comprises more than 400 transmitters, which provide greater than 97% coverage of UK homes [2]. The network was launched in 1995 with an initial 31 transmitters. Several subsequent phases of additional deployment have resulted in a dense medium-power network which covers most of the major roads and homes in the UK. The staggered deployment of equipment over this period has resulted in a range of equipment types and designs. Some early designs exhibit system electrical efficiency of approximately 10 to 20%, whilst more modern designs achieve 40% or better. In this paper the term efficiency refers to system electrical efficiency which is the conducted RF power provided to the antenna divided by the total energy used by the transmitter, including internal cooling.

A DAB OFDM signal without clipping typically exhibits >10 dB peak-to-average power ratio (PAPR). A lack of headroom in the amplifier can cause signal clipping, creating in-band and out-of-band (OOB) intermodulation products (IPs). In-band IPs reduce signal quality, whilst OOB IPs, if not removed, can exceed regulatory emissions limits. Therefore, initial Class-AB amplifier deployments were designed to operate with a large headroom in the amplifier power supply to satisfactorily manage the peaks, resulting in transmitter electrical efficiencies in the region of 10 to 20%. Because of coverage and regulatory requirements (on a site-by-site basis), many transmitters operate below their rated power. The combination of this with limited capabilities to optimise the power supply resulted in relatively poor efficiency. Later strategies were developed which allowed a small reduction in the PAPR in the modulator (or amplifier) combined with power supply optimisation in the amplifier, hence improving the efficiency of the transmitter. A step change in DAB transmitter efficiency was achieved with the introduction of Doherty designs [3], combined with flexible power supply optimisation, which led to efficiencies greater than 35%.

A measure of signal distortion is modulation error ratio (MER), described in more detail later in this paper. More aggressive clipping at the transmitter causes greater signal distortion, and lowers the value of the MER. The statistical nature of the time-domain waveform will result in the peaks occurring infrequently. Therefore, transmitter design now allows for a lowered PAPR, thus accepting a small increase in signal distortion. Modern power supply designs can be adapted to remove the unused headroom (even when the configured RF conducted power is significantly below the manufacturer's rated power), lowering the voltage power supply rails and improving efficiency.

Typically network operators target 25 to 30 dB MER for output signal quality to protect coverage. This study has investigated the relationship between MER and coverage and has confirmed that the MER can be lowered from earlier targets to 20 dB with negligible effect on service coverage. Allowing a lowering of MER enables a further reduction in power supply headroom and improvement in transmitter efficiency.

As described above, the BBC DAB network comprises a range of transmitter technologies jointly consuming approximately 10 GWhr/year. The aim of this study is to reduce the energy consumption of the already-installed Doherty transmitters (consuming 2.4 GWhr/year), with adaptable power supply designs, combined with a reduction of the transmitter MER signal quality to 20 dB. This approach cannot be practically replicated at the remainder of the older transmitter sites but could be implemented in new transmitter designs.

Signal Quality and Modulation ERROR Ratio

The DAB signal comprises 1536 carriers each modulated with quadrature phase-shift keying (QPSK). The notional carrier phase changes each symbol period. A constellation diagram showing the amplitude and phase of all carriers superimposed is shown in Figure 1 below, in the diagram on the left. This is the ideal situation; in practice, noise on the signal causes the constellation points to move away from the ideal.

The 'correct' carrier amplitude and phase is Uc, and the deviation is Ue for one particular carrier. As mentioned earlier, the overall signal quality is quantified by the modulation error ratio, or MER [4]. It is defined as the ratio Uc/Ue averaged over all N carriers, using RMS addition and expressed in dB:

$$MER_{dB} = 20\ log_{10}\sqrt{\left(\frac{1}{N}\right)\sum_{n=0}^{N-1}\left(\frac{U_c}{U_{en}}\right)^2}$$

Here, U_{en} is the error associated with the n^{th} carrier.

It appears that the MER defined by this equation is equivalent to the signal-to-noise ratio (SNR) of the signal. However, DAB uses differential QPSK, where the carrier phase of the previous symbol is used as the reference for the present symbol. Any noise on the previous symbol then gets added to that of the present symbol, so doubling the effective noise:

$$MER_{dB} = SNR_{dB} - 3\ dB$$

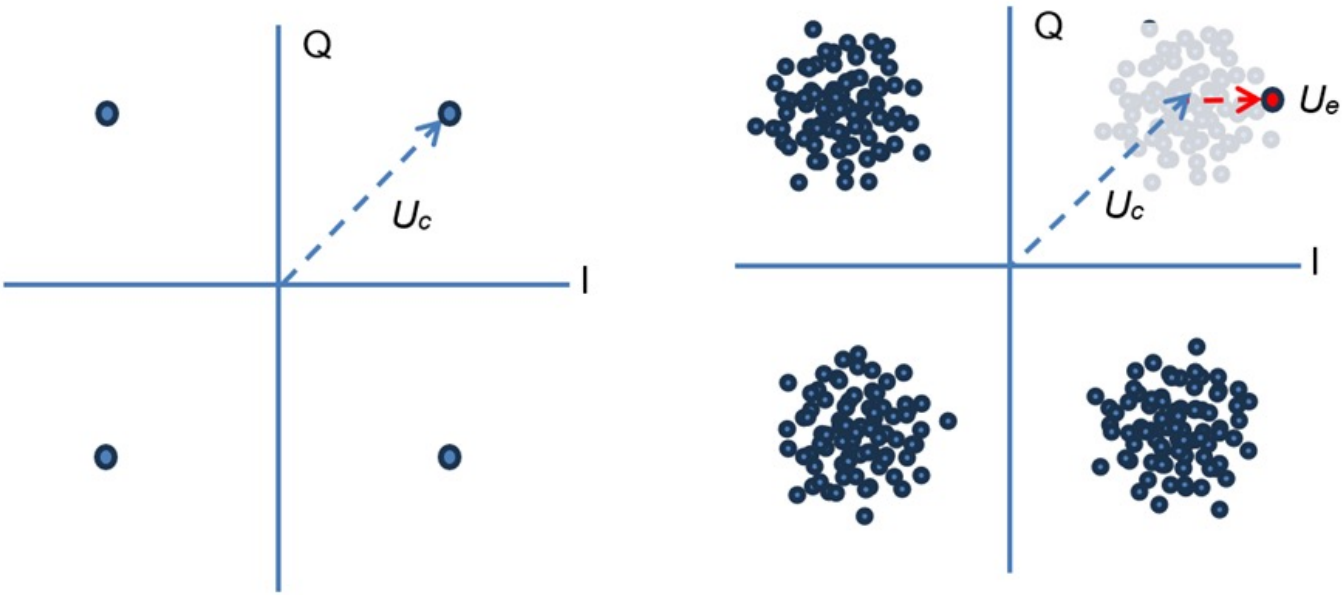


Figure 1: Idealised and practical DAB QPSK constellation

In practice, the receiver has its own noise contribution, as well as that introduced by the transmitter. The impact of the additional transmitter noise contribution can be quantified by calculating the increase in transmitter power which would be required to compensate for that additional noise. The model is shown in Figure 2 below:

The transmitter generates a wanted signal S_T and noise N_T. There is a path loss L between the transmitter and receiver, and we assume that the transmitter power needs to be boosted by a fraction k to overcome N_T. The receiver itself introduces noise N_R, and needs a minimum signal-to-noise ratio $(S/N)_F$ (the failure point) to decode the signal.

In the absence of transmitter noise, we have

$$(S_T/L)/N_R = (S/N)_F \quad (1)$$

at the failure point. (S_T/L) simply represents the transmitter power reaching the receiver. With transmitter noise NT present, and the transmitter power boosted by k,

$$\frac{kS_T/L}{kN_T/L + N_R} = (S/N)_F. \quad (2)$$

Although we have defined k as the increase in transmitter power needed to overcome the noise associated with decreased transmitter MER, we can equally use its reciprocal, 1/k, to characterise

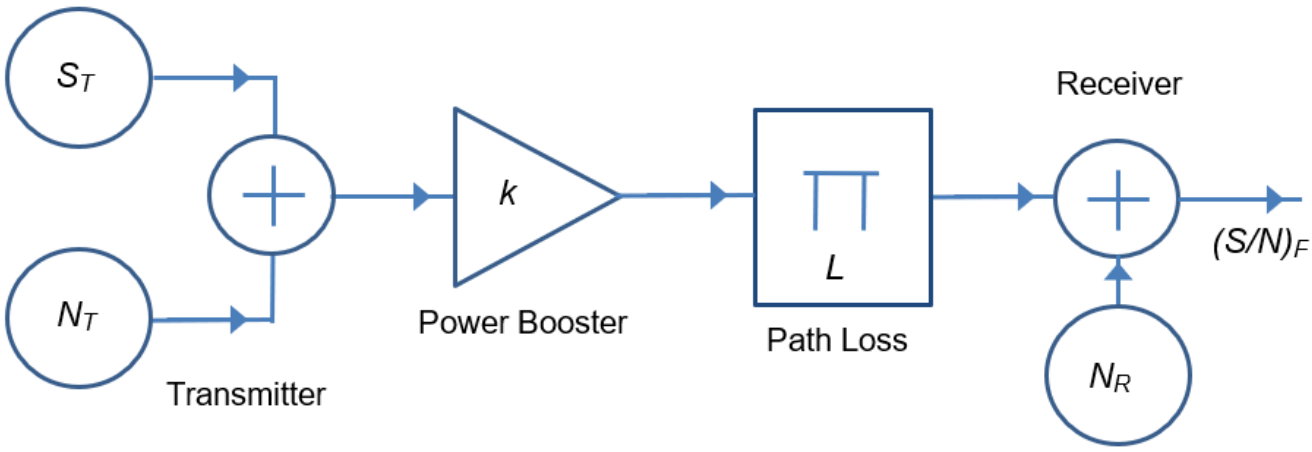


Figure 2: Transmitter power increase k to compensate for MER degradation.

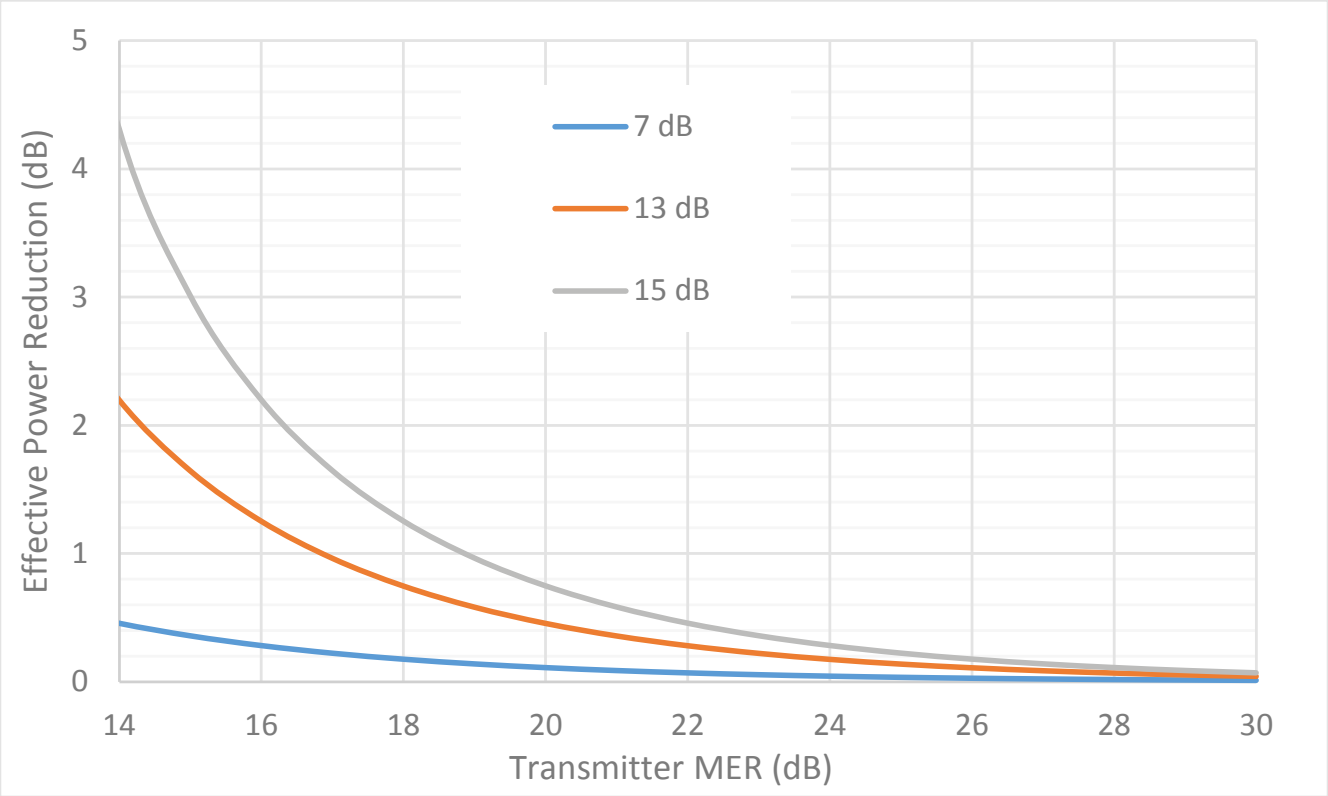


Figure 3: Calculated transmitter power increase to compensate for MER degradation

the effective loss of power if the actual transmitter power is kept constant. This effective power reduction represents a coverage reduction.

Rearranging the equation and noting that S_T/N_T = 2 × MER, and $(S_T/L)/N_R = (S/N)_F$ at the failure point:

$$\frac{1}{k} = 1 - \frac{(S/N)_F}{2MER} \qquad (3)$$

In a Gaussian channel, the $(S/N)_F$ can be taken as 7 dB [5]. More realistic fading channels, which include impairments such as multipath and Doppler, are used for network coverage planning, and for these $(S/N)_F$ in the range of 13 to 15 dB are used for different fading channel profiles. The plots below show the predicted effective loss of transmitter power for failure points of 7, 13 and 15 dB.

For the Gaussian case, the results are very encouraging. Assuming we start with baseline MER of 25 dB, then reducing this to 20 dB MER, causes the effective loss of power to be a negligible 0.1 dB. Further reduction in MER to 15 dB results in effective loss of power of less than 0.4 dB. However, the situation looks much worse for the fading channels, with the effective loss of power being 3 dB at 15 dB MER for the channel with $(S/N)_F$ of 15 dB.

Measuring the impact of changing MER in Simulated Channels

The effective power reductions under different channels were tested in laboratory experiments. DAB signal files were generated from baseband ETI reference files in the WorldDAB library [6], generating IQ samples using the open source ODR-DabMod modulator [7]. These were processed in software using the Rapp AM-AM compression model [8] to approximate the effects of linear amplifier compression. A series of files were created with reduced PAPR, ranging from 1.9 to 13 dB. The corresponding MER for these files ranged from 8.4 to 38 dB.

The DAB signal files were selected and played out using a Rohde & Schwarz Broadcast Test Centre signal generator. Fading channel characteristics were then applied, followed by the addition of noise, all within the same instrument. Three fading channel profiles were used: Gaussian channel (no fading), rural area (RA) and typical urban (TU) multipath [6]. The noise level was adjusted while the wanted signal level was kept constant, so that the signal-to-noise ratio at the receiver could be adjusted for testing.

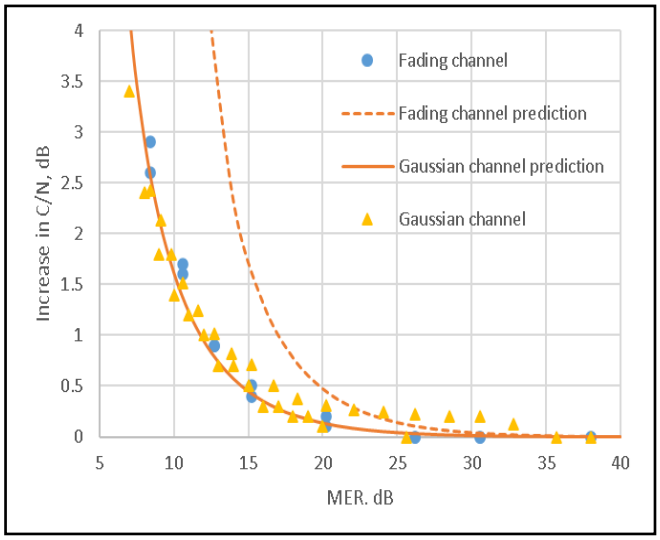


Figure 4: Experimental verification data for Gaussian and TU fading channel.

For each DAB signal file, the noise level was adjusted until the receiver BER is at the reference threshold. As used here, BER refers to pre-Viterbi MSC BER reported by a Rohde & Schwarz ETL Analyser. This raw BER is a pseudo-channel BER [9] since it relies on the output of the Viterbi FEC to detect errors rather than the audio source data stream. The specific value for the reference BER threshold is not critical in these tests but a value of $5x10^{-2}$ was chosen [10]. Measuring BER is complicated by the variability inherent in fading channels. The BER fluctuates widely with the channel gain and, to achieve a stable mean value, a very long averaging period is used. Because of the large number of points, measurement accuracy was traded against averaging time. This caused some scatter in the data.

The test results are shown in Figure 4 for the Gaussian and TU fading channels. Similar results were also found for the RA fading channel. This shows the increase in C/N needed when the transmitted MER is degraded. This is equivalent to the effective power difference calculated in the previous section, equation (3). The Gaussian channel results were as expected and showed good agreement with the calculation (solid line). However, the results for the fading channel do not follow the calculations based on the fading channel C/N (dashed line). Instead, they also follow the Gaussian channel calculation. This is a beneficial result since it means the MER can be reduced far lower than would be the case if the behaviour was aligned with that expected of the fading channel.

The impact of changing MER in a FADING channel

The practical test results above show that the fading channel behaviour is similar to that in a Gaussian channel. The explanation is as follows. Three noise sources contribute to the total noise at the receiver. The channel noise, made up of environmental noise and receiver thermal noise, can be assumed to be constant in all channel types. The transmitter noise is not constant at the receiver because it is subject to the loss in the channel, so it is directly proportional to the received signal level. This means that in a fading channel, when the signal level is low, the transmitter noise power is also proportionally reduced. When the fading channel signal level is high, the transmitter noise power is increased.

Figure 5 illustrates the Gaussian channel on the left, at the failure point C/NF, with the total noise being the sum of the channel noise and transmitter noise. On the right are three representations of the fading channel at three separate instants. In the first case, the signal has faded to a lower level, with the transmitter noise faded to the same extent. The BER is now poor, primarily because of the low signal strength relative to the channel noise, with the transmitter noise being less significant. In the second case, the signal is at the same level as for the Gaussian channel, and the noise contributions are the same; hence the impact on the BER is the same. In the third case, the signal level is greater. Although the transmitter noise and total noise are also greater, the signal is greater still, and the BER is better.

Most bit errors are associated with C/N values near their minimum because the BER has a very sharp dependence on C/N [11]. What happens at higher C/N values is largely irrelevant. In a fading channel a higher C/N is needed for a given average BER. In our present example, we can imagine raising the signal level by about 3 dB, so that our worst-case C/N now corresponds to that of the Gaussian channel at the failure point. Figure 5 now looks like Figure 6.

Since the fading channel at its worst looks just like the Gaussian channel at the failure point, we can assume that the effect of transmitter noise is the same in Gaussian and fading channels. Hence the same k-factor applies (the increase in transmitter power needed). A simulation of a more realistic Rayleigh distributed fading channel has been carried out for a range of MER from 10 to 40 dB. The transmitter power increase required was consistent with equation 3. In conclusion, as a convenient rule-of-thumb, the effective loss of transmitter power can be taken as 0.1 dB for an MER of 20 dB, in practical channels.

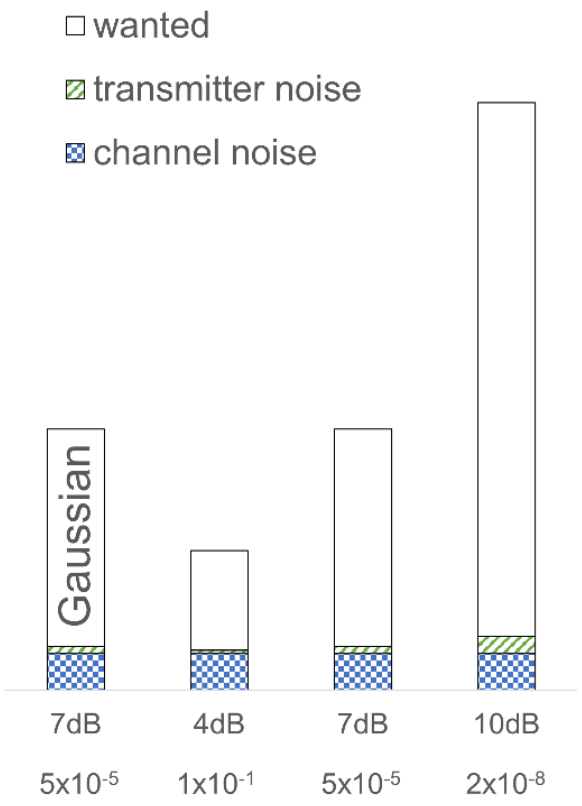


Figure 5: Noise in Gaussian and fading channels
BER is post-Viterbi BER

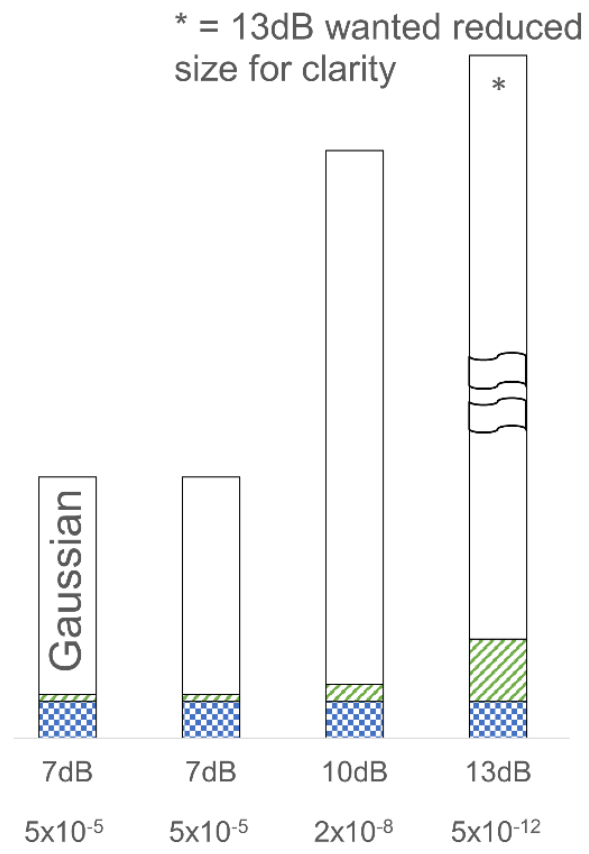


Figure 6: Noise in Gaussian and fading channels, increased power in fading channel
BER is post-Viterbi BER

Coverage Impact of Modulation Error Rate

The nominal coverage impact on a single isolated transmitter site can be estimated using the Hata propagation model [12]. The impact of the suggested 0.1dB effective change in transmitter power is to reduce the coverage radius to 99.2% of the original radius, for an effective transmitter antenna height of 100m. A circular coverage area is correspondingly reduced to 98.4% of the original area.

But this calculation assumes a cut-off threshold in coverage, where reception is either acceptable or unacceptable. In practice there is coverage variability due to propagation variation. For mobile outdoor reception the DAB network is typically planned for 99% locations served at the edge of coverage. With standard deviation assumed to be 4 dB [5] and effective change in transmitter power of 0.1 dB, then the coverage percentage at the edge is reduced from 99% to 98.93% locations. The overall average percentage locations served within the circular coverage area [13] reduces from 99.83% to 99.82% locations, a difference of only 0.01% locations.

In practice the propagation will be affected by larger scale variations in terrain and clutter, and there will be overlaps between neighbouring transmitters in the DAB single frequency network (SFN), so the practical coverage impact will deviate from these assumptions. A field trial was carried out to investigate the impact in a practical network environment.

An operational transmitter site at Clun in Shropshire, Figure 7, was selected, in the BBC UK national SFN. This allowed testing in both edge-of-coverage and transmitter overlap conditions. The dominant coverage area includes rural roads and village buildings. There is additional overlap coverage into the market town of Craven Arms. The transmitter (manufactured by COMMTIA) was operating on the BBC frequency 12B 225.648MHz with Effective Radiated Power (ERP) 600W. The transmission format is DAB Mode 1, UEP-3.

Coverage was evaluated using a series of drive tests in the coverage area. The test receiver configuration was a Factum Radioscape Observa Field Monitor with a ¼λ whip antenna magnetically mounted on the vehicle roof.

Figure 7: Clun Transmitter Site

The receiver can log useful parameters including signal strength, pre-Viterbi BER and GPS location. The number of transmitters identified, and their Transmitter Identifier Information (TII) codes, were used to confirm which transmitters contributed to the measurements. The TII codes were used to identify overlap areas where multiple transmitters were received in the SFN.

Three transmitter MER options were configured and tested to ensure compliance with the spectrum mask required by the ETSI standard [14].

The MER configurations were:
- 25dB MER, representative of existing network configurations
- 20dB MER, proposed reduced MER
- 15dB MER, extreme reduction for test purposes

Each of these MER configurations were tested in three reception scenarios:
- Strong dominant single-transmitter coverage
- Edge-of-coverage single transmitter
- Transition into overlap area with multiple transmitters

The routes were driven separately three times for the three different MER configurations. The measurement samples are averaged over grid pixels for analysis. The percentage of measurement samples where the BER is less than the 5x10-2 threshold is calculated for each pixel (before Viterbi error correction). Figure 8 shows the pixels identified as the baseline where percentage coverage is 99% for MER 25 dB, for 50-metre pixel size.

The results in Table 1 show the percentage coverage on the route, in terms of measurement sample points. This is restricted to pixels where the 25 dB MER results meet the 99% criterion. The overall average percentage points shows that the coverage degradation is <0.1% for both 20 dB and 15 dB MER. This is less than the variation between repeated measurements of the same route, so impact is not practically measurable.

For the above three cases, no measurable difference between the transmitter MER options was found for the scenarios with a dominant strong single transmitter (Clun) or in the scenarios where there was overlap of up to 6 transmitters. However, at the edge of the SFN coverage there were small differences.

A more detailed view of the edge of coverage is calculated for a 670m sub-section of the drive route at the western edge of the area (red rectangle in Figure 8). The vehicle is travelling east to west, from good reception to bad reception. BER is averaged with a 50m moving window. Table 2 shows the average BER on this edge of coverage segment, and the changes in the distance for which the BER exceeded the 0.035 and 0.05 thresholds, compared to the 25dB case. The differences in distances between the MER modes depends on where on the route the analysis is carried out, but the distances are small in all cases.

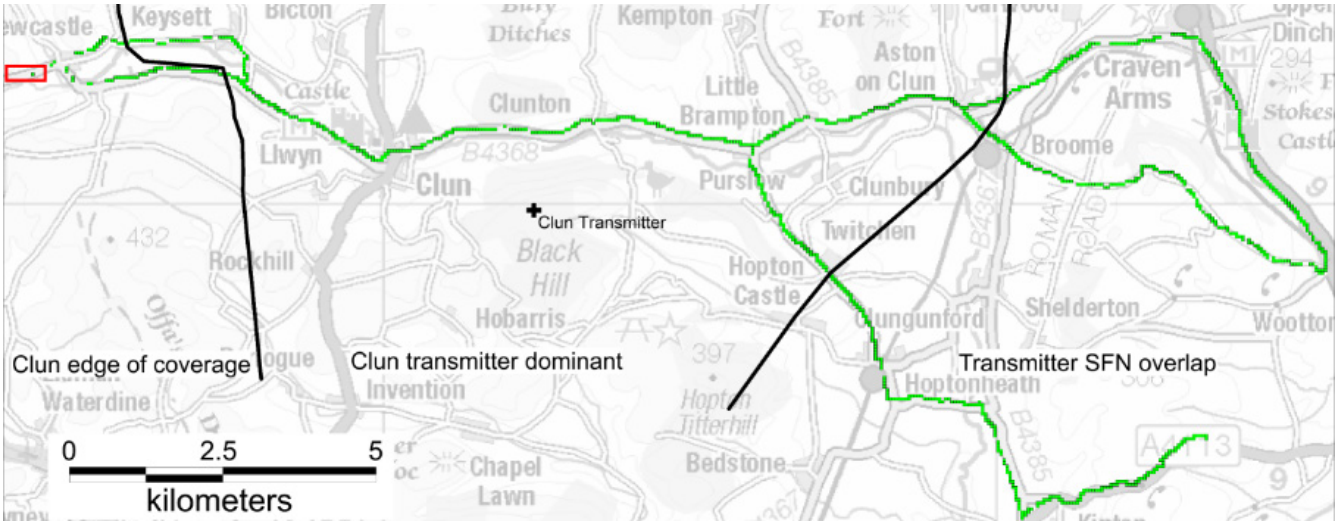


Figure 8: Clun drive route.

MER	Percentage sample points
15 dB	99.92%
20 dB	99.96%
25 dB	100.0%

Table 1: Percentage of sample points where pre-Viterbi BER is better than $5x10^{-2}$.

Figure 9 shows the moving average BER (over 50m) along this section of the route for the three transmitter MER options. There is some variability between the result for different MER options, even in areas of low BER. This will be due to propagation variability between successive drive runs, and the small variations in the path that the vehicle travels along the road. The BER is higher in some parts of this route for 15dB and 20dB MER, with 15dB MER being the worst.

The measurements show that the difference in BER between MER 25 and 20 dB is small. The differences are comparable with the variability between different drives on the same road. Subjective listener comments showed that the difference between 25 and 20 dB MER was not perceptible. The change from 25 to 15 dB was slightly perceptible at the extreme edge of coverage locations, with audio break-up earlier by 10 to 20 metres.

Implementation of energy savings

Using a combination of MER reduction from 25 to 20 dB, combined with optimising the power supply, can result in a worthwhile reduction in transmitter energy consumption. Careful analysis has been undertaken which has identified which types of transmitters already operating in the network can benefit from this optimisation approach. Evaluation of the BBC DAB network has confirmed that this approach is best suited to the modern Doherty design. The combined energy usage of the Doherty only transmitters in the network is approximately 2.3GWhr/year. This modern type of transmitter can be remotely optimised which will save transmitter site visits and can be implemented without a further carbon cost.

Overall reduction in energy consumption between 12 and 17% can be achieved, depending on the frequency of operation and other factors. Power supply changes resulting from the MER change should yield 5 to 11% reduction in consumption. Remaining benefits will be achieved due to other power supply optimisation carried out simultaneously. This approach will be rolled out to 165 present Doherty transmitters during 2024. Other types of transmitters in the existing network will not be modified using these techniques. This is because the changes to the power supply arrangement are far more intrusive and would require extensive site-by-site modification where in many cases the benefits could be negligible.

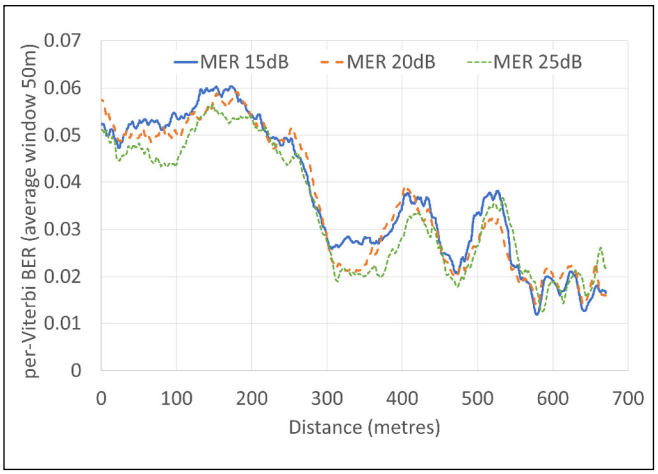


Figure 9: Clun edge of coverage average pre-Viterbi BER for different MER configurations, travelling east to west.

MER	15 dB	20 dB	25 dB
Average BER	0.0369	0.0351	0.0334
Distance BER>0.035	-44m	-13m	
Distance BER>0.05	-84m	-55m	

Table 2: Percentage of sample points where pre-Viterbi BER is better than 5x10-2.

Conclusions

In this paper we have presented an approach to reducing the energy consumption of DAB broadcast transmission, by improving the transmitter efficiency. A change in the transmitted signal quality, in the form of a small degradation in modulation error rate, combined with the optimisation of the power supply, enables the transmitters to operate in a more energy efficient configuration. This approach will be part of improvements to 165 BBC Doherty transmitters which should reduce energy usage by between 12 and 17%. This technique can also be considered for inclusion in future transmitter design. Theoretical analysis, laboratory simulation and field tests have shown that this approach has a negligible effect on practical reception. Compliance with regulatory limits can be maintained.

References

1. "BBC sets out path to Net Zero by 2030". Press Release. BBC. October 2021.

2. Digital radio and audio review, Department for Culture, Media & Sport. April 2022, Section 6.15.

3. Cripps, Steve C, RF power amplifiers for wireless communications. Vol. 250. Norwood: Artech house, 2006.

4. Rohde & Schwarz 2013. DAB Transmitter Measurements for Acceptance, Commissioning and Maintenance. Rohde & Schwarz Application Note. July 2013

5. EBU 2018.Guidelines for DAB Network Planning. EBU Tech 3391. May 2018.

6. ETI Reference Library. WorldDAB.

7. ODR-DabMod modulator. Opendigitalradio.org.

8. Rapp, C., 1991. Effects of HPA- nonlinearity on a 4DPSK/OFDM signal for a digital sound broadcasting system. Proceedings of the 2nd European Conference on Satellite Communication, 22-24.

9. Schramm, R., 1997. Pseudo channel BER-an objective quantity for assessing DAB coverage. EBU Technical Review, pp.23-30.

10. EBU 2020. Measuring Techniques for DAB Coverage Performance. EBU Tech Report 051. April 2020.

11. TR 101 758, Digital Audio Broadcasting (DAB); Signal strengths and receiver parameters; Targets for typical operation, ETSI, November 2000.

12. Hata, M 1980. Empirical formula for propagation loss in land mobile radio services. IEEE Transactions on Vehicular Technology 29.3: 317-325.

13. Reudink D O. Chapter 2 in Jakes, William C., and Donald C. Cox, eds. Microwave mobile communications. Wiley-IEEE press, 1994.

14. ETSI EN 302 077 Transmitting equipment for the Digital Audio Broadcasting (DAB) service; Harmonised Standard for access to radio spectrum.

Acknowledgements

The authors acknowledge the contribution of colleagues: Michael Clark, Marcin Rydlinski and Lindsay Cornell (BBC), Will Partridge, Thomas Smith, Mark Abbot, Julian Marlow (Arqiva)

Content Distribution at Mega Concurrency Scale

Girish Gopalakrishnan Nair
Amazon Web Services, India

Abstract
The rapid pace of content creation, proliferation of larger screen devices and easy internet access have led to an unprecedented surge in global user demand for the best-quality content streaming. When it comes to live sports streaming, unpredictable traffic patterns during a game can often overwhelm even the most well-designed systems. Reasons for this being:

- The inadequate capabilities of traditional autoscalers to keep pace with a traffic spike;
- Infrastructure limitations that are apparent only at mega-scale; and
- Suboptimal system configurations.

This paper explores architectural strategies, resources allocation guidelines and cloud-native workflows that are time-tested to support mega-scale live streams. This paper also examines common failure scenarios and anti-patterns, identifies infrastructure bottlenecks and proposes fallback strategies to enhance resilience and reliability at massive concurrency scales. Additionally, it shares our experience of supporting live cricket streaming on the AWS cloud for Indian audiences, where traffic can surge by over 1 million viewers per minute and occasionally drops from millions to a few hundred within seconds.

Introduction

The rise of video streaming platforms has revolutionized the way that the world consumes media today. This modern streaming landscape is characterized by a relentless demand for scalable, high-performance applications that can accommodate millions of concurrent users while maintaining the highest-quality user experiences.

Previous research on large-scale video streaming has primarily focused on aspects such as video encoding best practices [1], edge computing techniques [2], client behaviour analysis [3], infrastructure scaling strategies [4][5], or caching mechanisms [6].

However, live streaming at massive concurrency levels necessitates a holistic approach where all components seamlessly integrate and operate in tandem. When executed together, the cumulative resource requirements, including computing power, storage, and network bandwidth, may exceed the available infrastructure capacity within a given geographic region or country. Specially, in a country like India where cricket is not just a sport, but is considered an emotion, over-the-top (OTT) platforms must overcome the challenges posed by finite infrastructure, limited bandwidth, and other constraints, to deliver content to millions of concurrent viewers. This paper presents our experience in supporting large-scale cricket matches, where we witness challenges like traffic surge from 1.5 million to over 10 million concurrent viewers within the first 10 minutes of the match commencing, and rapid viewership drops from 13 million to less than 4 million viewers within seconds due to interruptions. Notably, we facilitated the creation of a world record with 59 million concurrent viewers and numerous other significant events in India.

Understanding Scalability Needs

The global video streaming market size is expected to grow from USD 11.0 billion in 2023 to USD 25.5 billion by 2028 at a compound annual growth rate of 18.3% during the forecast period. [7]

Concurrent User Demand Trends

59 Million In 2023: A record peak concurrent audience of 59 million viewers tuned in for Disney Hotstar's coverage of the 2023 Cricket World Cup final, setting a new streaming record [8].

32 Million In 2023: Over 500 million watched IPL in 2023, Jio saw 32.1 million peak concurrency [9].

7 Million In 2023: FOX Sports says Super Bowl LVII delivered an average of 7 million simultaneous streams, up +18% over the 2022 Super Bowl [10].

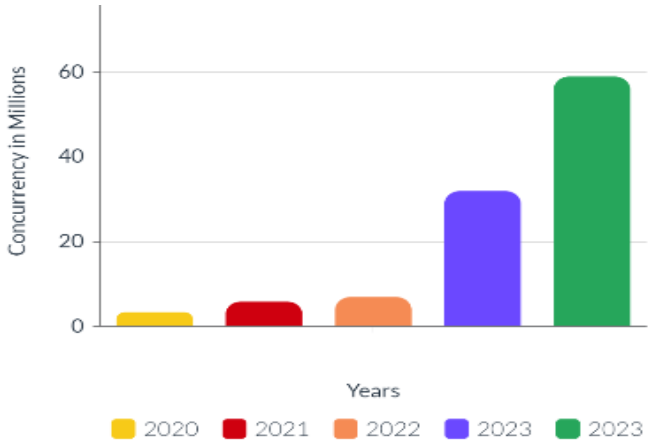


Figure 1: Concurrency Trends

5.9 Million In 2021: The 2021 UEFA European Championship final between Italy and England peaked at 5.9 million concurrent viewers on the BBC's digital platforms (BBC, 2021).

3.4 Million In 2020: The 2020 Tokyo Olympics saw a peak of 3.4 million concurrent viewers on the NBC Sports app, a record for the network (CNBC, 2021).

Technical Challenges

This section examines the practical challenges faced by streaming platforms when accommodating mega concurrent traffic. Consider a scenario where an organization aims to support a concurrency of 50 million viewers during an upcoming live event. Listed below are the challenges, potential bottlenecks and practical limitations that may arise when attempting to accommodate such a substantial concurrent user base.

Finite Infrastructure

Facilitating a massive concurrent user base demands a robust infrastructure capable of handling high traffic volumes while ensuring acceptable response times. Accommodating 50 million concurrent users necessitates substantial computational resources, upwards of thousands of CPU cores to ensure a seamless user experience. However, expecting an unlimited supply of CPUs from any infrastructure provider is unrealistic.

Limited Bandwidth

Live streaming events that scale up to a concurrency of approximately 50 million viewers end up consuming 40+Tbps of bandwidth on Content Delivery Networks (CDNs) for video delivery (including 4K videos) across India. Therefore, the assumption that a single CDN possesses the requisite bandwidth capacity, represents a critical oversight, necessitating careful consideration in planning delivery strategies.

ISP Peering

Customer diversity, encompassing geographical locations, devices and network providers, are critical factors in effective traffic distribution planning. If a substantial portion of the traffic is expected from a single Internet Service Provider (ISP), it may overwhelm the peering junction between the CDN and ISP. Such scenarios cannot be effectively simulated during testing.

Sudden Traffic Fluctuation

The occurrence of traffic spikes and drops are inevitable, presuming traditional cluster auto-scaling mechanisms are adept at managing such fluctuation presents a notable concern.

Scaling Strategies and Guidelines

Auto-scaling offers a dynamic approach for resource provisioning, enabling organizations to adjust their computing infrastructure based on demand fluctuations. Some technical challenges highlighted by this paper can potentially be addressed by determining the optimal timing for scaling actions (when to scale) and the appropriate magnitude of resource adjustments (how much to scale).

Scale Cube

The scale cube model talks about 3 dimensions of scaling:

Scale Out

The X-axis of the scale Cube (Fig. 2) represents horizontal scaling, achieved through a technique known as scale-out. This approach involves deploying multiple instances of the application behind a load balancer. The load balancer distributes incoming requests among these application instances based on a chosen algorithm (e.g., Round Robin, Least Outstanding Requests). However, a critical challenge associated with this approach is data access. Each application instance may require access to the entire data set, necessitating a large cache to maintain acceptable performance under high concurrency loads.

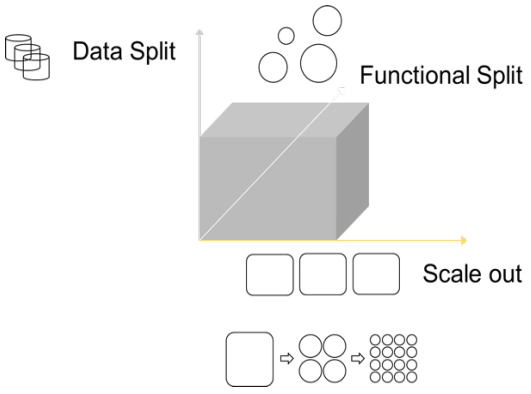


Figure 2:Cube Scale

Data Split

The Y-axis of the scale cube (Fig. 2) denotes data partitioning. This strategy often involves deploying micro services, with each service encapsulating a specific functionality and having access to its relevant data subset. Decomposition techniques such as verb-based or noun-based approaches can be employed to achieve this data partitioning. Combined with horizontal scaling (scale-out), this approach offers a robust scaling strategy for large-scale applications. Additionally, it serves as a fundamental unit for load estimation and subsequent scaling decisions. Hereafter, we will refer to this unit as a scaling domain for consistency throughout the paper.

Functional Split

The Z-axis of the scale cube (Fig. 2) represents functional partitioning. This approach leverages scaling by deploying multiple, identical application instances. However, each instance is assigned a specific data subset, ensuring data isolation. A dedicated routing component (e.g., load balancer or CDN) directs incoming requests to the appropriate cluster based on pre-defined criteria. Common routing strategies include utilizing request attributes like the primary key of the accessed entity or customer segmentation (e.g. routing requests from paying customers with higher Service Level Agreements (SLAs) to servers with enhanced capacity). This approach offers a significant advantage in failure isolation, as a malfunction within a single instance only impacts its associated data subset. When combined with other scaling techniques, functional partitioning contributes to enhanced platform scalability and reliability.

Cap Theorem

The **C**, A, P in CAP theorem stand for:

1. C (Consistency): Every node in a distributed cluster returns the same, most recent response after a successful write operation.

2. A (Availability): Every read or write request for a data item results in either a successful completion or a clear error message indicating the failure.

3. P (Partition Tolerance): The system continues to operate even when the network connecting the nodes experiences a failure that partitions the network. Nodes within each partition can still communicate with each other, but communication across partitions is disrupted.

In a distributed streaming ecosystem, achieving CAP simultaneously is not possible. Therefore, a clear comprehension of when and what to sacrifice, as well as the appropriate extent of such trade-offs, is a critical factor in facilitating the adoption of the scaling practices delineated within this paper.

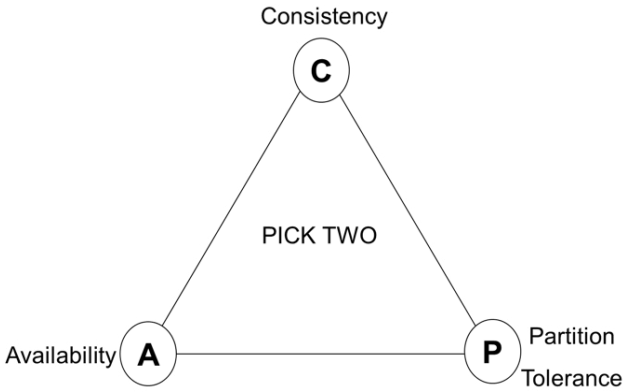


Figure 3: Cap Theorem

Algorithm: Concurrent-Viewer-based-Dynamic-Scaling

Having established the principal guidelines for building scalable streaming services, the paper proposes an algorithm to facilitate scaling beyond these spiky traffic patterns. Traditional cluster auto-scaling approaches rely on metrics such as **CPU** and **memory utilization**. While these auto-scalers can be effectively utilized for non-critical (non-P0) services or day-to-day traffic, but they are inadequate for scaling P0 services during spiky traffic situations, as they do not guarantee the completion of scaling before a traffic surge overwhelms the available server capacity. The paper evaluates the use of a **"Concurrent Viewers Metric"** (Cvm) as a key parameter to guide the auto-scaling mechanisms of a streaming platform.

The algorithm can effectively guide the scaler to auto-scale down but the approach for scaling down the infrastructure requires more careful consideration. Automatic triggers for downscaling based solely on a reduction in traffic can be risky, especially at massive scale, due to the possibility of false positives reported by the various sources. In most practical scenarios, a manual trigger is preferred to initiate the downscaling process. This manual approach is typically employed after the conclusion of the live event or when a sustained decrease in viewership is confirmed, rather than relying on immediate traffic reduction signals.

How We Do It

Having established a fundamental understanding of the core principles and algorithms, it is now crucial to delve into the practical aspects of running such a system successfully in the real world. This section will provide a detailed examination of the end-to-end process, from the preparatory stages to the execution of the live event.

1: Variables:

V_a: Current number of actively viewing users

V_n: Number of newly logged-in users

C_{vm}: Concurrent Viewers Metric, calculated as $C_{vm} = V_a + V_n$

V_i: Target initial traffic volume for the event

V_l: Ramp-up ladder of new viewers

V_b: Buffer capacity to accommodate new viewers during scaling

V_t: Scaling trigger threshold

S_f: Scaling factor, metric which will be used to determine the instances to be added

2: Initialize variables defined:

$V_a = C_{vm} = V_n = V_t = 0$

V_i, V_l, Vb = pre-configured values

3: During the event, update the variables in each time interval:

Measure V_a (current actively viewing users)

$V_a = \text{Max}(S_1, S_2, \ldots S_n)$ Maximum value of users detected from multiple sources.

Where:

S_1 represents the sum of unique users fetching a particular content segment from the CDN

S_2 represents a unique heartbeat sent by a mobile application

S_i represents unique user numbers obtained from multiple sources

Measure V_n (Realtime stream of logged-in users from Authentication Service)

Calculate $\mathbf{C_{vm}} = V_a + V_n$

4: Scaling Trigger Condition:

If $C_{vm} > V_t$ **then**

Calculate delta between time intervals of each trigger (ΔT)

$\Delta T = T_i - T_{i-1}$

- Where T_i is the timestamp of the i^{th} trigger.
- T_{i-1} is the timestamp of the $(i-1)^{th}$ trigger.

Calculate mean of time intervals (μT)

$\mu T = (T_0 + T_1 + \ldots + T_n)/n$

- Where n is the total number of triggers
- T_n is time of n the trigger

Calculate deviation of time intervals of each trigger from mean

S_f = deviation ($\Delta T - \mu T$)

Empower auto-scaler to calculate required capacity

Trigger Karpenter Scaling(S_f)

Update Scaling Trigger Threshold:

$V_t = V_t + V_l - V_b$

This adjusts the scaling trigger to account for the expected ramp-up and the buffer capacity

Planning and Prioritization

Large-scale event management necessitates meticulous planning. While there are no silver bullets to address all scaling challenges at once, effective strategies can be progressively refined through the experience gained from conducting such events over time. Cloud Service Providers (CSPs), like Amazon Web Services' IEM (Infrastructure Event Management) & MEM (Media Event Management) teams, are involved in planning and testing phases which typically ranges from 3 to 6 months for most platforms.
This experience is necessary to ensure that no critical aspect is overlooked, as even minor misconfigurations or miscalculations can adversely impact streaming performance. The challenges associated with large-scale video streaming are multifaceted, including bandwidth requirements, content delivery, scalability, redundancy, and fault tolerance. The MEM team enhances the operational reliability of business-critical video streaming events. They cover a suite of AWS media services and follow a structured four-phase approach to ensure the successful delivery of large-scale video streaming events:

1. **Pre-event Review and Preparation:** The MEM team conducts an Operation Readiness Review assessment to identify and mitigate potential risks in the video streaming architecture and configuration before the event.

2. **Event Preparation and Testing:** The team develops a comprehensive end-to-end test plan, creating operational runbook guidance and weekly status updates.

3. **Live Support during the Event:** During the live event, the MEM team engages subject matter experts (SMEs) to provide effective monitoring, troubleshooting, and triaging support.

4. **Post-event Analysis, Retrospection, and Summary:** After the event, the MEM team conducts a thorough analysis and retrospection. An event summary report is generated, which includes an assessment of learning and customer engagement evaluation.

Prioritization

The first step of prioritization is giving up aspirations to run everything; even under extreme circumstances. For OTT platforms, the top priority should be ensuring continuous content playback for viewers. The user flow and potential user paths associated with achieving smooth playback becomes your P0-Workflow (Priority Zero Workflow).

All services that directly enable viewers to complete this workflow and successfully view content are classified as P0 services (Priority Zero Services). The remaining services are categorized as Non-P0s. The prioritization of Non-P0 services (e.g. P1, P2, etc.) can be further refined based on their specific platform architectures. For the purposes of this paper, we will focus on the categorization of services into P0 and Non-P0 categories.

The P0 services underpinning the core user experience (P0 workflow) necessitate utmost reliability and elasticity to guarantee uninterrupted content delivery. In this context, the paper will examine a typical P0 workflow for live cricket streaming: the typical architectures employed, the specific parameters that facilitate optimal performance, and contingency measures in the event of unforeseen disruptions:

P0.1: Authentication Services

A user's interaction with the platform typically begins with the authentication process. The majority of users will be already signed-in or have an existing account, and mere authorisation will suffice. The usage patterns for the Registration API, Login API, and Authorization APIs will vary significantly, and each of these components falls under distinct scaling domains.

Reference architecture

The Authorization APIs play a crucial role in the video streaming platform, as they are invoked for every user request, regardless of whether the user is logged in or not. Every service within the platform leverages the Authorization APIs to validate the user's permissions and access rights. This high-frequency usage pattern necessitates a different architectural approach compared to the Sign-in/Sign-up APIs. A serverless set-up, backed by a NoSQL database or an in-memory cache, can be an effective solution for scaling the Authorization APIs seamlessly. This approach allows for automatic scaling based on demand, ensuring optimal performance and cost-efficiency. However, when operating at a large scale, certain parameters may require fine-tuning to maintain the desired level of performance and reliability.

Parameters & Considerations

The API Gateway allows configuring burst request limits and maximum request limits. These limits need to be increased in anticipation of high-traffic events. However when using an Amazon API Gateway, an often overlooked aspect is the ability to customize the Integration Request Timeout value for API integrations. The default timeout value is set to 29 seconds, but most API calls, even under scaled traffic, typically complete within 1 second unless the backend is

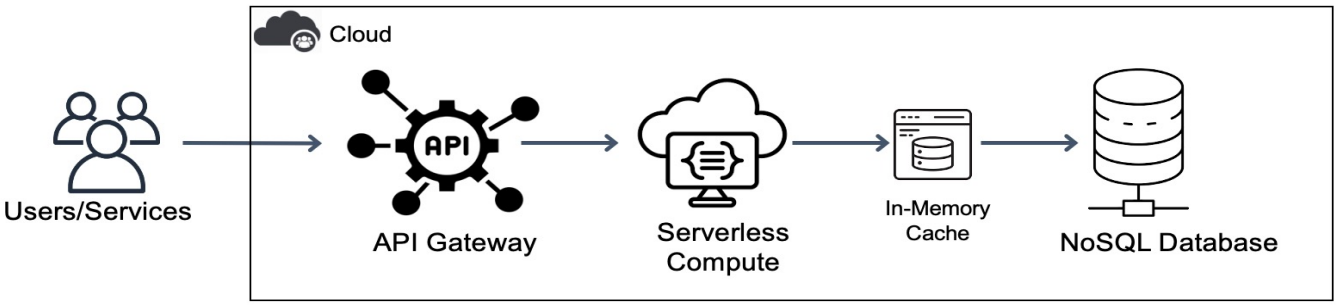


Figure 4: Authorisation Micro-service

throttled and unable to serve requests efficiently. By reducing the timeout value to a lower value, such as 1 or 2 seconds, the API Gateway can scale seamlessly and optimally utilize resources for the given workload. For fire-and-forget APIs, like those used to capture application heartbeats, the timeout can be set even lower, around 50 milliseconds.

To handle the inrush of hundreds of thousands of users joining a live stream, the platform can engage a NoSQL database like Amazon DynamoDB. DynamoDB's "on-demand" capacity mode, coupled with a well-designed key schema can help address this inrush. Additionally, to cater to the read-heavy nature of live events, deploying an in-memory cache like DAX (DynamoDB Accelerator) can provide increased throughput and guard against increased costs due to over-provisioned read capacity in DynamoDB.

Panic Modes

To ensure optimal latency and reliability, all APIs in the video streaming platform should be accessed via a Content Delivery Network (CDN) which helps with HTTP Cloning. Apart from that, in an event of increased requests per second (RPS) (more than the planned capacity), a "panic mode" can be enabled at the CDN level. This mode will respond with cached responses, allowing all users to gain access to the system and watch the live event. This will ensures that all customers are able to stream some content at any given point in time, prioritizing availability over consistency.

API Gateways provide a feature called "Usage Keys," which can be leveraged to implement rate-limiting and throttling seamlessly. This feature helps to safeguard the backend services from being overwhelmed by excessive traffic, ensuring that the platform remains responsive and stable even during periods of high demand.

P0.2: Dashboard/Home Page

After successful authentication, users land on the Home Page, where they can: explore available content, search for live matches, and start watching them. The Dashboard page is critical and architecturally complex as it is built from responses from multiple APIs, including:

1. Geo-location Service: This API fetches the user's location to provide location-specific content and personalization.
2. Personalization Service: Based on the user's location and other preferences, this API retrieves personalized content recommendations.
3. Continue-Watching Service: This API fetches partially watched or bookmarked content for the user, enabling them to resume their viewing experience seamlessly.

When a user clicks on specific content to watch, other APIs are invoked, such as:

1. Get CDN: This API determines the optimal CDN for streaming the requested content, ensuring low latency and high-quality playback; and
2. Entitlement API: This API verifies the user's access rights and entitlements to the requested content, ensuring that authorized users can access only the content they have subscribed to.

Thus the Dashboard page is built from responses from both critical (P0) and non-critical (non-P0) APIs, which introduces complexity in terms of performance, scalability, and fault tolerance.

Reference architecture

The backend APIs of the video streaming platform face significant load and stress during critical periods, such as the start and end of a live match. At the beginning of a match, a surge in traffic is expected as viewers attempt to access the platform and start streaming the content.

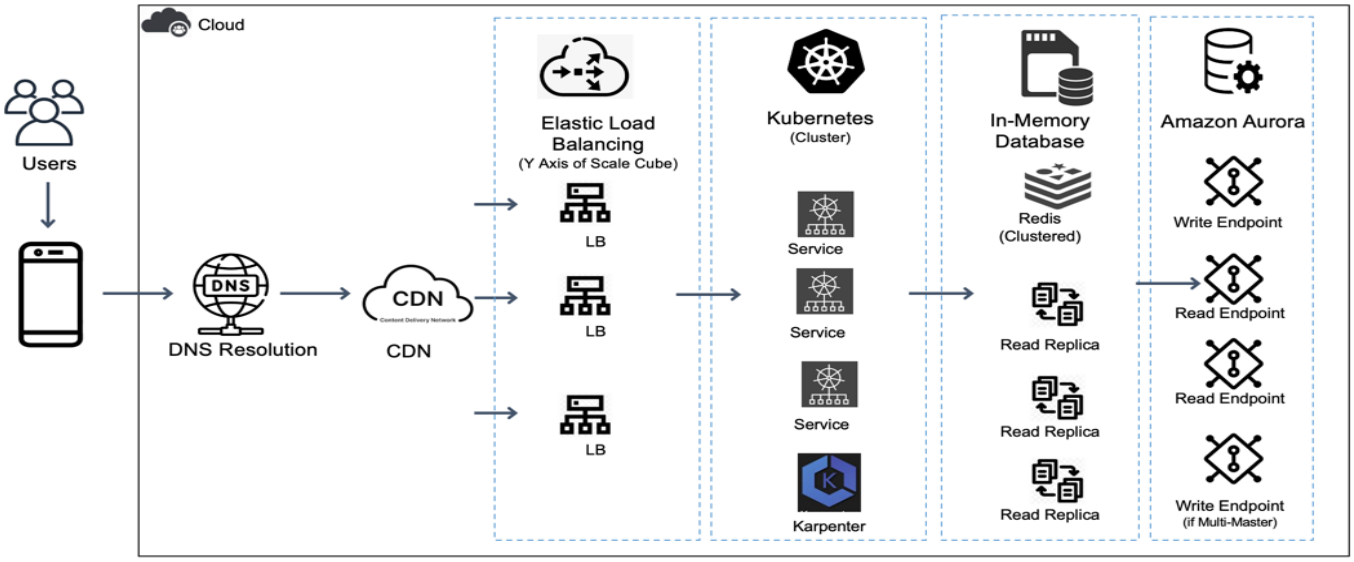


Figure 5: Backend Services

Similarly, when the match concludes, viewers tend to click on the home page before exiting the platform, potentially generating another spike in dashboard traffic.

If this scenario is not properly addressed, it can potentially overwhelm the infrastructure and lead to system failures. While the architecture components deployed for business-as-usual (BAU) traffic may be similar, the approach to handling large-scale events requires specific modifications and optimizations.

Parameters & Considerations

Some key optimizations here that can help you scale seamlessly during spiky events are:

Load balancer Sharding and Pre-warming: This architecture has multiple load-balancers, which implement sharding and distribute traffic across multiple Application Load Balancers (ALBs). This helps to mitigate potential throttling or resource exhaustion during traffic spikes by invoking Route53's weighted routing policy to route traffic to shards. ALBs behind a CDN are also employed for improved performance and reduced latency. The load balancer is also pre-warmed, for events expecting flash traffic. Whether to shard an ALB is a decision normally taken at the load testing phase by looking at following parameters:

1. Number of requests per second expected;
2. Average size of request (Headers + body); and
3. Average response time of request.

EKS Clusters: Carefully select the appropriate instance types for Kubernetes nodes. Larger instance sizes are a no-brainer when it comes to launching instances, as it reduces the number of instances to be launched and reduces the frequency of scaling. But this can lead to CPU underutilization, particularly when using Kubernetes that limit the number of pods that can be deployed on a single instance. Kubernetes features like resource requests limits and node affinity are employed to ensure efficient resource utilization and workload distribution. High-performance Kubernetes cluster auto-scalers, such as Karpenter, are used to offload heavy processing by scaling the underlying cloud provider's compute service (like Amazon EKS) achieving optimised computing in least possible of time

Panic Modes

In scenarios where resources get constrained, disabling Non-P0 services is an effective approach. This approach will free-up Non-P0 resources allowing p0 services to consume these and scale.

Another panic strategy is to implement static responses from the API. For instance, instead of invoking backend services to construct dynamic dashboard responses for each user request, a pre-built static dashboard response can be served. This approach allows users to view essential information on the dashboard and initiate live stream playback with minimal processing overhead. This panic mode can be enabled at the Content Delivery Network (CDN) level, ensuring minimal change at infrastructure.

Another panic behaviour could be serving a default. In cases where the computation of the optimal CDN URL for a 'GetCDN' request is unavailable, a primary or preconfigured CDN can be used as a fallback option. This ensures that users can access content from any one CDN even if that is not the best one to serve content to particular user.

P0.3: Video Streaming - Playback of secured content

Upon clicking the Live Match Banner, the video playback is initiated. For a video streaming platform, the core functionality is delivering smooth, uninterrupted streams to viewers. Playback must remain flawless at all times and for all viewers, regardless of the scaling event's magnitude. A viewer's experience is not contingent upon the platform's ability to handle a specific number of concurrent streams; rather, it hinges on a near-flawless viewing experience for their chosen content.

Reference architecture

The reference architecture (Figure 6) initiates requests for redundancy across all stages of the streaming pipeline – from ingestion, playout, encoder, packager to CDN. This redundancy ensures continued service availability and minimizes the impact of potential component failures.

Parameters & Considerations

Pipeline Locking

Redundancy in large-scale video streaming architectures ensures high reliability but introduces complexities in managing a seamless viewing experience. The key challenge arises from potential mismatches in chunk IDs and frame alignment across redundant delivery paths, leading to disruptions or inconsistent playback. The technique AWS deploys to mitigate this issue is MediaLive's "Pipeline Locking" feature. This feature provides frame-accurate outputs by synchronizing the video and audio pipelines.

Multi-CDN

Relying on a single CDN to serve a diverse customer base is often insufficient for achieving optimal performance. CDN performance can vary across different locations and time periods, as certain CDNs may perform better in specific regions or during certain times of the day. To address this challenge, the video streaming system should be designed to leverage the best-performing CDN for a given location and time, ensuring a superior Quality of Experience (QoE) for end-users. The intelligence in selecting the optimal CDN is derived from metrics such as Video Start-up Time (VST), Video Start-up Failures (VSF), rebuffering events and playback failures, etc. By analysing these metrics, areas where streaming performance is suboptimal and negatively impacting QoE can be identified and the root causes of these issues can be pinpointed. Utilizing multiple CDNs mitigates the risk of outages and bandwidth limitations of one CDN.

Timely corrective actions can be implemented by dynamically switching to a better-performing CDN.
Streaming Protocol considerations: HLS or DASH or Both?

The DASH (Dynamic Adaptive Streaming over HTTP) manifest, is an XML-based file that contains essential information for video playback. HLS (HTTP Live Streaming) manifest is comparatively easy to read and was developed much earlier than DASH. Both protocols are efficient in delivering adaptive bit rate streaming.

When deciding on the appropriate protocol for high-concurrency events, even minor factors can play a significant role. Although DASH is defined as an industry standard, with wide device support, our experience has shown it is not consistently supported across all platforms and devices. The Indian market encompasses a diverse range of alternative operating systems, including KaiOS, iOS, and Linux distributions, which currently lack DASH compatibility.

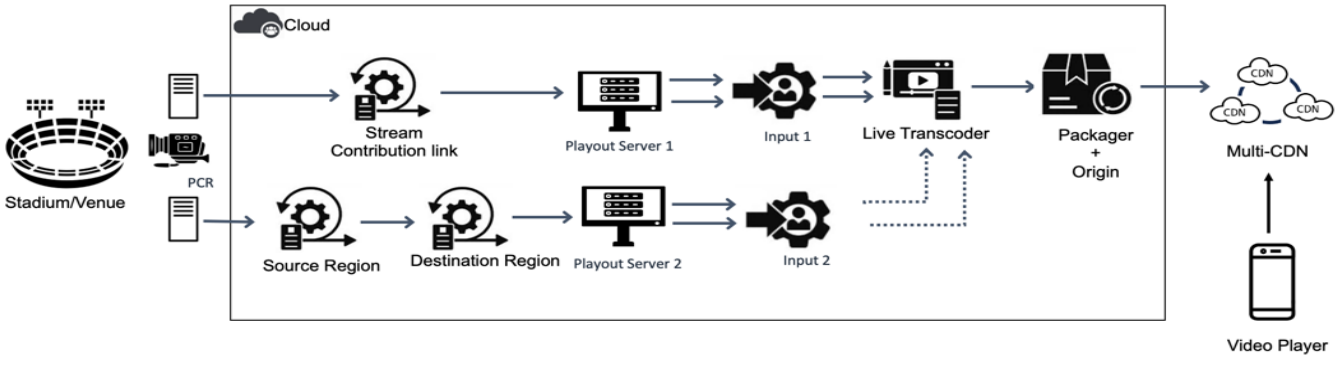


Figure 6: Video Pipeline Architecture

A specific challenge observed with DASH for live events when enabling Segment Timeline is variation in audio segment sizes. The Segment Timeline helps Server-Side Ad Insertion (SSAI) services to insert advertisements, which is crucial for generating revenue during the event. But the variation in segments complicates SSAI.

Enabling a segment timeline also result in larger manifests. At scale, large manifest sizes can become problematic. Especially when these are refreshed every two seconds. This can generate significant traffic on the CDN as well. Thus the reduced complexity and comparatively smaller size of HLS manifests have proven effective in the Indian device landscape.

Panic Modes

To ensure continuous playback and mitigate the impact of potential failures, a world-feed of the live match is provided as additional input to MediaLive. This redundancy measure ensures that viewers can continue watching the match in some language, even if the primary playout servers or the clean feed ingest fails, preventing playback errors and maintaining a continued viewing experience.

Developing plans for controlled degradation such as the intentional disabling of non-critical services during critical events, serves as an effective panic strategy. For example, if the CDN experiences bandwidth exhaustion, the ability to dynamically remove high-quality renditions, such as 4K and 1080p, becomes crucial. The 'Manifest Filtering' feature of a Packager (MediaPackage) helps to implement this graceful degradation technique. This prioritizing of lower resolutions enables a larger number of viewers to access the content at significantly less bandwidth. Without graceful degradation, only a limited audience would be able to access the content, potentially resulting in errors or suboptimal experiences for others.

P0.4: Monetisation - Ad insertion and reporting

In the context of cricket streaming in India, where most matches are free to watch and powered by advertisements, ad insertion is a P0 for organizations to ensure revenue generation from these streams. There are multiple options available for ad insertion, Server-Side Ad-Insertion (SSAI) being one of the most common approaches that allows seamless ad insertion without disturbing the playback.

SSAI involves the integration of an ad decision server (ADS) within the video delivery workflow. The ADS determines the relevant advertisements for a particular ad slot based on factors such as viewer demographics, content metadata, and ad campaign rules.

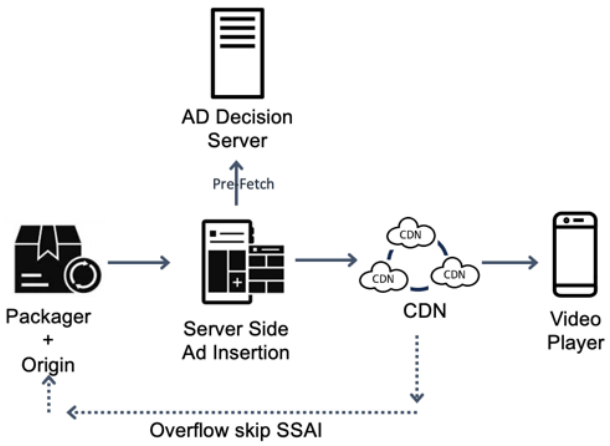


Figure 7: SSAI Architecture

The ad insertion server then seamlessly splices the selected advertisements into the main content manifest, creating a unified stream for delivery to the viewer. Furthermore, SSAI maintains a consistent viewing experience, which is crucial for user engagement and satisfaction.

Reference architecture

To implement server-side ad insertion (SSAI) at a large scale on the AWS cloud, AWS employs AWS MediaTailor. This acts as a manifest manipulator, responsible for dynamically inserting ad segments into the video manifest files as they are delivered to viewers. MediaTailor will front-end the video delivery pipeline to perform SSAI.

Parameters & Considerations

During large-scale live events, ad request and transcoding timeouts can lead to significant revenue leaks if not addressed properly. To mitigate this issue and maximize ad fill rates and monetization, the architecture employs ad prefetching. Ad prefetching is a proactive approach where MediaTailor fetches and transcodes ad assets in advance, before they are required for insertion. This technique serves two key purposes: it provides more time for programmatic ad trading, and it reduces ad insertion latency. Implementing ad prefetching is particularly crucial in scenarios with high concurrency, stringent latency requirements, or when dealing with a large volume of ad requests and assets. It helps to optimize ad delivery and revenue generation by minimizing the risk of timeouts and missed ad opportunities during live events with high viewer traffic. To reduce the ad transcoding time, "Transcode profile" is used which can only be configured via AWS Support as of today.

Panic Modes

Fail over and overflows

When implementing Server-Side Ad Insertion at large scales, it is crucial to consider the predefined limits of the SSAI service, as exceeding these limits can lead to request failures. To mitigate this issue, an overflow mechanism can be implemented. One approach is to configure the CDN, such as AWS CloudFront, to bypass the ad insertion process entirely when encountering errors like HTTP 429 (Too Many Requests) from the SSAI component. Alternatively, this can be achieved by configuring the GetCDN API to return the direct origin URL for the core video content, bypassing SSAI during overflow conditions. By implementing this overflow mechanism, the architecture ensures the uninterrupted delivery of core video content, even if ad insertion needs to be temporarily suspended for overflow traffic.

Conclusion

Successfully scaling video streaming applications to accommodate millions of concurrent viewers necessitates a multifaceted approach encompassing robust infrastructure, efficient content delivery mechanisms, effective data management strategies, and comprehensive security measures. By meticulously planning and implementing the best practices outlined throughout this paper, platform engineers can foster a resilient and scalable streaming ecosystem capable of delivering exceptional viewing experiences to a global audience.

References

1. Layered video streaming in large-scale networks - https://www.sciencedirect.com/science/article/abs/pii/S1084804514001441
2. Large-Scale Video Stream Concurrent Transmission for Edge - https://www.mdpi.com/2227-7390/11/12/2622
3. Modelling large-scale live video streaming client behaviour - https://link.springer.com/article/10.1007/s00530-021-00788-4
4. Auto Scaling API Gateway based for K8 - https://ieeexplore.ieee.org/document/8663784
5. Proactive-Reactive Auto-Scaling Mechanism for Unpredictable Load Change - https://ieeexplore.ieee.org/document/7557733
6. Large-Scale Video Streaming in Highly Heterogeneous Environment - https://ieeexplore.ieee.org/abstract/document/4359973
7. Markets and Markets, "Video Streaming Software Market by Component (Solutions, Services)" - https://www.marketsandmarkets.com/Market-Reports/video-streaming-market-181135120.html.
8. SportsProMedia: 2023 World Cup Final - https://www.sportspromedia.com/news/cricket-world-cup-final-2023-disney-hotstar-live-streaming-viewership-record/.
9. Mint - "Over 500 million watched IPL in 2023, Jio saw 3.21 crore peak concurrency" https://www.livemint.com/industry/media/over-500-million-watched-ipl-in-2023-jio-saw-3-21-crore-peak-concurrency-11686232243972.html.
10. Steaming media blog - "thirteen years of super bowl streaming viewership stats, 2012-2024" – february 22 2024 https://www.streamingmediablog.com/2024/02/superbowl-streaming-numbers.html
11. https://blog.hotstar.com/scaling-for-tsunami-traffic-2ec290c37504
12. https://www.ibc.org/technical-papers/ibc2023-tech-papers-implementing-hls/dash-content-steering-at-scale/10258.article

Acknowledgements

The author would like to thank his colleagues for their contributions to this work. He would also like to thank the International Broadcasting Convention for permission to publish this paper.

"Mr. Tony Thomas" Head of DevOps & Quality at JIO, for valuable insights and assistance during the course of research. "Mr. Alastair Cousins" Sr.Manager Solution Architect at AWS, for valuable insights and assistance during the course of research. "Mr. Maheshwaran G" Principal Solution Architect at AWS, for assistance and inputs. "Mr. Sanjay Singh" Sr Solution Architect, TMEGS at AWS Professional Services, for data points and metrics. "Mr. Ajay Bhardwaj Swami" Sr.Technical M&E Consultant at AWS, for constructive feedback and suggestions at review.

Novel Image Sensor with Area-Based Optimisation of Shooting Conditions for Immersive Content Productions

K. Kikuchi[1], K. Tomioka[1], T. Usui[1], A. Honji[1], K. Kitamura[1], and S. Kawahito[2]
[1]NHK, Japan and [2]Shizuoka university, Japan

Abstract
We present a novel scene-adaptive imaging technology designed to enhance the image quality of wide-angle immersive videos such as 360-degree videos. It addresses the challenge of balancing resolution, frame rate, and dynamic range due to sensor limitations by dynamically adjusting shooting conditions within a single frame on the basis local subject characteristics. This involves capturing still subjects at high resolution and moving subjects at increased frame rates, adjusting exposure time according to subject brightness while maintaining pixel readout rate. To validate this approach, we developed a block-wise-controlled image sensor prototype with 1.1 million pixels that enables flexible control of shooting conditions individually for 272 separated blocks. Real-time scene analysis and a feedback control system were also developed. Experimental results demonstrate that the proposed method improves subjective image quality compared with conventional imaging that captures the entire frame under a single shooting condition, even at the same data rate.

Introduction

The global demand for highly immersive video content, such as 360-degree videos and dome screen videos, is escalating. Accompanying this demand is an increasing need for cameras capable of capturing wide viewing angles effectively (e.g. panoramic cameras and omnidirectional multi-cameras). Wide-angle videos typically feature subjects exhibiting diverse textures, movements, and brightness on a single screen, requiring image sensors to meet rigorous performance quality, including not only resolution and frame rates exceeding ultra-high definition television levels (see [1]) but also excelling in dynamic range for incident light. However, developing an image sensor that fulfils all these requirements simultaneously is challenging. Traditional image sensors, such as Complementary Metal Oxide Semiconductor (CMOS) image sensors operating under constant shooting conditions across the entire pixel array, are limited by a trade-off between resolution, frame rate, and the noise performance related to dynamic range (El-Desouki et al [2] and Kawahito [3]). Moreover, higher pixel readout rates lead to increased data transfer streams and higher power consumption in image sensors.

On the other hand, from the perspective of improving subjective image quality, uniformity in the sensor's shooting conditions across the screen appears dispensable. For example, areas with still subjects may benefit from high resolution, whereas those with moving subjects may require ensuring temporal resolution rather than spatial resolution to minimise motion blur. In addition, those with high brightness

do not necessitate achieving precise dark gradation, and those with low brightness do not require high pixel saturation. Acknowledging this, we propose a new shooting approach, scene-adaptive imaging that facilitates dynamic control of the shooting conditions on the basis of local subject characteristics. To implement this method, a novel image sensor capable of flexibly controlling shooting conditions individually for each separated area on a pixel array is a requisite.

This paper presents a Block-Wise-Controlled CMOS Image Sensor (BWC-CIS) prototype with 1.1 million pixels separated into 272 shooting-condition-controllable blocks. We conducted a simulated shooting experience of the developed sensor by using a real-time scene analysis and feedback control system, demonstrating enhancement in subjective image quality achieved through area-based optimisation of shooting conditions.

Related Works

Several methods have been proposed to tackle the challenges involved in enhancing the dynamic range of CMOS image sensors. For instance, several single-shot technologies employ a sensor with a dedicated pixel structure (Miyauchi et al [4] or Sakano et al [5]), others utilise a pixel-wise digital conversion strategy (Ikeno et al [6]) or a block-wise exposure control logic (Hirata et al [7]) in a sensor. Although the latter is particularly aligned with our approach, these methods present manufacturing difficulties due to their complex pixel structures or logic circuits, and difficulties in scaling up their configurations. In contrast, our method enables the development of a sensor with relatively simple configurations.

Furthermore, it introduces the unique concept of optimising image quality relative to data-rate by flexibly controlling not only dynamic range, but also resolution and frame rate, which provides ease of scalability.

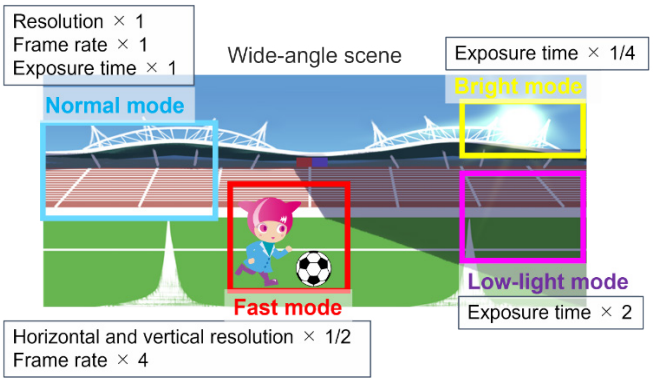


Figure 1: Overview of scene-adaptive imaging

Overview of Scene-Adaptive Imaging

The concept of scene-adaptive imaging is shown in Figure 1, which briefly describes how different areas of a scene can be captured with various shooting conditions (imaging modes) in a BWC-CIS, as following.

- **Normal mode** captures still and detailed subjects with standard resolution (maximum resolution based on the sensor's pixel array) and standard frame rate.
- **Fast mode** captures fast moving subjects with horizontally and vertically half-reduced resolution (reduced 1/4 in total) and four times faster frame rate.
- **Bright mode** captures highly bright subjects with one-fourth shorter exposure time.
- **Low-light mode** captures dark subjects with two times longer exposure time with two times lower frame rate.

The exposure time in Bright and Low-light modes exemplifies our current implementation, where both shorter and longer exposure times are adjustable. Note that the pixel readout rate remains consistent in Normal and Fast modes. Consequently, the proposed method addresses the sensor's limitation challenge by optimising local shooting conditions of resolution, frame rate, and dynamic range, which improves image quality without escalating data rates. To fully harness the potential of the BWC-CIS, the concurrent utilisation of a scene analysis and feedback control system dedicated for the sensor is needed. As shown in Figure 2, this system determines the optimal imaging modes of the sensor by analysing the subject's brightness distribution and movement with low latency and subsequently feeds back the information of determined imaging modes to the sensor. In the following sections, we provide details of the developed experimental imaging systems.

Block-Wise-Controlled Image Sensor

To evaluate the proposed method, we developed a monochrome BWC-CIS prototype grounded in CMOS image sensor technology (Tomioka et al (8)). As shown in Figure 3, our sensor comprises 1,024 × 1,088 pixels with a pixel pitch of 2.6 μm, a mode control circuit, pixel drive circuit, Analogue-to-Digital Converter (ADC) circuit, and output circuit. The right part of Figure 3 shows the pixel structure of our sensor, where each element circuit shares a pixel amplifier for four pixels (A, B, C, and D). Although this structure closely resembles typical CMOS image sensors, our sensor distinguishes itself by incorporating the mode control circuit and selection switches. These additional components can manage controlling pixel signal charge, reset, and readout of four pixels by the control signal regulated individually for each block of 64 × 64

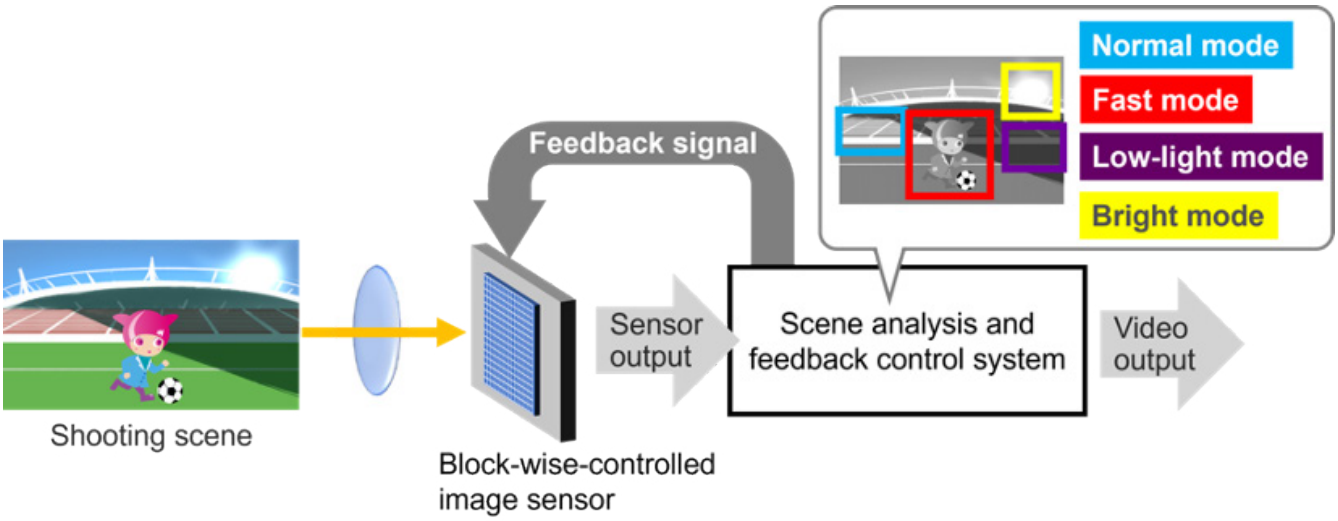


Figure 2: System configuration of scene-adaptive imaging

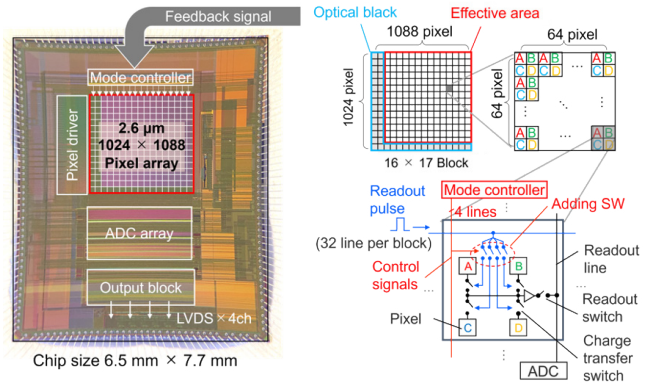


Figure 3: Die photograph (left) and pixel architecture (right) of our sensor

pixels (control block). This configuration results in 16 × 17 control blocks across the pixel array, facilitating flexible control of imaging modes in 272 areas of the sensor, where the effective counts of pixels and control blocks are 960 × 960 pixels and 15 × 15 blocks, respectively, due to the presence of light shielding (optical black) pixels.

Figure 4 presents two types of pixel scanning methods employed in our sensor: (a) sub-pixel readout and (b) pixel binning readout. The former sequentially applies readout pulses to charge transfer switches for each A, B, C, and D pixel within a control block at 240 frames per second (fps) and reads out the pixel signal through a pixel amplifier, resulting in full resolution at 60 fps (64 × 64 pixels at 60 fps). Conversely, the latter reads out the signal after combining signals from all four pixels (2 × 2 pixel binning) in a single scan.

Although the exposure time per pixel scan is one-fourth compared with sub-pixel readout, the four-pixel binning compensates for the reduction in signal value. With this offsetting outcome, the frame rate can be quadrupled while the horizontal and vertical resolution is halved (32 × 32 pixels at 240 fps).

Table 1 and Figure 5 display the four imaging modes assignable in our sensor and their corresponding operational flow within a horizontal block line, respectively. Normal mode operation is achieved by using the sub-pixel readout, whereas Fast mode

Mode	Resolution (per control block)	Frame rate	Exposure time
Normal	64 x 64 pixels	60 fps	1/60 s
Fast	32 x 32 pixels	120 fps	1/240 s
Bright	64 x 64 pixels	60 fps	1/240 s
Low-light	64 x 64 pixels	30 fps	1/30 s

Table 1: Imaging modes assignable in our sensor

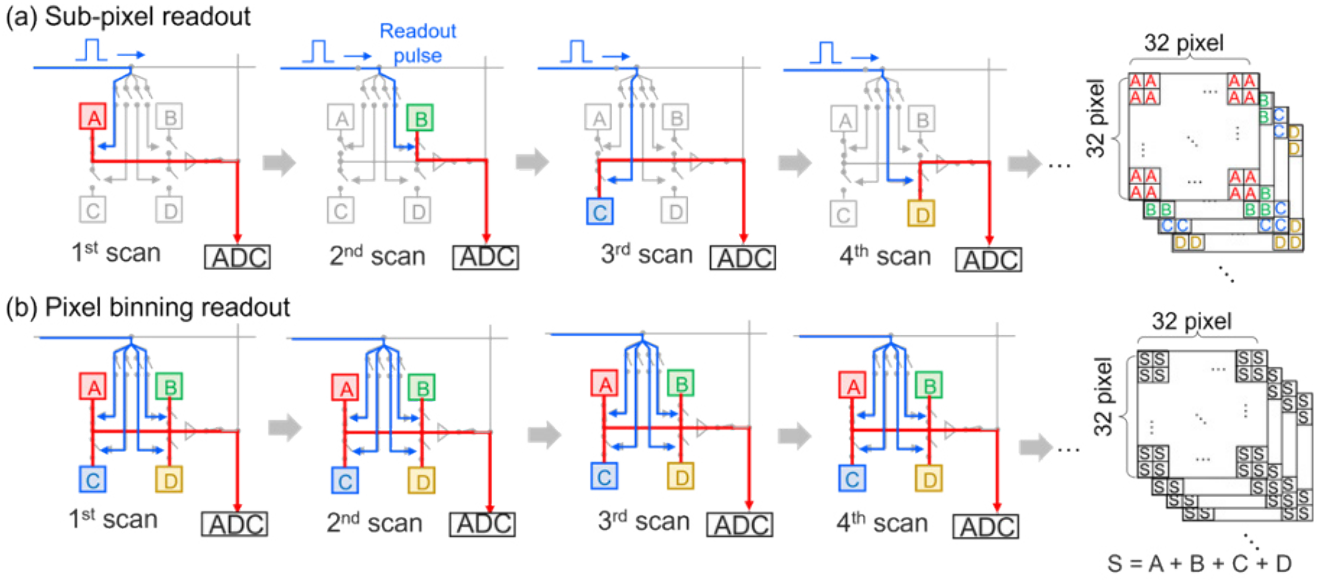


Figure 4: Readout methods of our sensor

employs the comparable pixel binning readout. The sub-pixel readout also facilitates in Bright mode, albeit with a sequential reset of A, B, C, and D pixel charges per scan, constraining each pixel's exposure time to 1/240 second to avoid pixel saturation. Low-light mode alternates sub-frame readout with a four-scan pause to extend exposure to 1/30 seconds (30 fps) to enhance pixel sensitivity. The scene-adaptive imaging is achieved by switching these operations for the entire control blocks in response to feedback signals.

Scene Analyses System

By integrating the fundamental imaging functions of the BWC-CIS prototype into Field Programmable Gate Array (FPGA) devices for real-time processing, we developed a scene-adaptive imaging experimental system. Figure 6 shows the comprehensive set-up of the system on the upper side and its photographs on the lower side, comprising the developed sensor installed on a sensor driving board, optics including a beam splitter, a sub-sensor (the Event Vision Sensor, described below), and two FPGA boards designed for scene analysis and feedback control and for video signal processing. This system was devised to analyse the brightness distribution and motion detection for subjects independently, and then merge the results to provide feedback signal to the BWC-CIS. Figure 7 shows the processing pipeline for the scene analysis processing. The followings detail specific functionalities of each component.

Brightness Distribution Analysis

The brightness distribution analysis is conducted by directly examining the BWC-CIS's output. The approach is outlined in the following, maximizing the exploitation

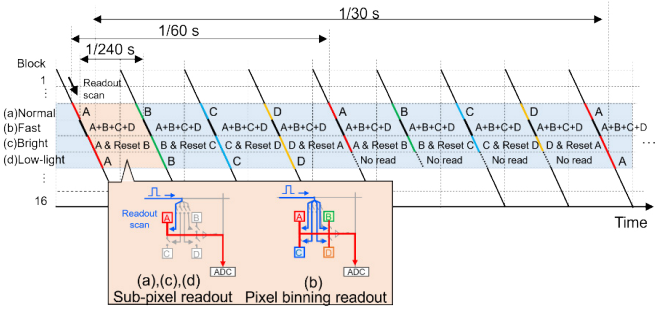


Figure 5: Operation flow corresponding imaging modes in our sensor

of the sensor's scanning method to achieve analysis with minimal latency.

1. Subdividing the control block signals from 3rd scan (sub-pixel C or S in Figure 4, 32 × 32 pixels) into smaller segments (8 × 8 pixels, totalling 16 segments).
2. Averaging and applying thresholding to each segment, resulting a brightness distribution map that categorises the segments into three labels (high, middle, and low).
3. Determining the optimal mode regarding exposure time (i.e. Normal, Bright, or Low-light mode) for each control block through a mode filter applied to the map (e.g., Bright mode is assigned to the block exhibiting the highest frequency occurrence of "high").

These series of processes are designed to be completed within 1/240 seconds in the FPGA after acquiring the signal from the 3rd scan.

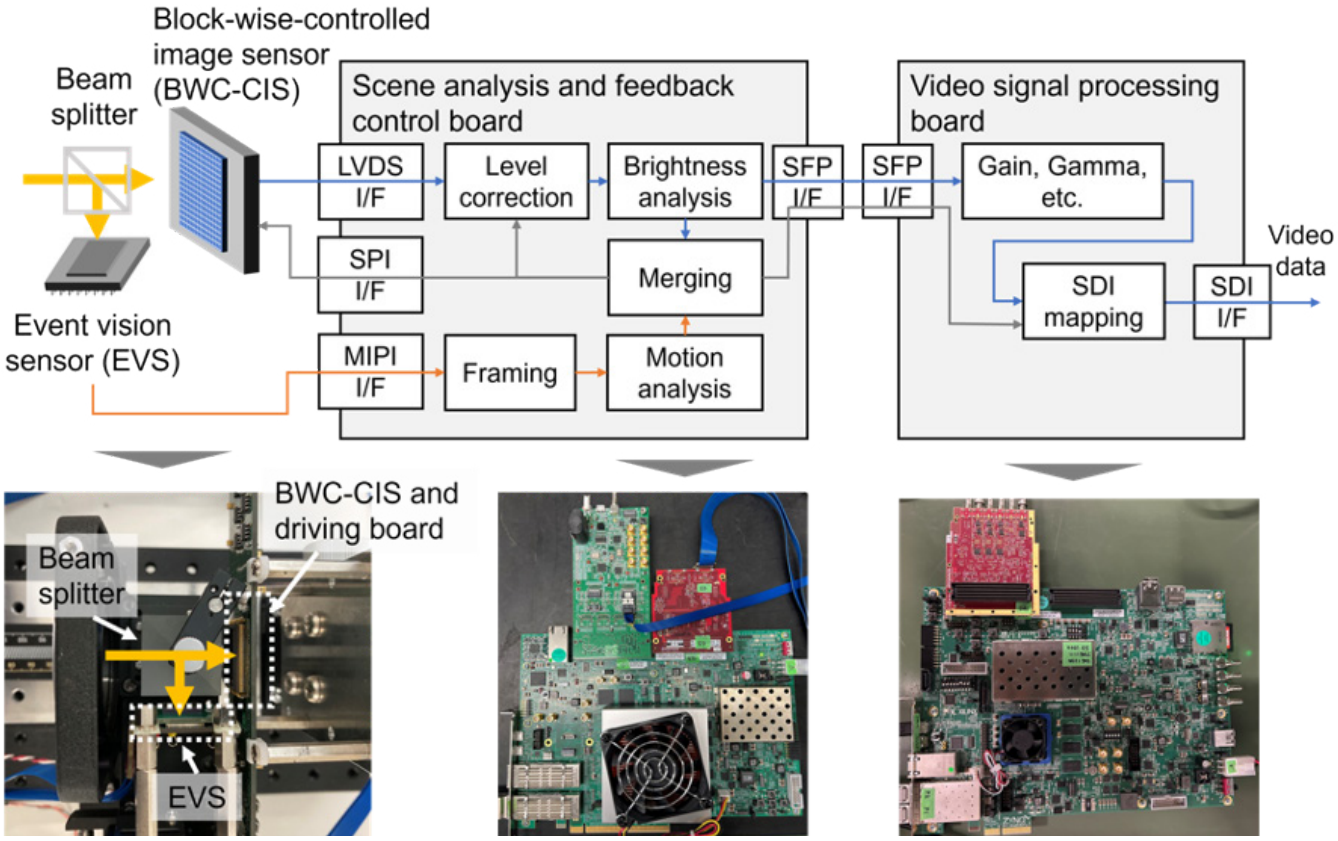


Figure 6: Block diagram (upper) and photographs (lower) of the experimental system

Motion Detection Analysis

Detecting the subject's motion domain with minimal delay poses a significant challenge. The widely-used motion detection methods, such as background separation (Piccardi [9]), require buffering multiple images for processing, resulting in delayed detection output. Given the system's need to consistently determine the subject's motion domain during shooting, minimising processing delays is crucial.

Therefore, as shown in the optics of Figure 6, we aimed to detect motion with the aid of the Event Vision Sensor (EVS, Finateu et al [10]) that is incorporated into the same light path of the BWC-CIS by using the beam splitter. An EVS detects pixel-level changes as events and transmits data in ultra-low latency, enabling us to promptly obtain cues for the subject's motion. The processing flow with the acquired event follows the steps following:

1. Accumulating events from the EVS for a 3rd scan period and generating binary images (1: with event, 0: without event).
2. Filtering out isolated events (eliminating false events arising from shot noise, etc.).
3. Calculating the total number of events for each segment corresponding to the control block of the BWC-CIS.
4. Applying thresholding to the control blocks and classifying them as a motion area (preferred area for Fast mode) or not.

Similar to the brightness distribution analysis, this procedure is executed within 1/240 seconds after accumulating events on an FPGA.

Finally, combining the preceding two analyses establishes imaging modes for the subsequent scans of A, B, C, and D pixels in the entire sensor, where the priority of selecting modes regarding brightness (Normal, Bright, and Low-light modes) or motion (Fast mode) can be set arbitrarily. This mode information is promptly transmitted to the sensor via feedback signals at the appropriate timing. The comprehensive operation enables continuous real-time updates of the sensor's shooting conditions every 1/60 seconds and enables it to respond to transitions in the local brightness or movement of the scene with a minimal delay. Additionally, our system performs a signal level correction of data from the BWC-CIS, compensating

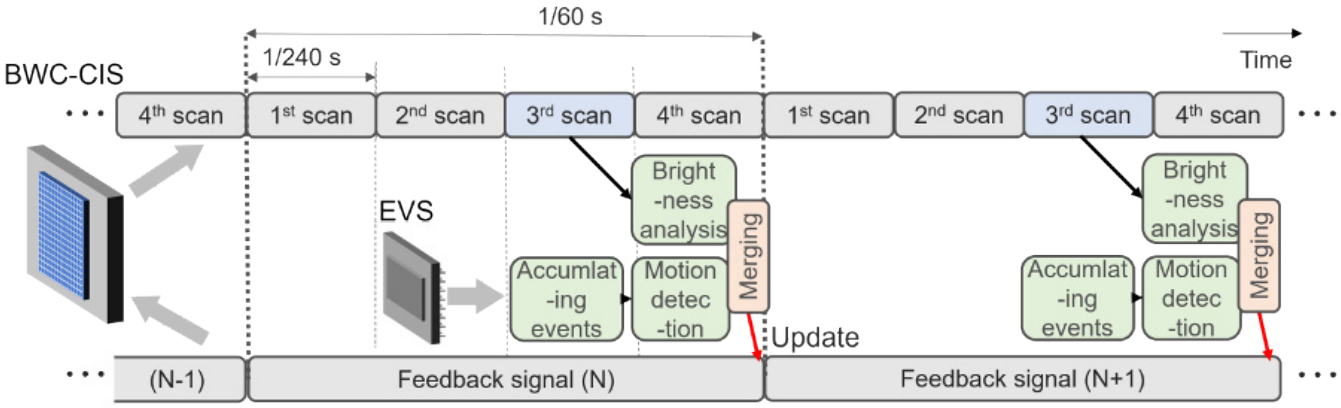


Figure 7: Processing pipeline of scene analysis and feedback control system

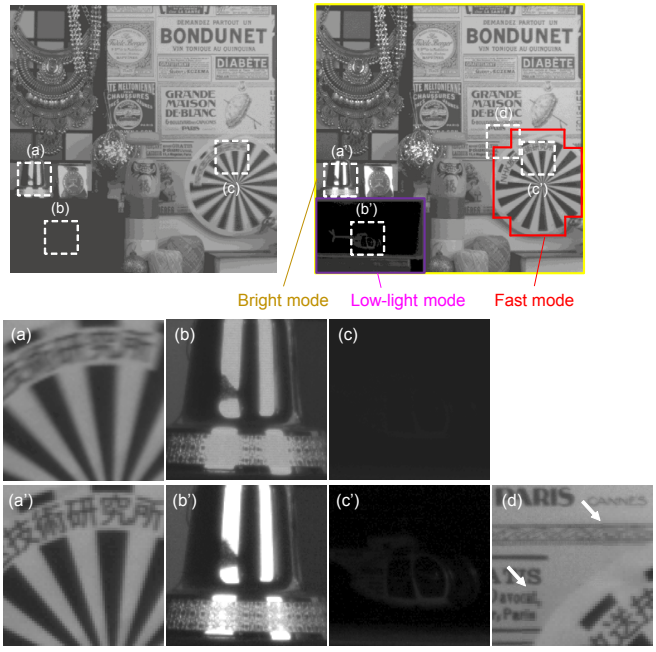


Figure 8: Comparison of images acquired by the proposed and conventional methods and their mode settings (upper) and comparison of enlarged images (lower).

for sensitivity differences due to variations in exposure times between Normal, Bright, and Low-light modes by multiplying the pre-calculated correction coefficients.

Experimental Results

Shooting experiments were conducted using the developed system to assess its performance. The experiment involved still subjects with varying local luminance and rotating radial charts. Images were acquired using a lens with an f-number of 4.0 and a focal length of 17 mm, and were recorded using an uncompressed serial digital interface (SDI) recorder.

Since the spatial-temporal mixed sample structures of the raw data from our sensor made it difficult to handle as a conventional raster image, it was rendered into a final output format of 1K × 1K pixels at 240 fps video, wherein areas captured in Fast mode had their horizontal and vertical resolution doubled, while those in other modes were interpolated four times finer in temporal resolution. For comparison, the experiment also included images captured when all control blocks were set in Normal mode to simulate the conventional imaging method.

Figure 8 shows the imaging results. With the conventional method, a reduction in spatial resolution due to motion blur was observed in area (a), where the rotating subject was captured. Additionally, clipped whites and crushed shadows were observed in areas (b) and (c), where bright specular reflection subjects or subjects hidden in the shadows were located, respectively. On the other hand, the results of proposed method demonstrated that shooting conditions were suitably adjusted for each subject, as depicted in Figure 8. This resulted in improvements in motion blur in region (a'), even though the spatial resolution is reduced in Fast mode, as well as improvements in overexposure and underexposure in regions (b') and (c'), respectively. However, it can be observed that the proposed method may result in blocking artifacts at the imaging mode boundaries, as shown in region (d). To address this issue, potential solutions include developing CIS with finer control blocks or implementing post-processing techniques to smooth the boundaries.

It is essential to emphasise that the sensor's output rates are the same for both the proposed and conventional methods. Consequently, although there is room for improvement in resolution and the granularity

of control block size in our sensors, the experimental results indicate its advantages in enhancing subjective image quality. This also indicates that the performance of CMOS image sensors can be drastically extended in a relatively straightforward manner to meet the growing demand for stringent specification requirements in immersive video content production.

Conclusions

We proposed a scene-adaptive imaging method designed to overcome the limitations of conventional image sensor technology by adjusting shooting conditions on the basis of local subject features, with the aim of achieving high resolution, high frame rate, and high dynamic range simultaneously. To validate the principle of our approach, a block-wise-controlled image sensor was developed, capable of selecting different resolutions, frame rates, and exposure times in units of 64 × 64 pixels over a 1K × 1K pixel array.

Through imaging experiments combining this sensor with a real-time scene analysis system, we confirmed the enhancement in subjective image quality achieved by locally optimising shooting conditions on the basis of subjects. Our next objectives include developing a practical sensor with higher resolution and more flexible imaging adjustability, as well as constructing a wide-field camera system for high-quality immersive video production.

References

1. ITU-R, 2019. Video parameter values for advanced immersive audio-visual systems for production and international programme exchange in broadcasting. Recommendation ITU-R BT.2123-0.

2. El-Desouki M., Deen M. J., Qiyin F., Liu L., Tse F., and Armstrong D., 2009. CMOS Image Sensors for High Speed Applications. Sensors 2009. 9(1). pp. 430 to 444.

3. Kawahito S., 2018. Column-Parallel ADCs for CMOS Image Sensors and Their FoM-Based Evaluations. IEICE Trans. Electron., vol. E101-C, no. 7. pp. 444 to 456.

4. Miyauchi K., Okura S., Takayanagi I., Nakamura J., and Sugawa S., 2019. A High Optical Performance 2.8μm BSI LOFIC Pixel with 120ke-FWC and 160 μV/e- Conversion Gain. Int. Image Sensor Workshop, 2019. R30.

5. Sakano Y., Toyoshima T., Nakamura R., Asatsuma T., Hattori Y., Yamanaka T., Yoshikawa R., Kawazu N., Matsuura T., Iinuma T., Takahiro T., Watanabe T., Suzuki A., Motohashi Y., Asami J., Tateshita Y., and Haruta T., 2020. A 132dB Single-Exposure-Dynamic-Range CMOS Image Sensor with High Temperature Tolerance. IEEE Int. Solid-State Circuits Conf., 2020. pp. 106 to 108.

6. Ikeno R., Mori K., Uno M., Miyauchi K., Isozaki T., Takayanagi I., Nakamura J., Wuu S., Bainbridge L., Berkovich A., Chen S., Chilukuri R., Gao W., Tsai T., and Kiu C., 2022. A 4.6-μm, 127-dB Dynamic Range, Ultra-Low Power Stacked Digital Pixel Sensor With Overlapped Triple Quantization. IEEE Trans. Electron Devices, vol. 69 no. 6. pp. 2943 to 2950.

7. Hirata T., Murata H., Matsuda H., Tezuka Y., and Tsunai S., 2021. A 1-inch 17Mpixel 1000fps Block-Controlled Coded-Exposure Back-Illuminated Stacked CMOS Image Sensor for Computational Imaging and Adaptive Dynamic Range Control. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers, Vol. 64. pp.120 to 122.

8. Tomioka K., Kikuchi K., Usui T., Kitamura K., and Kawahito S., 2023. Feedback Control of a Block-Wise-Controlled Image Sensor Based on Brightness Distribution Analysis. Int. Image Sensor Workshop, 2023. p.25.

9. Piccardi M., 2004. Background subtraction techniques: a review. Proc. IEEE Int. Conf. Systems, Man and Cybernetics, vol. 4. pp. 3099 to 3104.

10. Finateu T., Niwa A., Matolin D., Tsuchimoto K., Mascheroni A., Reynaud E., Mostafalu. P., Brady F., Chotard L., LeGoff F., Takahashi H., Wakabayashi H., Oike Y., and Posch C., 2020. A 1280×720 Back-Illuminated Stacked Temporal Contrast Event-Based Vision Sensor with 4.86μm Pixels, 1.066GEPS Readout, Programmable Event-Rate Controller and Compressive Data-Formatting Pipeline. IEEE Int. Solid-State Circuits Conf., 2020. pp. 112 to 114.

Project Timbre: How well do Mobile Networks Work for Live Audio Streaming?

A. J. Murphy, S. D. Elliott
British Broadcasting Corporation, United Kingdom

Abstract
In Project Timbre, data is being collected from around twenty standard, off-the-shelf mobile phones to examine the performance of live audio streaming over today's mobile networks. The project focusses on assessing the real-world quality of experience and investigates how this correlates with existing definitions of mobile coverage, which may be more general in nature.

The analysis presented uses a statistical approach based on the collected data to examine the probability of a given level of service, much as is done today for broadcast networks. It shows that there can be a marked difference between so-called 'signal coverage' and 'service coverage' and has identified the 'Time to First Byte' as a key metric of interest for live audio distribution over mobile networks.

This paper also makes the case for using better data to enable continued dialogue between content providers, the mobile industry and regulators to both optimise live audio streaming and to better inform listener expectations.

Introduction

Audio apps from broadcasters give access to the widest possible range of both live radio stations and on-demand content. Listening live and on the move, for example as a pedestrian or in a car, is almost certainly delivered by mobile networks.

But how well do mobile networks work for live streaming?

In the European Union, 4G (LTE) coverage is already termed 'near-universal' [1], reaching 99.8% of all households. However, this headline figure can belie the real-world experience when further away from home. In the UK, 93% of the landmass and 98% of motorways & main ('A') roads have 4G coverage from at least one Mobile Network Operator (MNO) [2], reducing to 71% and 69% respectively from all operators. Schemes such as the Shared Rural Network [3] are expected to see this rise to 95% of the landmass from at least one operator (84% from all operators) by the end of 2025.

But what does 'coverage' mean in the real world for listeners' Quality of Experience (QoE) in the context of audio apps?

Illustrating a wider held desire to find out more about listeners' QoE over mobile networks, the UK Government's 2022 Digital Radio and Audio Review [4] made recommendations for further work on mobile networks and radio streaming, specifically:

R32: Industry should work closely with Mobile Network Operators to promote the build-out of robust mobile data networks (5G) and deliver on-demand, streamed listener services focused on in-car listening.

R33: … radio broadcasters, transmission providers and Ofcom should initiate a programme of field-testing and trials to review and validate the … findings on 4G/5G coverage. The results of this testing should be discussed with Ofcom to ensure they include in their Connected Nations reporting, a measure appropriate for reliable radio/audio streaming.

MNO	Hours	Roads, km (% of total)		Rail, km (% of total)
		Motorways	A + B	
A	3,300	2,900 (77%)	13,800 (26%)	1,850 (14%)
B	1,950	2,200 (59%)	7,450 (14%)	1,250 (10%)
C	3,350	2,500 (66%)	10,500 (20%)	1,350 (10%)
D	1,500	1,500 (40%)	5,000 (10%)	1,000 (7.5%)

Table 1: The long and winding roads: length of transport routes surveyed (and as a percentage of UK total) over 32 months

Project Timbre is starting to investigate these issues by focussing on the real-world Quality of Experience (QoE) for listeners using the BBC's own audio product, BBC Sounds. It concentrates on live radio as the most challenging use case, since this requires a constant, reliable internet connection to successively download audio segments as they are created i.e. it is not possible to download live audio via another means before it has happened. However, the concepts being explored can also apply to streaming of on-demand content.

This paper sets out the work carried out in the project so far and presents some preliminary thoughts and findings as well as ideas for future work.

Quality of Experience

Broadcasters have a high confidence in the QoE delivered by conventional broadcast radio transmitter networks. These are downlink-only and dedicated to live audio delivery, being built for the specific requirements of broadcast. They deliver high audio quality with low latency and few interruptions i.e. 'broadcast quality'.

Mobile networks, in contrast, are multi-purpose and built to satisfy a broad range of simultaneous uses and requirements. These bi-directional IP networks enable the full range of content – both live and on-demand – to be made available, while offering the potential for new experiences and personalisation. Here, it is harder to have confidence in the QoE since it varies depending on many factors such as time of day (e.g. rush hour vs. night-time) and location (e.g. city centre train stations vs. countryside). It is also affected by the complex interaction between the dynamic nature of the network and the response of the algorithms in the playback client to those varying characteristics.

However, it can be anticipated that listeners will expect a similar quality of experience for streamed live audio as that provided by broadcast.

A plethora of sources of mobile coverage information is already available to the public, such as coverage checkers, maps based on predictions and signal strength thresholds or speed-test data. However, the variability outlined above means that these existing sources of information may describe the signal coverage but do not always accurately convey the QoE for audio streaming or indeed the service coverage for any specific application. This can make it difficult for the user to know what to expect.

In comparison to broadcast networks, the bi-directional nature of mobile networks offers an opportunity to collect feedback on performance and to harness data to better understand the real-world performance of services in relation to predictions of mobile coverage as well as to optimise products like BBC Sounds to improve QoE in the mobile environment.

Data Collection

In *Project Timbre*, an augmented, private prototype of the Android version of BBC Sounds has been developed. This has been deployed internally to engineers who are using it on standard, off-the-shelf handsets to collect real-world QoE data over mobile networks as they are today. With no need for special test equipment or targeted drive-test campaigns, a significant amount of data has been cost-effectively logged from the everyday trips of a small number of engineering staff (Table 1).

While the phones are not calibrated, they capture real-world experience data, which is used to explore variations due to location and time of day and to identify metrics having the greatest impact on QoE. The data also helps to enable better collaboration with wider industry to improve QoE.

Data is collected at various points in the distribution chain (Figure 1).The live radio stream originates at the BBC, where it is encoded and packaged as a sequence of MPEG-DASH segments. These are delivered using HTTP via, Content Delivery Networks (CDNs) and a mobile network to the BBC Sounds app running on a smartphone (or, in future, perhaps a connected car). The sequence of discrete audio segments is reassembled into a continuous stream of audio to be played back to the listener. Four categories of data are being collected:

1. **QoE metrics:** These are derived from the audio playback client (the BBC's in-house Standard Media Player, which in turn is built on ExoPlayer);

2. **Audio Delivery metrics:** These are provided by the underlying HTTP library used to download each of the audio segments that constitutes the live stream;

3. **Network Quality metrics:** The Android Telephony APIs enable the collection of detailed information about the mobile network in use, its signal strength, signal quality and the primary serving cell identifier; and

4. **CDN View:** Logs from the CDNs give a server-side view of what is happening with the connection and the delivery of each individual audio segment.

This data is recorded second-by-second and logged against location and time.

The client-side data (1, 2 and 3 above) is collected using Message Queue Telemetry Transport (MQTT) messages that are delivered to a back-end Influx database. Local storage in the client is used to ensure that data is captured everywhere, including in areas with no, or insufficient mobile signal, with messages sent to the database once connectivity returns.

Grafana dashboards monitor and debug the real-time system. A snapshot of this (Figure 2) depicts playback state (top row, map left & graph centre), behind live latency (bottom row, graph centre), a timer to measure the delivery of individual audio segments (top row, map right) and mobile signal metrics (bottom row, map left & graph right).

Data Insights

The ultimate listener-experienced QoE is defined by a number of factors such as latency, audio bit rate and both the frequency and duration of any audio interruptions. In order to better understand one of these areas, a simple proxy for QoE has been defined as the fraction of playing time (T_p) to overall listening time (T_l), or Playing to Listening Ratio (PLR):

$$PLR_{[\%]} = \frac{100\,T_p}{T_l} = \frac{100\,T_p}{T_b + T_p}$$

where T_b is the buffering duration when playback was interrupted.

Figure 1: Data collection in Project Timbre

As broadcasters and audiences typically seek uninterrupted playback for extended periods, a suitably high PLR target needs to be considered. The relationship between $PLR_{[\%]}$ and T_b for an hour's listening is shown in Table 2.

$PLR_{[\%]}$	T_b (s)
97.5	90
99.0	36
99.5	18
99.8	7.2
99.9	3.6

Table 2: $PLR_{[\%]}$ vs T_b for one hour of listening

Figure 3 shows that the $PLR_{[\%]}$ in the beginning of 2024 has remained above 97.5% across all MNOs. However, there is distinct variation over both time and between different networks.

This high-level view of QoE based purely on a time-series analysis inevitably disguises the variations according to the changing locations and environments of the population of devices, and the network performance experienced while logging. For example, the QoE is often poorer in trains than elsewhere due to greater signal attenuation of train carriages and railway cuttings compared with typical in-car reception.

To consider location, multiple data points are aggregated into 100m-by-100m pixels. Figure 4 shows the $PLR_{[\%]}$ for two different networks in west London. In general, the QoE has been very high. However, in some locations the QoE has clearly been consistently poorer than in others. Areas of poor QoE common across networks often indicate challenging local topography such as railway lines in cuttings while poor QoE occurring in different places from one network to another are more likely due to differences in network deployments. QoE can therefore vary from location to location and from network to network.

Variations in QoE between different mobile networks are important to be aware of as they create greater complications and potential confusion for both listeners and content providers alike. These are in sharp contrast to the QoE offered by a single broadcast network providing a near-universal service.

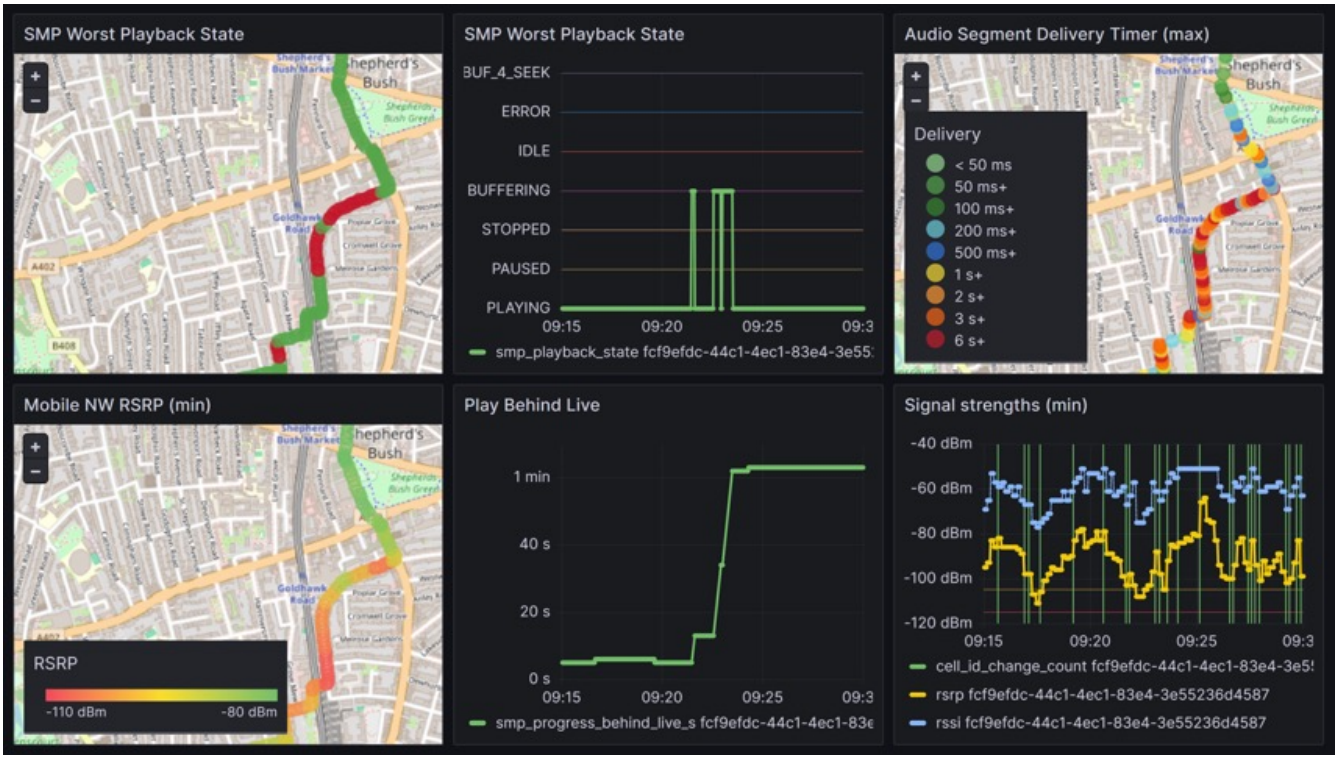


Figure 2: Real-time monitoring of a single measurement session

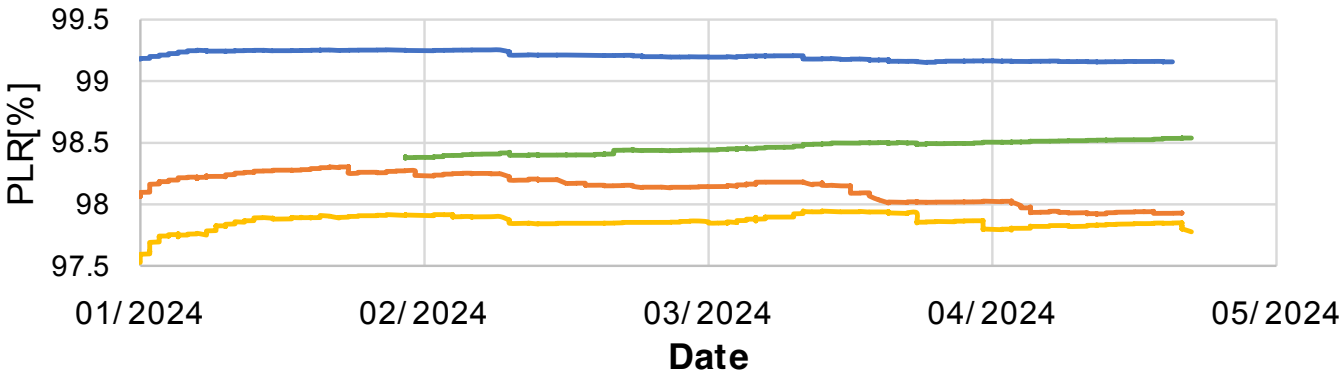


Figure 3: Overall $PLR_{[\%]}$ for the four physical UK mobile networks, 2024 to date

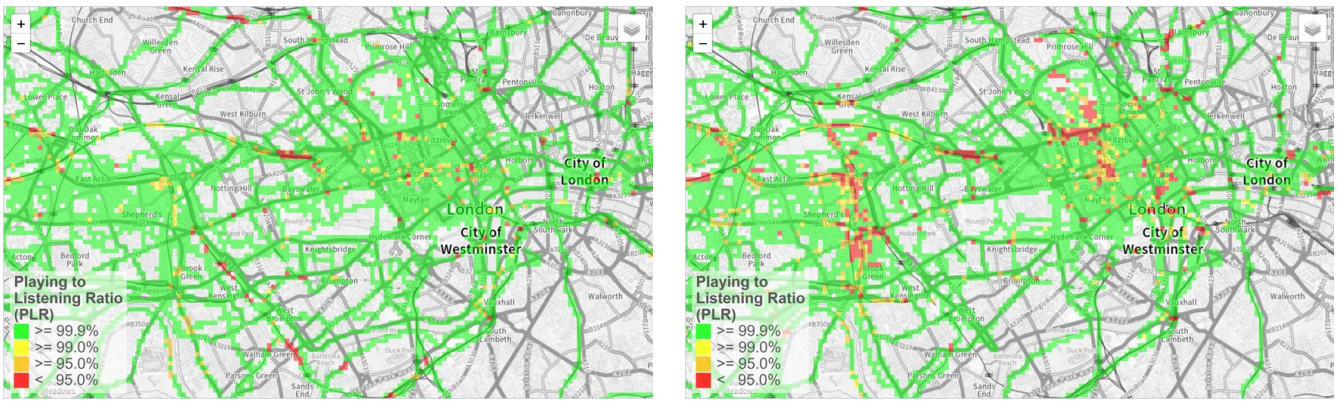


Figure 4: $PLR_{[\%]}$ variation over location for two different MNOs

When analysing delivery over mobile networks, both temporal and spatial data are therefore needed and can be used to promote realistic consumer expectations and for broadcasters in planning, optimising and promoting their app-based products and services, as well as informing broadcasters around the provisioning and deployment of future experiences.

Quality of Experience and buffering

While $PLR_{[\%]}$ is a useful proxy for the Quality of Experience, it is hard to measure objectively. The public version of BBC Sounds normally makes use of adaptive bit rate (ABR) streaming, choosing the most appropriate stream quality and codec depending on its view of a user's network bandwidth at any given time. However, to reduce the number of variables and to concentrate on the delivery of a minimum service, in Project Timbre the bit rate and codec are fixed at the lowest bit rate representation, namely HE-AACv1 at 48 kbit/s.

The audio segment duration on BBC Sounds is 6.4 seconds, chosen as a compromise between minimum latency, compatibility with early playback clients and the need to encapsulate an integer number of audio frames for seamless ABR switching. Since there is no variation of the playback speed, to play live audio uninterrupted – and to keep up with the live edge of the broadcast – the client player needs to request and download successive audio segments every 6.4 seconds, on average.

Fluctuations in the time it takes for each segment to be delivered over the network (the delivery time, $T_{delivery}$) are smoothed out by a buffer in the client app, the length of which effectively sets the averaging period over which the above 6.4 second criterion must be met.

Under 'normal' conditions, where the entire distribution chain (including, for example, CDN, mobile network and client) performs sufficiently well, successive audio segments will typically arrive within the order of tens of milliseconds. Such conditions enable uninterrupted playback with a very short buffer and low resulting latency. However, playback interruptions ('BUFFERING' events) do happen in practice, with the result that, when playback resumes, the listener is subject to increased latency; they don't miss that vital goal being scored, but rather, they hear it later.

Figure 5: Impact of $T_{delivery}$ on QoE for four handsets, across four MNOs, same journey

Barring any erroneous operation of the BBC Sounds app itself – or being in an area without any mobile coverage – playback interruptions will be the result of excessive delay (i.e. more than 6.4 seconds) in the requested audio segment data being delivered.

The client's audio buffer decreases the chance of a delayed segment interrupting playback, and there is a relationship between the length of this buffer and the tolerance to excessively long delivery times. A longer buffer means that the client can wait longer for a segment to be delivered before an audio interruption occurs. However, this increased resilience comes at the cost of increased latency for the listener; increasing the risk of hearing about that goal from their friends before they've heard it for themselves.

The initial length of the client buffer at which playback starts is a design decision that needs to balance the less desirable aspect of increased latency on the listener experience against the initial resilience to excessive segment delivery times.

Further increases in latency due to playback interruptions allow the client to maintain an even longer buffer. Client algorithms are typically 'greedy', with any increase in latency an opportunity for players to gorge themselves on audio segments up to the latest available at the live edge, subject ultimately only to memory constraints on the device.

As a result, any prior interruptions to playback that result in increased latency for the listener – such as those caused by excessive segment delivery times brought about by poor network coverage – result in a longer buffer and hence more resilience to subsequent segment delivery delays. There is further dependence of buffer length on how the user has controlled playback (e.g. pausing or rewinding).

The net effect is that different users will report different values of $PLR_{[\%]}$ depending on their past behaviour. A single user may even experience a different $PLR_{[\%]}$ in the same location on the return vs. the outward leg of their journey.

In summary, the listener's true QoE is a complicated product of many factors, initial client buffer length, including user interactions, exposure to previous network outages and segment delivery delays that introduce latency as well as the behaviour of the playback client in reaction to those. QoE may even vary depending on the programme content, with for example, someone listening to live sport being most sensitive to latency as outlined above.

Of interest, therefore, are occasions when the delivery time ($T_{delivery}$) of the current audio segment is excessively long since this has the potential to cause an audio interruption and hence a reduction in $PLR_{[\%]}$ (Figure 5). It also acts as a common currency, being neither dependent upon, nor impacted by, variation in the instantaneous length of the client audio buffer.

The 'BUF_4_SEEK' playback state is used to distinguish between buffering that occurs at the behest of the listener (i.e. when changing station, fast-forwarding, etc.) and buffering caused by delay in delivery somewhere in the underlying distribution chain.

Over the course of the thirty minutes depicted here, several 'BUFFERING' instances can be seen to occur, causing pauses in the audio. These are a direct result of certain audio segments having an excessively long delivery time and happen at different times on different mobile networks. The net effect is a spread of buffer lengths and resulting variation in audio latencies from around 7 to 24 seconds across the four handsets.

Segment Delivery Time In Detail

The time taken for the HTTP transaction for each segment is the sum of two components:

$$T_{delivery} = T_{TFB} + T_{transfer}$$

T_{TFB} is the time taken between the initial request and reception of the first byte of the response from the server (typically a CDN end-point), the so-called Time To First Byte and $T_{transfer}$ is the time taken to transfer the data itself, in this case the audio.

Figure 6 takes a closer look at the distribution of this segment delivery time ($T_{delivery}$) of 4G-only data logged for two networks in a single 100m-by-100m pixel. The median (50th percentile) $T_{delivery}$ for both networks is approximately 150 ms; even a very short client buffer would 'mop up' latency of this order, seemingly preventing audio interruptions.

However, $PLR_{[\%]}$ observed for the network depicted in blue exceeded 99% while it was less than 95% for the network depicted in amber. This is due to the amber distribution's long tails extending to the right, revealing that significantly longer latencies do occur. As the 99.8th percentile for the blue network was around 1.2 s, very few audio segments would have been delayed by more than a typical buffer length, enabling a high QoE. On the other hand, long segment delivery times were more common for the amber network, where the 99.8th percentile has been around 11.5 s. These more frequent, longer $T_{delivery}$ values, more often exceeded the buffer length, resulting in more frequent audio interruptions, and lower QoE.

Segment delivery time is therefore a key indicator of QoE, with the 99th or higher percentiles being of primary interest.

Figure 7 shows the distribution of a larger number of segment delivery times across multiple pixels, as the constituent T_{TFB} and $T_{transfer}$ elements. The occasional instances of complete download failure are captured in the bin labelled 'FAIL'.

Somewhat surprisingly, T_{TFB} dominates over $T_{transfer}$ i.e. it often takes longer for the first byte of the data request to arrive than it takes for the rest of the payload to be transferred. This may be due to the small audio segment size but requires further investigation.

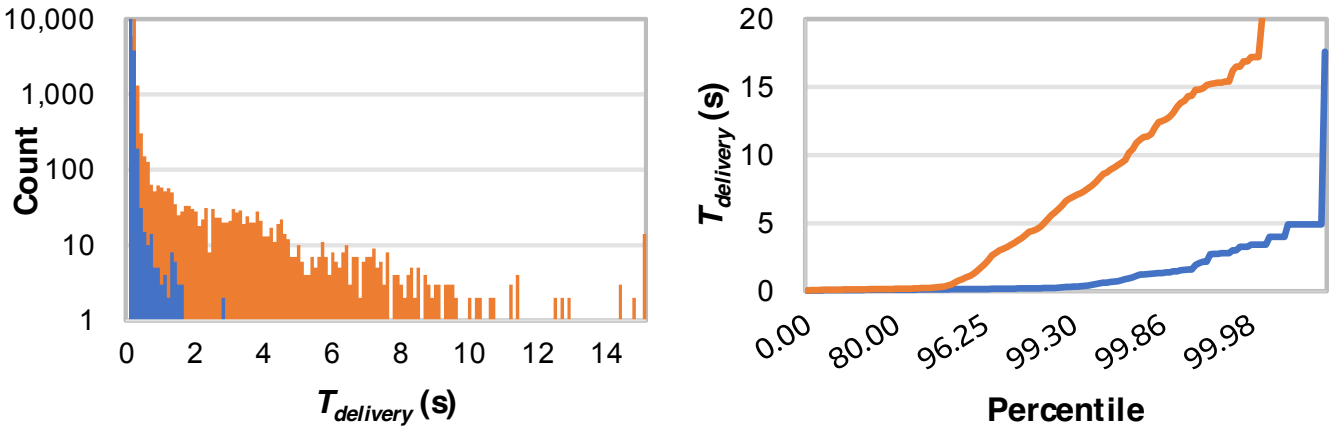


Figure 6: T_delivery (histogram left, cumulative right) in a single pixel, two MNOs, 4G only

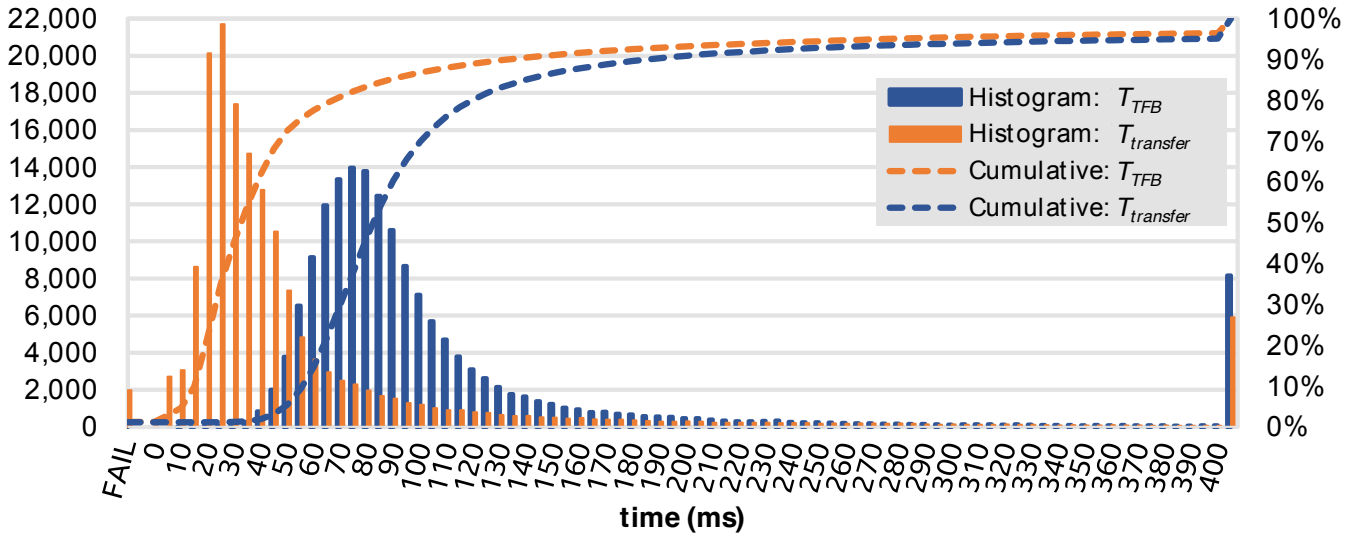


Figure 7: Distribution of T_{TFB} and $T_{transfer}$ for one MNO

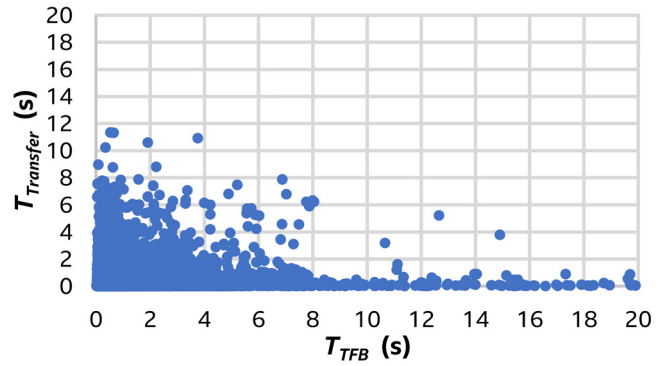


Figure 8: T_TFB vs T_transfer for one MNO, all data

Furthermore, there appears little correlation between them (Figure 8) i.e. longer T_{TFB} does not always imply longer $T_{transfer}$. This may be important for adaptive bit rate services where clients rely on accurate assessments of network bandwidth. Simply measuring transfer time ($T_{transfer}$) may not give a true assessment of the overall segment delivery time for audio and can result in a significant overestimate of connection bandwidth.

Network Metrics and QoE

Determining the causes of any poor QoE is of interest. In the UK both 2G and 3G remain in use, with the latter being phased out. Both older generations often show long segment delivery times compared with 4G (Figure 9) resulting in audio interruptions.

Although considerably better than 2G and 3G, unusually long segment delivery times are also observed on the 4G networks. The cause of many of these delays is as yet unclear.

To stream audio, a device must, at a minimum, be connected to a network with sufficient signal. Logging signal strength (RSRP) and network connection status has enabled the identification of locations where either one of these is untrue. Insufficient signal for a network connection has, as expected, been found to be the cause of poor QoE in some locations, particularly in remote rural areas with rugged terrain, but also in urban areas, albeit less frequently. Other locations, however, appear to have sufficient signal yet frequently suffer long segment delivery times. The expectation is that these areas are congestion limited.

Figure 10 depicts the relationship between signal strength and segment delivery times by mapping RSRP against the observed downlink bit rate (calculated over $T_{delivery}$) for two different networks. Although not shown, similar variation has been observed for the RSRQ. It demonstrates that increased signal strength is no guarantee of improved throughput (reduced $T_{delivery}$), especially at lower percentiles. For example, at the 2nd percentile the network depicted in the bottom chart provides a throughput of 132 kbit/s at -60 dBm, compared with a higher throughput of 536 kbit/s at -71 dBm RSRP.

These findings suggest that other factors not captured by these metrics – such as the aforementioned network congestion (in either the up- and/or downlinks), handover delays, and potentially operational issues – contribute to the performance observed.

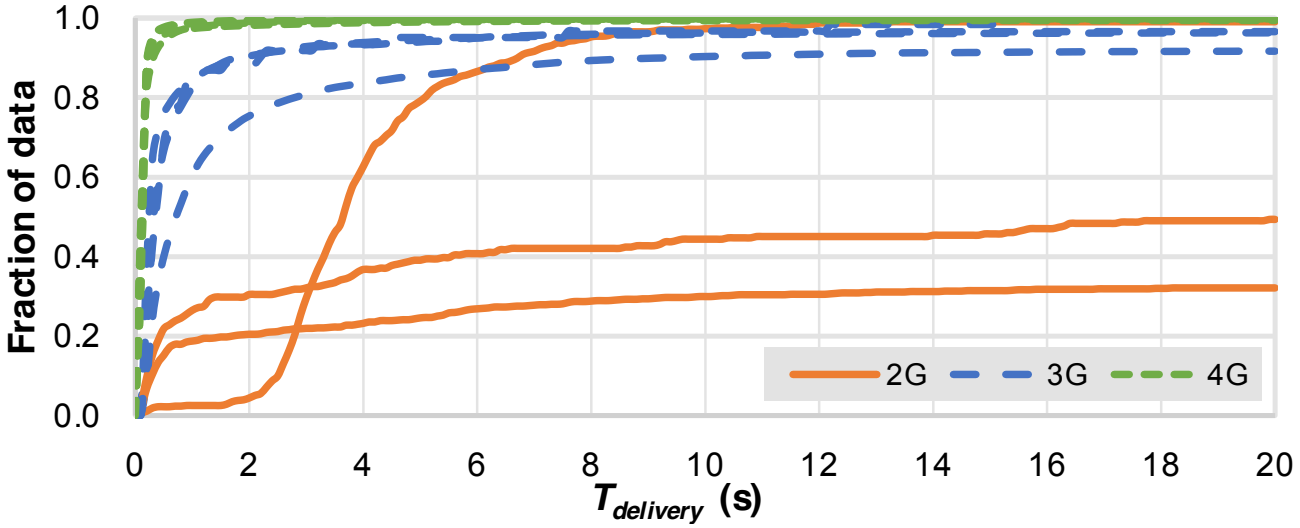


Figure 9: $T_{delivery}$ for the different generations of mobile technology

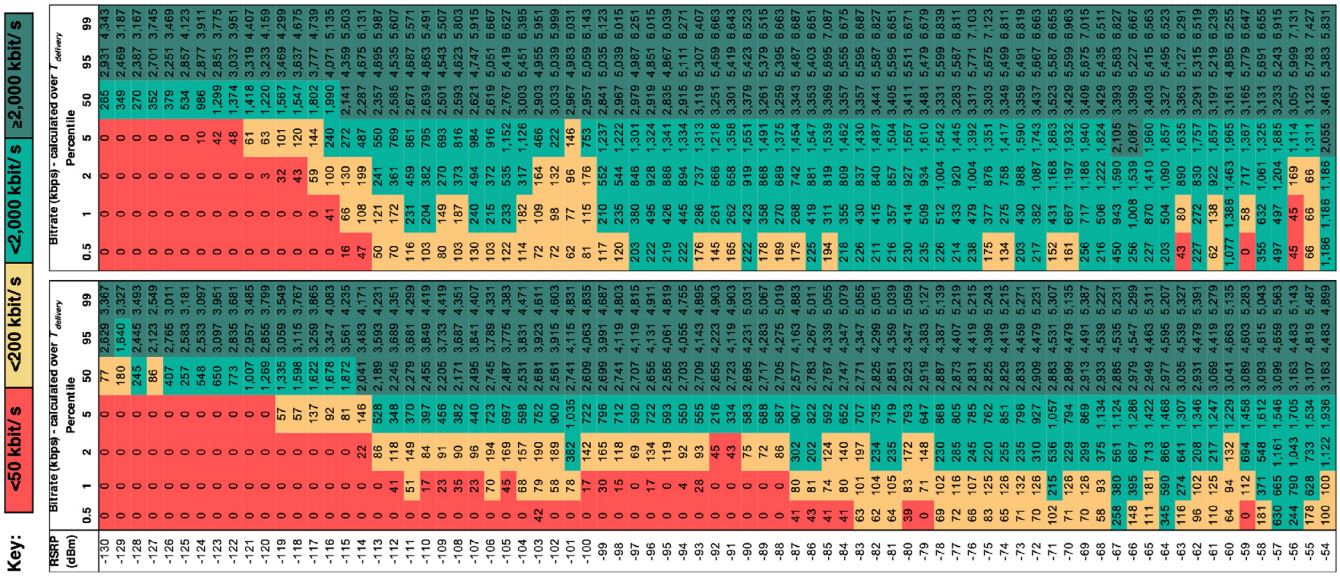


Figure 10: Measured Downlink speed vs. RSRP for two MNOs, RSRP <= -54 dBm

All that can be stated with any certainty is that signal strength is a necessary but not always sufficient requirement for adequately fast segment delivery times. This underlines the point that correlating the coverage available for live audio delivery with a given QoE – i.e. 'service coverage' – to maps or data based on signal strength alone – i.e. 'signal coverage' – appears fraught with error.

Of the other signal quality measures, SINR would be expected to provide the best insight into network performance. However, limitations of the telephony APIs available on Android has not made it possible to log SINR with enough regularity to test the relationship.

Conclusions

Listeners are consuming more and more content through audio streaming apps such as BBC Sounds. On the move, these are reliant on mobile networks for the delivery of the content and services. However, broadcasters don't yet have a detailed understanding of the delivered QoE in the same way they do for conventional broadcast networks; listeners meanwhile face the uncertainty of not knowing what level of expectation to have as to how well these services will work.

Project Timbre has begun to explore the role that real-world measurements and subsequent data analysis can have in addressing these issues, harnessing the inherent two-way nature of mobile IP networks to capture metrics directly from the test handsets in the project.

Some initial findings have been presented, including:

- Segment delivery time appears to be a useful proxy for QoE, that is independent of the length of the client buffer;
- The Time to First Byte is a significant element of the overall delivery time of each audio segment, and should be taken into account when estimating the effective bandwidth of the connection;
- Spatial and temporal aggregation of data reveals that QoE varies in both these dimensions, as well as from one network to another;
- The segment delivery time shows significant peak to mean variation. High QoE (of the type associated with broadcast network delivery) is associated with the tails of the associated distributions e.g. the 99th percentiles and above;
- It is possible to build-up a more complete picture of network performance, and to make comparisons with existing data sets, by the aggregation of measurements taken over an extended period of time; and
- Signal strength alone does not appear to reliably correlate with segment delivery time and, as such, there is a distinction between the 'signal coverage' and 'service coverage' offered by mobile networks.

More work is needed to identify the causes of, and to suggest solutions for, the issues set out in this paper. This is likely to be beyond both the resource and expertise of content providers alone. Similarly, mobile network operators do not have access to all the information required, including the performance of audio streaming services, themselves.

By using data to identify areas of concern, and to give better understanding of the real-world performance of mobile networks today, it is hoped to inform and encourage dialogue with the mobile industry – and other interested parties such as regulators – around how expectations of coverage for a given service can be better communicated to the user.

Future Work

The work on Project Timbre is ongoing and a number of areas of further work have already been identified. The current analysis and results are solely based on the use of HTTP/1.1 over TCP for the delivery of audio segments over the network to the client. While limitations such as head of line blocking are likely to be less significant for the serial delivery of live audio segments, the alternative congestion algorithms available with HTTP/3 over QUIC are worthy of study as well as Low-Latency DASH techniques.

There is also the opportunity to continue to optimise and monitor future performance improvements based on further analysis of the dataset, such as the choice of initial client buffer size and its impact on service resilience.

Finally, there is a need to develop a better understanding of the role that techniques in the latest 3GPP standards such as the 5G Media Streaming (5GMS) System [5] could play in improving QoE for the listener.

5GMS, specified by 3GPP in TS 26.512 [6], allows Content Service Provider applications to actively collaborate with 5G networks to jointly achieve better outcomes for all users of the network. For example, differing Quality of Service envelopes to support real-time streaming or background downloads can be provisioned within the 5GMS system for a media session. These are then referenced by a Media Session Handler running as a background service on the handset.

In its simplest form, an app such as BBC Sounds could initiate media session handling by requesting just a 3GPP Service URL at the start of the streaming session. Meanwhile, a more sophisticated 5GMS-aware application could take advantage of network assistance during the media streaming session by, for example, receiving asynchronous recommendations of the bit rate that can currently be delivered reliably by the 5G network. This could be fed by the UE application into the media player's ABR algorithm to vary the streaming quality or by asking for a short bit rate 'boost' to allow the client to replenish its playback buffer following a short drop in network capacity or a loss of signal.

References

1. European Commission. Broadband Coverage in Europe 2022: Mapping progress towards the coverage. September 2023. Clause 4.3.4.3.
2. House of Commons Library Research Briefing. Rural mobile coverage in the UK: Not-spots and partial not-spots. March 2024. Clause 2.2.
3. UK. Shared Rural Network.
4. UK Government. Digital radio and audio review. April 2022.
5. Gabin, F., Lohmar, T., Heikkilä, G., D'Acunto, L. and Stockhammer, T. 5G Media Streaming Architecture. Proceedings of the IBC. October 2019.
6. 3GPP. 5G; 5G Media Streaming (5GMS); Protocols. TS 26.512.

Acknowledgements

The authors would like to thank their colleagues across the BBC for their significant contributions to this work. They would also like to thank the International Broadcasting Convention for permission to publish this paper.

Large Multimodal Model-Based Video Encoding Optimization

Z. Duanmu, M. Jiang (zduanmu@imax.com, mjiang@imax.com)
IMAX Streaming and Consumer Technology (SCT), Canada

Abstract
In the realm of video encoding, achieving the optimal balance between encoding efficiency and computational complexity remains a formidable challenge. This paper introduces a groundbreaking framework that utilizes a Large Multi-modal Model (LMM) to revolutionize the process of per-title video encoding optimization. By harnessing the predictive capabilities of LMMs, our framework estimates the encoding complexity of video content with unprecedented accuracy, enabling the dynamic selection of encoding configurations tailored to each video's unique characteristics. The proposed framework marks a significant departure from traditional per-title encoding methods, which often rely on expensive and time-consuming sampling in the rate-distortion space. Through a comprehensive set of experiments, we demonstrate that our LMM-based approach not only significantly reduces the computational complexity required for sampling-based per-title video encoding – by an astounding 13 times – but also maintains the same level of bitrate saving. These findings not only pave the way for more efficient and adaptive video encoding strategies but also highlight the potential of multi-modal models in enhancing multimedia processing tasks. The implications of this research extend beyond the immediate improvements in encoding efficiency, offering a glimpse into the future of multimedia content distribution and consumption in an increasingly video-centric digital landscape.

Introduction

Adaptive streaming [1] has become the cornerstone of modern video delivery, enabling content providers to offer a seamless viewing experience across a wide range of devices and network conditions. This technology dynamically adjusts video quality during playback, based on the user's bandwidth and device capabilities, utilizing a predefined set of bitrate-quality pairs known as a bitrate ladder.

However, the traditional "one-size-fits-all" approach [2-4] to constructing these bitrate ladders often falls short. It fails to account for the unique characteristics of each video, leading to suboptimal use of bandwidth and a compromised viewing experience.

In response to these limitations, per-title encoding optimization [2] has emerged as a solution that tailors the encoding settings for each video title based on its content complexity. This method promises to significantly enhance the viewer's experience by optimizing the balance between video quality and file size. However, per-title optimization is computationally expensive [5]. It involves analyzing each video to determine its optimal bitrate ladder, a process that requires extensive computational resources and time. This complexity limits the scalability of per-title encoding, making it a challenge for content providers with large libraries of video content.

Recent efforts to streamline the per-title encoding process have explored the use of low-dimensional hand-crafted features such as Spatial Information and Temporal Information (SI/TI) [6] and regression models to predict the rate-distortion (RD) function [7-11], a key factor in determining optimal encoding settings. However, this approach suffers from two limitations. Figure 1 illustrates the R curves of two videos with similar SI/TI, from which we have two observations. First, low-level features are not a good representation of the encoding complexity, as they often overlook complex interplays of visual elements that significantly impact perceived quality. Second, multiple intersections between the two curves suggest that the encoding complexity lies in a high dimensional space. These limitations underscore the need for more sophisticated models that can better understand video content and predict encoding parameters.

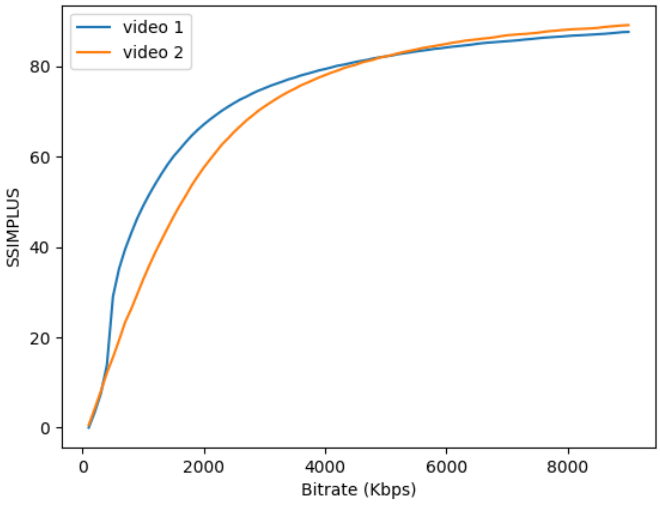


Figure 1: Rate-distortion curves of two videos with similar Spatial Information and Temporal Information

Against this backdrop, the promise of Large Multi-modal Models (LMMs) [12-14] offers a compelling solution. LMMs leverage advances in artificial intelligence to analyze video content across multiple modalities—combining visual, audio, and textual information—to understand content complexity comprehensively. This paper introduces a novel framework that utilizes LMMs for video encoding optimization, aiming to overcome the shortcomings of traditional per-title optimization methods. By harnessing the predictive power of LMMs, our proposed solution not only aims to reduce the computational expense associated with per-title encoding but also improves the accuracy of RD function prediction, leading to more efficient and viewer-centric adaptive streaming experiences. This approach signifies a paradigm shift in how video content is delivered, promising substantial improvements in streaming quality and resource utilization.

Background

Per-title Encoding Optimization

Per-title encoding optimization [1] represents a targeted approach in the realm of video processing, designed to tailor encoding parameters specifically for each video based on its unique content characteristics. This method manipulates additional encoding dimensions such as spatial resolution, ensuring that each video is encoded in a way that delivers the highest perceptual quality within a fixed bitrate budget. However, despite its effectiveness in enhancing viewer experience, per-title encoding optimization is notoriously expensive and time-consuming [2]. The process involves extensive sampling and analysis within the RD space for each video title, requiring significant computational resources.

This intensive approach, while beneficial for achieving optimal encoding settings, places a substantial burden on resources, making it a challenging endeavour for content providers who must manage large libraries of video content. Recent advancements in per-title encoding optimization have led to two notable approaches: one using curve fitting to reconstruct RD functions [15-18] and another predicting these functions based on low-level video features with regression models [7-11]. While these methods offer more efficient alternatives to traditional exhaustive sampling, they come with limitations. The computational complexity of curve fitting techniques, for instance, increases exponentially with respect to the dimensionality of encoding configuration. Similarly, predicting RD functions using hand-crafted low-level features such as spatial information and temporal information may overlook the impact of higher-level content attributes, such as narrative elements, objects, and texture type, on encoding efficiency. Furthermore, hand-crafted features may fail to capture all the nuances of the data, leading to suboptimal performance in video encoding optimization. These limitations underscore the ongoing need for more sophisticated models that can holistically account for the multifaceted nature of video content in the encoding process.

Large Multimodal Model

Large Multi-modal Models (LMMs) [12-14] have emerged as a transformative force in video understanding, harnessing the power of integrating multiple data modalities—text, images, and audio—to achieve a comprehensive analysis of video content. Their success is largely attributed to their ability to discern intricate details and contextual nuances within videos, which traditional single-modality approaches might miss [12]. This capability enables LMMs to perform exceptionally well in various video understanding tasks, including video retrieval [19], content classification [20], and activity recognition [21], thereby setting new benchmarks in the field.

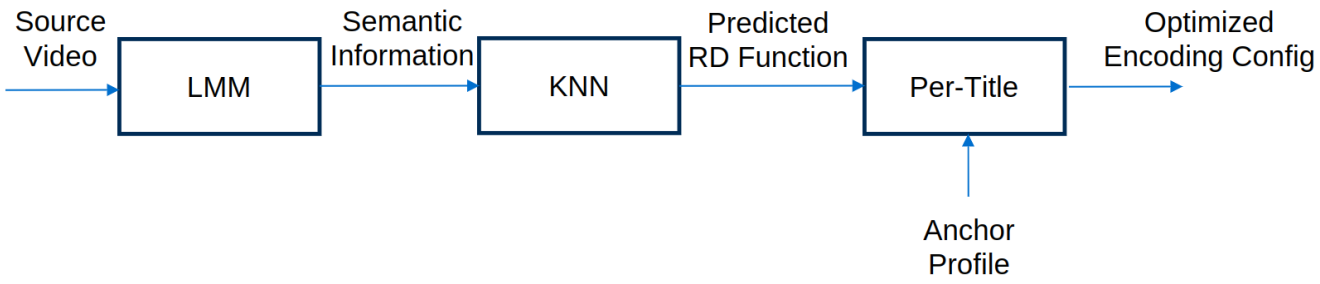


Figure 2. Proposed LMM-based video encoding optimization system

Proposed Framework

We initiate our discussion by delineating the foundational assumptions that underpin our framework. The first of these is predicated on the notion that videos sharing congruent characteristics will elicit analogous verbal descriptions. This assumption is deeply rooted in the theoretical understanding that language functions as an information compression heuristic [22], a concept that is well-documented within the research field. In practical scenarios, this is exemplified by numerous subjective video quality assessment datasets [23-25], which often classify video content based on observable characteristics such as the level of motion, texture, and camera movement.

Our second assumption extends from the premise that videos with akin complexity profiles are likely to demonstrate comparable RD behaviours. This underlying hypothesis forms the bedrock of regression-based RD models, where it is implicitly inferred that the intricacies of a video's content correlate directly with its RD function. The explicit recognition not only provides a more transparent foundation for the mathematical soundness, but also implies a more deliberate and methodical approach in the RD function modelling, as we will see in the subsequent section.

Building upon our foundational assumptions, we introduce an innovative framework, as depicted in Figure 2. At the genesis of this process stands the source video, which is ingested by a Large Multi-modal Model (LMM). The LMM is engineered to distil content characteristics from the video, translating complex visual and auditory data into a structured semantic representation. This semantic information, echoing our first assumption, encapsulates the notion that videos with shared characteristics can be uniformly described, compressing intricate content details into a descriptive language that mirrors human categorization.

Leveraging our second assumption—that videos with similar complexities will exhibit analogous RD functions—the semantic information is then fed into a k-nearest neighbours (KNN) algorithm. The KNN serves as a predictive tool that, using the distilled semantic descriptors, forecasts the RD function specific to the video. This function aims to map the relationship between bitrate and perceived quality, serving as a pivotal factor in the subsequent optimization steps.

The predicted RD function is then utilized in the Per-Title encoding stage. Here, the RD prediction is harmonized with an anchor profile, a predefined baseline of encoding parameters, to fine-tune the encoding process to satisfy the requirement from content distributors. This tailored encoding is crucial for transcoding the source video into the optimized encoding configuration, which is fitted to the video's unique content characteristics.

This framework embodies a strategic blend of linguistic theory and video analytics, transcending traditional exhaustive search and feature engineering and paving the way for a new era of content-specific encoding strategies that are both efficient and viewer-centric.

Experimental Results

Experiment Setup

Dataset

We use the Waterloo Generalized Rate-Distortion Dataset (WaterlooGRD) [18] in our experiment. The dataset contains 1000 pristine semantically coherent videos. All the sequences are downsampled to FHD (1920x1080 pixels), converted to 4:2:0 chroma subsampling, and temporally cropped to 10 seconds. The source videos are then encoded at 90 bitrate levels at each of the following spatial resolutions 1920×1080, 1280×720, 720×480, 512×384, 384×288, and 320×240 according to the list of Netflix certified devices [2]. We evaluate the quality of each video representation using SSIMPLUS [26] due to its demonstrated effectiveness in perceptual video quality prediction [19]. In the

end, we obtained 1000 generalized RD functions. The dataset is further segregated into three non-overlapping subsets: 490 for training, 210 for validation, and 300 for testing.

Competing Models

Our evaluation framework pits the proposed method against two predominant approaches in the realm of RD function reconstruction: sampling-based and feature-based methods. Within the sampling-based category, we benchmark against methods such as the piece-wise cubic Hermite interpolating polynomial (PCHIP), reciprocal regression [16], and logarithmic regression [15]. To boost the performance of these sampling-based models, our experiment utilizes an information-theoretic approach to sampling [18], designed to produce a sequence of samples that strategically reduces the uncertainty inherent in the RD function.

In the arena of feature-based approaches, our comparison extends to methodologies like SI/TI [6], and the Video Complexity Analyzer (VCA) [7]. For these methods, we employ an array of multi-layer perceptrons, honing them through the gradient descent algorithm on a designated training dataset. The optimal architecture is then chosen based on its superior performance against a set of pre-determined criteria on the validation dataset. This meticulous training and selection process ensures that our feature-based approach is finely tuned for accurately modelling the RD function.

Implementation Details

In our study, we selected the CLIP model [12] for its demonstrated robustness, efficiency, and adaptability to serve as the LMM. However, it is noteworthy that other LMMs could also be integrated into this framework. We specifically employ CLIP's vision component to distil semantic features from test videos. These extracted features, when derived from identical segments, are subjected to average pooling to consolidate them into segment-level features.

Consistent with the guidelines provided in [12], we employ cosine similarity as our metric for quantifying the likeness between feature sets. For the KNN classifier, we have determined that setting K to 3 yields the most favourable outcomes, as evidenced by the enhanced performance metrics observed within our validation dataset.

Evaluation Criteria

Our assessment of the RD function models encompasses two critical performance dimensions: the accuracy of the prediction and the bitrate savings achieved within a per-title encoding optimization framework. To gauge prediction accuracy, we calculate the Mean Absolute Error (MAE) by comparing the model-estimated RD functions against the ground truth for each piece of source content. Regarding bitrate savings, we utilize the predicted RD functions to inform and guide the per-title optimization process. This involves constructing an actual RD function reflective of the predicted convex hull at various bitrates and then calculating the Bjøntegaard Delta rate (BD-rate) [26] to quantify the bitrate efficiency relative to Apple's established bitrate ladder [3]. The performance is averaged across all content in the test set. The experimental procedure is repeated for 50 times with different training/validation split, and we report the median performance.

Performance in Prediction Accuracy

Table 1 details the performance of various competing models in predicting rate-distortion functions, as measured by the Mean Absolute Error (MAE). This measure quantifies the average magnitude of errors in the predictions, with a lower MAE indicating higher predictive accuracy.

The competing models are evaluated with different number of samples in the rate-distortion space: 0, 18, and 24. Two key observations can be made from the table:

Sample #	0	18	24
PCHIP	N.A.	7.81 ± 0.05	5.12 ± 0.03
Reciprocal	N.A.	9.67 ± 0.08	5.83 ± 0.07
Logarithmic	N.A.	3.61 ± 0.04	**2.24 ± 0.04**
SI/TI	2.49 ± 0.05	2.49 ± 0.05	2.49 ± 0.05
VCA	2.56 ± 0.04	2.56 ± 0.04	2.56 ± 0.04
Proposed	**2.32 ± 0.07**	**2.32 ± 0.07**	2.32 ± 0.07

Table 1: Performance of Competing Models with Different Number of Samples on Rate-Distortion Functions in Terms of Mean Absolute Error

Sample #	0	18	24
PCHIP	N.A.	15.1% ± 0.04%	16.3% ± 0.02%
Reciprocal	N.A.	15.3% ± 0.04%	15.9% ± 0.04%
Logarithmic	N.A.	**18.6% ± 0.04%**	**20.2% ± 0.04%**
SI/TI	13.5% ± 0.06%	13.5% ± 0.06%	13.5% ± 0.06%
VCA	17.2% ± 0.05%	17.2% ± 0.05%	17.2% ± 0.05%
Proposed	**18.6% ± 0.07%**	**18.6% ± 0.07%**	18.6% ± 0.07%
Offline Optimal	28.4%	28.4%	28.4%

Table 2: Performance of Competing Models with Different Number of Samples on Rate-Distortion Functions in Terms of Bitrate Saving

- The proposed model exhibits superior performance over all regression-based counterparts, underscoring the enhanced predictive power of LMM features. Most importantly, the improvement in performance presented by the proposed model is statistically significant.

- Remarkably, the proposed model achieves the best performance among all competing models even with 18 quality analysed encoding samples are available to the sampling-based models. At the same time, our model maintains parity with the best-performing sampling-based method, the Logarithmic model, even when 24 additional RD samples.

Overall, the table underscores the superiority of the proposed model in terms of prediction accuracy for RD functions, which is a pivotal aspect of optimizing video encoding processes.

Performance in Bitrate Saving

Table 2 provides a detailed comparative analysis of RD models in terms of bitrate saving in the context of per-title optimization. Alongside the competing models—PCHIP, Reciprocal, Logarithmic, SI/TI, VCA, and the proposed method – the table also introduces the 'Offline Optimal' result. This result represents an ideal scenario where each rate-distortion function in the dataset is known in advance, serving as a benchmark for the utmost bitrate saving achievable.

The results align with the patterns identified in the previous table, confirming the general trend observed earlier. Two key insights emerge from the analysis of the table. Firstly, the proposed method, even with zero RD samples, matches the bitrate saving performance of the state-of-the-art sampling-based algorithmic model at a sampling size of 18. This indicates that the proposed model, when integrated into a per-title optimization system, can drastically reduce computational demands while attaining equivalent levels of bitrate saving. Such efficiency suggests that the LMM method leverages its predictive capabilities to streamline the optimization process without compromising on performance outcomes.

Secondly, the proposed method demonstrates a statistically significant improvement over all regression-

Models	Computation Time (s)
PCHIP	84.003 ± 1.88
Reciprocal	83.512 ± 1.83
Logarithmic	83.027 ± 1.81
SI/TI	4.613 ± 0.50
VCA	1.126 ± 0.41
Proposed	6.047 ± 0.72

Table 3: Computation Time in Seconds

based models in terms of bitrate saving. This enhancement not only underscores the robustness and effectiveness of the proposed method but also highlights its superiority in optimizing video encoding parameters.

Computation Complexity

We evaluate the processing efficiency of various competing algorithms by examining their computational complexity. This is quantified by the average computation time in the context of per-title optimization. In the case of sampling-based approaches, the computation time encompasses the encoding at three different bitrate levels for each resolution, objective quality assessment tasks, and the fitting of RD curves. Conversely, for the regression-based method, the computation involves the feature extraction, and the estimation of the RD function. The assessment is conducted using 300 10-second 1080p videos as the source material, with an Amazon EC2 g5.2xlarge instance serving as the platform for benchmarking.

Table 3 presents the computation times for various competing models, measured in seconds, and includes a margin of error for each measurement. From the data, several observations stand out:

- Feature-based approaches demonstrate significantly higher speed compared to sampling-based methods, with the VCA model being the quickest among them. This distinction underscores the efficiency of feature-based models in processing video content.
- The proposed method showcases the capability to operate in real-time, as indicated by its computation time. This attribute makes it a viable option for applications requiring immediate video processing and encoding decisions.
- When juxtaposed with the results from the previous table, it is evident that the proposed method not only matches the Logarithmic model in terms of bitrate saving but also drastically reduces the computation time by approximately 13 times. This efficiency gain highlights the proposed method's advantage in offering substantial bitrate savings with significantly lower computational demand.

Overall, the table illustrates the computational efficiency of the proposed method compared to traditional sampling and feature-based approaches, establishing its potential for real-time applications and substantial computational savings without compromising on performance.

Discussion

Advantage of LMM

The integration of LMM into per-title encoding optimization presents a transformative approach to video processing, offering substantial advantages over traditional methods. At the heart of its benefit is the LMM's unparalleled ability to analyze and interpret complex video content at a granular level. Unlike conventional models that rely on basic features or extensive sampling, LMMs leverage deep learning to understand the nuances of video data, including visual elements, audio cues, and textual context. This comprehensive understanding enables the LMM-based framework to make more accurate predictions about the optimal encoding parameters for each video title. As a result, videos are encoded in a way that maximizes quality and efficiency, tailored to the specific characteristics of the content.

Moreover, the use of LMMs in per-title encoding optimization significantly reduces the computational overhead traditionally associated with video processing. By accurately predicting rate-distortion functions and encoding parameters from a deep, semantic understanding of the content, LMMs eliminate the need for brute-force sampling and testing across multiple bitrate and resolution settings.

Limitations and Challenges

A noteworthy observation is that while the proposed method's savings are impressive, they fall short of the 'Offline Optimal' result, which stands at a significant 28.4%. This gap indicates the scope of potential improvement and the ceiling of performance that could be aimed for in future iterations or enhancements of the model.

Another notable limitation of the current work is its performance relative to state-of-the-art sampling-based approaches, particularly when a high number of RD samples are available. In scenarios where extensive RD sample data can be utilized, sampling-based methods tend to outperform our feature-based approach, capturing nuances in video encoding optimization that our current model may overlook. This gap underscores the need for a more holistic framework that integrates the precision and depth of feature-based approaches, like the one presented here, with the comprehensive data utilization of sampling-based methods. Such a combined approach would ideally leverage the strengths of both methodologies, ensuring that the predictive accuracy and efficiency of the encoding optimization process are maximized across all scenarios.

Conclusion

Our work represents a significant step forward in the pursuit of more efficient and adaptive video encoding technologies. By harnessing the power of Large Multimodal Models, we have not only achieved substantial improvements in encoding efficiency but have also laid the groundwork for future innovations in the field of multimedia processing. As the digital landscape continues to evolve, we are confident that the insights and methodologies presented in this paper will contribute to the development of more sustainable, efficient, and user-centric video content delivery solutions.

References

[1] T. Stockhammer. Dynamic adaptive streaming over HTTP: Standards and design principles. In Proceedings of the ACM Multimedia Systems Conference, pages133–144, San Jose, CA, USA, Feb. 2011.

[2] A. Aaron, Z. Li, M. Manohara, D. J. Cock, and D. Ronca. (2015) Per-Title encode optimization. [Online]. Available: https://medium.com/netflix-techblog/per-title-encode-optimization-7e99442b62a2.

[3] Apple. (2016) Best practices for creating and deploying HTTP live streaming media for iPhone and iPad. [Online]. Available: http://is.gd/LBOdpz.

[4] G. Michael, T. Christian, H. Hermann, C. Wael, N. Daniel, and B. Stefano. (2013) Combined bitrate suggestions for multirate streaming of industry solutions. [Online]. Available: http: //alicante.itec.aau.at/am1.html.

[5] J. Dahl. (2018) Instant per-title encoding. [Online]. Available: https://www.mux.com/blog/instant-per-title-encoding.

[6] ITU-T P. 910. 1999. Recommendation: Subjective video quality assessment methods for multimedia applications. [Online]. Available: https://www.itu.int/rec/T-REC-P. 910-199909-S.

[7] V. V. Menon, C. Feldmann, H. Amirpour, M. Ghanbari, and C. Timmerer. VCA: Video complexity analyzer. In Proceedings of the ACM Multimedia Systems Conference, New York, NY, USA, pp. 259–264. Aug. 2022.

[8] H. Amirpour, P. T. Rajendran, V. V. Menon, M. Ghanbari, and C. Timmerer. Light-weight video encoding complexity prediction using spatio temporal features. In Proceedings of the IEEE International Workshop on Multimedia Signal Processing, Shanghai, China, pp. 1-6. Sep. 2022.

[9] A. V. Katsenou, M. Afonso, D. Agrafiotis and D. R. Bull. Predicting video rate-distortion curves using textural features. In Picture Coding Symposium, Nuremberg, Germany, pp. 1-5, Dec. 2016.

[10] A. Telili, W. Hamidouche, S. A. Fezza, and L. Morin. Benchmarking learning-based bitrate ladder prediction methods for adaptive video streaming. In Picture Coding Symposium, San Jose, CA, USA, December, pp. 325-329, Dec. 2022.

[11] K. S. Durbha, H. Tmar, C. Stejerean, I. Katsavounidis, and A. C. Bovik. Bitrate ladder construction using visual information fidelity. arXiv preprint arXiv:2312.07780. Dec. 2023.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and L. Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of International Conference on Machine Learning, Virtual Event, pp. 8748-8763. Jun. 2021.

[13] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. VideoBERT: A joint model for video and language representation learning. In Proceedings of International Conference on Computer Vision, Long Beach, CA, USA, pp. 7464-7473. Jun. 2019.

[14] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems. Vancouver, BC, Canada, pp. 13-23. Dec. 2019.

[15] C. Chen, S. Inguva, A. Rankin, and A. Kokaram. A subjective study for the design of multi-resolution ABR video streams with the VP9 codec. In Electronic Imaging, pp.1–5. Nov. 2016.

[16] C. Kreuzberger, B. Rainer, H. Hellwagner, L. Toni, P. Frossard. A comparative study of DASH representation sets using real user characteristics. In Proceedings of International Workshop on Network and Operating Systems Support for Digital Audio and Video, pp. 1-6. May 2016.

[17] Z. Duanmu, W. Liu, Z. Li, and Z. Wang. Modelling generalized rate-distortion functions. IEEE Transactions on Image Processing. vol. 23, no. 29, pp. 7331-7344. Jun. 2020.

[18] Z. Duanmu, W. Liu, Z. Li, K. Ma, and Z. Wang. Characterizing generalized rate-distortion performance of video coding: An eigen analysis approach. IEEE Transactions on Image Processing. vol. 29, pp. 6180-6193. Apr. 2020.

[19] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, Y. Qiao. Unmasked teacher: Towards training-efficient video foundation models. arXiv preprint arXiv:2303.16058. Mar. 2023.

[20] Z. Wang, K. Kuan, M. Ravaut, G. Manek, S. Song, Y. Fang, S. Kim, N. Chen, L. F. D'Haro, L. A. Tuan, H. Zhu. Truly multi-modal youtube-8M video classification with video, audio, and text. arXiv preprint arXiv:1706.05461. Jun. 2017.

[21] S. S. Kalakonda, S. Maheshwari and R. K. Sarvadevabhatla, Action-GPT: Leveraging large-scale language models for improved and generalized action generation. In Proceedings of the IEEE International Conference on Multimedia and Expo, Brisbane, Australia, pp. 31-36. Jul. 2023.

[22] G. Delétang, A. Ruoss, P. A. Duquenne, E. Catt, T. Genewein, C. Mattern, J. Grau-Moya, L. K. Wenliang, M. Aitchison, L. Orseau, and M. Hutter. Language modeling is compression. arXiv preprint arXiv:2309.10668. Sep. 2023.

[23] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, Z. Wang. A quality-of-experience index for streaming video. IEEE Journal of Selected Topics in Signal Processing, vol.11, no. 1, pp. 154-166. Sep. 2016

[24] S. Wang, A. Rehman, Z. Wang, S. Ma, W. Gao. SSIM-motivated rate-distortion optimization for video coding. IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 4, pp. 516-529. Sep. 2011.

[25] Z. Li, Z. Duanmu, W. Liu, Z. Wang, AVC, HEVC, VP9, AVS2 or AV1? – A comparative study of state-of-the-art video encoders on 4K videos. In Proceedings of the International Conference on Image Analysis and Recognition, Waterloo, ON, Canada, pp. 162-173. Aug. 2019.

[26] A. Rehman, K. Zeng, and Z. Wang, Display device-adapted video Quality-of-Experience assessment. in Proceedings of SPIE, San Francisco, CA, USA, pp. 939406.1-11. Mar. 2015.

[27] G. Bjøntegaard. Calculation of average PSNR differences between RD curves. Tech. Rep. VCEGM-33, ITU-TSG16/Q6, 13th VCEG Meeting. Telenor Satellite Services, Oslo, Norway. Apr. 2001.

Multi-Label Indexing Technology for News with Ai-Based Text Processing

Y. Yasuda[1], S. C. Clippingdale[2], T. Miyazaki1, J. Goto[1]
[1]NHK (Japan Broadcasting Corporation), Japan and [2]NHK Foundation, Japan

Abstract
Broadcast media organisations produce many news scripts every day for dissemination as content. Such text data is often reused in the process of producing TV programmes and web news. To efficiently utilise this much data, it is necessary to accurately attach metadata such as labels that indicate the content of the text. However, manually assigning labels takes an enormous amount of time and effort. With the aim of reducing costs, we have developed a system that automatically labels news articles. A major challenge in the multi-label text classification task in the news domain is known as 'imbalanced learning.' We proposed a novel loss function that utilises some weights and a label-smoothing technique to suppress label imbalance. Experimental results show that our method outperforms baselines. We introduce a prototype system based on our method as a test bed for content creation and discuss some of the results that it achieves.

Introduction

Much text data is utilised in the process of producing TV programmes and web news. To create media content efficiently, it is necessary to attach accurate metadata, such as labels that indicate the content, to large amounts of text. Metadata attached to text can enable producers to efficiently retrieve and use past material in the creation of new content and enable viewers to easily access articles that they want.

Conventionally, metadata indicating content has been added to text data manually. In recent years, the amount of text data that needs to be handled has grown very large, driving a need to develop technology that supports the task of metadata generation to reduce costs, particularly for producers and broadcast stations with limited numbers of staff.

We have developed a system that uses AI technology to automatically assign multiple labels representing genre and content to news articles. This system performs multi-label text classification based on neural networks and makes it possible to add accurate metadata to large amounts of text in less time than is required for conventional manual metadata addition.

The system was introduced and tested at two local broadcast stations. In one instance it was used to analyse the topics of news articles over a one-year period, and in the other instance, to add labels to news articles published online.

News articles cover a wide variety of topics in general, and the system must assign multiple labels to them. Since the labels vary from major genre areas down to quite detailed minor topics, their frequencies of appearance vary widely, and the labels for minor topics in the training data can appear extremely infrequently. It is known that when AI models are trained with datasets including labels that occur infrequently, the classification accuracy deteriorates. We therefore developed a learning algorithm that suppresses the influence of wide imbalances in label appearance frequencies and aimed to improve its performance.

Multi-Label Text Classification

Multi-label text classification is a key task in natural language processing that is applied to various situations in the real world [1], [2], [3]. Examples of real-world usage include classification of legal

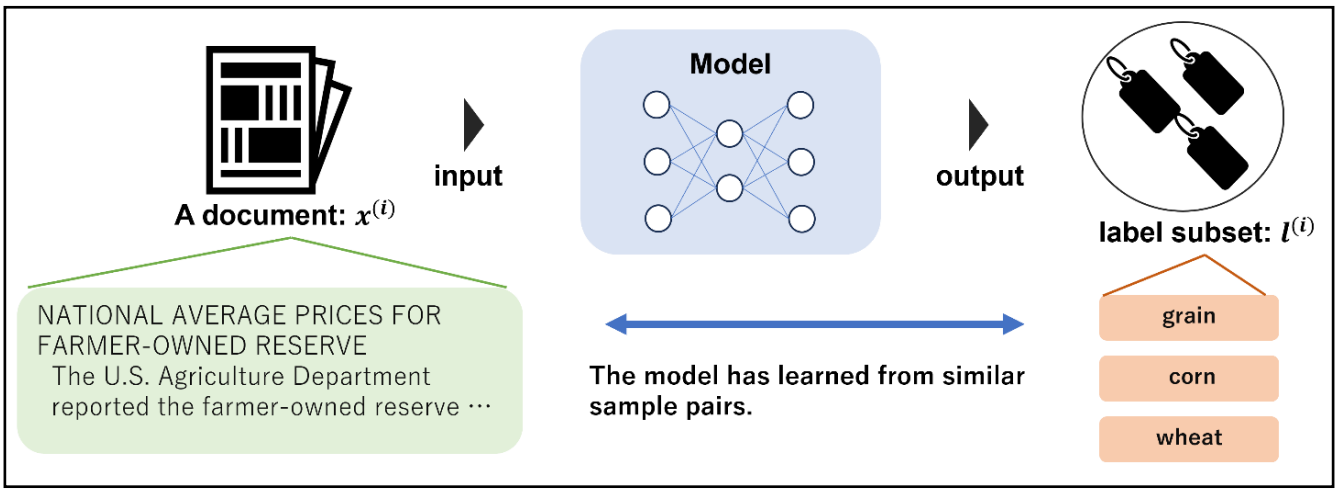


Figure 1: Conceptual image of the multi-label text classification task

documents [4] and automatic diagnosis through medical records [5]. In multi-label text classification, the task is to assign an appropriate label subset $l^{(i)}$ ($l^{(i)} \in \mathrm{L}$) to document $x^{(i)}$ ($x^{(i)} \in X$), where L is the set of predefined labels, X is the set of documents and i is the index of input samples. Fig. 1 shows the conceptual image of multi-label text classification.

Following rapid advances in machine learning technology in recent years, multi-label text classification is often tackled with neural networks. We also approach this task using a neural network paradigm.

Learning of A Model in Multi-label Text Classification

The task of multi-label learning is to learn from training data a function f: $\chi^{(i)} \rightarrow y^{(i)}$, where $\chi^{(i)}$ ($\chi^{(i)} \subset \mathbb{R}^d$) denotes a d-dimensional feature vector representing a document and $y_{\square}{}^{(i)} = \{y_1{}^{(i)}, y_2{}^{(i)}, \ldots, y_N{}^{(i)}\} \in \{0, 1\}^N$ denotes ground-truth values as a multi-hot vector of labels. A model outputs the estimated probability p_n of the n-th label, and the probabilities are fed into a loss function, which measures the loss as the difference between the model's output pn and the ground-truth value $y_n{}^{(i)}$. For the sake of simplicity, we omit (i), the index of a sample, in the rest of this paper.

Binary cross entropy (BCE) is a widely used loss function in the field of multi-label text classification. BCE is defined for one sample as

$$\mathrm{BCE} = -\sum_{n=1}^{N} [y_n \log p_n + (1 - y_n) \log(1 - p_n)] \quad Eq.(1)$$

Fig. 2 shows a conceptual image of the model learning with BCE loss for one sample. The probabilities p_n as outputs of the model and the ground-truth values y_n are input to the BCE computation. The ground-truth values y_n comprise a multi-hot vector defined on the label subset. If the label corresponding to the n-th position is assigned to the input document, y_n becomes 1; otherwise it is 0. Therefore, if the label is positive for the sample, only the term log p_n is summed into the final loss value BCE in Eq. 1. On the other hand, if $y_n = 0$, the term log $(1 - p_n)$ is summed into the final loss value. The terms log p_n and log $(1 - p_n)$ represent the loss associated with differences between the probabilities produced by the model and the ground-truth values. BCE calculates the loss for each label and then calculates the sum for one sample. The model learns to reduce the loss by making the output probabilities $\{p_n\}$ closer to the ground truth values $\{y_n\}$.

The Challenge of the News-Domain Text Classification Task

In real-world applications, multi-label text classification is a very important technology for automatic metadata assignment to content. Generally, a document in the real world contains multiple concepts, so single-label text classification tasks are insufficient for relating documents using metadata. For example, simply labelling an article about a tennis match as 'sports' would not be a very useful metadata addition. Instead, we can organise the data more usefully by adding more detailed labels such as 'tennis' or 'Wimbledon' as metadata.

In multi-label text classification, many predefined labels suitable for the domain could be constructed.

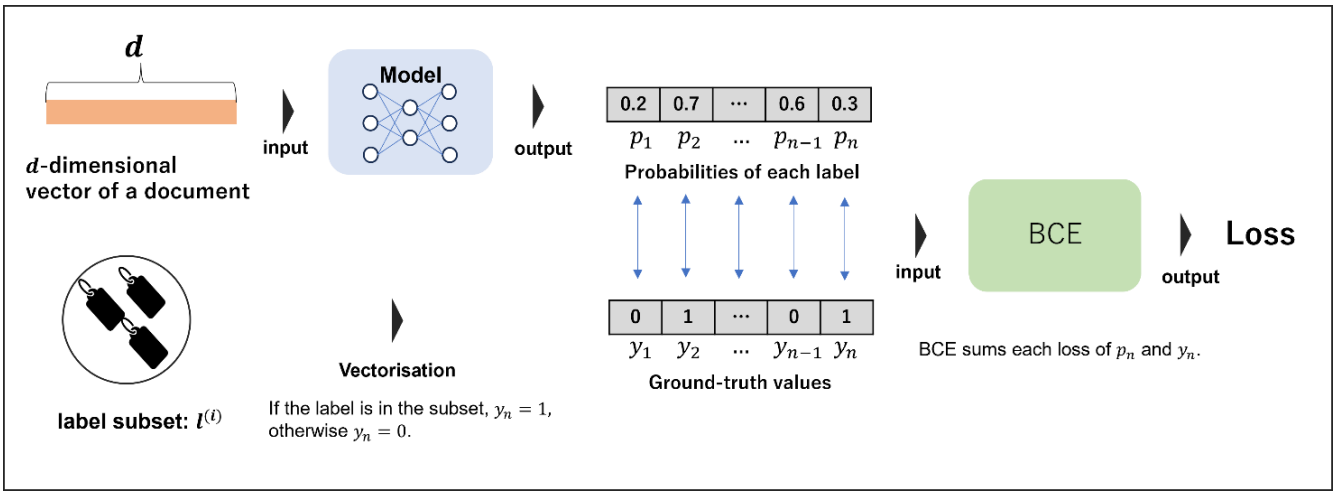


Figure 2: Model learning with BCE loss in the multi-label text classification task.

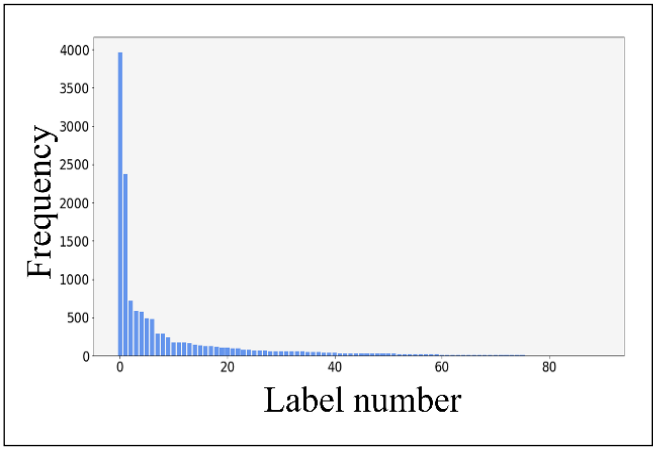


Figure 3 – Label distribution in Reuters-21578. The vertical axis represents the frequency of labels on the dataset, and the horizontal axis represents the index *n* assigned to the labels in frequency order.

Especially in the news domain, a very wide range of topics must be covered. However, it is difficult to train neural network models to output accurate labels on training data with a wide variety of detailed topics. This is because there are significant differences in the frequencies of occurrence of labels in the training data [6], [7]. As an example, Fig. 3 shows label frequencies for the news-based benchmark dataset Reuters-21578 for multi-label text classification. Reuters-21578 consists of 21,578 documents taken from the original Reuters-22173 corpus after the removal of 595 duplicate documents by Lynch and Lewis in 1996 [8], [9], [10]. As the figure shows, the label frequencies follow a long-tailed distribution.

Such a long-tailed label distribution leads to a decrease in the accuracy of neural networks. Major (high-frequency) labels that appear frequently in the training data have many documents to train on, while minor (low-frequency) labels that appear rarely have a far smaller number of document types to train on. As a result, the neural network may become overtrained to specific document features and minor labels, and will often not output minor labels on documents that do not contain an exact match with the learned sentences or words. This is a major problem in multi-label text classification in general, but is particularly prevalent in the news domain due to the wide range of topics covered.

Automatic News Labelling System

In this section, we describe a prototype Automatic News Labelling System developed using multi-label text classification. The system automatically assigns labels indicating the genre and content of news articles entered through a web-based interface. Fig. 4 shows the process flow of the assignment of labels to a news article.

First, the article provided by the user is tokenised by the tokeniser. To process natural language using a neural network, it is necessary to segment the language into 'tokens.' This is because neural network models cannot process language per se, but use a dictionary to map short input word sequences (tokens) into corresponding vectors (i.e. a numerical representation) for processing. However, unlike English, where tokens may correspond to individual words, Japanese words are not separated by spaces, so we split sentences into tokens using other sentence parsing techniques [11].

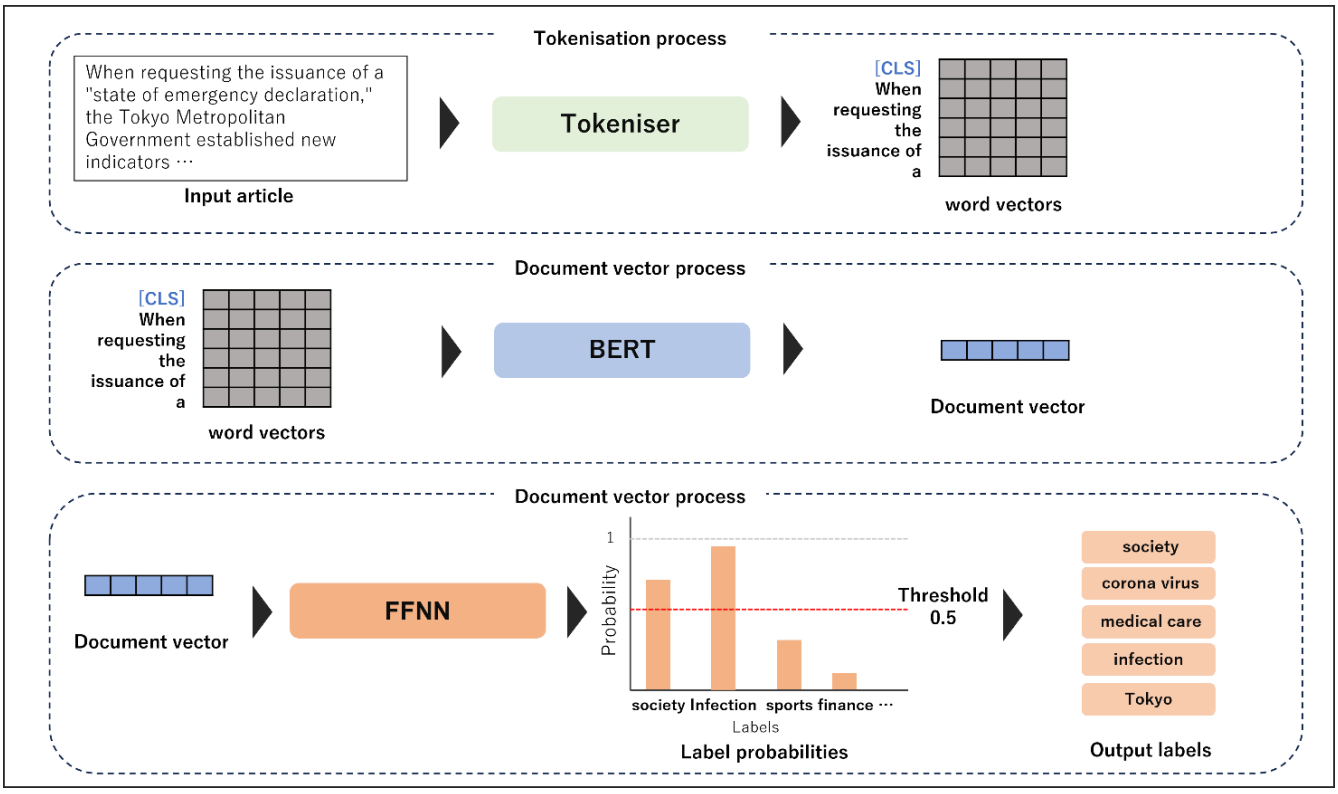


Figure 4: Process flow in labelling a news article in our system

The split tokens are then fed into a BERT ('Bidirectional Encoder Representations from Transformers') network [12]. BERT is a type of Transformer [13] neural network that uses attention techniques and is known to deliver stable performance in a variety of natural language processing tasks. The BERT network in our system was pre-trained using NHK news articles and Twitter data. The first token output as a feature of BERT is determined to be a special token used for classification tasks. The system treats the output vector of this special token as a vector representing the meaning of the entire document and extracts it.

Finally, the system feeds the extracted vector of special tokens into a feed-forward neural network (FFNN). The vector dimensionality is fixed at 768 for BERT. However, the number of labels we want to classify is not necessarily 768. A FFNN is responsible for converting vectors that represent the meaning of a sentence into score vectors whose dimension is the number of labels we want to classify. The system calculates a score (probability) for each label and outputs those labels whose score exceeds a threshold (0.5 in this case).

Fine-tuning of the Model Utilised in the System

In recent years, most language models such as BERT used in various applications have been pre-trained to acquire general linguistic information using a large amount of text. However, for downstream tasks such as multi-label text classification, the model must be trained with more task-specific data ('fine-tuning') to adapt the model better to the task at hand.

In order to fine-tune the language model used in this system, we constructed a multi-label text classification dataset based on NHK news articles as the domain. Based on past articles published on NHK's web news site, NHK NEWS WEB, we assigned labels indicating the content to about 50,000 articles in the data set. The total number of labels in the dataset was 1,015, and the labels were defined as a set of nouns that can generically represent the content of the articles. All labels were determined by consensus between annotators and supervisors by reading published articles. Subsets of labels were assigned to articles according to consensus between two annotators.

Since we did not intentionally eliminate label imbalance in the dataset we constructed, the model in this system is also expected to have low accuracy for low-frequency labels. We therefore constructed a new loss function to reduce the effect of label imbalance.

Suppression of Effect of Label Imbalance

BCE is often used in multi-label text classification, but it is vulnerable to label imbalance effects. The ultimate loss value for one sample is the sum of the loss values from each label. But across the whole dataset, most of the loss from low-frequency labels is dominated by the terms log $(1 - p_n)$ in Eq. (1) contributed by the many negative samples (the many articles n for which the label is absent).

In addition, an imbalance between positive and negative labels in a single input sample may have a negative impact on model training. In news article classification tasks, a large number of labels are predefined. For example, in our prototype system, we predefined 1,015 labels (N=1015). On the other hand, a single news article tends to be assigned no more than 10 labels. The ultimate loss is constructed by the summation of the loss values from all (1015) labels, so a large proportion of the loss for a single sample is dominated by losses from the negative labels. The imbalance disturbs meaningful learning of the positive labels for which the model should learn to output higher probabilities.

In this paper, we propose a loss function that combines a weight based on label frequencies, a weight that reduces the influence of negative samples, and a weight based on label co-occurrence information.

This loss function aims to suppress the impact of imbalanced datasets including many low-frequency labels on the training of a model.

We propose a method called weighted asymmetric loss (WASL), inspired by the asymmetric loss (ASL) [14] that has been proposed in the image classification domain. We aim to appropriately select loss values from negative labels to suppress, and we adjust the suppression in accordance with the label frequencies.

Furthermore, by performing label smoothing (LS) [15] based on co-occurrence of (correlations between) the labels, we compensate for the small number of samples of low-frequency labels and suppress the overfitting of the model. We define WASL as

$$\mathrm{WASL} = -\sum_{n=1}^{N} w_n [y'_n \log L_+ + (1 - y'_n) \log L_-] \quad Eq.(2)$$

$$L_+ = (1 - p_n)^{\gamma_+} \log p_n \quad Eq.(3)$$

$$L_- = (p'_n)^{\gamma'^{(n)}_-} \log(1 - p'_n) \quad Eq.(4)$$

$$p'_n = \max(p_n - \mathrm{m}, 0) \quad Eq.(5)$$

$$\gamma'^{(n)}_- = w_n \gamma_-, \quad Eq.(6)$$

$$w_n = \frac{\frac{(1-\beta)}{1-\beta^{c_n}} N}{\sum_n^N \frac{(1-\beta)}{1-\beta^{c_n}}} \quad Eq.(7)$$

where β (β∈ [0,1]) is a hyperparameter that determines the strength of the class-balanced weights and m is a hyperparameter that is sufficiently small to discard the loss values from negative samples for which learning has sufficiently progressed. γ_+ and γ_ (0≤ γ+ <γ -) are hyperparameters that adjust the balance between loss values from positive or negative labels. c_n represents the frequency of the n-th label in the dataset.

The three principal concepts underlying our proposed method are as follows: suppression of loss values from negative labels by $y'^{(n)}_-$; re-sampling of label frequencies by w_n; and the use of the smoothed ground-truth values y'_n in Eq. (2) in place of the original y_n (see below).

In multi-label text classification, for any given sample (article), most labels are negative, meaning that most of the y_n are 0. As a result, the total loss value is dominated by $(1\text{-}y_n')\log L_-$ in Eq. (2), representing the loss derived from negative labels. During the training process, the model can reduce the loss value by making the output probability of the corresponding label closer to the ground-truth value.

This dominance of negative labels prevents the model from learning meaningfully. Thus, we introduce in our method the weight $\gamma'^{(n)}_-$ in Eq. (4). $\gamma'^{(n)}_-$ is determined by the hyperparameter γ_- and the weight w_n to adjust for the effect of label frequency. It can selectively reduce the loss values derived from already well-learned negative labels.

We introduce the class-balanced weight utilised in class-balanced loss (CBL) w_n in Eq. (7) so that the model can more appropriately consider the differences in label frequencies [16]. As a label y_n appears more frequently in a dataset, w_n becomes smaller. This reduces loss values associated with high-frequency

labels and increases those associated with low-frequency labels. CBL was proposed by Cui et al. [16] for correcting label imbalance in a single-label classification task but has also proven successful in text classification (17) so in this work, we use this extension of CBL for multi-label text classification.

Label smoothing (LS) techniques proposed by Szegedy et al [15] can suppress overfitting of a model and calibrate the model. We perform LS with the label co-occurrence information set as prior probabilities to suppress overfitting of the model. LS here is defined as

$$y'_n = (1-\alpha)y_n + \alpha \,\mathrm{norm}(o_n) \quad Eq.(8)$$

$$\boldsymbol{o} = \boldsymbol{y} \cdot \boldsymbol{P} \quad Eq.(9)$$

where α ($\alpha \in [0,1]$) is a hyperparameter that determines the degree of smoothing and norm $(\cdot)$ is the min-max normalisation function. Also, $\boldsymbol{y}$ ($y_n \in \{0,1\}^N$) is a multi-hot vector for the input sample and $\boldsymbol{P}$ indicates a square matrix of positive pointwise mutual information (PPMI) calculated from the number of co-occurrences of each label [18], [19]. Fig. 5 shows an example smoothed target vector produced by our method, where the labels 'wheat', 'flour', and 'grain' tend to co-occur with each other in the training data. If 'wheat' and 'grain' are positive labels for some input sample, that sample is likely to also be associated with the label 'flour.' Some value is re-distributed to related labels for some sample (from 'wheat' and 'grain' to 'flour' in this case) according to the strength of the co-occurrence (PPMI score) across the whole dataset. On the other hand, no value is distributed to the label 'gold' because it does not co-occur with 'wheat' and 'grain' in the training data. The model can learn from input sample - relevant label pairs, even if the labels are not truly positive labels.

Experiment

We conducted a comparative experiment with baseline methods to investigate the effectiveness of the proposed method. We used macro-f1 and micro-f1 as evaluation metrics. Reuters-21578 and NHK NEWS WEB were utilised in this experiment as benchmark datasets. The model was implemented using PyTorch [20] and Transformers [21], and the optimisation method was AdamW [22]. The common hyperparameters in both datasets and all methods were as follows. The number of learning epochs was set to 50, dropout rate was 0.1, and output threshold was unified to 0.5. All hyperparameters (except the common ones) were determined by using grid search according to micro-f1 on validation data. Table 1 lists the results of this experiment. As we can see in Table 1, the proposed method significantly outperformed the baselines on both datasets. These results suggest that the proposed method is effective for training a model on an imbalanced dataset.

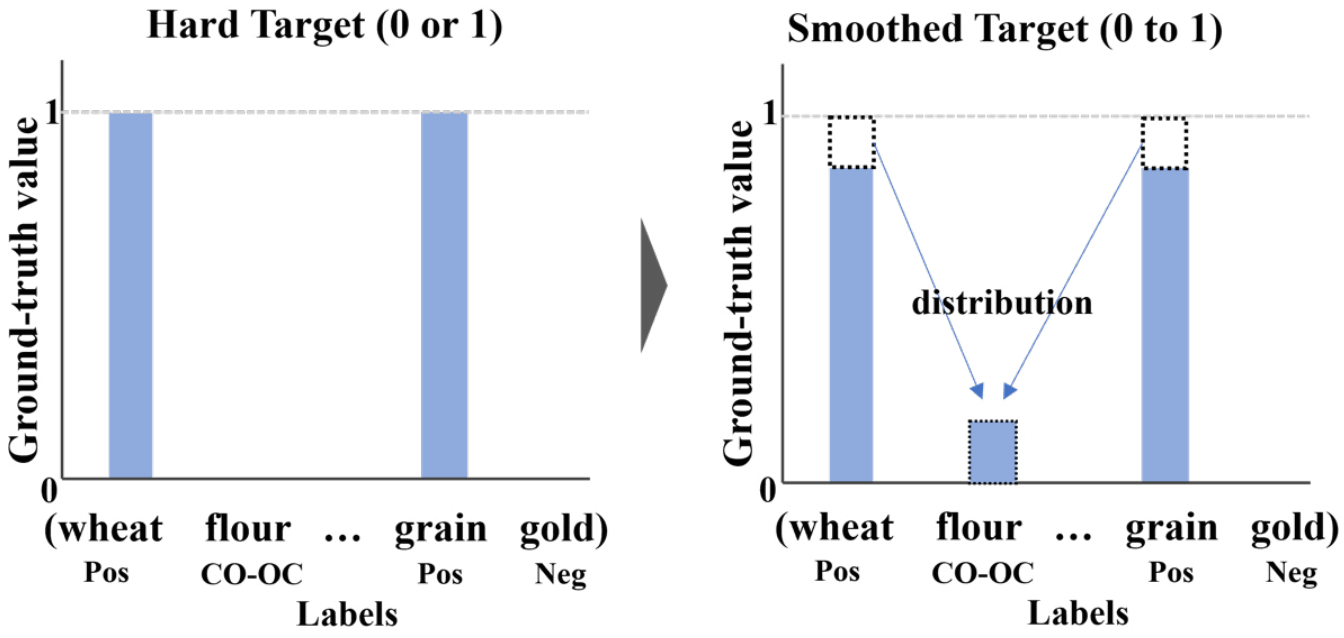


Figure 5: Hard target $\boldsymbol{y}$ (multi-hot vector) and smoothed target $\boldsymbol{y'}$ produced by label smoothing in our method.

Dataset Methods / Metrics	Reuters-21578 Macro-f1 / Micro-f1	NHK NEWS WEB Macro-f1 / Micro-f1
BCE	0.586 / 0.903	0.488 / 0.732
ASL	0.644 / 0.904	0.503 / 0.745
WASL (proposed)	**0.669** / **0.911**	**0.506** / **0.754**

Table 1: Experimental results. Means of five trials are shown for the proposed method and baselines

Auxiliary Modules other than Automatic Labelling

In addition to the automatic labelling function, the system also contains modules for extracting keywords and listing related articles. These modules are implemented with the aim of facilitating content creation assistance based on automatic labelling.

The keyword extraction function highlights place names and proper nouns in a news article provided by the user. This can assist the user in creating and applying new fine-grained labels. In multi-label text classification using neural networks, all the labels to be assigned must be defined in advance, and it can be difficult to add topical news article labels corresponding to the latest popular words or place names. However, there is a considerable need among content producers for such functions that can add metadata such as labels for proper nouns and place names. By utilising named entity extraction technology, likely proper nouns and place names in the text can be extracted, listed, and highlighted to assist producers in considering the creation of new labels.

The system can also list published articles in a news database that are related to an article input or specified by the user. Some news topics have a high degree of continuity from previous articles, and when publishing new articles, including links to related prior articles increases reader satisfaction and engagement.

This module is implemented by comparing vectorised news articles using cosine similarity scores. By using cosine similarity, we eliminate the effect of the indeterminate number of labels assigned to a news article, making it possible to measure the relevance of an article according to its content.

Articles in the database have labels like those assigned to the input article by the system. The system can automatically label large amounts of old article data that does not yet have metadata, making it easier to reuse large amounts of old articles that have not thus far been organised for content creation.

Use Cases

NHK has local broadcasting stations in each Japanese prefecture providing content that is specialised for each region. Such stations have fewer staff and need more efficient content production, so we introduced our prototype Automatic News Labelling System in related but different use cases at two local stations.

Programme Creation Reusing News Articles

Here we explain an example where the system was used for content reuse. Broadcasters often need to reuse published content to help create new content. However, it may be the case that the content they have created does not have the metadata necessary for reuse.

At the end of 2023, producers at a local broadcast station wished to create a segment in a news programme about ranking news articles published on the local news website over the course of a year by topic. Unfortunately, the news articles published over the one-year period were not annotated with topic metadata. We used our system to automatically assign labels and re-aggregate the published article data by topic. Specifically, we reaggregated data from 121 articles, and the labels automatically assigned by the system were manually reviewed and confirmed. The rankings were then shown on news programmes to help viewers visually grasp local events from the past year.

Automatic Labelling of News Articles

In NHK, each regional station publishes local news on its website for the prefecture it serves. Unlike the national news site, the local news sites did not label news articles due to limited staffing. It is hoped that the introduction of our system will allow efficient metadata addition even with a small number of staff, enabling viewers also to access such information associated with local services.

In the article publishing workflow, producers input completed pre-publication news articles into the system. They then check and correct the output labels, and manually enter new labels if necessary. Labels constructed in this way are then published on the website alongside the news articles and displayed as hyperlinks to related articles.

Operational Issues

We found that using this system can help broadcasters with limited staff to produce rich content. However, some practical issues were also identified.

News articles on local news websites tend to be focused on unique traits of the region, and producers in local regions want more specific labels expressing these unique regional features. The labels defined in this system, however, are intentionally constructed to be common words. To attract the interest of local viewers, more specific proper nouns should be assigned to news articles instead of generic words. It is difficult to collect a large number of labels that represent the characteristics of the region. Even if we collect a few samples, such labels are trained even less frequently than other minor labels, making it difficult to solve this problem using only the method in this paper. Therefore, it will be necessary to prototype new algorithms, such as generative methods.

The user interface of the system was also identified as a major practical issue. Our system was initially conceived and intended to be usable nationally by NHK, and we built the interactive interface accordingly. However, in practice, producers needed to deal with many articles at once, for example for aggregation work. When using the system for content reuse, it was reported that the system's interface was not a good fit with the production workflow, resulting in inefficient use of time. We thus concluded that we should implement an interface more suitable for individual workflows, including for example a batch processing interface.

Related Work

Many studies have proposed model architectures to improve the quality of multi-label classification. For example, Chen et al [23] developed a basic encoder-decoder model using a convolutional neural network and recurrent neural network for multi-label classification. Adhikari et al [24] reported that a well-tuned simple bidirectional-LSTM model can outperform some complex models. Models considering the relationship between labels have also been proposed. Yang et al [25] proposed a sequence generation model that treats the multi-label text classification task as a sequence generation problem with the aim of considering label relations. Also, Xiao et al [26] achieved a state-of-the-art performance on the AAPD dataset of Yang et al [25]. While many methods focusing on architecture have been proposed, such complex approaches tend to require extensive computing resources. Furthermore, even simple pre-trained models like BERT [12] still maintain a state-of-the-art performance on some datasets. Most existing models for multi-label text classification are trained with BCE, which makes them susceptible to label distribution, especially for imbalanced learning. Therefore, an approach that focuses on the loss function is required.

From a different perspective, methods that utilise dependencies between labels can also help improve the accuracy on multi-label text classification tasks. Dembczyński et al [27] pointed out the importance of considering label correlations in the multi-label text classification domain. Pal et al [28] proposed a model that treats the relationships between labels by means of graph attention networks [29]. Zhao et al [30] developed a model that creates clusters of labels and extracts the correlations between clusters. Song et al [31] achieved state-of-the-art performance on some datasets with a method combining a cloze task and multi-label text classification. On the other hand, many methods that take into account the relationship between labels have been proposed in the image classification domain [32], [33], [34]. Many methods based on label correlation utilise label embedding techniques [35] and graph convolutional networks [36]. Methods based on adjusting ground-truth values directly, for example LS, are not discussed so much. LS can suppress overfitting of a model and calibrate the model [37]. We believe that LS can be helpful for improving the accuracy of models for multi-label text classification.

Conclusions

Broadcasting stations produce many news scripts every day for TV programmes and web content. In order to efficiently use and re-use this large amount of text data, metadata labels that indicate the contents of the manuscripts must be accurately assigned. However, manually labelling takes a huge amount of time and effort, and in order to improve efficiency and significantly reduce the burden and cost of news production, we developed a prototype system that uses AI to automatically label news manuscripts.

To realise this system, we utilise multi-label classification, a task in which a computer outputs the applicable subset of labels for a given input manuscript. The model learns from training pairs consisting of an article and its associated label subset.

Since news articles cover a vast range of topics, label classification requires an enormous number of words to be prepared as label candidates, but in most cases only a few appropriate labels are assigned to any single article. As a result, previous research has shown that the adjustment of probabilities during learning is biased towards the process of reducing the probability of labels that do not match the content of the article being wrongly output (rather than increasing the probability of those labels that do appear being correctly output). This has a negative impact on learning and leads to a decline in classification performance.

To tackle this issue, we developed a novel loss function to suppress the effects of this 'label imbalance'. The loss function we developed utilises some weights that reduce the loss values from negative labels and high-frequency labels. Additionally, label smoothing based on label co-occurrences is introduced to suppress overfitting of the model to low-frequency labels. Evaluation experiments confirmed that our method improved classification performance on imbalanced data.

The prototype Automatic News Labelling System we developed has been introduced for testing at two local broadcasting stations and has achieved a reduction in the effort required for content production, while identifying some workflow issues that remain to be tackled. We expect subsequent versions of this and other similar systems to allow producers to create richer content in the future with less cost and effort.

References

1. Zhang, M.-L. and Zhou, Z.-H., 2014. A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering. Vol.26, No.8, pp. 1819 to 1837.

2. Tsoumakas, G. and Katakis, I., 2007. Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining, Vol.3, No.3, pp. 1 to 13.

3. Ueda, N. and Saito, K., 2002. Parametric Mixture Models for Multi-Mabeled Text. Advances in Neural Information Processing Systems. Vol.15.

4. Chalkidis, I., Fergadiotis, E., Malakasiotis, P., Aletras, N., and Androutsopoulos, I., 2019. I. Extreme Multi-Label Legal Text Classification: A case study in EU Legislation. In Proceedings of the Natural Legal Language Processing Workshop 2019. pp. 78 to 87.

5. Yao, L., Mao, C., and Luo, Y., 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Medical Informatics and Decision Making, Vol.19, No.3, pp. 31 to 39.

6. He, H. and Garcia, E. A., 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, Vol.21, No.9, pp. 1263 to 1284.

7. Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M., 2020. Generalizing from a Few Examples: A Survey on Few-shot Learning. ACM Computing Surveys (csur), Vol.53, No.3, pp. 1 to 34.

8. Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of the 10th European Conference on Machine Learning, pp. 137 to 142.

9. Lewis, D. D., 1992. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 37 to 50.

10. Yang, Y. and Liu, X., 1999. A re-examination of text categorization methods. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42 to 49.

11. Kudo, T. and Richardson, J., 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66 to 71.

12. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp. 4171 to 4186.

13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I., 2017. Attention is All you Need. Advances in Neural Information Processing Systems. Vol.30.

14. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L., 2021. Asymmetric Loss for Multi-Label Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision. pp.82 to 91.

15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818 to 2826.

16. Cui, Y., Jia, M., Lin, T.-Y., Song, Y. and Belongie, S., 2019. Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9268 to 9277.

17. Huang, Y., Giledereli, B., K¨oksal, A., ¨Ozg¨ur, A., and Ozkirimli, E., 2021. Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 8153 to 8161.

18. Church, K. W. and Hanks, P., 1990. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, Vol.16, No.1, pp. 22 to 29.

19. Niwa, Y. and Nitta, Y., 1994. Co-Occurrence Vectors From Corpora vs. Distance Vectors From Dictionaries. In COLING The 15th International Conference on Computational Linguistics. Vol.1.

20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems, Vol. 32.

21. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38 to 45.

22. Loshchilov, I. and Hutter, F., 2019. Decoupled Weight Decay Regularization. In International Conference on Learning Representations.

23. Chen, G., Ye, D., Xing, Z., Chen, J., and Cambria, E., 2017. Ensemble Application of Convolutional and Recurrent Neural Networks for Multi-label Text Categorization. In 2017 International Joint Conference on Neural Networks, pp. 2377 to 2383.

24. Adhikari, A., Ram, A., Tang, R., and Lin, J., 2019. Rethinking Complex Neural Network Architectures for Document Classification. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp. 4046 to 4051.

25. Yang, P., Sun, X., Li, W., Ma, S., Wu, W., and Wang, H., 2018. SGM: Sequence Generation Model for Multi-label Classification. In Proceedings of the 27th International Conference on Computational Linguistics, pp. 3915 to 3926.

26. Xiao, L., Huang, X., Chen, B., and Jing, L., 2019. Label-Specific Document Representation for Multi-Label Text Classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 466 to 475.

27. Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E., 2012. On label dependence and loss minimization in multi-label classification. Machine Learning, Vol.88, pp. 5 to 45.

28. Pal, A., Selvakumar, M., and Sankarasubbu, M., 2020. Multi-Label Text Classification using Attention-based Graph Neural Network. arXiv preprint ArXiv:2003.11644.

29. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. 2018. Graph Attention Networks. arXiv preprint arXiv:1710.10903.

30. Zhao, F., Ai, Q., Li, X., Wang, W., Gao, Q., and Liu, Y. 2024. TLC-XML: Transformer with Label Correlation for Extreme Multi-label Text Classification. Neural Processing Letters, Vol.56, No.1, pp. 1 to 25.

31. Song, R., Liu, Z., Chen, X., An, H., Zhang, Z., Wang, X., and Xu, H., 2023. Label prompt for multi-label text classification. Applied Intelligence, Vol.53, No.8, pp. 8761–8775.

32. Chen, Z.-M., Wei, X.-S., Wang, P., and Guo, Y., 2021. Learning Graph Convolutional Networks for Multi-Label Recognition and Applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.45, No.6, pp. 6969 to 6983.

33. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W., 2016. CNN-RNN: A Unified Framework for Multi-Label Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285 to 2294.

34. Ye, J., He, J., Peng, X., Wu, W., and Qiao, Y., 2020. Attention-Driven Dynamic Graph Convolutional Network for Multi-Label Image Recognition. In Proceedings of Computer Vision–ECCV 2020: 16th European Conference, pp. 649 to 665.

35. Zhang, H., Xiao, L., Chen, W., Wang, Y., and Jin, Y., 2018. Multi-Task Label Embedding for Text Classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4545 to 4553.

36. Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., and Welling, M., 2018. Modelling Relational Data with Graph Convolutional Networks. In Proceedings of the Semantic Web: 15th International Conference, pp. 593 to 607.

37. Müller, R., Kornblith, S., and Hinton, G. E., 2019. When Does Label Smoothing Help? Advances in Neural Information Processing Systems, Vol.32.

Selection of Papers from *Electronics Letters*

Depression recognition from facial videos: Preprocessing and scheduling choices hide the architectural contributions

Authors: Manuel Lage Cañellas, Constantino Álvarez Casado, Le Nguyen, Miguel Bordallo López

Abstract: Deep learning models have been widely applied in video-based depression detection. It is observed that the diversity of preprocessing, data augmentation, and optimization techniques makes it difficult to fairly compare model architectures. In this study, the typical ResNet-50 model is enhanced by using specific face alignment methods, improved data augmentation, optimization, and scheduling techniques.

The extensive experiments on two popular benchmark datasets (AVEC2013 and AVEC2014) obtained competitive results, compared to sophisticated spatio-temporal models for single streams. Moreover, the score-level fusion approach based on two texture streams outperformed the state-of-the-art methods. It achieved mean square errors of 5.82 and 5.50 on AVEC2013 and AVEC2014, respectively. These findings suggest that the preprocessing and training configurations result in noticeable improvements, which have been originally attributed to the network architectures.

https://doi.org/10.1049/ell2.12992

An efficient geometric partitioning mode by adaptive blending method for screen content coding

Authors: Minhun Lee, Seoung-Jun Oh, Donggyu Sim

Abstract: Geometric partitioning mode (GPM) was newly adopted in Versatile Video Coding (VVC) to increase the partitioning precision of the arbitrary boundaries of actual objects. Note that the blending process of the GPM is not suitable for screen content with various sharp and strong edges. To alleviate this drawback, three new GPM matrices are introduced, and an efficient GPM is proposed to utilize four different GPM matrices and implicitly determine one of them depending on the discontinuity and gradients of the two predictors. The proposed method achieves an average BD-rate reduction of 1.2% and 1.6% without additional computational complexity for random access and low-delay B configurations, respectively, over the VVC test model (VTM18.0).

https://doi.org/10.1049/ell2.12961

Global strategy for robust air-light estimation in dehazing

Authors: Zhu Zhu, Xiaoguo Zhang

Abstract: This letter proposes a novel approach to enhance the robustness of air-light estimation for dehazing. Unlike most existing methods, it employs a global strategy, considering all pixels instead of specific individual ones, for recovering air-light. Through an iterative algorithm via the Gray World (GW), the authors extract the air-light orientation from the entire image. Next, a global detail-preserving algorithm is designed to determine the optimal magnitude of air-light. Experimental results on a diverse set of hazy images reveal that the authors' method outperforms other state-of-the-art alternatives, highlighting the advantage of air-light estimation using the entire image information.

https://doi.org/10.1049/ell2.13079

Local color display technology

Authors: Weijian Wu

Abstract: The local color display technology utilizes a high refresh rate LCD open-cell without a color filter and enables color adjustment through independent RGB backlight zones isolated by an anti-scatter grid (ASG). Synchronized calibration between the open-cell and the backlight zones is achieved by aligning the backlight controller with the vertical clock pulse (CPV)/vertical start pulse (STV) signals in the open-cell gate driver, thereby determining the open-cell scan line position. Once synchronous calibration is completed, the backlight controller refreshes the corresponding RGB backlight zones whenever the open-cell scan line passes through them. The local color display technology can reduce the cost of open-cell by 20%, improve resolution by three times, enhance brightness by eight times, expand the color gamut to BT2020, and improve grayscale to 32-bits.

https://doi.org/10.1049/ell2.13306

Superpixel-guided locality preserving projection and spatial–spectral classification for hyperspectral image

Authors: Hailong Song, Shuzhen Zhang

Abstract: Locality preserving projection (LPP) is a typical feature extraction method based on spectral information for hyperspectral image (HSI) classification. Recently, to improve the classification performance, the spatial information of HSI has been applied in the LPP method. However, for most of spatial–spectral-based LPP methods, they explore the spatial–spectral information within a fixed local window, which cannot be appropriate to the irregular-shape ground objects in HSI. To over this issue, an effective superpixel-guided LPP and spatial–spectral classification method are proposed, in which the spatial–adaptive structure information is fully excavated for HSI classification. Specifically, superpixel segmentation is first conducted on the HSI to generate shape-adaptive homogeneous subregions.

Then, to learn more discriminative projection, the neighbourhood graph for LPP is constructed based on spatial–spectral similarity, in which pixels within the same superpixel are connected. Finally, the obtained projection feature is input a classifier to yield the initial classification result, and the edge information of ground objects captured by superpixels is utilized to optimize the initial classification result. Experiments on two real hyperspectral datasets demonstrate that the proposed superpixel-guided and spatial–spectral classification method significantly outperforms the other well-known techniques for HSI classification.

https://doi.org/10.1049/ell2.13293

Nighttime wildlife object detection based on YOLOv8-night

Authors: Tianyu Wang, Siyu Ren, Haiyan Zhang

Abstract: Monitoring nocturnal animals in the field is an important task in ecological research and wildlife conservation, but the complexity of nocturnal images and low light conditions make it difficult to cope with traditional image processing methods. To address this problem, researchers have introduced infrared cameras to improve the accuracy of nocturnal animal behaviour observations. Object detection in nighttime images captured by infrared cameras faces several challenges, including low image quality, animal scale variations, occlusion, and pose changes. This study proposes the YOLOv8-night model, which effectively overcomes these challenges by introducing a channel attention mechanism in YOLOv8.

The model is more focused on capturing animal-related features by dynamically adjusting the channel weights, which improves the saliency of key features and increases the accuracy rate in complex backgrounds. The main contribution of this study is the introduction of the channel attention mechanism into the YOLOv8 framework to create a YOLOv8-night model suitable for object detection in nighttime images. When tested on nighttime images, the model performs well with a significantly higher mAP (0.854) than YOLOv8 (0.831), and YOLOv8-night scores 0.856 on mAP_l, which is obviously better than YOLOv8 (0.833) in terms of processing large objects. The study provides a reliable technical tool for ecological research, wildlife conservation and environmental monitoring, and offers new methods and insights for the study of nocturnal animal behaviour.

https://doi.org/10.1049/ell2.13305

PCQD-AR: Subjective quality assessment of compressed point clouds with head-mounted augmented reality

Authors: Chunling Fan, Yun Zhang, Linwei Zhu, Xinju Wu

Abstract: This letter fully studies the coloured point cloud quality assessment in augmented reality (AR) environment through subjective test. Firstly, a point cloud dataset, named point cloud quality dataset-AR, including ten reference point clouds and their 90 distorted versions is presented, which were generated using the reference software of video-based point cloud compression across various combinations of geometry and texture quantization parameters. Then, the impact of geometry and texture distortions on perceived quality of point clouds in AR environment was discussed in detail. Moreover, the performance of existing objective point cloud quality assessment metrics on the proposed dataset is evaluated. The subjective dataset including the values of mean opinion score were released to public.

https://doi.org/10.1049/ell2.13134

Empowering lightweight video transformer via the kernel learning

Authors: Xiaoxi Liu, Ju Liu, Lingchen Gu

Abstract: Video transformers achieve superior performance in video recognition. Despite the recent advances in video transformers, they still require substantial computation and memory resources. To cater for the computation efficiency, a kernel-based video transformer is proposed, including: (1) a new formulation of the video transformer via the kernel learning is presented to better understand the individual components of it; (2) a lightweight Kernel-based spatial–temporal multi-head self-attention block is explored to learn the compact joint spatial–temporal video feature; (3) an adaptive-score position embedding method is conducted to promote the flexibility of video transformer. Experimental results on several action recognition datasets demonstrate the effectiveness of the proposed method. Only pretrained on ImageNet-1K, the method achieves the preferable balance between computation and accuracy, while requiring 7 x fewer parameters and 13 x fewer floating point operations than other comparable methods.

https://doi.org/10.1049/ell2.13215

Speaker front-back disambiguity using multi-channel speech signals

Authors: Xinyuan Qian, Jichen Yang, Alessio Brutti

Abstract: This paper tackles the front-back disambiguity problem in speaker localization when the audio signals are captured by a symmetric microphone array. To this end, a deep neural network is proposed with an attention-based mechanism designed to assign different weights to features obtained from individual microphones. For support, a real dataset with synchronized multichannel audio signals captured by a large linear microphone array is introduced, along with manual annotations. The experimental results demonstrate the effectiveness of the proposed method over the other approaches. In particular, more than 50% reduction in Equal Error Rate (EER) is achieved when comparing with the single-channel case. The designed multi-channel self-attention mechanism also brings further improvements. The dataset and source code will be released.

https://doi.org/10.1049/ell2.12666

Super-resolution using deep residual network with spectral normalization

Authors: Yogendra Rao Musunuri, Oh-Seol Kwon

Abstract: In this letter, the authors present a single-image super-resolution method based on introducing a novel spectral normalization to the convolution of a deep residual network. Moreover, the authors construct a new residual block (RB) and assemble it in a cascade form. The new RB was restructured by the spectrally normalized convolution layers and activated function. In addition, the RB allows the spectral normalization introduced on the trained network to update the additional weights. Furthermore, it minimizes pixel loss, thereby helping to obtain enhanced reconstruction results, limiting the number of parameters, and facilitating a low computational cost. The experimental results demonstrate that the proposed model shows the superior performance over that of state-of-the-art methods in terms of visual quality metrics such as UQI and PIQE.

https://doi.org/10.1049/ell2.12734

The Institution of Engineering and Technology

iet.tv

iet.tv Productions

In a world where visual content is key and your audience is always switched on, iet.tv Productions will help you produce outstanding video content. iet.tv Productions is the in-house film and video unit for the Institution of Engineering and Technology.

We offer:

Full-Service Video Production:
In-house producers, managers, camera teams, video technicians, and editors for all project stages.

Film Crews and Equipment:
State-of-the-art cameras and recording equipment tailored to project needs.

Dedicated Post-Production Services:
Editing suites for high-quality video production, global distribution, and web embedding.

Visual Effects Creation:
Imaginative visual effects and animations using After Effects software.

Webcasts and Webinars:
Planning, setup, and hosting of web-based presentations, training sessions, and interactive conferences.

Web Conferencing:
Facilitation of group meetings with collaborative document editing and presentation sharing.

Drone and UAV footage:
CAA approved filming using state of the art UAV's (drones).

More than 30 years' experience in creating engaging and inspiring, award winning video content.

Find out more. Visit: iet.tv

The Institution of Engineering and Technology

Contact information

London, UK
T +44 (0)20 7344 8460
E faradaycentre@ietvenues.co.uk

Stevenage, UK
T +44 (0)1438 313311
E postmaster@theiet.org

Beijing, China*
T +86 10 6566 4687
E china@theiet.org
W theiet.org.cn

Hong Kong SAR
T +852 2521 2140
E infoAP@theiet.org

Bengaluru, India
T +91 80 4089 2222
E india@theiet.in
W theiet.in

New Jersey, USA
T +1 (732) 321 5575
E ietusa@theiet.org
W americas.theiet.org

@TheIET theiet.org