1. Multimedia Services, 5G Technology, and XR: An In-depth Exploration of Technologies, Standards, and Challenges

Tiziana Cattai, Stefania Colonnese

Dept. of Information Engineering, Electronics and Telecommunications, Sapienza University of Rome, Italy

Abstract: *This chapter provides a comprehensive overview of multimedia services and their boost in 5G and beyond networks. The chapter will cover the key technologies, challenges, and ongoing developments, on multimedia communications with special attention to Extended Reality service provisioning.*

1.1 Introduction

This chapter presents a comprehensive overview of the most important topics related to Multimedia communication services. To this aim, we present the fundamental Multimedia Services with their associated requirements. We also explore the developments of those services enabled by the advanced in 5G technology. This presentation describes standard solutions highlighting the possibility of generalization to different proprietary solutions. In the last sections, we discuss interesting applications, with a specific focus on the potential of extended reality (XR).

Multimedia communication services are systems that allows the transmission of different types of multimedia information, such as voices, videos, real time-images. Herein we identify few services categories, based on the specific system requirements; several applications and commercial solutions derive from the main solutions described here.

A coarse service categorization is as follows: i) broadcasting, ii) video-telephony, video-conferences, iii) streaming, iv) messaging.

The above defined service categories differ under many respects, but the key feature for the system design can be found in delay tolerance; broadcasting is typically a one-way service, with relatively loose delay requirements; video-telephony, video-conferences are two-way services, with strict delay requirements (up to 50-80 ms); streaming, which is characterized by receiving data from a packet network while decoding and presentation take place, is typically sensitive not to the absolute delay value, but to its fluctuation with respect to its average value (jitter). Messaging services can take place either by downloading all the data before presenting them, with very loose or no delay requirement at all, or by actually streaming the data; in this latter case they can be regarded as streaming service in the system design stage; thereby, we will not discuss further the messaging services, which do not pose specific design challenges.

For each service, we will review

2. Visual Compression: New Paradigms?

Nicola Adami, Marco Dalai, Alessandro Gnutti, Fabrizio Guerrini, Riccardo Leonardi, Pierangelo Migliorati, Alberto Signoroni

Department of Information Engineering, University of Brescia, Italy

Abstract: *The massive production and sharing of visual digital data, i.e., images and video, is having a tremendous impact on human activities, both in business and societal terms. Image and video perceptual compression is one of the key enabler technologies that allowed such immense growth in digital imagery consumption to take place, and it has a storied tradition among the signal processing community. Current image and video compression de facto standards rely on conventional, highly refined signal processing methods, developed over the last 50 years, and it is a vital research topic to the present day. Recently, Artificial Intelligence (AI) based techniques, due to their remarkable performance with respect to more traditional signal processing tasks, are bringing a paradigm shift for visual compression as well. Accordingly a so-called learnt compression is being considered for inclusion in future image and video compression standards. In this chapter, we will address the challenges that both the new and old compression paradigms must face in the modern digital world. First, we will briefly describe in general terms the characteristics and the requirements of a visual compression system. Then, conventional and AI-based compression frameworks will be discussed and compared, and relevant examples to particular facets for both will be provided demonstrating that given the widespread need of visual information exchange the field remains a very active research arena. Finally, considering the emergence of the novel AI-based paradigm, we will provide a critical discussion on the present and future impact of video compression on modern applications.*

2.1 Introduction

In today's digital era, the consumption and creation of visual content, comprising images and videos, have become ubiquitous across various platforms, including social media, entertainment, communication, education, healthcare, and beyond. This surge in visual data availability has underscored the critical role of image and video compression in managing, transmitting, storing, and delivering these multimedia assets efficiently.

Image and video compression techniques play an indispensable role in addressing the inherent challenges posed by the vast amounts of data associated with visual content. These techniques are essential for reducing the storage space required for multimedia files, facilitating faster transmission over networks, enabling efficient streaming on bandwidth-limited platforms, and ensuring optimal utilization of storage resources across devices or between cashing data centers.

The importance of image and video compression extends far beyond mere data size reduction. It significantly impacts various facets of modern digital experiences. For instance, in the realm of online streaming services, efficient compression techniques allow for

3. Gaussian Class-Conditional Training for Secure and Robust Deep Neural Networks

Tiziano Bianchi, Andrea Migliorati, Enrico Magli

Department of Electronics and Telecommunications, Politecnico di Torino, Italy

Abstract: *This chapter presents a Gaussian Class-Conditional (GCC) training strategy for deep neural networks. The approach is based on a novel loss that maps the input data onto Gaussian target distributions in the latent space, where the parameters of the target distributions can be optimized for the specific task. For multiclass classification, the mean values of the learned distributions are centered on the vertices of a simplex such that each class is at the same distance from every other class. For metric learning, the distances between similar and dissimilar instance pairs are mapped on distributions with well-separated means. The proposed strategy has many advantages compared to conventional training. First, the optimal decision surface in the latent space is always a hyperplane, yielding a simple and interpretable decision rule. Second, the regularization of the latent space enforces high inter-class separation and low intra-class separation, minimizing the presence of short paths toward neighboring decision regions. The GCC training strategy is applied to two different multimedia problems. In image classification, GCC training provides both improved accuracy and robustness against adversarial perturbations, outperforming models trained with conventional cross-entropy loss and adversarial training. In biometric verification, GCC training yields lower error rates than other state-of-the-art approaches even on challenging and unconstrained datasets.*

3.1 Introduction

Deep neural networks have become the state-of-the-art solution for several multimedia, scientific, and industrial applications thanks to their ability to provide even greater accuracy levels than humans in multiple and complex visual and classification tasks [1]. Deep networks have also shown remarkable performance at learning complex mappings for image translation and segmentation [2]. However, the ever-growing use of deep neural networks in our society raises serious security concerns as they can be targeted by malicious adversaries. This is particularly relevant in applications where security is of key importance, such as medical diagnostics and autonomous driving [3, 4]. One of the most severe threats to deep neural networks is represented by *adversarial examples*. Legitimate inputs can be altered even very slightly, often in a way undetectable to the human eye, in such a way that the expected behavior of the network is completely disrupted [5].

Despite many advancements, the community has not yet been able to establish a conclusive defense mechanism against adversarial examples. There is general agreement on the importance of performing correct evaluations of adversarial attacks and defense algorithms, introducing a rigorous benchmark method to evaluate adversarial robustness

4. Graph Neural Networks for Inverse Problems in Multimedia

Diego Valsesia, Emanuele Aiello

Department of Electronics and Telecommunications, Politecnico di Torino, Italy

Abstract: *Graph Neural Networks (GNNs) have emerged as state-of-the-art models to address inverse problems arising from multimedia data. In inverse problems, one deals with partial and corrupted observations of some data and seeks to reconstruct the original with the highest possible fidelity, among the infinitely many possible solutions. This requires learning a good model of the data, and it is particularly challenging for multimedia, due to the complexity of data such as images, video, point clouds, etc.. Graph neural networks can capture powerful models, due to their ability to model non-local interactions such as non-local self-similarity in images, and the ability to naturally process data defined on irregular domains such as point clouds. In this chapter, we present recent advances on problems such as image denoising and restoration, point cloud denoising and shape completion and more.*

4.1 Introduction

Inverse problems are ubiquitous in imaging and in the processing of visual data at large [1]. An inverse problem is faced whenever a signal must be reconstructed from a set of measurements or observations. Such observations may be degraded or insufficient to uniquely recover the original signal, rendering the task ill-posed. A famous subset of inverse problems is constituted by linear inverse problems. For such, a linear forward model $\mathbf{A}$ describes how measurements $\mathbf{y}$ are obtained from the original signal:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{4.1}$$

being $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{y} \in \mathbb{R}^M$ with $M \leq N$ and being $\mathbf{n}$ some noise. $\mathbf{A}$ might be known or unknown, the latter corresponding to *blind* inverse problems. Its structure determines the specific kind of problem one has to solve. Notable inverse problems involving visual signals include denoising [2], deblurring [3], super-resolution [4], inpainting and data completion [5], dequantization [6], compressive sensing reconstruction [7], and many more. While several inverse problems involve images, novel visual data types are emerging as relevant for applications thanks to technology developments. In particular, point clouds and 3D data are gaining attention due to increased availability of instruments such as LiDARs and time-of-flight cameras and disruptive applications such as autonomous driving.

Being ill-posed, effective solutions to inverse problems require regularization to enforce prior knowledge about the signal of interest. In underdetermined problems, infinitely many solutions fit the measurements, and only prior knowledge can filter out "unrealistic" solutions and lead to a unique "realistic" one. A classic approach to address inverse

5. Distributed Learning for Adaptive Multimedia Processing

Marco Carpentiero[1], **Vincenzo Matta**[1], **Ali H. Sayed**[2]

[1] DIEM, University of Salerno - Italy
[2] School of Engineering, EPFL, Switzerland

Abstract: *Current and next-generation networks will feature an unprecedented growth of multimedia services involving distributed signal processing implementations. Notable examples are IoT networks of cameras, or distributed cloud systems deployed to provide multimedia content delivery. Distributed learning algorithms are an essential part of these types of systems. Two fundamental requirements these algorithms must cope with are data compression and adaptation. Data compression is critical to guarantee proper management of the energy/bandwidth/latency constraints that are particularly important in multimedia applications. Adaptation is critical to face the highly dynamic environment where the learning networks are called to operate. This article illustrates the fundamentals and the most recent trends in distributed learning over graphs (such as distributed regression). The focus will be on algorithms working with compressed data and infused with strong adaptation capabilities. An example of application to distributed image classification of traffic signs is provided.*

5.1 Introduction

Inference and learning algorithms are critical in extracting information hidden in data measurements [1]. The learning potential is nowadays magnified by the steady exchange of data made possible by the ubiquitous presence of *communication networks*.

In their earlier implementations, communication networks were essentially means for voice/text communication, with very limited data rates (few kbps in 2G, up to few Mbps in 3G). New service requirements, especially focused on multimedia applications, forced the evolution to data-oriented network infrastructures in 4G. The ongoing shift to 5G marks a transition towards a number of heterogeneous applications, ranging from ultra-low latency tasks (e.g., robotics, industrial automation, unmanned vehicles driving, remote healthcare, and smart grids) to massive bandwidth tasks (e.g., high-quality multimedia streaming, holographic communication, real-time gaming, and virtual reality).

In this evolving scenario, the two seemingly separate tasks of learning and communications are starting to blend. In fact, learning is becoming a critical part of current and future networks. Standardization committees and research groups are working to integrate machine learning and artificial intelligence tools at different levels of the network architecture, from the physical and medium access layers to the application layer. Numerous tasks, both in the network control and user planes, require sophisticated inferential algorithms to handle the increasingly stringent performance targets and the increasing

6. Multimodal Deepfake Detection Using Audio-Visual Cues

Davide Salvi[1], **Sara Mandelli**[1], **Honggu Liu**[2], **Paolo Bestagini**[1], **Stefano Tubaro**[1]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy
[2] School of Cyber Science and Technology, University of Science and Technology of China, China

Abstract: *The widespread application of deep learning techniques for generating authentic-looking synthetic media poses a significant threat to individuals, organizations, and society at large. Among these synthetic media forms, deepfakes, which involve swapping the identity of an individual in videos at either the visual (i.e., face) or audio (i.e., speech) level, present a particular concern. Given the potential for malicious use leading to undesirable consequences, it has become increasingly crucial to differentiate between authentic and fake media. Despite the ability of deepfake generation systems to create compelling audio-visual content, they often struggle to maintain consistency across different data modalities. This includes challenges in producing realistic video sequences where both visual and audio components align seamlessly. Additionally, these systems may struggle in accurately replicating semantic and temporally accurate aspects. Exploiting these inconsistencies becomes critical to developing robust methods for detecting fake content. In this study, we demonstrate the feasibility of detecting deepfake video sequences by capitalizing on data multimodality. Our approach involves extracting audio-visual features from input videos over time and analyzing them using time-aware neural networks. By leveraging both video and audio modalities, we exploit inconsistencies between and within them, thereby improving the overall detection performance. Notably, our method is distinct in that it is never trained on multimodal deepfake data but instead relies on disjoint monomodal datasets containing visual-only or audio-only deepfakes. This design choice enables us to sidestep the need for multimodal datasets during training, a notable gap in the existing literature. Furthermore, during testing, it allows us to evaluate the robustness of our proposed detector on previously unseen multimodal deepfakes.*

6.1 Introduction

Recent advances in deep learning and new media technologies have made the creation and sharing of multimedia content more accessible than ever. Users can now generate super realistic synthetic images, videos and speech tracks with minimal effort and without requiring any particular skill. The growth of these technologies can have a twofold effect. On one side, such techniques allow consumers to explore new creative and artistic possibilities and introduce applications that make everyday life easier. On the other hand, they can also lead to dangers and threats when misused. An example of the latter case

7. Uncertainty-driven detection and localization of image forgeries

Fabrizio Guillaro, Davide Cozzolino, Giovanni Poggi, Luisa Verdoliva

Università degli Studi di Napoli Federico II

Abstract: *In the last ten years, remarkable advancements have been made in synthetic media generation, largely due to the development of powerful deep learning-based methods. These new opportunities stimulate the creativity of benign and malicious users alike. Previously, creating a multimedia disinformation campaign demanded advanced skills, limiting attackers to basic manipulations such as copying, replicating, or removing objects, also known as "cheapfakes". However, with the rapid advancement of deep learning, image manipulation tools have not only become more user-friendly but also significantly more powerful. This has enabled users to easily modify images using textual descriptions. This chapter presents part of the research activities on image forensics developed at the University of Napoli Federico II (in collaboration with Google Research) and describes a trustworthy detection and localization strategy that can cope with a wide variety of manipulation methods, from classic cheapfakes to more recent manipulations based on deep learning. We exploit both high-level and low-level traces by properly combining the RGB image and a learned noise-sensitive fingerprint. Forgeries are detected as deviations from the expected regular pattern that characterizes real images. Beyond a pixel-level localization map, the algorithm outputs also a reliability map that highlights areas where localization predictions may be erroneous, and a global integrity score. This latter contributes significantly to reduce false alarms. Experiments on several datasets demonstrate that our method can reliably detect and localize both cheapfakes and deepfakes and achieves state-of-the-art results. The code and trained network of our work are publicly available*[1].

7.1 Introduction

In the last few years, multimedia forensics has been drawing ever-increasing attention in the scientific community. In fact, editing software tools have become easier and more powerful and can be used to create realistic image manipulations even using natural language prompts and adapting the inserted content to the existing context (see Fig. 1). In the wrong hands, this capability may represent a major threat and a powerful asset for spreading disinformation. This justifies the interest of the research community and of governments and funding agencies to design tools for the detection and localization of such manipulated content.

A large number of methods have been proposed so far. These approaches typically exploit the low-level artifacts that are caused by the in-camera acquisition process, such as the sensor, the lens, the color filter array or the JPEG compression algorithm [1], These

[1] https://grip-unina.github.io/TruFor/

8. On-the-move Multimodal Biometric Recognition

Emanuele Maiorana[1], **Lorenzo Giusti**[2], **Patrizio Campisi**[1]

[1] Department of Industrial, Electronic and Mechanical Engineering, Roma Tre University, Rome, Italy
[2] Department of Computer, Control and Management Engineering, University La Sapienza, Rome, Italy

Abstract: *Biometric recognition systems are nowadays one of the most used ways to validate an individual's identity. As such, they are naturally among the main candidates to be used in contexts with high-security requirements, such as those associated with border control, as in airports. Traditionally, the traits employed for such purpose include fingerprints and facial images, with the recent addition of iris. Yet, to ensure the ever greater security of such systems and greater ease of use, solutions based on novel traits and innovative modalities of acquisition of the already employed characteristics are highly desired. Within this context, this chapter illustrates a reliable, low-cost, and convenient solution designed and deployed to perform security checks within an airport. The presented approach relies on a multimodal biometric recognition system exploiting face traits to trigger the recognition process at a distance and subcutaneous vein patterns to guarantee the desired robustness against malicious spoofing attempts. Furthermore, the implemented system requires users to have only minimal interaction with it, having been designed to enable an on-the-move recognition modality, where subjects do not need to stop at a gate to provide the data upon which the recognition process has to be carried out since the interested characteristics can be captured. In contrast, users continue to walk, thus increasing the system throughput and the passengers' convenience. The design of the proposed acquisition device, the characteristics of employed processing boards, and the algorithms developed to perform the acquisition of the interested biometric traits, as well as to carry out the recognition process, are outlined in the chapter.*

8.1 Introduction

Border management has always been a topic of significant interest for the security of countries and for ensuring the controlled transit of travelers worldwide. Given the importance of this context, over the years, various methods have been used to monitor the flows of people in transit areas such as ports, airports, or border customs, all with the primary objective of allowing rapid and effective identification of anyone wishing to leave or enter a particular country [1]. Traditionally, and still today, these activities have been carried out through a passport issued by the competent authorities of the countries where a person resides and, therefore, certified by them, containing the information necessary to carry out secure recognition. The modern concept of a passport, with the functions

9. Digital Twins for Data Generation: the use case of Crowd Modeling and Understanding

Antonio Luigi Stefani, Niccolò Bisagno, Nicola Garau, Nicola Conci

Dipartimento di Ingegneria e Scienza dell'Informazione, Università di Trento, Italy

Abstract: *The development of so-called Digital Twins has emerged as a relevant research topic due to their ability to adequately represent real-world phenomena in a virtualized fashion. The applications are countless and span from industry-related domains for predictive maintenance purposes to more complex scenarios, such as the simulation of traffic scenes and human-related events, for planning, and adverse events prevention. This contribution aims to investigate the role of Digital Twins in the study and development of solutions capable of resembling real-world scenes fulfilling the requirements of visual and behavioural fidelity, also outlining viable validation pipelines. To this goal, effective tools have been proposed at both research and commercial levels, including realistic engines for visual rendering and sophisticated solutions capable of resembling the behaviour of moving objects, including cars and humans. In this work, we provide a holistic view of the world of Digital Twins, highlighting the benefits and potential issues that might arise when dealing with synthetically generated data. We also present the use case of crowd modelling and understanding as an integrated solution that deals with the simulation, rendering, and analysis of the relevant events.*

9.1 Introduction

The incessant need for data to propel the advancement of modern machine learning and deep learning algorithms has driven the quest for innovative solutions. Yet, as successes emerge in fields like image classification [1], natural language processing [2], speech recognition [3], medical imaging [4], and remote sensing [5], fresh challenges continuously arise, exposing the constraints of established architectures and their associated datasets.

To address this perpetual challenge, researchers explored multiple solutions, encompassing both supervised and unsupervised learning paradigms. Supervised learning relies on vast labelled datasets, but the lack of annotated data often leads to overfitting and diminished generalization capabilities. The literature has shown that it is possible to mitigate this problem with various regularization techniques, such as dropout [6], batch normalization [7], transfer learning between distinct datasets [8], pre-training the network on different datasets [9], or implementing few-shot [10] and zero-shot learning [11].

An alternative strategy to mitigate the data scarcity problem involves a focus on data-centric solutions, with fine-tuning and data augmentation emerging as prevalent strategies [12]. In the fine-tuning process, a network previously tailored for a specific task is adapted to recalibrate itself for a different problem. With data augmentation, instead, we recognize the inadequacy of both the quantity and quality of available training data;

10. Affective-based Modelling Approaches for Quality of Experience-based Management Systems

Alessandro Floris, Simone Porcu, Matteo Anedda, Daniele Giusto

Department of Electrical and Electronic Engineering, University of Cagliari, Italy

Abstract: *The ubiquity of multimedia content consumption in human digital lives has made Quality of Experience (QoE) a pivotal factor in user satisfaction and engagement. However, to build personalised QoE models, extensive, costly, and time-consuming subjective studies are required, which are affected by several limitations (e.g., the need to ask for explicit feedback from the user) and are unsuitable for real-time management systems. Thus, novel QoE models based on the analysis of the affective state of the user while consuming multimedia content have gathered particular attention in recent years. Machine learning techniques are essential in this regard for supporting the analysis of vast amounts of user emotional data (e.g., features extracted from facial expressions, speech, and heart rate) and indicating a possible correlation between these features and the perceived QoE. In this Chapter, affective-based modelling approaches for QoE estimation are discussed. In particular, two QoE estimation systems based on facial-related features are presented, which are developed for video streaming and WebRTC-based applications, respectively. The QoE estimation performance achieved by these models demonstrates the potential of affective-related indicators in estimating the QoE for multimedia services.*

10.1 Introduction

In recent years, Internet usage has surged significantly, driven largely by the popularity of video streaming platforms (such as YouTube, Amazon Prime Video, and Netflix) and telemeeting services (such as Microsoft Teams, Google Meet, and Zoom). Currently, it is estimated that video traffic constitutes nearly 66% of the global Internet traffic [1]. The increasing demand for high-quality multimedia content, combined with the rising user expectations for interactive applications, emphasizes the need for user-centred approaches in both application and network management. Indeed, efficient resource allocation on the network and server side is crucial to maintain desired quality levels and ensure a positive user experience.

The evaluation of the Quality of Experience (QoE) plays a crucial role in this context. The concept of QoE, which is defined as "*the degree of delight or annoyance of the user of an application or service*", has emerged because the Quality of Service (QoS) was not robust enough to comprehensively encapsulate all the elements involved in current and emerging interactive multimedia services [2]. The QoE is a user-centred measure that aims to consider the user's personal quality perception of multimedia services that depends