PhD Thesis Proposal 2025 – MESR/AAP Région NA Funding

Coding for Machines and Humans for Visual Data

Summary and Context

With the explosion of visual data, image and video compression methods must adapt to diverse needs: maximizing compression while minimizing perceived quality loss for humans and ensuring compact yet usable representations for machines. Until now, research has explored two distinct approaches: compression for humans, based on visual perception, and compression for machines, optimized for artificial vision tasks. However, these two paradigms are often developed separately and rely on contradictory objectives. Some studies have attempted to develop hybrid approaches combining both aspects. Recent advances in deep learning and perceptual modeling open the way for hybrid compression, capable of dynamically adapting to the specific needs of both machines and humans.

State of the Art and Current Limitations

Historically, image and video compression methods have been divided into two main categories: those optimized for human perception and those dedicated to artificial vision algorithms.

- Compression for Humans: Classical codecs like JPEG or H.26x aim to maximize visual fidelity while minimizing the amount of stored or transmitted data. These approaches can be optimized using principles from cognitive psychology, such as Just Noticeable Difference (JND) or Satisfied User Ratio (SUR) [6], to eliminate information deemed non-essential by the human visual system [5]. However, these codecs do not consider the needs of AI models, which can introduce undesirable artifacts that disrupt the automatic analysis of images and videos.
- Compression for Machines: With the rise of computer vision algorithms and AI applied to images, machine-oriented compression has become a crucial research area. Standards like Video Coding for Machines (VCM), recently promoted by MPEG [1][3], aim to optimize video compression for automated analysis tasks such as classification, segmentation, and object detection. Additionally, variational autoencoders (VAE), generative adversarial networks (GANs), and visual transformers [7] have enabled the generation of compact latent representations while maintaining high performance for these tasks [10]. However, these methods lack interpretability, as they produce representations that are not necessarily readable by humans and do not allow faithful reconstruction of the original image.

Given these limitations, a new research direction is to unify these two approaches by developing hybrid compression models capable of dynamically adapting to the needs of both machines and humans. Recent methods like TransTIC [4] have attempted to transfer codecs optimized for human perception to artificial vision tasks without requiring retraining. Moreover, generative compression models [2], based on diffusion models and GANs, offer promising perspectives by enabling optimized image reconstruction based on the type of user. However, these approaches remain computationally expensive and are not suited to the energy constraints of embedded systems.

Another promising advancement is the integration of neuromorphic computing into compression systems. Inspired by biological neural networks, these architectures allow for dynamic information management and adaptive transmission based on visual stimuli and

computational needs. Specifically, spiking neural networks (SNNs) have shown great potential for energy-efficient compression and event-driven data transmission. Furthermore, bio-inspired hierarchical coding models, inspired by the functioning of the human visual cortex, could enable adaptive compression, where resource allocation is modulated based on perceptual importance and analytical requirements.

PhD Objectives

The goal of this PhD project is to design a hybrid compression model capable of simultaneously meeting the requirements of artificial vision systems and human perception, while ensuring explainability and dynamic adaptability. Unlike traditional approaches that separately optimize compression for machines or humans, this project aims to unify both paradigms by leveraging recent advances in signal processing, computer vision, and explainable artificial intelligence (XAI).

The proposed approach will be based on:

- Deep learning models and adaptive coding strategies integrating explainability mechanisms to make compressed representations interpretable for both human users and artificial vision algorithms.
- Explainable attention models and perceptual regularization strategies, improving both visual fidelity and coding efficiency for automated tasks, while enhancing control over information loss due to compression.
- Neuromorphic architectures, particularly spiking neural networks (SNNs) and adaptive associative memories, to explore bio-inspired circuits that reduce energy consumption and enhance computational scalability in compression.

By combining deep learning, neuromorphic computing, and XAI, this PhD will contribute to next-generation image and video compression techniques, optimizing for both human perception and machine analysis while remaining suitable for embedded systems and massive video streams.

References

[1] Zhang, Q., Mei, J., Guan, T., Sun, Z., Zhang, Z., & Yu, L. (2024). Recent Advances in Video Coding for Machines Standard and Technologies. ZTE Communications, 22(1), 62.

[2] Chen, B., Yin, S., Chen, P., Wang, S., & Ye, Y. (2024). Generative Visual Compression: A Review. arXiv preprint arXiv:2402.02140.

[3] Lee, D., Jeon, S., Jeong, Y., Kim, J., & Seo, J. (2023). Exploring the Video Coding for Machines Standard: Current Status and Future Directions. Journal of Broadcast Engineering, 28(7), 888-903.

[4] Chen, Y. H., Weng, Y. C., Kao, C. H., Chien, C., Chiu, W. C., & Peng, W. H. (2023). Transtic: Transferring transformer-based image compression from human perception to machine perception. Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 23297-23307).

[5] Ballé, J., Laparra, V., & Simoncelli, E. P. (2016). End-to-end optimized image compression. arXiv preprint arXiv:1611.01704.

[6] Zhang, Q., Wang, S., Zhang, X., Jia, C., Wang, Z., Ma, S., & Gao, W. (2024). Perceptual Video Coding for Machines via Satisfied Machine Ratio Modeling. IEEE Trans. on Pattern Analysis and Machine Intelligence.

[7] Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. IEEE Int. Conf. on Computer Vision.

[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should I trust you? Explaining the predictions of any classifier. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (pp. 1135-1144).

[10] Cheng, Z., Sun, H., Takeuchi, M., & Katto, J. (2020). Learned image compression with discretized Gaussian mixture likelihoods and attention modules. IEEE/CVF Conf. on Computer Vision and Pattern Recognition.