Affective Communication of Sensorimotor Emotion Synthesis over URLLC

Ibrahim M. Amer*, Sharief M. A. Oteafy†*, and Hossam S. Hassanein*
*School of Computing, Queen's University, Kingston, ON, Canada
‡School of Computing, DePaul University, Chicago, Illinois, USA
ibrahim.amer@queensu.ca, soteafy@depaul.edu, hossam@cs.queensu.ca

2023 IEEE 48th Conference on Local Computer Networks (LCN) | 979-8-3503-0073-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/LCN58197.2023.10223329

***Abstract*— Affective computing is an emerging field that aims to develop technologies capable of recognizing and responding to human emotions. However, during communication sessions, the exchange of a high volume of data can cause high latency. One approach to mitigating this issue is semantic communication, which may reduce the amount of data exchanged. Hereby, we propose a novel idea that utilizes semantic communication in affective computing by minimizing the amount of information exchanged between endpoints. Specifically, we examine a use case of a remote doctor application, where a patient's emotions are captured, and their vital signs are obtained using wearable devices, with this information reported to a remote doctor. To reduce data exchange, we utilize semantic communication to extract the meaning of the conveyed information, rather than transmitting the raw information itself. This approach can enhance the efficiency of communication in URLLC applications and has the potential to improve patient outcomes.**

***Index Terms*—Affective Computing, Semantic Communication, Artificial Intelligence, Edge Intelligence, Ultra-Low Latency, Tactile Internet, Generative Networks**

I. Introduction

Affective Computing (AC) is a multidisciplinary field of research that focuses on developing computational models and systems that can detect, interpret, and respond to human emotions, moods, and other affective states. AC was introduced by Picard et al. [1], and there has been a lot of research going on ever since. The goal of AC is to enable machines to interact with humans in a more natural and empathetic way, by understanding and responding to their emotions and affective cues. AC draws on a range of disciplines, including computer science, psychology, neuroscience, and engineering [2], and it has applications in a wide range of fields, including healthcare, education, entertainment, and marketing. Some of the key technologies used in AC include natural language processing, facial expression analysis, and physiological sensing. Human emotions can be recognized using facial expressions, text, voice, or even a combination of these.

Shannon and Weaver have categorized Semantic Communication (SC) into three levels [3]. 1) transmission of symbols; which is concerned with the transmission of the symbols from the transmitter to the receiver successfully, 2) exchanging of transmitted symbols semantically; which handles the semantic information sent from the transmitter and the interpreted meaning at the receiver, 3) and a level that deals with the effects of communication which are translated into the ability of the receiver to perform tasks as instructed by the transmitter [4]. SC focuses on the second category where the meaning of the transmitted information is the main concern.

The focus of communication systems nowadays is directed toward dealing with the transmission of data with the lowest possible latency while in the past the focus was on how accurate the transmitted data are. Ultra-Reliable and Low-Latency (URLLC) Communication applications require high network reliability, more than 99.999%, and extremely low latency of approximately 1 millisecond for data transmission. Applications such as Tactile Internet (TI), augmented and virtual reality (AR/VR) applications, real-time traffic density estimation, remote surgery, autonomous truck platooning, and factory automation are examples that follow URLLC standards.

Delay-sensitive applications, like remote surgery, necessitate minimizing roundtrip time between transmitter and receiver. This low latency requirement led to the development of measures to avert potential catastrophic situations. In the context of Tactile Internet (TI), exchanging data between endpoints is particularly challenging as it requires sending and receiving data with minimal latency while ensuring high quality. Tactile Haptic codecs emerged as a research direction focused on the codecs for signals transmitted in a TI session [5]. The IEEE P1918.1.1 ongoing standardization activity defines codecs that facilitate the exchange of haptic information (kinesthetic and tactile) by employing data reduction and compression algorithms. Moreover, the advancement of 5G technology has enabled high transmission rates, approaching the Shannon limit and expanding system capacity.

In this paper, we leverage the benefits of semantic communication to reduce the amount of information bouncing between the transmitter and the receiver which will result in a reduction of communication latency. The application we are addressing is targeted to patients and physicians to help physicians diagnose patients more effectively. We combine the techniques of affective computing and semantic communication accompanied by Generative Adversarial Networks (GANs) to achieve our goal.

The remainder of the paper is organized as follows. Section II highlights some of the related works. Section III presents the

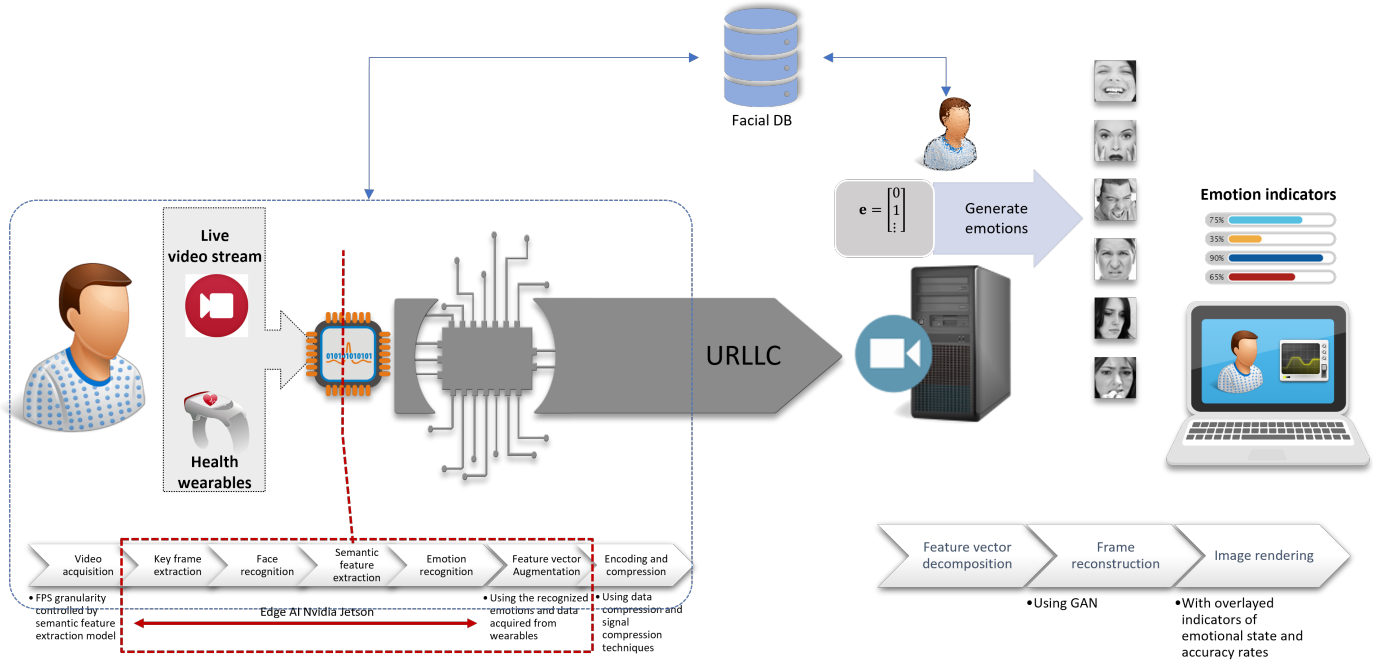

Figure 1: A system overview of the proposed scheme.

proposed AFFSEM scheme. Section IV concludes the paper.

II. Related Work

This section provides a review of state-of-the-art techniques in affective computing and semantic communication relevant to our proposed scheme. The reviewed techniques are chosen based on their prominence and applicability to our work.

Lu et al. [6] introduced a semantics-aware communication framework that incorporates semantic encoding and the semantic communication problem. The authors pursue high bit-level precision and reduced bits while transmitting the information. The proposed method establishes a Joint Semantic -Noise Coding (JSNC) solution for the semantic coding problem and RL-based similarity-targeted semantic communication mechanism.

Yang et al. [7] investigate how edge intelligence can be enhanced with semantic communication by reducing the communication overhead of information exchange and improving the capabilities of the intelligent agents at lower computation overheads. The authors address the costly implementation overheads of training, maintaining, and executing Semantic Extraction (SE) by introducing federated learning-enabled SE for semantic management.

Li et al. [8] proposed a scheme for facial Emotion Recognition (ER) that fully leverages all unlabeled data for semi-supervised deep facial expression recognition, not only parts of the unlabeled data. The authors evaluated their scheme on four commonly used datasets: RAF-DB, SFEW, AffectNet, and CK+. The scheme outperforms state-of-the-art schemes available in the literature.

Zhan et al. [9] presented a scheme that exploits zero-shot learning mechanisms in ER tasks. The goal is to recognize the new unseen emotions. The authors utilized Adjective-Noun Pairs (ANP) features that utilize mid-level semantic representation to construct the intermediate embedding space. The learned intermediate space is used to bridge the gap between the low-level visual features and the high-level semantics.

Roich [10] proposed a Pivotal Tuning Inversion (PTI) method for editing real-life facial images using latent-based editing techniques. The method addresses the inherent tradeoff between distortion and editability in StyleGAN's latent space, enabling effective ID-preserving facial latent-space editing for out-of-domain faces. PTI slightly alters the generator, allowing faithful mapping of out-of-domain images to in-domain latent codes through a brief training process. This technique preserves editing quality while changing portrayed identity and appearance. It demonstrates superior performance compared to state-of-the-art methods in inversion and editing metrics and showcases its ability to handle challenging cases, such as heavy makeup or elaborate hairstyles.

In this work, we propose the conjoining of SC and AC and explore the research field where AC and SC meet. This research field is called Affective Communication (AffCom).

III. Affective Semantics (AffSem)

In this section, we discuss our proposed idea, namely AFFSEM. Emotions play a critical role in our day-to-day communication, as they allow us to convey our feelings, intentions, and desires to others. They provide vital cues to help us understand how others are feeling, enabling us to respond appropriately and effectively. Emotions also help us establish and maintain social connections, allowing us to bond with others, express empathy, and build trust. Therefore, understanding and being aware of our emotions, as well as those of others, is essential to successful communication and building healthy relationships.

In this paper, we are mainly focusing on facial emotions but other kinds of emotions are also applicable. We focus on the seven basic facial emotions that most work in affective computing has focused on: fear, anger, happiness, sadness, disgust, surprise, and neutral [11]. We apply our proposed scheme to a case study related to telehealth where a patient needs a diagnosis.

The proposed scheme has some steps that will be done on the patient's side and some other steps that will be done on the physician's side. The main steps on the patient's side can be summarized as follows: 1) Data acquisition, 2) data preprocessing, 3) semantic feature extraction, 4) emotion recognition, 5) features vector construction, and 6) encoding and compression The steps required to display the patient's information on the physician's side can be summarized into the following: 1) feature vector decomposition, 2) frames reconstruction using emotion synthesis, and 3) image rendering these steps are thoroughly illustrated in Figure 1 and will be further discussed in the following subsections.

A. Patient's Side (Transmitter)

In this subsection, we will elaborate more on how the different components of the system are interacting on the patient's side. We will use an NVIDIA Jetson Kit as an edge processing server instead of relying on the remote cloud servers that could negatively affect the latency. Key frames extraction, emotion recognition, semantic features extraction, and data encoding and compression will all be performed on the Kit.

1) Data acquisition

The data acquisition involves capturing a stream of video of the patient to capture their traits, facial emotions, and voice. We will use a camera with a resolution of 720×480. We note that the camera in this step could be of any resolution as we will not be sending any actual frames through the communication link. We can also capture the patient's vital signs using wearable devices such as wristbands that can measure a variety of things that could be beneficial in the diagnosis; it can measure the body temperature, IBI or systolic peaks, blood volume pulse, electrodermal activity, and respiratory rate.

2) Key Frames Extraction

As not all frames of a captured video contain pertinent information, selecting only the meaningful frames, or key frames is essential. Keyframes represent a small subset of all the frames included in the video. In our study, we consider keyframes to be those that capture a patient's facial expressions during an abrupt or gradual change in emotional state. For example, if a patient's facial expression remains static over 100 frames, it would be inefficient to process all 100 frames, and instead, a single frame would suffice. Key frame extraction, commonly used in video summarization schemes [12], can also be applied to our proposed approach.

3) Semantic Feature Extraction

This module will extract features that will be most helpful for identifying emotions for the next stage. This will be performed using deep learning architectures pre-trained and fine-tuned for this task.

4) Emotion Recognition

Captured patient's emotions can be used in the diagnosis. For instance, a droopy face can be detected through facial emotions which can help diagnose stroke. The droopy face is one of the key signs of ischemic attacks according to the FAST assessment [13]. Deep learning models will use learnt semantic features for emotion prediction.

5) Feature Vector Construction

This module involves creating the feature vector that will be sent to the remote physician. The first k elements of the feature vector will contain the classified emotions of the k key frames extracted from the captured video. The classified emotions are denoted by $\mathbf{e}$. The next 720×480 elements will represent a compressed patient's image for identification purposes. The patient's facial image is denoted by $\mathbf{f}$. The next elements of the feature vector represent the features acquired by the wristband and are denoted by the vector $\mathbf{w}$. The whole feature vector is given by the column vector $\mathbf{v} = \left[\mathbf{e}\ \mathbf{f}\ \mathbf{w}\right]^{\intercal}$.

6) Data Encoding and Compression

In this step, we aim to reduce the required data rate and perceive the signal information as much as possible. We plan to use a technique that is being used in Vibrotactile signal encoding that is used to transmit information in Tactile Internet applications. We will use the method adopted by Noll et al. [14]. The encoded data will be then sent through a URLLC communication link to ensure yet reliable and low latency transmission of bits.

B. Physician's Side (Receiver)

After acquiring and transmitting the core semantic information extracted from the last step, the physician's side (receiver) will process the input and interpret it in a way that could help the physician diagnose the patient. In this subsection, we will illustrate the steps required to achieve this.

1) Feature Vector Decomposition

As depicted in Figure 1, The feature vector received will be processed to extract the relevant information. First, the first k

elements of the vector $\mathbf{v}$ representing the classified emotions in the k key frames are procured and assigned to vector $\mathbf{e}$. Then, the subsequent 720×480 elements that represent the patient's facial image will be extracted and assigned to vector $\mathbf{f}$. Finally, the remaining elements of the vector that represent the data acquired by the wristband are extracted and assigned to the vector $\mathbf{w}$.

2) Frame Reconstruction

In this step, the video server will strive to reconstruct the video frames from the classified emotions in vector $\mathbf{e}$. We can utilize Generative Adversarial Networks (GANs) to achieve this task. We can use the methods adopted by the work done in [15], [16] can be used to synthesize facial expressions.

3) Image Rendering

The ultimate step of the system entails presenting data on the patient, which includes indicators of their emotional state and a live synthetic video feed of their facial expressions. Such information can be utilized by the attending physician to facilitate the diagnosis process. By providing a continuous stream of data on the patient's emotional and physical state, the system enables the physician to gain a comprehensive understanding of the patient's condition, thereby maintaining the accuracy of the diagnosis and reducing the communication load.

IV. Conclusions & Future Work

In this paper, we combine the techniques of Affective Computing (AC) and Semantic Communication (SC). AC is concerned with how human emotions are detected, interpreted, and responded to, while SC is concerned, mainly, with the exchanging of transmitted information semantically. Our system architecture shows a case study of a remote doctor application. At the patient's side, we first detect and recognize human emotions acquired from a real-time video stream. Next, the semantic features are extracted from the video's frames that do not include any visual information. After that, the extracted features are sent to the physician. Finally, at the physician's side, the video frames are reconstructed using Generative Adversarial Networks (GANs) with the extracted semantic features. The goal of our proposed method is to reduce the latency imposed by sending visual information using only the information's semantics.

Acknowledgment

This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant number ALLRP 549919-20.

References

[1] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[2] J. Tao and T. Tan, "Affective computing: A review," *Lecture Notes in Computer Science*, vol. 3784 LNCS, pp. 981–995, 2005.

[3] C. E. Shannon and W. Weaver, *The mathematical theory of communication, by CE Shannon, W. Weaver*. University of illinois Press, 1949.

[4] H. Xie, Z. Qin, G. Y. Li, and B. H. Juang, "Deep Learning Enabled Semantic Communication Systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[5] E. Steinbach, M. Strese, M. Eid, X. Liu, A. Bhardwaj, Q. Liu, M. Al-Ja'Afreh, T. Mahmoodi, R. Hassen, A. El Saddik, and O. Holland, "Haptic codecs for the tactile internet," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 447–470, 2 2019.

[6] K. Lu, Q. Zhou, R. Li, Z. Zhao, X. Chen, J. Wu, and H. Zhang, "Rethinking Modern Communication from Semantic Coding to Semantic Communication," *IEEE Wireless Communications*, 2 2022.

[7] W. Yang, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Cao, and K. B. Letaief, "Semantic Communication Meets Edge Intelligence," *IEEE Wireless Communications*, vol. 29, no. 5, pp. 28–35, 10 2022.

[8] H. Li, N. Wang, X. Yang, X. Wang, and X. Gao, "Towards Semi-Supervised Deep Facial Expression Recognition with An Adaptive Confidence Margin," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 4156–4165, 2022.

[9] C. Zhan, D. She, S. Zhao, M. M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 1151–1160, 10 2019.

[10] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal Tuning for Latent-based Editing of Real Images," *ACM Transactions on Graphics*, vol. 42, no. 1, 6 2021. [Online]. Available: https://arxiv.org/abs/2106.05744v1

[11] S. D'Mello, A. Graesser, and R. W. Picard, "Toward an affect-sensitive autotutor," *IEEE Intelligent Systems*, vol. 22, no. 4, pp. 53–61, 7 2007.

[12] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 7894–7903, 6 2019.

[13] J. Kennedy, M. D. Hill, K. J. Ryckborst, M. Eliasziw, A. M. Demchuk, and A. M. Buchan, "Fast assessment of stroke and transient ischaemic attack to prevent early recurrence (FASTER): a randomised controlled pilot trial," *The Lancet Neurology*, vol. 6, no. 11, pp. 961–969, 2007.

[14] A. Noll, L. Nockenberg, B. Gulecyuz, and E. Steinbach, "VC-PWQ: Vibrotactile Signal Compression based on Perceptual Wavelet Quantization," *2021 IEEE World Haptics Conference, WHC 2021*, pp. 427–432, 7 2021.

[15] R. Bodur, B. Bhattarai, and T. K. Kim, "3D dense geometry-guided facial expression synthesis by adversarial learning," *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, pp. 2391–2400, 1 2021.

[16] F. T. Hong, L. Zhang, L. Shen, and D. Xu, "Depth-Aware Generative Adversarial Network for Talking Head Video Generation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 3387–3396, 2022.