Demo: Leveraging Edge Intelligence for Affective Communication over URLLC

Ibrahim M. Amer*, Sarah Adel Bargal†, Sharief M. A. Oteafy*‡, and Hossam S. Hassanein*
*School of Computing, Queen's University, Kingston, ON, Canada
†Department of Computer Science, Georgetown University, Washington, D.C., USA
‡School of Computing, DePaul University, Chicago, Illinois, USA
ibrahim.amer@queensu.ca, sarah.bargal@georgetown.edu, 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/LCN58197.2023.10223383

***Abstract*— IoT systems are advancing to enable higher levels of engagement and omnipresence. A critical yet uncharted domain lies in communicating *affect* across participants, especially in medical settings where emotions and expressions are pivotal, and eXtended Reality (XR) systems that rely on immersive virtualization of the participating parties. However, IoT systems seldom have the bandwidth or reliability to enable such services. In this demo, we present an experiment that leverages Edge Intelligence and Artificial Intelligence to extract and encode emotions at one edge, and communicate a low-footprint encapsulation of such emotions at the other edge. The proposed architecture is designed to reduce overall traffic and build on low-power video and display equipment, to realize Affective Semantic Communication (AffSeC). This demonstration shall represent AffSeC in a medical setting, where a patient interacts with a physician over a low-BW E2E route. The proposed scheme will be contrasted to standard video compression to demonstrate the efficacy and promise of this model.**

***Index Terms*—Affective Computing, AffSeC, Emotion Recognition, Artificial Intelligence, Edge Intelligence, Semantic Communication, Ultra-Low Latency, Tactile Internet**

I. Introduction and Motiviation

Currently, the emphasis in communication systems lies in achieving the shortest possible delay when transmitting data, whereas in the past, the main focus was on the accuracy of the transmitted data. Ultra-Reliable and Low-Latency Communication (URLLC) applications necessitate a highly reliable network, exceeding 99.999% reliability, and extremely low latency of approximately 1 millisecond for data transmission. Applications such as Tactile Internet (TI), augmented and virtual reality (AR/VR), real-time traffic density estimation, remote surgery, autonomous truck platooning, and factory automation exemplify domains that adhere to URLLC standards.

Applications sensitive to delays, like remote surgery, cannot tolerate the time it takes for data to travel between transmitter and receiver. To address this, novel approaches have been developed to prevent disastrous situations. In teleoperation interfaces, the exchange of data between endpoints poses challenges, requiring minimal delay and high-quality transmission. One area of research focuses on developing Tactile Haptic codecs for transmitting tactile and haptic signals in a Teleoperation Interface (TI) session. The developing IEEE P1918.1.1 standard establishes haptic codecs for exchanging kinesthetic and tactile information [1]. It includes algorithms for reducing and compressing data to enable the communication of such information. Additionally, the progress of 5G technology has increased transmission rates, nearing the theoretical Shannon limit.

In this work, we propose AffSeC, which focuses on the semantic communication of human emotion captured from video via affective computing from one edge to the other. Shannon and Weaver classified Semantic Communication (SC) into three levels [2]. 1) The first level involves the successful transmission of symbols from the transmitter to the receiver. 2) The second level pertains to the exchange of semantically meaningful symbols, encompassing the semantic information transmitted by the transmitter and its interpreted meaning at the receiver. 3) The third level addresses the impact of communication, which translates into the receiver's ability to carry out tasks as instructed by the transmitter [3]. Within SC, the primary focus of this work lies in the second category, which emphasizes the significance of the conveyed information's meaning. SC transmits only the relevant information to the receiver which can lead to a significant reduction in data volume [4]. Delivering the meaning of the transmitted signal depends on the physical content and the intention of the message and any other factor that could affect the Quality of Experience (QoE) [5].

The field of affective computing aims to bridge the gap between humans and machines by developing computational models and systems that can detect, interpret, and respond to human emotions, moods, and other affective states. Driven by the pioneering work of Picard et al. [6], extensive research has been conducted in this interdisciplinary domain. The primary objective of affective computing is to enable machines to engage with humans more naturally and empathetically, comprehending and reacting to their emotional cues and affective signals.

By drawing upon diverse fields such as computer science, psychology, neuroscience, and engineering [7], Affective Computing presents a foundational block for applications in various domains including healthcare, education, entertainment, and marketing. Their success draws from key technologies

Authorized licensed use limited to: Queen's University. Downloaded on January 09,2024 at 15:30:55 UTC from IEEE Xplore. Restrictions apply.

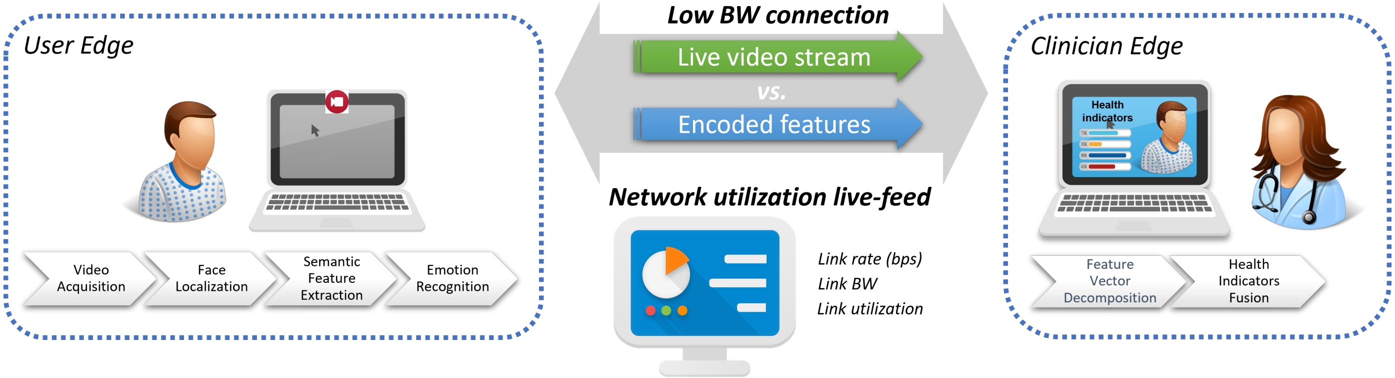


Figure 1: A functional diagram of AffSeC. Affective and identity features are communicated from the user edge to the clinician edge across a low bandwidth connection. In our demo we will provide live feedback that contrasts the network utilization for (a) sending a compressed video, and (b) sending affect features, in terms of line rate, link bandwidth, and link utilization. Steps at the user edge include recognizing the face from the captured video, and recognizing the emotion based on semantically extracted features. Such emotions are communicated with identity information to the clinician side where feature decomposition and health indicators fusion happens.

employed in affective computing, including natural language processing [8], analysis of facial expressions [9], [10], and physiological sensing [11]. These techniques enable the recognition of human emotions through multiple modalities, such as facial expressions, text, voice, and more recently a combination thereof. At its core, affective computing seeks to enable machines to perceive and respond to the rich tapestry of human emotions. By comprehending and adapting to emotional cues and affective signals, machines can engage with humans in a more natural and empathetic manner. The ultimate goal is to create a seamless interaction where machines understand not only the explicit commands but also the underlying emotional context. This empathetic understanding can enhance the user experience and foster more meaningful connections between humans and machines.

This paper presents a demo that builds on semantic communication in order to minimize the information exchange between the sender and the receiver, thereby reducing communication delay. The objective is to enhance the diagnostic process for physicians by developing an application for patients and doctors that is able to encode human emotions on one edge and communicate them to the other edge targeting a low communication footprint. Our approach involves integrating affective computing and semantic communication techniques, to successfully achieve our objective. This builds on foundational architectures for leveraging Edge intelligence and contextualizing TI interactions, as outlined by Oteafy and Hassanein in [12].

The remainder of the paper is organized as follows. Section II presents the scope of the demo. Section III previews a sample demo for our scheme. Section IV concludes the paper.

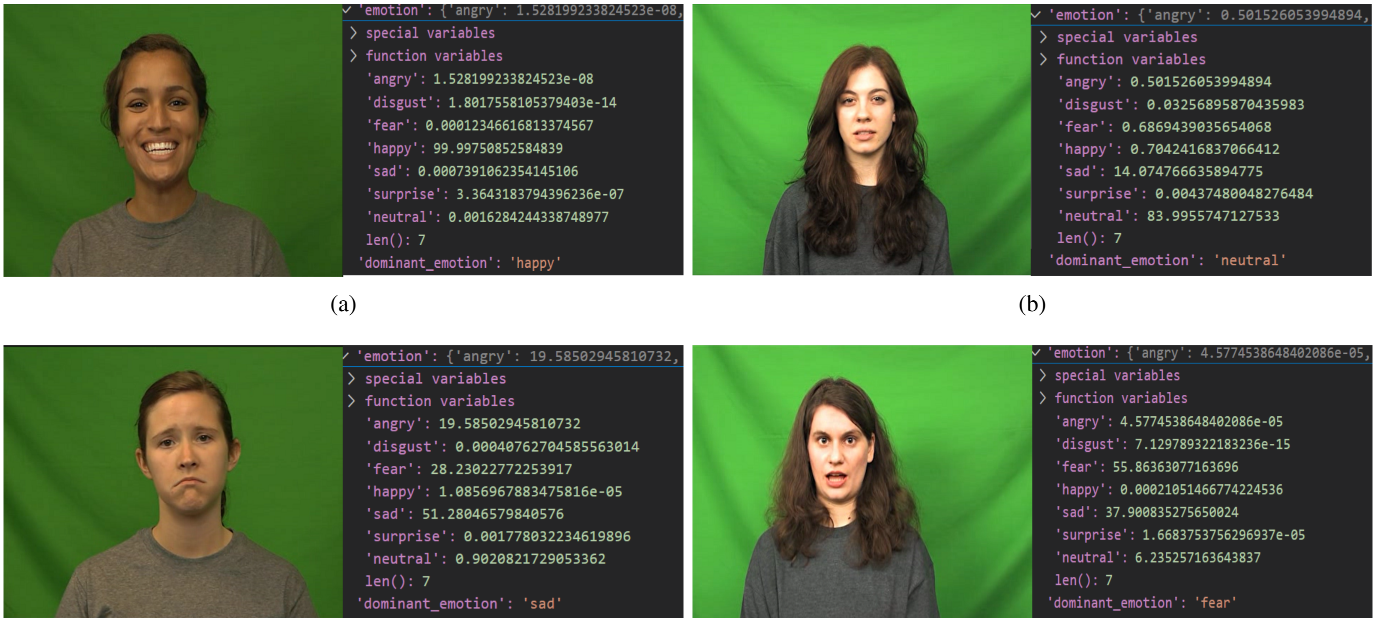
II. Scope and Implementation of AffSeC

We propose AffSeC, a scheme that transmits the emotional states of a person acquired using a video stream. Emotion labels are solely transmitted, No images or video frames are sent down the communication link. Transmitting visual information typically encompasses a high communication footprint being a function of the frame rate and the number of per-frame pixels. AffSeC focuses on the seven basic emotions: *Happiness, Sadness, Disgust, Anger, Surprise, Fear, Neutral*.

Figure 1 illustrates the functional block diagram of AffSeC, showcasing the different stages involved in the system. The left-hand side of the diagram presents the steps conducted at the transmitter's or patient's side, while the right-hand side represents the steps carried out at the receiver's or clinician's side. At the patient's side, a video stream is captured using any camera of varying resolution. The captured video undergoes processing to extract keyframes, which are essential frames containing pertinent information. These keyframes are then subjected to further processing to recognize the patient's emotions. One downsampled image of the patient's face is transmitted to the clinician's edge for additional identification purposes. The emotions detected and the patient's identity are encapsulated as a feature vector, transmitted over a low-bandwidth route to the physician's side. This demo paper highlights the processes involved in capturing, extracting, recognizing emotions, and transmitting the feature vector, enabling the communication of affect to the physician's side.

Automated emotion prediction from facial expressions in a video stream started using classical machine learning techniques such as Support Vector Machines [9]. Later on, deep learning was employed for emotion recognition from images and videos [10]. SDKs and Toolboxes have been published for emotion recognition from visual data [14]–[16]. In this work,

Authorized licensed use limited to: Queen's University. Downloaded on January 09,2024 at 15:30:55 UTC from IEEE Xplore. Restrictions apply.



```
✓ 'emotion': {'angry': 1.528199233824523e-08,
 > special variables
 > function variables
   'angry': 1.528199233824523e-08
   'disgust': 1.8017558105379403e-14
   'fear': 0.00012346616813374567
   'happy': 99.99750852584839
   'sad': 0.0007391062354145106
   'surprise': 3.3643183794396236e-07
   'neutral': 0.0016284244338748977
   len(): 7
  'dominant_emotion': 'happy'
```

(a)

```
✓ 'emotion': {'angry': 0.501526053994894,
 > special variables
 > function variables
   'angry': 0.501526053994894
   'disgust': 0.03256895870435983
   'fear': 0.6869439035654068
   'happy': 0.7042416837066412
   'sad': 14.074766635894775
   'surprise': 0.00437480048276484
   'neutral': 83.9955747127533
   len(): 7
  'dominant_emotion': 'neutral'
```

(b)

```
✓ 'emotion': {'angry': 19.58502945810732,
 > special variables
 > function variables
   'angry': 19.58502945810732
   'disgust': 0.00040762704585563014
   'fear': 28.23022772253917
   'happy': 1.0856967883475816e-05
   'sad': 51.28046579840576
   'surprise': 0.001778032234619896
   'neutral': 0.9020821729053362
   len(): 7
  'dominant_emotion': 'sad'
```

(c)

```
✓ 'emotion': {'angry': 4.5774538648402086e-05,
 > special variables
 > function variables
   'angry': 4.5774538648402086e-05
   'disgust': 7.129789322183236e-15
   'fear': 55.86363077163696
   'happy': 0.00021051466774224536
   'sad': 37.900835275650024
   'surprise': 1.6683753756296937e-05
   'neutral': 6.235257163643837
   len(): 7
  'dominant_emotion': 'fear'
```

(d)

Figure 2: Screenshots depicting keyframes extracted from video and processed for emotion recognition. The sample videos are from the CREMA-D dataset [13]. The numbers presented next to each keyframe communicate the likelihood of the presence of each of the seven emotion classes in each frame. The ground-truth labels for the four video frames are (a) *Happiness*, (b) *Neutral*, (c) *Sadness*, and (d) *Fear*.

we will use deep learning models for the prediction of emotion from facial expressions in the video using the DeepFace open-source library [17]. This demo is only scoped to perform keyframe extraction and communicate emotion recognition. The next section provides a preview of sample screenshots of our scheme.

Our scheme starts with video acquisition using a resolution of 1920×1080 pixels. This is then downsized to be input to the architecture used for both the face localization and emotion recognition steps. Not all frames of the captured video contain pertinent information, so it is essential to select a subset of keyframes. Keyframes represent a small subset of all the frames included in the video. Therefore, we extract the keyframes using the method adopted in [18]. The next step is to localize and extract faces in the video frames . We use the DeepFace open-source library to localize the face on the patient edge [19].

The extracted faces are then downsized to 48×48 pixels to be input into the Convolutional Neural Network (CNN) deep learning neural network model. This model consists of three convolutional layers and one fully connected layer to compute the prediction results, and was pretrained on FER2013 dataset [20]. The deep learning model outputs a feature vector $v \in \mathbb{R}^7$ that computes the likelihood of the input frame to possess each of the seven emotions. We label the video frame with the most likely emotion, also called the 'network prediction.'

The identity and emotions of the patients will be used to construct the feature vector that will be sent over the communication link to the physician's side. We use low-bandwidth communication links to facilitate the advantage of sending the semantic features instead of the entire video's visual content.

The demo setup will involve two machines: one for the patient and the other for the physician. The nature of the scheme pertains to chatting applications; thus, we establish a continuous connection using socket programming to allow Inter-Process Communication (IPC) between the patient and the physician. WebSockets can also be used but, for the sake of this demo, we rely on sockets only. We establish two socket connections: one to send the visual content and the other to send only the semantic features. The reason for instantiating two connections is to use one for sending the visual content of the actual captured frames, while using the other to send only the semantic features to assess the efficiency of both methods.

In addition, the demo setup will include a tablet device used to display performance metrics to assess the efficacy of AffSec. The device will display the communication links' bit rate/second (bps), bandwidths, and utilization.

III. Controlled Demo of AffSec

For the purpose of demonstrating our scheme, we utilize the CREMA-D dataset [13]. This dataset is suitable for exploring the expression and perception of multi-modal emotions. It encompasses facial and vocal emotional expressions conveyed through sentences that portray various basic emotional states,

Authorized licensed use limited to: Queen's University. Downloaded on January 09,2024 at 15:30:55 UTC from IEEE Xplore. Restrictions apply.

including happiness, sadness, anger, fear, disgust, and neutrality. The dataset consists of 7,442 clips featuring 91 actors from diverse ethnic backgrounds. Multiple raters evaluated these clips across three modalities: audio, visual, and audio-visual but in this work, we are only concerned with the visual modality.

All modules were implemented using Python and OpenCV. Figure 2 illustrates keyframes extracted from a sample video in the CREMA-D dataset, comprising a total of 100 heavily sampled frames, showcasing multiple individuals expressing happiness, neutral, sadness, and fearfulness emotions together with a score conveying the likelihood of each emotion. Figure 2a displays an extracted keyframe, accompanied by the corresponding emotion likelihood scores, indicating the individual's dominant emotion as *Happiness*. Similarly, Figure 2b displays another extracted keyframe, along with the emotion likelihood scores, highlighting the dominant emotion of the individual as *Neutral*. Figures 2c and 2d display the outputs for *Sadness* and *Fear* emotions, respectively. While this section presents a controlled demo based on input from a dataset stored on disk, the live demo will capture input directly from a camera feed.

IV. Conclusion

In this demo paper, we present the integration of Affective Computing (AC) and Semantic Communication (SC) techniques. AC focuses on the detection, interpretation, and response to human emotions, while SC primarily focuses on the exchange of semantically transmitted information.

Our system architecture serves as a case study for a remote physician application. At the patient's end, we employ real-time semantic feature extraction from the input video stream to detect and recognize human emotions. These extracted features are then transmitted to the physician's end. Our proposed scheme aims to reduce latency by exclusively relying on lightweight semantic information instead of transmitting heavy-weight visual data. This demo will be presented as an interactive real-time demo, highlighting the importance of semantic features to enhance URLLC.

References

[1] 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.

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

[3] 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.

[4] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic Communications: Principles and Challenges," 12 2021. [Online]. Available: https://arxiv.org/abs/2201.01389v5

[5] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From Semantic Communication to Semantic-Aware Networking: Model, Architecture, and Open Problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 8 2021.

[6] 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.

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

[8] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 579–586.

[9] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proceedings of the 5th international conference on Multimodal interfaces*, 2003, pp. 258–264.

[10] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 433–436, 10 2016.

[11] S. Pal, S. Mukhopadhyay, and N. Suryadevara, "Development and progress in sensors and technologies for human emotion recognition," *Sensors*, vol. 21, no. 16, p. 5554, 2021.

[12] S. M. Oteafy and H. S. Hassanein, "Leveraging tactile internet cognizance and operation via IoT and edge technologies," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 364–375, 2 2019.

[13] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 10 2014.

[14] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, 2016, pp. 3723–3726.

[15] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 2018, pp. 59–66.

[16] E. Jolly, J. H. Cheong, T. Xie, S. Byrne, M. Kenny, and L. J. Chang, "Py-feat: Python facial expression analysis toolbox," *arXiv preprint arXiv:2104.03509*, 2021.

[17] S. I. Serengil and A. Ozpinar, "HyperExtended LightFace: A Facial Attribute Analysis Framework," *7th International Conference on Engineering and Emerging Technologies, ICEET 2021*, 2021.

[18] 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.

[19] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," *Proceedings - 2020 Innovations in Intelligent Systems and Applications Conference, ASYU 2020*, 10 2020.

[20] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, and others, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, 2013, pp. 117–124.

Authorized licensed use limited to: Queen's University. Downloaded on January 09,2024 at 15:30:55 UTC from IEEE Xplore. Restrictions apply.