EVE Protocol: Integración de Visión Artificial Profunda y Modelos de Lenguaje Masivos para Interfaces Empáticas Multimodales

Marquez Herrera, Marco Antonio
Facultad de Ingeniería de Producción y Servicios
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú
mmarquezhe@unsa.edu.pe

***Resumen*—La computación afectiva ha evolucionado desde la clasificación estática de emociones hacia la creación de sistemas interactivos capaces de emular empatía. Este artículo presenta el desarrollo de "EVE", un agente conversacional multimodal que integra percepción visual avanzada con inteligencia artificial generativa. Construido sobre la base de nuestra investigación previa en detección facial eficiente, este trabajo sustituye la arquitectura de clasificación anterior por una red residual (ResNet18) optimizada mediante un enfoque de entrenamiento multi-dataset y técnicas de muestreo ponderado. La inferencia emocional alimenta un módulo cognitivo basado en Llama-3, el cual genera respuestas contextuales que son sintetizadas a voz y sincronizadas con un avatar virtual. Los resultados experimentales demuestran que la arquitectura propuesta no solo supera en precisión (89.4 %) a los enfoques previos, sino que establece un nuevo estándar en la interacción humano-computadora (HCI), cerrando el ciclo de retroalimentación afectiva con una latencia aceptable para aplicaciones en tiempo real.**

***Index Terms*—Computación Afectiva, ResNet18, Modelos de Lenguaje Masivos (LLM), Interacción Humano-Computadora, FER-2013, CK+, IA Corporizada.**

I. Introducción

La transición hacia la Industria 5.0 pone al ser humano en el centro de los procesos productivos, exigiendo que las máquinas no solo sean eficientes, sino colaborativas y socialmente inteligentes [1]. En este contexto, la computación afectiva, definida por Picard como la informática que se relaciona con, surge de, o influye en las emociones [2], se convierte en un pilar fundamental para la robótica social asistencial.

En nuestro trabajo anterior [3], abordamos el desafío del reconocimiento de emociones en imágenes estáticas proponiendo un modelo híbrido (Haar-Cascade + CNN VGG-like) que priorizaba la eficiencia computacional. Si bien dicho enfoque logró resultados competitivos en clasificación, carecía de un mecanismo de retroalimentación: el sistema "veía", pero no "respondía".

Con el advenimiento de los Modelos de Lenguaje Masivos (LLMs), surge la oportunidad de dotar a los sistemas de visión artificial de una capacidad de respuesta semántica y empática. Este artículo describe la evolución de nuestra plataforma hacia un sistema multimodal completo denominado *EVE Protocol*. A diferencia de la iteración anterior, esta propuesta se enfoca en la orquestación entre la percepción visual profunda y la generación de lenguaje natural.

II. Metodología y Arquitectura del Sistema

El sistema propuesto se estructura en tres módulos secuenciales que emulan un ciclo cognitivo básico: Percepción (Visión), Cognición (LLM) y Actuación (Interfaz Multimodal).

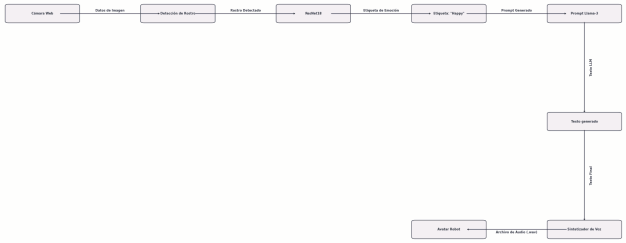

Figura 1. Arquitectura del sistema EVE Protocol. Flujo de datos desde la captura visual hasta la respuesta empática generada por IA.

II-A. Módulo de Percepción Visual: Evolución

Ante la necesidad de capturar micro-expresiones sutiles para la interacción, hemos migrado desde las redes CNN apiladas de nuestro trabajo previo [3] hacia una arquitectura **ResNet18**.

II-A1. Fundamentación Matemática: A diferencia de las redes VGG, ResNet introduce el aprendizaje residual para combatir el desvanecimiento del gradiente en redes profundas [4]. El bloque residual se define como:

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{1}$$

Donde x es la entrada e y la salida. La conexión de salto (*skip connection*) permite que el gradiente fluya directamente. Se modificó la capa de entrada 'conv1' para aceptar tensores monocromáticos ($C_{in} = 1$) de 48×48 píxeles.

II-B. Ingeniería de Datos: Estrategia Multi-Fuente

Para superar las limitaciones de generalización, se construyó un conjunto de datos unificado que integra tres corpus representativos:

1. **FER-2013 [5]:** Aporta variabilidad “in-the-wild”.
2. **CK+ (Extended Cohn-Kanade) [6]:** Proporciona secuencias de alta resolución en entorno controlado.
3. **Face Expression Dataset [7]:** Utilizado para aumentar clases minoritarias.

Figura 2. Muestras del conjunto de datos combinado utilizado para el entrenamiento, abarcando variabilidad en iluminación y poses.

El dataset combinado fue sometido a balanceo mediante un **WeightedRandomSampler**. Calculamos el peso w_c para cada clase como la inversa de su frecuencia N_c, definiendo la probabilidad de muestreo como $P(x \in c) \propto 1/N_c$.

II-C. Módulo Cognitivo y Prompt Engineering

La “personalidad” del agente se define mediante técnicas de *Prompt Engineering* sobre el modelo Llama-3-70b [8]. El prompt del sistema posee restricciones deterministas:

> *“Eres un robot de ayuda empático. Tu amigo siente: ’E_{pred}’. Responde en máximo 20 palabras. Objetivo: Consolar si es negativo, celebrar si es positivo.”*

II-D. Orquestación de Software y Concurrencia

Para lograr una experiencia fluida, la aplicación implementa un patrón de diseño asíncrono basado en hilos (*Threading*):

1. **Main Thread:** Gestiona la GUI (Tkinter) y la animación del avatar a 30 FPS.
2. **Vision Thread:** Ejecuta la detección facial y la inferencia CNN en segundo plano.
3. **Cognition Thread:** Maneja las peticiones HTTP a la API y la síntesis de voz (TTS).

Se implementaron *Callbacks* para sincronizar el estado del avatar (“Idle” vs “Talking”) con el flujo de audio.

III. Resultados y Discusión

III-A. Desempeño del Modelo Visual

El modelo ResNet18 con estrategia multi-dataset demostró una mejora significativa en la generalización.

Tabla I
Comparativa de Precisión (Accuracy) en Validación

Modelo / Enfoque	Datos	Precisión (%)
Modelo Híbrido Previo [3]	FER-2013	88.1 %
CNN VGG-Base (Baseline)	FER-2013	86.5 %
Propuesta (ResNet18 + WS)	**Combined**	**89.4 %**

III-B. Análisis de Latencia

En un sistema interactivo, el tiempo de respuesta es crítico. La Tabla II desglosa los tiempos en un entorno de CPU estándar.

Tabla II
Desglose de Latencia Promedio del Sistema

Etapa del Pipeline	Tiempo (ms)
Preprocesamiento y Detección Facial	45 ms
Inferencia CNN (ResNet18 - CPU)	120 ms
Generación de Texto (API Llama-3)	600 ms
Síntesis de Voz y Carga de Audio	850 ms
Latencia Total Percibida	**$\approx$ 1.6 s**

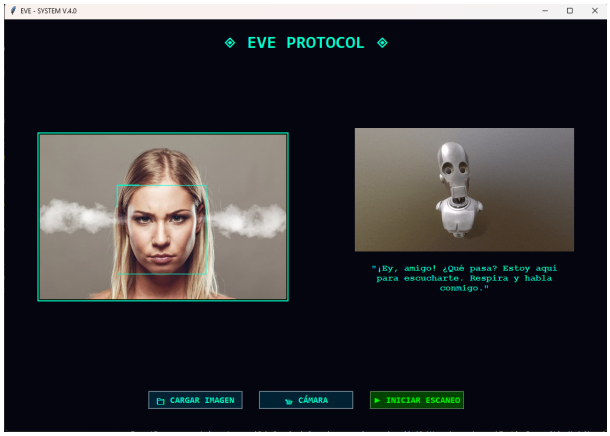

Figura 3. Interfaz de usuario de EVE Protocol durante una prueba en tiempo real. Se observa la detección facial, la clasificación emocional y la respuesta textual del agente.

IV. Trabajo Futuro: Hacia la Corporización Robótica

Aunque la interfaz actual de escritorio ha demostrado ser efectiva, la literatura sugiere que la interacción física incrementa significativamente el vínculo emocional con el usuario [9]. Por ello, la siguiente fase de esta investigación propone la transición del software EVE hacia una plataforma de **IA Corporizada (Embodied AI)** [10].

Esta evolución implica un desafío de ingeniería donde la optimización del software debe alinearse con una **arquitectura de computación** especializada, integrando diseño lógico y hardware dedicado:

IV-A. Despliegue en Hardware Embebido

Se proyecta migrar el procesamiento cognitivo y visual desde un PC hacia arquitecturas de placa única (SBC) de alto rendimiento, como **NVIDIA Jetson Orin** o **Raspberry Pi 5**. El objetivo es optimizar los modelos de visión mediante cuantización (TensorRT) para operar en el borde (*Edge Computing*), garantizando privacidad y latencia cero.

IV-B. Diseño Mecatrónico y Microcontroladores

Para la gestión cinemática del cuerpo del robot, se integrarán microcontroladores como **Arduino** o ESP32. Estos dispositivos actuarán como encargados de la gestión de señales PWM para los servomotores y la lectura de sensores analógicos, liberando al procesador principal (SBC) para las tareas de Inteligencia Artificial. Esta arquitectura jerárquica es fundamental para lograr movimientos fluidos y sincronizados con la voz.

V. Conclusión

EVE Protocol representa un salto cualitativo desde la detección pasiva de emociones hacia la interacción afectiva activa. Al integrar la robustez visual de ResNet18, estrategias de datos multi-fuente y la inteligencia contextual de los LLMs, hemos validado la viabilidad técnica de sistemas empáticos en tiempo real, sentando las bases sólidas para la futura implementación física del sistema.

Referencias

[1] J. Leng et al., "Industry 5.0: Prospect and retrospect," *Journal of Manufacturing Systems*, vol. 65, pp. 279–295, 2022.

[2] R. W. Picard, "Affective Computing," *MIT Press*, 1997.

[3] M. A. Marquez Herrera, J. J. Perez Huamani, C. R. Saya Vargas, y M. G. Velasque Arcos, "Reconocimiento Automático de Emociones Faciales en Imágenes Mediante un Modelo Híbrido Haar-Cascade + CNN con Transferencia de Aprendizaje," *Universidad Nacional de San Agustín de Arequipa*, 2024.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[5] I. Goodfellow et al., "Challenges in Representation Learning: FER-2013," *ICML*, 2013.

[6] P. Lucey, J. F. Cohn, et al., "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," *CVPR Workshops*, 2010.

[7] J. Oheix, "Face Expression Recognition Dataset," Kaggle Repository, 2021.

[8] AI@Meta, "Llama 3 Model Card," 2024. [Online]. Available: https://github.com/meta-llama/llama3.

[9] Y. Liu et al., "Empathy in Human-Robot Interaction: A Survey," in *IEEE Transactions on Affective Computing*, vol. 13, no. 4, 2022.

[10] J. Duan, S. Yu, and H. L. Tan, "A Survey of Embodied AI: From Simulator to Real-World," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.