UNIVERSIDAD DEL VALLE DE GUATEMALA
Visión por Computadora
Sección 10
Dr. Alan Gerardo Reyes

Proyecto #4

José Pablo Orellana - 21970
Diego Alberto Leiva – 21752
María Marta Ramírez - 21342
Gustavo González - 21438

Guatemala 6 de junio, 2025

Introducción

El reconocimiento automático de emociones faciales tiene aplicaciones importantes en interfaces humano-computadora, análisis de comportamiento, sistemas de seguridad y aplicaciones de salud mental. La capacidad de procesar emociones en tiempo real abre posibilidades para sistemas interactivos más naturales e intuitivos.

Objetivo general

Desarrollar un sistema de detección de emociones faciales en tiempo real con alta precisión y rendimiento.

Objetivos específicos

- Implementar un pipeline de procesamiento eficiente usando MediaPipe
- Entrenar un modelo de clasificación robusto con el dataset FER+
- Lograr inferencia en tiempo real con visualización interactiva
- Evaluar el rendimiento del sistema en condiciones reales

Procesamiento del DATASET FER+

Descripción del DATASET

El modelo fue entrenado utilizando el dataset FER+ (Facial Expression Recognition Plus), una mejora del conocido FER2013. Este dataset contiene imágenes en escala de grises de 48x48 píxeles, etiquetadas por múltiples anotadores en ocho categorías emocionales. Para este proyecto se seleccionaron únicamente cinco clases: anger, happiness, neutral, sadness, surprise, descartando las demás por baja representación o ambigüedad.

Preprocesamiento de Datos

Los datos fueron preparados de la siguiente manera:

- Recorte del rostro: Se utilizó MediaPipe Face Detection en modo VIDEO con sus parámetros por defecto, para detectar y recortar cada imagen al tamaño del rostro.
- Extracción de landmarks: A partir del recorte, se aplicó MediaPipe Face Mesh también en modo VIDEO, extrayendo 478 puntos faciales en 3 dimensiones (X, Y, Z) por rostro.
- Normalización: Los vectores resultantes fueron normalizados para estandarizar las entradas al modelo.
- División del dataset: Se respetó la división original de FER+, utilizando un 80% para entrenamiento y un 20% para prueba.

- No se aplicó validación cruzada ni validación separada, ya que el enfoque estuvo orientado a rapidez de implementación en un entorno real.

Entrenamiento del Modelo

Se entrenó un modelo de tipo Multilayer Perceptron (MLP) simple, enfocado en la eficiencia en tiempo real. Se utilizaron las siguientes técnicas:

- Función de pérdida: CrossEntropyLoss con pesos ajustados para compensar la confusión frecuente entre las clases neutral y sadness.
- Optimizador: AdamW, que mejora la estabilidad del entrenamiento al manejar el decaimiento de los pesos de forma más eficiente que Adam clásico.
- Scheduler: Se aplicó un StepLR para reducir la tasa de aprendizaje a intervalos constantes, facilitando la convergencia.
- Objetivo de rendimiento: Se priorizó la velocidad de predicción en dispositivos móviles o integraciones en tiempo real, aceptando un leve sacrificio en precisión a cambio de latencia mínima.

Resultados

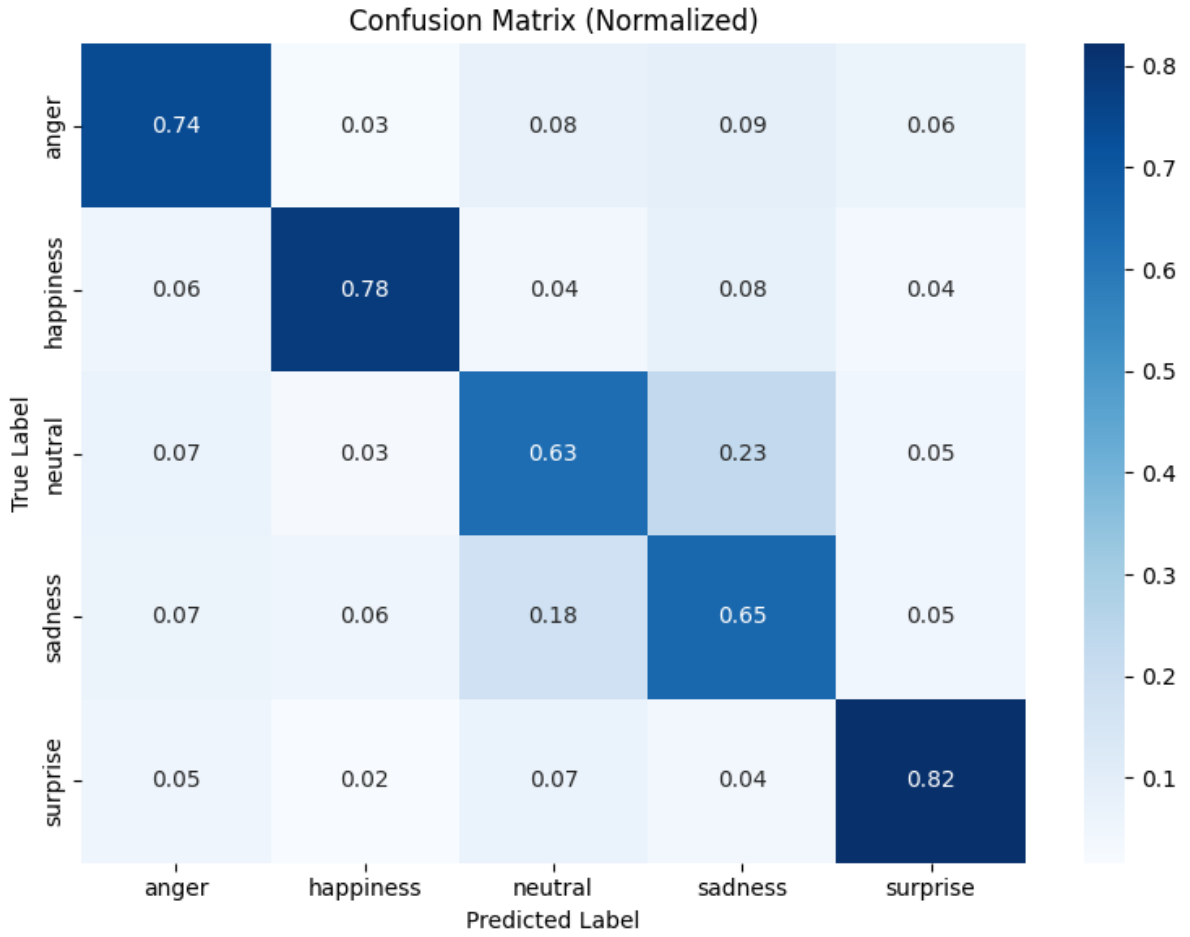

Figura 1. Matriz de confusión resultante del entrenamiento del modelo

Esta matriz muestra el desempeño del modelo por clase:

- Diagonal principal: representa las predicciones correctas (por ejemplo, la clase surprise fue clasificada correctamente el 82% de las veces).
- Valores fuera de la diagonal: indican confusiones del modelo entre clases.

Observaciones clave:

- La emoción surprise presenta la mayor precisión individual (82%).
- Happiness y anger también muestran buen rendimiento (78% y 74% respectivamente).
- Neutral y sadness presentan más confusión entre sí, lo cual es consistente con su cercanía semántica y expresiva.
- Las confusiones más frecuentes ocurren entre neutral y sadness (23%) y entre sadness y neutral (18%).

Este análisis reafirma la necesidad de considerar características faciales más sutiles y posibles mejoras con modelos secuenciales o análisis de contexto.

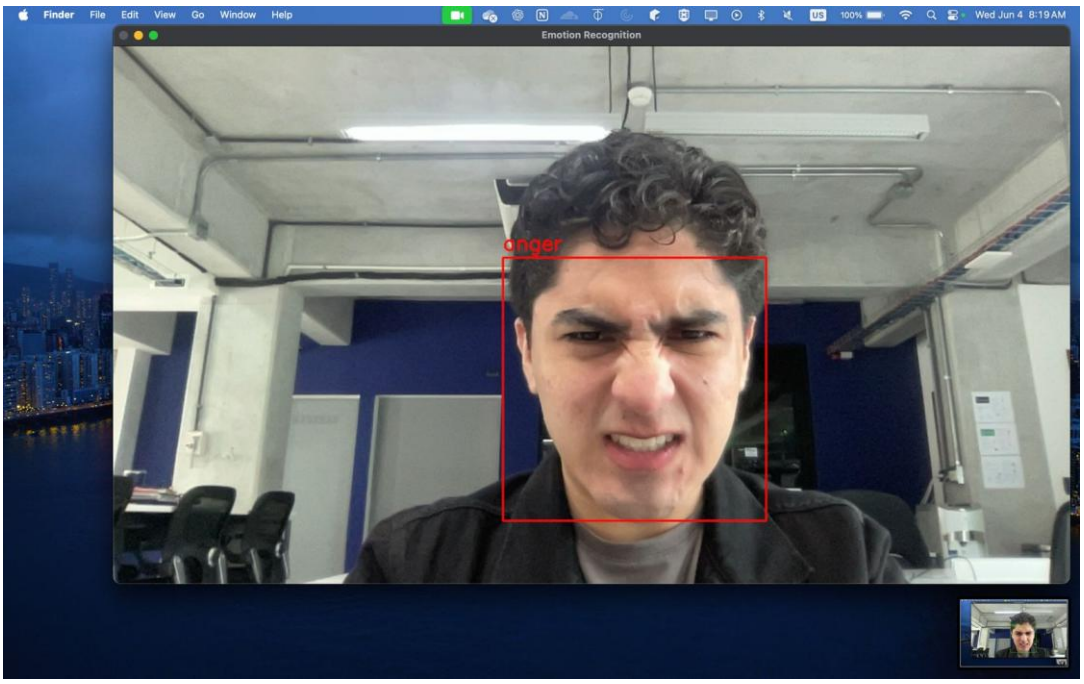

Figura 2. Reconocimiento de la expresión facial de ira.

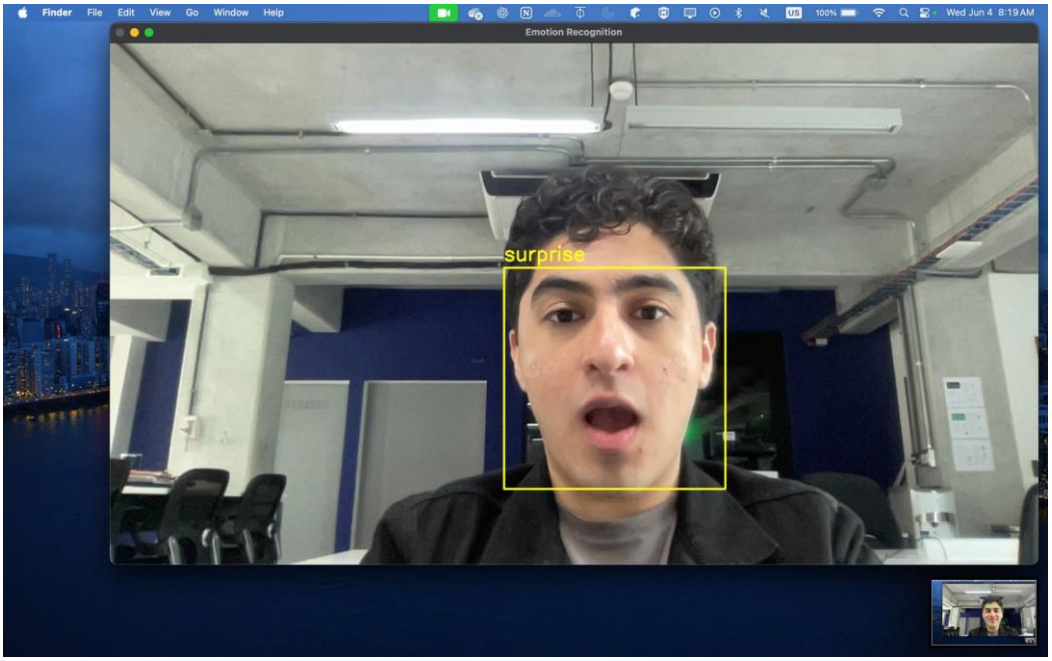

Figura 3. Reconocimiento de la expresión facial de sorpresa.

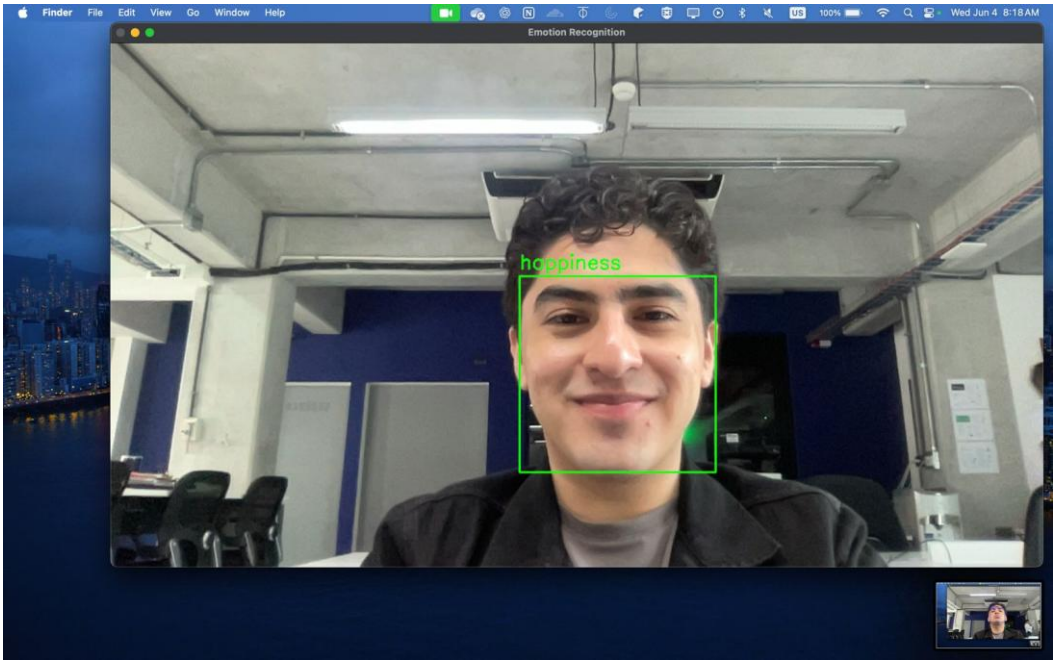

Figura 4. Reconocimiento de la expresión facial de alegría.

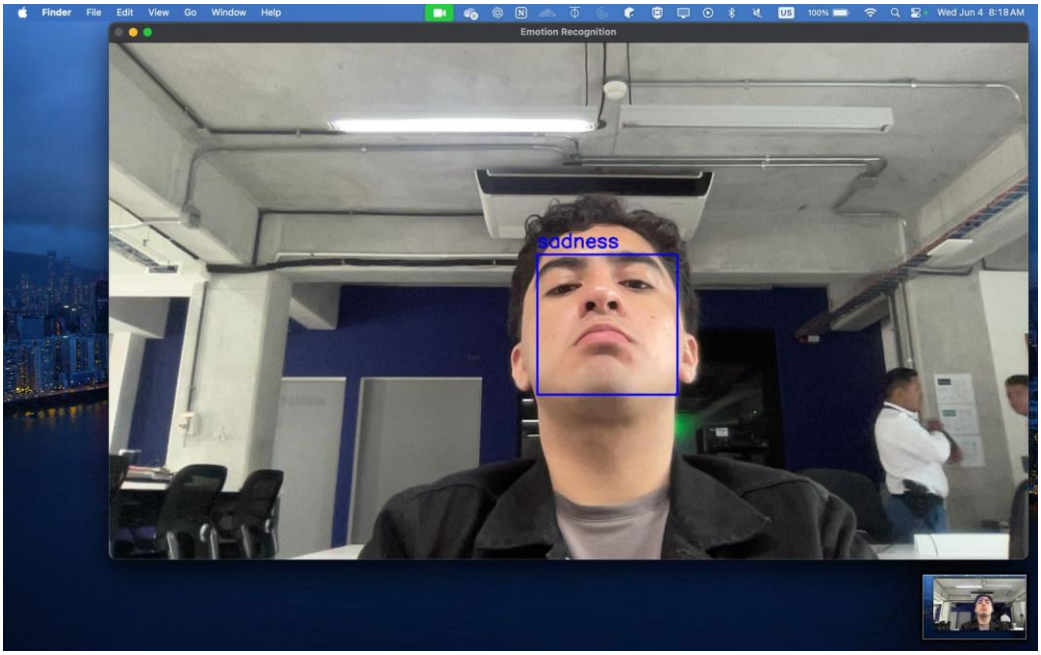

Figura 5. Reconocimiento de la expresión facial de tristeza.

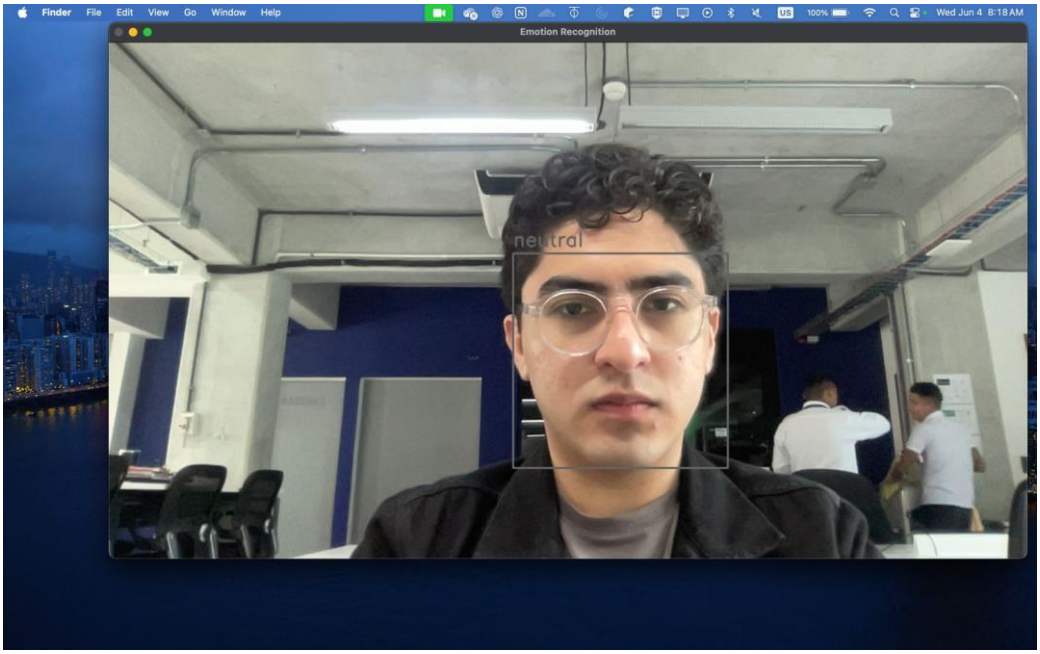

Figura 6. Reconocimiento de la expresión facial neutral.

Rendimiento del Sistema

- FPS: ~30 frames por segundo en GPU
- Latencia: < 33ms por frame
- Resolución: Soporta múltiples resoluciones de entrada

Calidad de Detección
El sistema demuestra robustez en:

- Condiciones de iluminación variables
- Múltiples ángulos faciales
- Diferentes distancias de la cámara
- Expresiones sutiles y pronunciadas

Emociones Mejor Clasificadas

Basado en observaciones durante las pruebas:

- Felicidad: Alta precisión debido a características distintivas (sonrisa)
- Sorpresa: Buena detección por apertura ocular y bucal
- Neutral: Clasificación estable como estado base
- Ira: Detección moderada en expresiones pronunciadas
- Tristeza: Desafiante debido a sutileza de la expresión

Discusión

Los resultados obtenidos reflejan un desempeño sólido del sistema, destacando la viabilidad del uso de landmarks 3D y MLP para clasificación emocional en tiempo real. La precisión observada en emociones como felicidad y sorpresa es atribuible a sus rasgos faciales distintivos, mientras que emociones como tristeza e ira presentan mayor variabilidad interindividual, lo cual afecta su clasificación.

La elección de MediaPipe permitió reducir significativamente el costo computacional, aunque persisten limitaciones con imágenes donde no se detectan correctamente los landmarks. Las pruebas en condiciones reales mostraron que el sistema mantiene su rendimiento con buena iluminación y posición frontal, pero disminuye en ángulos extremos o expresiones muy sutiles.

El balanceo del dataset resultó crucial para evitar sesgos en el modelo. La modularidad del pipeline abre oportunidades para futuras mejoras, como el uso de arquitecturas temporales y clasificación multimodal.

Conclusiones

- Implementación exitosa de pipeline completo de detección a clasificación
- Procesamiento eficiente de 13,432 muestras del dataset FER+
- Sistema funcional con inferencia en tiempo real a 30 FPS
- Arquitectura modular y extensible

Referencias

- MediaPipe Solutions. Google AI. https://mediapipe.dev/
- FER+ Dataset. Microsoft Research. https://www.kaggle.com/datasets/deadskull7/fer2013
- Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. Proceedings of the 18th ACM International Conference on Multimodal Interaction.
- Bazarevsky, V., Kartynnik, Y., Vakunov, A., et al. (2019). BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. arXiv preprint arXiv:1907.05047.