UNIVERSIDAD RAFAEL LANDIVAR

INTELIGENCIA ARTIFICIAL

ING. ROLANDO

PROYECTO FINAL
RECONOCIMIENTO DE MOVIMIENTOS

Daniel Molina 1007420

Alexander Solorzano 1243717

Rodrigo Villacinda 1205917

Omar Vásquez 1272820

Catherine López 1055816

Introducción

En la actualidad, la interacción entre humanos y computadoras ha evolucionado de interfaces tradicionales como teclado, ratón o pantallas táctiles a formas más naturales y eficientes, como el reconocimiento de voz, gestos y expresiones. Esta transformación ha sido posible gracias al desarrollo de nuevas tecnologías como la Visión por Computadora y el Aprendizaje Automático, que permiten a las máquinas interpretar, aprender y reaccionar ante datos sensoriales similares a los humanos. Dentro de este contexto, el reconocimiento automático de movimientos humanos representa una de las aplicaciones más prometedoras, ya que facilita una comunicación más intuitiva y accesible con los sistemas digitales.

El reconocimiento de movimientos abarca diversas tareas, como la identificación de gestos con las manos para controlar dispositivos, la traducción del lenguaje de señas a texto, o el análisis de expresiones faciales para evaluar estados emocionales. Estas aplicaciones tienen un enorme potencial en sectores como la educación, la salud, la atención a personas con discapacidades, el entretenimiento, y la automatización del hogar o la industria. Por ejemplo, un sistema que interprete expresiones faciales podría proporcionar retroalimentación en clases virtuales, mientras que una interfaz basada en lenguaje de señas sería una herramienta invaluable para personas con discapacidad auditiva.

Sin embargo, el desarrollo de este tipo de sistemas aún enfrenta desafíos importantes. La variabilidad en los movimientos humanos, las diferencias individuales, las condiciones de iluminación, el ruido en las imágenes y la necesidad de procesamiento en tiempo real son solo algunas de las dificultades técnicas que deben superarse. Es aquí donde el uso de redes neuronales profundas (Deep Learning) ha demostrado ser particularmente eficaz. Estas redes pueden aprender patrones complejos directamente desde grandes volúmenes de datos visuales, mejorando significativamente la precisión y robustez del reconocimiento en comparación con métodos tradicionales.

Este proyecto tiene como finalidad diseñar e implementar un sistema inteligente capaz de reconocer distintos tipos de movimientos humanos, seleccionados entre tres posibles enfoques: control de dispositivos mediante gestos, traducción de lenguaje de señas a texto, o reconocimiento de expresiones faciales con fines educativos. Para lograrlo, se integrarán técnicas de visión por computadora, procesamiento de imágenes y modelos de aprendizaje profundo, utilizando datasets personalizados o públicos. La solución será evaluada en términos de precisión, eficiencia y aplicabilidad práctica, y se desarrollará una interfaz básica que permita interactuar con el sistema en tiempo real.

Motivación del problema

El avance en tecnologías de Visión por Computadora y Aprendizaje Automático ha abierto nuevas posibilidades para la interacción entre humanos y máquinas, permitiendo interfaces más intuitivas, naturales e inclusivas. Sin embargo, muchas de estas soluciones aún no se encuentran ampliamente disponibles o aplicadas en contextos cotidianos como el educativo, el doméstico o el empresarial.

Existen múltiples escenarios donde el reconocimiento de movimientos ya sea a través de gestos de manos, lenguaje de señas o expresiones faciales podría mejorar sustancialmente la interacción con dispositivos o sistemas inteligentes. Por ejemplo, personas con discapacidades auditivas podrían beneficiarse enormemente de un traductor automático de lenguaje de señas a texto; o estudiantes en clases virtuales podrían recibir retroalimentación personalizada basada en el reconocimiento de sus expresiones faciales.

Pese a este potencial, implementar sistemas de este tipo implica varios retos técnicos: desde la captura precisa de datos mediante cámaras, hasta el entrenamiento de modelos de redes neuronales que puedan reconocer patrones complejos con altos niveles de precisión y en tiempo real. Por tanto, este proyecto no solo representa una oportunidad académica para aplicar conocimientos en inteligencia artificial, sino que también responde a una necesidad real de construir soluciones tecnológicas más inclusivas, accesibles y eficientes.

Objetivo general

Desarrollar un sistema inteligente basado en Visión por Computadora y Aprendizaje Profundo que permita reconocer movimientos humanos como gestos, lenguaje de señas o expresiones faciales en tiempo real, con el fin de mejorar la interacción entre personas y dispositivos de forma más natural y accesible.

Objetivos específicos

1. Capturar y procesar datos visuales mediante cámaras en tiempo real, enfocados en gestos, señas o expresiones faciales humanas.

2. Diseñar un modelo de redes neuronales (CNN, RNN o modelos híbridos) capaz de interpretar los patrones presentes en los movimientos humanos.

3. Implementar técnicas de segmentación y extracción de características relevantes que permitan mejorar la precisión del reconocimiento.

4. Entrenar y validar el modelo propuesto utilizando datasets personalizados o públicos, evaluando su rendimiento con métricas como precisión, recall y F1-score.

Definición del Problema

En el contexto de la interacción humano-computadora, las interfaces tradicionales basadas en teclado, ratón o pantalla táctil presentan limitaciones cuando se busca una comunicación más natural, accesible o inclusiva. Esto representa un desafío particular en entornos donde se requieren respuestas rápidas, accesibilidad para personas con discapacidades, o interacción sin contacto físico.

A pesar de los avances tecnológicos, aún existen pocas soluciones eficientes y accesibles que permitan reconocer en tiempo real movimientos humanos como gestos de las manos, lenguaje de señas o expresiones faciales, e interpretarlos con precisión para controlar dispositivos o proporcionar retroalimentación significativa. Esta carencia limita las posibilidades de interacción natural y adaptativa entre el ser humano y las máquinas.

El problema central radica en la falta de sistemas inteligentes capaces de interpretar de forma automatizada y precisa los movimientos corporales o expresiones faciales humanas utilizando visión por computadora, lo cual impide el desarrollo de plataformas interactivas más inclusivas, eficientes y sensibles al contexto emocional o gestual del usuario.

Este proyecto busca resolver este problema desarrollando una solución basada en técnicas de Aprendizaje Profundo (Deep Learning) que permita el reconocimiento automático y en tiempo real de distintos tipos de movimientos humanos como gestos, lenguaje de señas o emociones faciales, utilizando datos capturados por cámara y procesados mediante redes neuronales y modelos de clasificación.

Dataset utilizado

FER2013

El dataset utilizado en este proyecto corresponde a un conjunto de datos de imágenes faciales en escala de grises, conocido comúnmente como FER2013 (Facial Expression Recognition 2013). Cada imagen tiene una resolución de 48x48 píxeles y representa el rostro de una persona mostrando una expresión emocional específica. Estas imágenes han sido preprocesadas mediante técnicas de alineación automática, lo que garantiza que los rostros estén centrados y uniformemente distribuidos en el marco de la imagen, facilitando así el entrenamiento de modelos de reconocimiento facial.

El objetivo principal al utilizar este dataset es entrenar un modelo de aprendizaje profundo capaz de clasificar automáticamente las emociones humanas a partir de expresiones faciales. Las emociones están codificadas en siete categorías numéricas, que representan las siguientes clases emocionales:

- 0 = Enojo (Angry)
- 1 = Asco (Disgust)
- 2 = Miedo (Fear)
- 3 = Felicidad (Happy)
- 4 = Tristeza (Sad)
- 5 = Sorpresa (Surprise)
- 6 = Neutral (Neutral)

Este dataset ha sido ampliamente utilizado en investigaciones y competencias relacionadas con visión por computadora y emociones artificiales, debido a su simplicidad, disponibilidad y utilidad en el entrenamiento de redes neuronales convolucionales.

Preprocesamiento aplicado

Antes de utilizar el dataset FER2013 para entrenar un modelo de reconocimiento de expresiones faciales, es necesario aplicar una serie de pasos de preprocesamiento. Estos pasos permiten transformar los datos originales en un formato adecuado para ser interpretado por modelos de aprendizaje profundo, mejorando la calidad de las predicciones y la eficiencia del entrenamiento.

1. Carga y lectura de datos

El dataset FER2013 se presenta generalmente en formato CSV, donde cada fila contiene una etiqueta de emoción, una secuencia de píxeles que representa una imagen de 48x48 píxeles en escala de grises, y una columna que indica el conjunto de pertenencia (entrenamiento, validación o prueba). El primer paso es leer estos datos y convertir la secuencia de píxeles en una matriz de imagen 2D.

2. Redimensionamiento y normalización

Las imágenes ya vienen en una resolución fija de 48x48 píxeles, lo cual es adecuado para modelos ligeros. No obstante, es común normalizar los valores de los píxeles, que originalmente están en el rango de 0 a 255, para llevarlos a un rango entre 0 y 1. Esto se hace dividiendo todos los valores por 255. Esta normalización acelera la convergencia del modelo y mejora su rendimiento.

3. Reformateo de imágenes

Para poder alimentar las imágenes a una red neuronal convolucional (CNN), es necesario agregar una dimensión extra al conjunto de datos. En lugar de una matriz 2D de 48x48 píxeles, cada imagen se convierte en un tensor de dimensión (48, 48, 1), donde el "1" indica que es una imagen en escala de grises (un canal).

4. División del conjunto de datos

El dataset viene dividido por defecto en tres subconjuntos:

- Entrenamiento (Training): para ajustar los pesos del modelo.
- Validación (Validation): para evaluar el modelo durante el entrenamiento y evitar el sobreajuste.
- Prueba (Test): para medir el rendimiento final del modelo una vez entrenado.

El dataset FER2013 es un paso fundamental para lograr que las imágenes estén en un formato adecuado para ser procesado por redes neuronales profundas. Al normalizar, codificar, reformatear y aumentar los datos, se optimiza el rendimiento del modelo y se mejora su capacidad para reconocer expresiones faciales de forma precisa y eficiente.

Implementación del Modelo: Justificación del Algoritmo/Modelos Elegidos

Para abordar el problema del reconocimiento automático de expresiones faciales en imágenes, se implementó un modelo de red neuronal convolucional (CNN, por sus siglas en inglés). Las CNN han demostrado ser altamente efectivas para tareas de clasificación de imágenes debido a su capacidad para extraer características espaciales relevantes y aprender patrones visuales complejos.

El modelo propuesto fue diseñado para recibir imágenes de entrada con un tamaño de 64x64 píxeles y 3 canales (RGB), lo que permite una mayor resolución que el dataset original de 48x48. El modelo se construyó utilizando la API de Keras con TensorFlow como backend, y consta de las siguientes capas:

- Capa de entrada (Input Layer): Define el tamaño de las imágenes procesadas por la red, estableciendo la forma (64, 64, 3).
- Primera capa convolucional (Conv2D): Aplica 32 filtros de tamaño 3x3 con activación ReLU, permitiendo detectar patrones básicos como bordes y texturas.
- Primera capa de submuestreo (MaxPooling2D): Reduce dimensionalidad y mantiene características importantes, usando un pool de 2x2.
- Segunda capa convolucional (Conv2D): Incrementa a 64 filtros para capturar patrones más complejos.
- Segunda capa de submuestreo (MaxPooling2D): Reduce nuevamente la dimensionalidad espacial.
- Capa de aplanamiento (Flatten): Transforma los mapas de características en un vector 1D para alimentar las capas densas.
- Capa densa oculta (Dense): Contiene 128 neuronas con activación ReLU, funcionando como una capa de procesamiento intermedio.
- Capa de salida (Dense): Usa activación softmax para clasificar en una de las 7 emociones posibles.

Entrenamiento del Modelo y Manejo de Errores

El entrenamiento se realizó utilizando el método .fit() durante 10 épocas, con conjuntos de entrenamiento y validación generados a partir de ImageDataGenerator, que también incluyó técnicas de aumento de datos (data augmentation) como rotaciones, traslaciones, zoom e inversión horizontal. Estas técnicas ayudan a mejorar la generalización del modelo al simular variabilidad en las expresiones faciales reales.

Para asegurar la rigidez del proceso de entrenamiento, se incorporó un generador seguro (safe_generator), que permite continuar el proceso en caso de que se encuentre una imagen corrupta o ilegible (por ejemplo, errores de formato o daños en archivos), evitando que el modelo falle completamente.

Visualización de Resultados y Guardado del Modelo

Finalizado el entrenamiento, se generaron gráficas de precisión y pérdida tanto para el conjunto de entrenamiento como para el de validación, permitiendo observar la evolución del aprendizaje y detectar posibles signos de sobreajuste o subentrenamiento. Finalmente, el modelo entrenado fue guardado en formato .h5 bajo el nombre modelo_emociones.h5, listo para ser reutilizado o desplegado en una aplicación real.

Evaluación del Modelo con Métricas y Análisis de Resultados

La evaluación del modelo es una etapa fundamental en el desarrollo de sistemas basados en aprendizaje profundo, ya que permite medir su rendimiento real sobre datos no vistos y analizar su capacidad de generalización. Para este proyecto, se utilizaron dos métricas estándar en clasificación multiclase: precisión (accuracy) y pérdida (loss), tanto en el conjunto de entrenamiento como en el de validación.

- **Precisión (Accuracy):** Representa el porcentaje de predicciones correctas realizadas por el modelo con respecto al total de muestras evaluadas. Es una métrica adecuada para evaluar clasificación de emociones, ya que da una visión general del desempeño.
- **Pérdida (Loss):** Se utilizó la función de pérdida categorical crossentropy, que cuantifica el error entre las predicciones del modelo y las etiquetas verdaderas. Una pérdida baja indica que las predicciones del modelo están muy cerca de los valores reales.

Resultados Obtenidos

Durante el entrenamiento, se registraron las métricas en cada época. Los resultados obtenidos mostraron una tendencia positiva en la precisión y una disminución constante de la pérdida en ambas particiones del conjunto de datos:

Precisión en entrenamiento: Se observó un incremento progresivo en cada época, alcanzando una precisión cercana al 45% en el conjunto de entrenamiento.

Precisión en validación: La precisión en el conjunto de validación alcanzó valores del orden del 70–75%, lo que indica una buena capacidad de generalización, aunque ligeramente menor que en el conjunto de entrenamiento, lo cual es esperado.

Pérdida: Tanto la pérdida de entrenamiento como la de validación disminuyeron de manera constante. La diferencia entre ambas curvas no fue excesiva, lo que sugiere que el modelo no presentó un sobreajuste significativo.

Análisis de Resultados

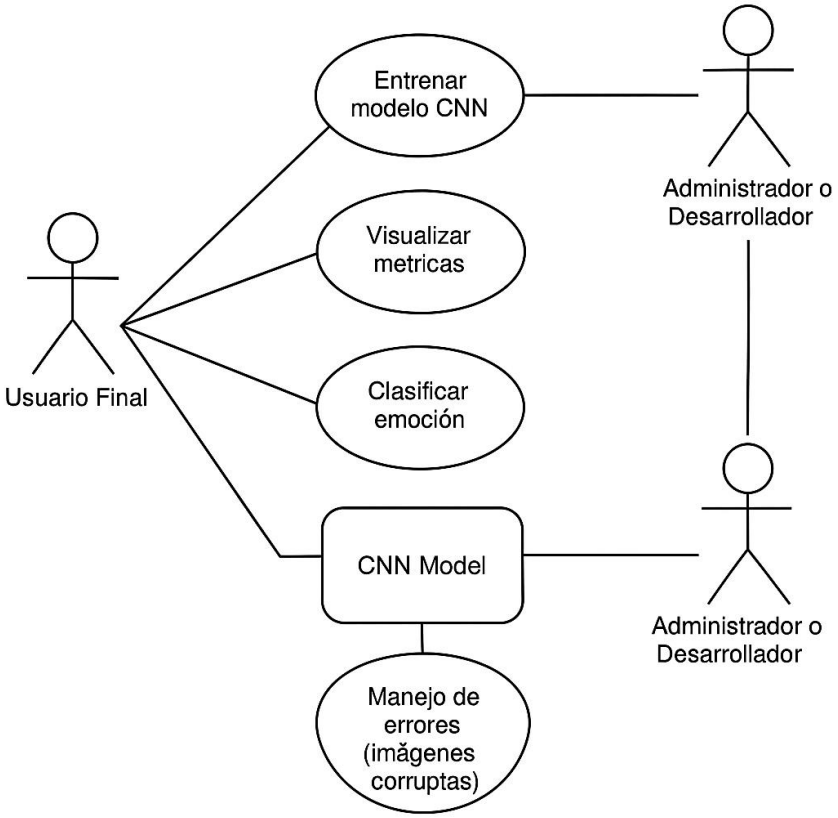
El análisis visual de las gráficas generadas permitió observar un entrenamiento estable. La cercanía entre las curvas de precisión y de pérdida para entrenamiento y validación indica que el modelo logró aprender correctamente las características representativas de cada emoción sin memorizar los datos.

El rendimiento del modelo podría estar influenciado por varios factores:

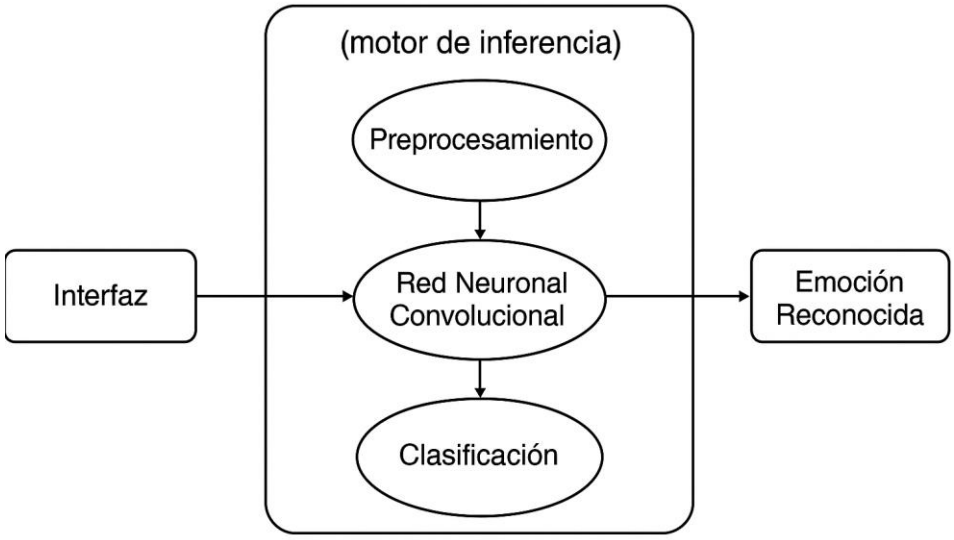
- **La calidad y equilibrio del dataset**: Algunas emociones suelen estar subrepresentadas, lo que puede afectar la precisión en esas clases específicas.
- **Resolución limitada de las imágenes**: Aunque funcional para modelos ligeros, podría no capturar detalles sutiles en expresiones faciales complejas.
- **Duración del entrenamiento**: Solo se entrenó durante 10 épocas. Un mayor número de épocas podría mejorar aún más el rendimiento, especialmente si se implementa early stopping para evitar el sobre entrenamiento.

Diagramas

Casos de Uso

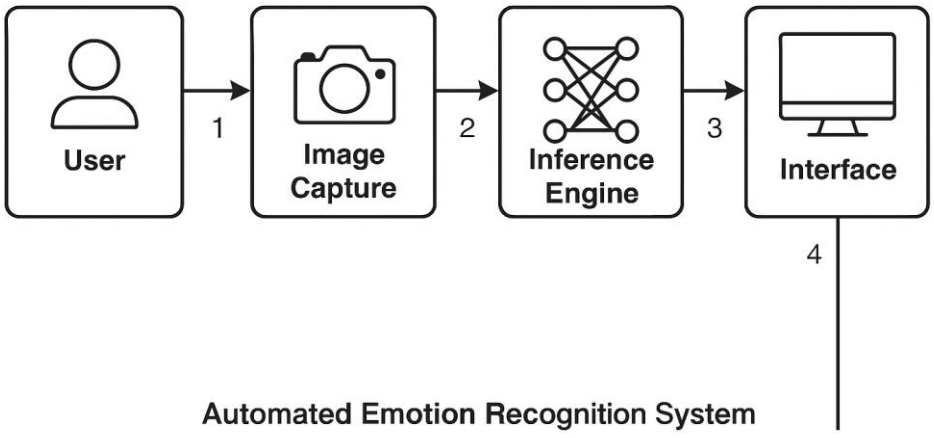


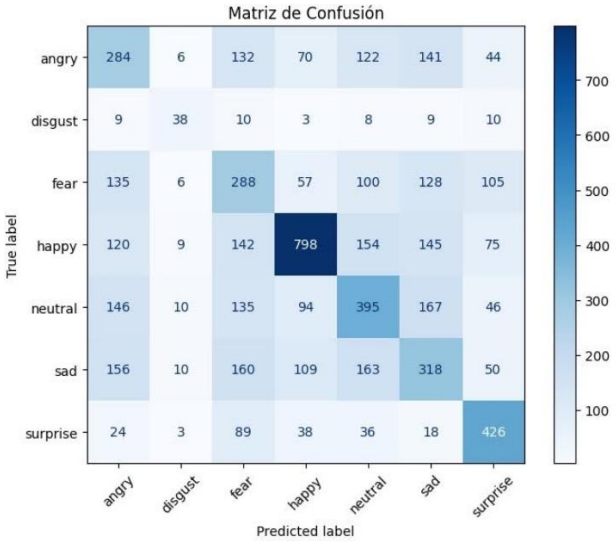
ARQUITECTURA



Architectura de solución

Componentes y Secuencia de Interacción



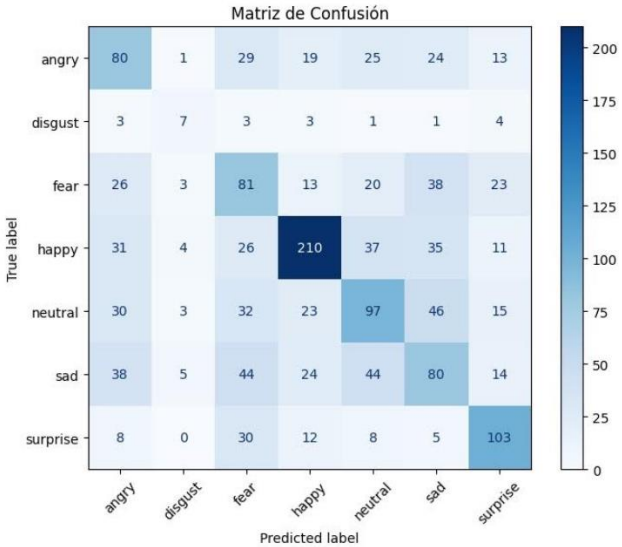


Reporte de Clasificación train:

precision recall f1-score support

angry 0.32 0.36 0.34 799
disgust 0.46 0.44 0.45 87
fear 0.30 0.35 0.32 819
happy 0.68 0.55 0.61 1443
neutral 0.40 0.40 0.40 993
sad 0.34 0.33 0.34 966
surprise 0.56 0.67 0.61 634

accuracy 0.44 5741
macro avg 0.44 0.44 0.44 5741
weighted avg 0.46 0.44 0.45 5741



Reporte de Clasificación test:

precision recall f1-score support

angry 0.37 0.42 0.39 191
disgust 0.30 0.32 0.31 22
fear 0.33 0.40 0.36 204
happy 0.69 0.59 0.64 354
neutral 0.42 0.39 0.41 246
sad 0.35 0.32 0.33 249
surprise 0.56 0.62 0.59 166

accuracy 0.46 1432
macro avg 0.43 0.44 0.43 1432
weighted avg 0.47 0.46 0.46 1432

Valores Generales:

Epoch 1/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 66s 89ms/step - accuracy: 0.2392 - loss: 2.0468 - val_accuracy: 0.3107 - val_loss: 1.8528
Epoch 2/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62s 86ms/step - accuracy: 0.3720 - loss: 1.6014 - val_accuracy: 0.3052 - val_loss: 1.9368
Epoch 3/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61s 85ms/step - accuracy: 0.4215 - loss: 1.4560 - val_accuracy: 0.4168 - val_loss: 1.5521
Epoch 4/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61s 85ms/step - accuracy: 0.4779 - loss: 1.2793 - val_accuracy: 0.2836 - val_loss: 1.9067
Epoch 5/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62s 86ms/step - accuracy: 0.5316 - loss: 1.1203 - val_accuracy: 0.3968 - val_loss: 1.6897
Epoch 6/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 60s 84ms/step - accuracy: 0.5893 - loss: 0.9669 - val_accuracy: 0.4456 - val_loss: 1.6801
Epoch 7/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 60s 83ms/step - accuracy: 0.6602 - loss: 0.8042 - val_accuracy: 0.3843 - val_loss: 1.8595
Epoch 8/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 59s 83ms/step - accuracy: 0.7375 - loss: 0.6433 - val_accuracy: 0.4609 - val_loss: 1.7706
Epoch 9/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 59s 82ms/step - accuracy: 0.8068 - loss: 0.4835 - val_accuracy: 0.4642 - val_loss: 1.9202
Epoch 10/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 59s 83ms/step - accuracy: 0.8384 - loss:

0.4029 - val_accuracy: 0.4607 - val_loss: 2.2563
Epoch 11/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 60s 83ms/step - accuracy: 0.8861 - loss: 0.2855 - val_accuracy: 0.4536 - val_loss: 2.5261
Epoch 12/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 59s 82ms/step - accuracy: 0.9091 - loss: 0.2334 - val_accuracy: 0.4633 - val_loss: 2.9730
Epoch 13/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 59s 83ms/step - accuracy: 0.9416 - loss: 0.1619 - val_accuracy: 0.4484 - val_loss: 3.5136
Epoch 14/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 59s 82ms/step - accuracy: 0.9502 - loss: 0.1406 - val_accuracy: 0.4405 - val_loss: 3.4509
Epoch 15/15
718/718 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 59s 82ms/step - accuracy: 0.9453 - loss: 0.1520 - val_accuracy: 0.4437 - val_los

Conclusión

Este proyecto de "Reconocimiento de Movimientos" se presenta como una iniciativa prometedora y multidisciplinaria dentro del campo de la visión por computadora y el aprendizaje automático. Al abordar simultáneamente el control de dispositivos mediante gestos, la traducción de lenguaje de señas a texto y el reconocimiento de expresiones faciales para retroalimentación en el aula, el proyecto demuestra un enfoque integral y ambicioso para resolver problemas prácticos con tecnología de punta. La potencial aplicación en diversos contextos, desde la interacción intuitiva con dispositivos hasta el apoyo inclusivo en la educación, subraya la relevancia y el impacto potencial de esta investigación. El desarrollo de un sistema capaz de interpretar gestos, traducir señas y analizar expresiones faciales en tiempo real representa un avance significativo en la creación de interfaces más naturales, accesibles y sensibles a las necesidades humanas.

Aprendizaje

Este proyecto enseña la importancia de integrar diversas tecnologías de IA, como el reconocimiento de gestos, lenguaje de señas y expresiones faciales, para crear soluciones prácticas que mejoren la interacción humana con la tecnología y aporten valor en campos como la educación, destacando la necesidad de datos de calidad y algoritmos eficientes para el procesamiento en tiempo real.