Simposio CEA de Robótica, Bioingeniería, Visión por Computador y Automática Marina 2026

Reconocimiento de emociones faciales en vídeo mediante deep–learning para interacción humano-robot

Falcone, A.[a], Cisnal, A.[a,*], Pérez Turiel, J.[a], Fraile, J. C.[a]

[a] *Instituto de las Tecnologías Avanzadas de la Producción, Escuela de Ingenierías Industriales, Universidad de Valladolid, Paseo Prado de la Magdalena 3-5, 47011, Valladolid, España.*

Resumen

El reconocimiento automático de emociones a partir de expresiones faciales en vídeo constituye un problema relevante en el ámbito de la inteligencia artificial afectiva. En particular, se considera su aplicación como módulo perceptivo para adaptar la interacción y el comportamiento de un robot en función del estado afectivo estimado. En este trabajo se presenta el desarrollo y la evaluación de dos enfoques complementarios para el análisis dinámico de emociones faciales. Por un lado, se emplea un modelo basado en Vision Transformers (ViT-B/32), entrenado sobre el conjunto de datos DFEW y validado de forma externa en MAFW, permitiendo una evaluación cuantitativa mediante métricas globales y por clase. Por otro lado, se implementa un sistema híbrido CNN+LSTM (EMO-AffectNet), orientado a ejecución en tiempo real y evaluado de forma cualitativa. Los resultados muestran un rendimiento sólido del modelo ViT-B/32, con una degradación moderada en validación externa, y una elevada capacidad de respuesta en tiempo real del modelo CNN+LSTM, alcanzando mayores tasas de procesamiento. El estudio pone de manifiesto el compromiso entre precisión, generalización y operatividad en aplicaciones reales.

Palabras clave: Control Basado en Datos, Inteligencia Artificial y Control, Inteligencia Computacional en Control, Sistemas Biológicos y Médicos, Sistemas Hombre–Máquina.

Deep Learning-Based Facial Emotion Recognition in Video for Human–Robot Interaction

Abstract

Automatic emotion recognition from facial expressions in video is a relevant problem in affective computing. In particular, its application is considered as a perceptual module to adapt the interaction and the robot's behavior based on the estimated affective state. This work presents the development and evaluation of two complementary approaches for dynamic facial emotion analysis. First, a Vision Transformer–based model (ViT-B/32) is trained on the DFEW dataset and externally validated on MAFW, enabling a quantitative evaluation using global and class-wise performance metrics. Second, a hybrid CNN+LSTM system (EMO-AffectNet) is implemented for real-time operation and evaluated qualitatively. The results show solid performance of the ViT-B/32 model, with a moderate degradation under external validation, and a high real-time responsiveness of the CNN+LSTM approach, achieving higher processing rates. Overall, the study highlights the trade-off between accuracy, generalization capability, and real-time operability in practical emotion recognition systems.

Keywords: AI and Control, Biological and Medical Systems, Computational Intelligence in Control, Data-Driven Control, Human Machine Systems.

1. Introducción

El reconocimiento automático de emociones a partir de expresiones faciales en vídeo constituye un área de creciente interés en el ámbito de la inteligencia artificial afectiva y la interacción humano-robot. Las expresiones faciales representan una fuente de información no invasiva, no verbal, relativamente universal y presente en la interacción humana (Graziotin et al., 2015), lo que las convierte en un canal privilegiado para inferir el estado emocional de una persona en aplicaciones como la robótica asistencial, la salud digital, la educación emocional y los

*Autor para correspondencia: ana.cisnal@uva.es
Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

sistemas interactivos inteligentes.

Desde los trabajos fundacionales de Darwin (Charles, 1872) y la posterior sistematización de las emociones básicas propuesta por Ekman y Fiesen (Ekman and Friesen, 1971), el estudio de las expresiones faciales ha evolucionado desde enfoques manuales y basados en codificación experta hacia métodos automáticos impulsados por el aprendizaje automático y, más recientemente, por el aprendizaje profundo. En este contexto, las redes neuronales convolucionales (CNN) han demostrado una elevada capacidad para la extracción jerárquica de características espaciales (Li and Deng, 2022), mientras que las redes neuronales recurrentes (RNN) y los modelos basados en Transformers han permitido incorporar de forma explícita la dimensión temporal inherente a las expresiones faciales dinámicas (Mohana et al., 2023).

Los avances recientes en *Vision Transformers* y modelos híbridos CNN–RNN han marcado el estado del arte en el análisis emocional a partir de vídeo, especialmente en escenarios no controlados caracterizados por variabilidad en la orientación y posición de la cara, iluminación y contexto (Xiong et al., 2024). Sin embargo, el reconocimiento automático de emociones en condiciones reales sigue enfrentando importantes desafíos: variabilidad interindividual, ambigüedad expresiva, cambios culturales, iluminación no controlada y desbalance de clases en los conjuntos de datos (Patel et al., 2020; Kopalidis et al., 2024).

En el ámbito de la robótica, la estimación del estado emocional puede emplearse como entrada de un módulo perceptivo para mejorar la naturalidad, seguridad y aceptación del robot, ajustando su conducta ante señales de estrés, confusión o estados emocionales positivos. En este sentido, el paradigma *human-in-the-loop* permite cerrar el lazo percepción–acción, integrando la información afectiva en la toma de decisiones del sistema robótico y manteniendo al usuario como supervisor implícito de la interacción. Este enfoque resulta especialmente relevante en escenarios reales, donde la ambigüedad emocional y la variabilidad interindividual exigen sistemas perceptivos robustos y comprensibles para el usuario final (Mosqueira-Rey et al., 2022).

Ante este contexto, el presente trabajo aborda el desarrollo y la evaluación de un sistema computacional para el reconocimiento dinámico de emociones humanas a partir de secuencias de vídeo facial, empleando arquitecturas avanzadas de *deep learning*. Se analizan tanto su rendimiento cuantitativo como su operatividad en escenarios prácticos, incluyendo inferencia en tiempo real. El sistema presentado se concibe como el módulo perceptivo de una arquitectura robótica basada en *human-in-the-loop*, tal como se ilustra en la Fig. 1.

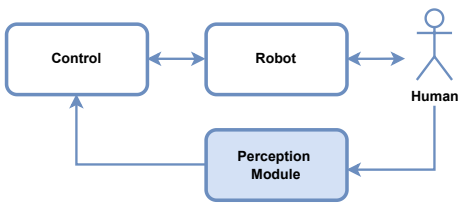


Figura 1: Esquema conceptual de una arquitectura *human-in-the-loop* en la que el módulo de percepción emocional extrae información afectiva del usuario y la integra en el lazo de control para adaptar el comportamiento del robot.

2. Materiales y métodos

En esta sección se describen las bases de datos utilizadas, los modelos empleados para la inferencia de categorías emocionales aparentes a partir de expresiones faciales, las métricas de evaluación y la configuración experimental.

2.1. Bases de datos

En este trabajo se emplearon dos bases de datos públicas para el reconocimiento de emociones faciales dinámicas en escenarios no controlados: *Dynamic Facial Expressions in-the-Wild* (DFEW) y *Multimodal Affective Faces in the Wild* (MAFW), ambas ampliamente utilizadas en la evaluación de sistemas de análisis emocional en condiciones reales.

La base de datos DFEW (ver Figura 2(a)), está compuesta por 16 372 clips de vídeo de corta duración (2–5 s), extraídos de aproximadamente 1 500 películas comerciales de distintas regiones y culturas (Jiang et al., 2020). Los clips presentan una elevada variabilidad en iluminación, pose, oclusiones, resolución y contexto narrativo, lo que los hace representativos de escenarios *in-the-wild*. Cada clip contiene una única expresión emocional dominante y está anotado en siete emociones básicas según la clasificación de Ekman: alegría (20,6 %), tristeza (16,7 %), ira (18,5 %), sorpresa (12,4 %), asco/repulsión (1,2 %), miedo (8,1 %) y neutral (22,5 %). Las anotaciones fueron realizadas de manera independiente por 10 evaluadores humanos, alcanzando una concordancia inter-anotador elevada, con valores del coeficiente Kappa de Fleiss entre 0,61 y 0,80, lo que garantiza la fiabilidad de las etiquetas.

Por su parte, la base de datos MAFW (ver Figura 2(b)) fue diseñada para capturar emociones espontáneas y de mayor complejidad semántica en contextos audiovisuales altamente diversos (Liu et al., 2022). Consta de 10 045 clips de vídeo procedentes de películas, series de televisión, programas de entrevistas, noticiarios y otros contenidos multimedia reales. A diferencia de DFEW, MAFW incluye 11 categorías emocionales discretas (ira, asco, miedo, felicidad, tristeza, sorpresa, neutral, desprecio, ansiedad, indefensión y decepción), incorporando emociones de baja intensidad y estados afectivos más sutiles. Esta mayor diversidad emocional y contextual incrementa significativamente la dificultad del conjunto de datos y lo convierte en un banco de pruebas especialmente exigente para evaluar la generalización de los modelos.

En este trabajo, DFEW se emplea como conjunto principal de entrenamiento y prueba, mientras que MAFW se utiliza exclusivamente como conjunto de validación externa, permitiendo analizar el comportamiento de los modelos ante un cambio de dominio pronunciado y evaluar su robustez en escenarios realistas.

2.2. Módulo perceptivo

Desde el punto de vista de la interacción humano-robot (HRI, *human–robot interaction*), el sistema se concibe como un módulo perceptivo que produce, a cada ventana temporal, una etiqueta emocional y una confianza asociada, susceptibles de consumirse por un controlador de interacción para adaptar el comportamiento del robot en tiempo real. De este modo, el módulo permite cerrar el lazo percepción–acción, integrando información afectiva en estrategias de control de alto nivel.

(a) DFEW (Jiang et al., 2020)

(b) MAFW (Liu et al., 2022)

Figura 2: Ejemplo de expresiones faciales dinámicas en los conjuntos de datos (a) DFEW y (b) MAFW.

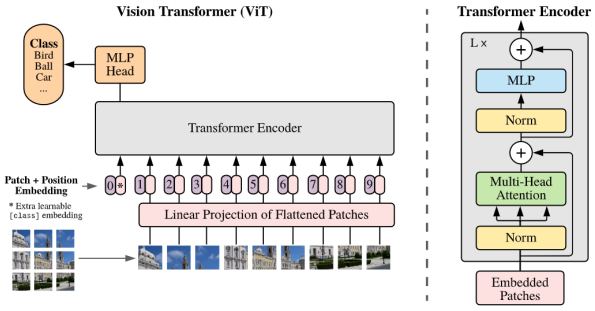
2.2.1. *Vision Transformer ViT-B/32*

El modelo Vision Transformer (ViT) (Dosovitskiy et al., 2020) aborda el reconocimiento visual tratando una imagen como una secuencia de parches independientes (ver figura 3(a)). En primer lugar, la imagen de entrada se divide en parches de tamaño fijo, que se aplanan y se proyectan linealmente a un espacio de características de dimensión constante. En este trabajo se emplea un tamaño de parche de 32×32 píxeles, lo que reduce la longitud de la secuencia de entrada y permite un compromiso adecuado entre coste computacional y capacidad representacional (modelo *ViT-B/32*). A estas representaciones se les añaden embeddings posicionales, así como un token de clasificación aprendible.

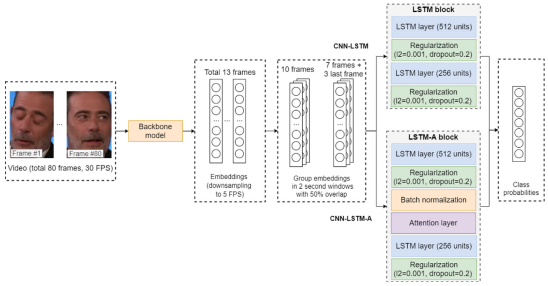
La secuencia resultante se procesa mediante un encoder Transformer estándar, compuesto por capas apiladas de autoatención multi-cabeza y bloques feed-forward, con normalización y conexiones residuales. Finalmente, la representación asociada al token de clasificación se utiliza para la predicción de la clase mediante una capa totalmente conectada, permitiendo capturar relaciones globales entre regiones espaciales de la imagen.

2.2.2. *EMO-AffectNet*

El modelo EMO-AffectNet (Ryumina et al., 2022) se basa en una arquitectura end-to-end CNN-LSTM para el reconocimiento de emociones faciales dinámicas, compuesta por dos bloques principales (ver figura 3(b)). En primer lugar, un backbone convolucional preentrenado (ResNet-50 sobre VGG-Face2) actúa como extractor robusto de características espaciales a partir de imágenes faciales individuales. Este backbone se afina mediante técnicas de regularización y aumento de datos, incluyendo Mixup, con el objetivo de mejorar la capacidad de generalización del sistema.



(a) Arquitectura basada en Vision Transformer (ViT) (Dosovitskiy et al., 2020)



(b) Arquitectura CNN-LSTM (EMO-AffectNet) (Ryumina et al., 2022)

Figura 3: Diagramas conceptuales de los dos modelos empleados: (a) Vision Transformer ViT-B/32 y (b) arquitectura CNN-LSTM EMO-AffectNet.

En segundo lugar, las características profundas extraídas de secuencias de vídeo se agrupan temporalmente y se procesan mediante una red recurrente LSTM, encargada de modelar la evolución temporal de las expresiones faciales. El bloque LSTM consta de dos capas recurrentes con 512 y 256 unidades, respectivamente, e incorpora regularización L2 y dropout para evitar el sobreajuste. La salida del sistema se expresa como una distribución de probabilidad sobre las emociones categóricas consideradas.

2.3. *Métricas*

Para la evaluación del rendimiento se emplearon métricas individuales por clase y métricas globales. En primer lugar, se calcularon la exactitud, la precisión, la sensibilidad y el F1-Score por clase, con el fin de analizar de manera detallada el comportamiento del sistema en cada categoría emocional. La precisión permite evaluar la proporción de predicciones correctas entre las muestras clasificadas como pertenecientes a una emoción determinada, mientras que la sensibilidad mide la capacidad del modelo para detectar correctamente las muestras

Tabla 1: Desempeño de la arquitectura Vision Transformer ViT-B/32 por emoción en los conjuntos de entrenamiento, prueba y validación.

	Exactitud (%)			**Precisión (%)**			**Sensibilidad (%)**			**F1-score (%)**		
Emoción	Train	Test	Val	Train	Test	Val	Train	Test	Val	Train	Test	Val
Felicidad	92,7	96,2	93,6	79,3	88,7	76,2	88,1	93,5	88,8	83,5	91,0	82,0
Tristeza	91,3	91,8	89,6	72,5	75,9	72,2	74,1	72,3	75,7	73,3	74,0	73,9
Neutral	83,6	83,9	84,8	61,7	62,8	49,6	74,1	72,0	72,3	67,3	67,0	58,9
Ira	89,1	90,3	87,3	74,7	78,0	68,0	62,6	66,7	58,1	68,1	71,9	62,7
Sorpresa	90,3	88,9	85,8	62,8	55,6	50,0	56,3	58,8	60,7	59,4	57,2	54,8
Repulsión	98,9	98,3	91,6	66,7	0,0	53,1	17,2	0,0	5,3	27,4	0,0	9,7
Miedo	92,7	91,3	91,8	53,9	42,8	50,4	39,8	37,6	30,2	45,8	40,0	37,8

Train = entrenamiento; Test = prueba; Val = validación.

reales de dicha emoción. El F1-Score, como media armónica entre precisión y sensibilidad, resulta especialmente útil en escenarios con desbalance de clases.

Adicionalmente, se emplearon métricas globales para resumir el rendimiento del sistema de forma agregada. En particular, se calcularon la sensibilidad media ponderada (Weighted Average Recall, WAR), que pondera el sensibilidad de cada clase en función del número de muestras, y la sensibilidad promedio no ponderada (Unweighted Average Recall, UAR), que asigna el mismo peso a todas las clases y permite evaluar el comportamiento del modelo en presencia de clases minoritarias. Asimismo, se incluyeron la precisión macro y el F1 macro, obtenidas como el promedio no ponderado de dichas métricas entre todas las clases, proporcionando una visión equilibrada del rendimiento global.

Por último, se analizaron las matrices de confusión con el objetivo de identificar patrones de error y posibles solapamientos entre emociones, y se evaluó el rendimiento operativo del sistema en tiempo real mediante la tasa de procesamiento en fotogramas por segundo (FPS).

2.4. Configuración experimental

Los modelos propuestos se implementaron y evaluaron en una estación de trabajo equipada con un procesador AMD Ryzen 5 5600G, una GPU NVIDIA GeForce GTX 1050 Ti y 16 GB de memoria RAM DDR4.

3. Resultados

3.1. Vision Transformer ViT-B/32

En esta sección se analiza el rendimiento del sistema basado en el modelo Vision Transformer ViT-B/32. El flujo experimental se basa en la carga de un modelo previamente entrenado y su aplicación secuencial sobre los conjuntos de entrenamiento (*train*), prueba (*test*) y validación externa (*validation*). En el caso de DFEW, se empleó una partición *train/test* del 80/20, sobre un total de 11 697 clips de vídeo. Para evaluar la generalización del modelo, se llevó a cabo una validación externa utilizando la base de datos MAFW, compuesta por 7576 clips con mayor variabilidad contextual y expresiva.

La Tabla 1 presenta los resultados detallados por emoción y conjunto de datos. Las emociones de *felicidad* y *tristeza* presentan el mejor desempeño global, indicando que el modelo identifica con mayor fiabilidad expresiones bien definidas. Por el contrario, emociones como *miedo* y *repulsión* muestran un rendimiento notablemente inferior, especialmente en validación, lo que sugiere dificultades asociadas al desbalance de clases y a la ambigüedad expresiva. Estos patrones se reflejan también en las matrices de confusión normalizadas mostradas en la Fig. 4, donde se aprecian confusiones recurrentes entre emociones de características similares.

La Tabla 2 resume las métricas globales obtenidas en las distintas particiones. Se observa un rendimiento consistente entre los conjuntos de entrenamiento y prueba, con valores de WAR del 69,31 % y 70,31 %, respectivamente. En la validación externa sobre MAFW se aprecia una reducción moderada del rendimiento (WAR = 62,30 %, UAR = 55,88 %), atribuible a la mayor complejidad del conjunto de datos y a las diferencias de dominio. Las métricas de Precisión macro y F1 macro refuerzan esta tendencia, mostrando una ligera degradación en validación, coherente con un escenario de generalización realista.

Tabla 2: Desempeño global de la arquitectura Vision Transformer ViT-B/32 en los conjuntos de entrenamiento, prueba y validación.

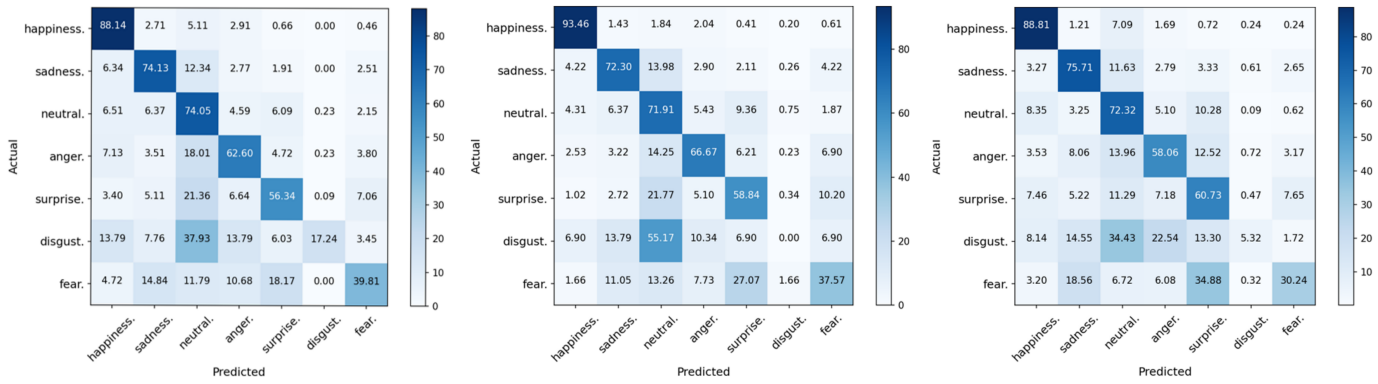
	Train	**Test**	**Val**
WAR (%)	69,31	70,31	62,30
UAR (%)	58,90	57,25	55,88
Precisión macro (%)	67,37	57,68	59,93
F1 macro (%)	60,69	57,31	54,25

Train = entrenamiento; Test = prueba; Val = validación.

Además, se desarrolló un sistema integral para el análisis dinámico de emociones faciales a partir de secuencias de vídeo, tanto en tiempo real mediante webcam como a partir de vídeos pregrabados. El flujo de análisis se estructura en varias etapas: detección facial mediante Haar Cascades, recorte de la región facial, redimensionado a 224×224 píxeles, conversión a tensor y normalización. Los fotogramas procesados se almacenan en un buffer temporal que actúa como una ventana deslizante. Al alcanzarse un número fijo de fotogramas consecutivos (8 o 16, según la configuración), estos se agrupan como un clip y se introducen al modelo para su análisis conjunto.

El modelo genera como salida una distribución de probabilidad sobre las emociones básicas consideradas. La emoción con mayor probabilidad se selecciona como predicción final y se visualiza en pantalla junto con su nivel de confianza y la tasa de procesamiento medida en FPS. La Fig. 5(a) presenta capturas del sistema en funcionamiento en tiempo real utilizando una webcam para el registro de las imágenes.

Durante las pruebas experimentales, el sistema mantuvo una tasa media de procesamiento comprendida entre 4,8 y 5,1 FPS, valor suficiente para interacción humano–robot básica,



(a) Resultados entrenamiento (b) Resultados prueba (c) Resultados validación

Figura 4: Matrices de confusión normalizadas en los conjuntos de (a) entrenamiento, (b) prueba, y (c) validación.

donde la latencia y la continuidad visual condicionan la respuesta del sistema interactivo. Se observó una reducción adicional de aproximadamente 2–3 FPS cuando se superpone información textual sobre la imagen en cada iteración, efecto atribuible a la sobrecarga de renderizado en la CPU empleada.

3.2. Modelo EMO-AffectNet

La validación del modelo EMO-AffectNet se llevó a cabo mediante un protocolo cualitativo en tiempo real, utilizando una webcam para capturar secuencias de vídeo facial. Durante estas pruebas, se simularon distintas expresiones faciales correspondientes a emociones básicas (Figura 5(b)), evaluando la coherencia temporal de las predicciones, la estabilidad del sistema ante cambios dinámicos y la latencia de respuesta. Este análisis permitió valorar el comportamiento del modelo en escenarios interactivos, aunque sin disponer de anotaciones de referencia que posibiliten una evaluación cuantitativa estándar. El sistema clasificó las emociones de forma coherente, mostrando una elevada estabilidad temporal, y mantuvo una tasa de procesamiento de 10–13 FPS.

(a) Modelo ViT-B/32. (b) Modelo EMO-Affectnet.

Figura 5: Fotogramas a tiempo real detectando emociones con los modelos (a) ViT-B/32 y (b) EMO-Affectnet.

4. Discusión

Los resultados obtenidos permiten analizar de forma complementaria el comportamiento de las dos arquitecturas evaluadas, diferenciando entre el rendimiento cuantitativo del modelo ViT-B/32 y la capacidad de respuesta en tiempo real de ambos modelos.

El modelo basado en Vision Transformers (ViT-B/32) presenta un desempeño sólido en términos de métricas globales y por clase cuando se evalúa sobre el conjunto DFEW, alcanzando valores de WAR del 69,31 % en entrenamiento y 70,31 % en test, con un F1 macro de hasta 60,69 %. Sin embargo, al validarse de forma externa sobre MAFW, se observa una degradación moderada del rendimiento (WAR = 62,30 %, UAR = 55,88 %, F1 macro = 54,25 %), atribuible a las diferencias de dominio y a la mayor complejidad contextual del conjunto de validación. Este comportamiento es consistente con observaciones previas en la literatura (Ryumina et al., 2022), donde los modelos entrenados en entornos controlados o semiestructurados tienden a ver reducido su rendimiento al enfrentarse a dominios más heterogéneos y emocionalmente ambiguos.

El análisis por emoción revela que las expresiones con patrones faciales bien definidos, como felicidad y tristeza, son reconocidas con mayor fiabilidad, mientras que emociones como miedo o repulsión muestran un rendimiento notablemente inferior, especialmente en validación externa. Este fenómeno ha sido ampliamente documentado y suele atribuirse tanto al fuerte desbalance de clases presente en los conjuntos de datos como a la naturaleza más sutil y dependiente del contexto de dichas emociones. Las matrices de confusión confirman además la existencia de solapamientos frecuentes entre emociones de características similares, lo que refuerza la dificultad inherente del reconocimiento emocional en escenarios reales.

En cuanto al sistema basado en la arquitectura híbrida CNN–LSTM (EMO-AffectNet), los resultados cualitativos evidencian una elevada estabilidad temporal y una mayor capacidad de respuesta en tiempo real, alcanzando tasas de procesamiento significativamente superiores a las del modelo ViT-B/32 (entre 10–13 FPS frente a 4,8–5,1 FPS). Este comportamiento confirma la idoneidad de las arquitecturas recurrentes para aplicaciones de robótica interactiva, donde la baja latencia es prioritaria para garantizar una interacción fluida y segura con el usuario. No obstante, las pruebas de generalización ponen de manifiesto limitaciones en la detección de micro-expresiones y emociones de baja intensidad, que tienden a ser clasificadas como neutras o reconocidas con baja confianza, un problema

recurrente en sistemas basados exclusivamente en información visual. Cabe señalar que el rendimiento en tiempo real del modelo ViT-B/32 está condicionado por la plataforma de ejecución empleada, y podría mejorarse significativamente mediante el uso de hardware más potente y la optimización del pipeline de inferencia y renderizado, ampliando así su viabilidad en aplicaciones de interacción humano-robot.

Una limitación relevante del modelo EMO-AffectNet en el contexto de este trabajo es la imposibilidad de realizar una evaluación cuantitativa estándar mediante métricas globales y por clase. Esta restricción deriva de su diseño modular, que combina modelos preentrenados independientes para el análisis espacial y temporal, dificultando una validación integrada sin una reimplementación completa del pipeline de entrenamiento y evaluación. En contraste, el modelo ViT-B/32, al constituir una arquitectura unificada y completamente trazable, permite una evaluación más rigurosa y comparable desde el punto de vista experimental.

Los resultados sugieren que ambos enfoques deben interpretarse como complementarios. Por tanto, no se plantea una comparación directa entre ambos modelos en términos cuantitativos, sino un análisis que pone de manifiesto el compromiso entre precisión y operatividad en tiempo real. El modelo ViT-B/32 resulta más adecuado para estudios cuantitativos y análisis comparativos a gran escala, mientras que la arquitectura CNN-LSTM ofrece ventajas claras en aplicaciones en tiempo real y sistemas interactivos, donde la baja latencia es un factor clave. Estas observaciones refuerzan la necesidad, ampliamente reconocida en la literatura, de diseñar sistemas que equilibren precisión, capacidad de generalización y viabilidad computacional para su despliegue en entornos reales.

5. Conclusiones

En este trabajo se ha abordado el reconocimiento automático de emociones faciales dinámicas en vídeo desde una perspectiva aplicada a la interacción humano–robot. Se han evaluado dos enfoques complementarios basados en aprendizaje profundo: un modelo unificado basado en Vision Transformers (ViT-B/32) y una arquitectura híbrida CNN-LSTM (EMO-AffectNet) orientada a ejecución en tiempo real. Los resultados muestran que el modelo ViT-B/32 ofrece un rendimiento cuantitativo sólido y una capacidad razonable de generalización al validarse sobre un conjunto externo, mientras que la arquitectura CNN-LSTM destaca por su elevada estabilidad temporal y mayor tasa de procesamiento, resultando más adecuada para aplicaciones interactivas.

El análisis por emoción pone de manifiesto la dificultad inherente del reconocimiento emocional en escenarios no controlados, especialmente para emociones de baja intensidad o con fuerte desbalance de clases. En conjunto, el estudio evidencia el compromiso existente entre precisión, generalización y operatividad en tiempo real, y sugiere que la combinación de ambos enfoques constituye una solución prometedora para sistemas de percepción emocional en entornos reales.

Como líneas futuras, se plantea la optimización del pipeline para su despliegue en plataformas robóticas. Además, también se explora la integración de información multimodal, incorporando señales fisiológicas como electrocardiografía (ECG), respuesta galvánica de la piel (GSR) y temperatura cutánea (SKT), con el objetivo de complementar la información visual y mejorar la robustez y fiabilidad del reconocimiento emocional.

En conjunto, el trabajo sienta una base para integrar percepción afectiva en arquitecturas de interacción humano–robot, apoyando estrategias adaptativas centradas en el usuario.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia, Innovación y Universidades de España a través del proyecto PID2022-138206OB-C33.

Referencias

Charles, D., 1872. The Expression of the Emotions in Man and Animals. John Murray, London.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16X16 Words: Transformers for Image Recognition at Scale. ArXiv 2010.11929, 1–22.
DOI: 10.48550/arXiv.2010.11929

Ekman, P., Friesen, W. V., 2 1971. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology 17 (2), 124–129.
DOI: 10.1037/H0030377

Graziotin, D., Wang, X., Abrahamsson, P., 7 2015. Do feelings matter? On the correlation of affects and the self-assessed productivity in software engineering. Journal of Software: Evolution and Process 27, 467–487.
DOI: 10.1002/smr.1673

Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J., 2020. DFEW: A large-scale database for recognizing dynamic facial expressions in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, New York, USA, p. 2881–2889.
DOI: 10.1145/3394171.3413620

Kopalidis, T., Solachidis, V., Vretos, N., Daras, P., 2 2024. Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. Information 15 (3), 135.
DOI: 10.3390/INF015030135

Li, S., Deng, W., 2022. Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing 13 (3), 1195–1215.
DOI: 10.1109/TAFFC.2020.2981446

Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., Shan, S., 10 2022. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. Proceedings of the 30th ACM International Conference on Multimedia, 24–32.
DOI: 10.1145/3503161.3548190

Mohana, M., Subashini, P., Krishnaveni, M., 2023. Emotion Recognition from Facial Expression Using Hybrid CNN–LSTM Network. International Journal of Pattern Recognition and Artificial Intelligence 37 (08), 2356008.
DOI: 10.1142/S0218001423560086

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, A., 8 2022. Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review 2022 56:4 56 (4), 3005–3054.
DOI: 10.1007/S10462-022-10246-W

Patel, K., Mehta, D., Mistry, C., Gupta, R., Tanwar, S., Kumar, N., Alazab, M., 2020. Facial Sentiment Analysis Using AI Techniques: State-of-the-Art, Taxonomies, and Challenges. IEEE Access 8, 90495–90519.
DOI: 10.1109/ACCESS.2020.2993803

Ryumina, E., Dresvyanskiy, D., Karpov, A., 12 2022. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. Neurocomputing 514, 435–450.
DOI: 10.1016/J.NEUCOM.2022.10.013

Xiong, L., Zhang, J., Zheng, X., Wang, Y., 2 2024. Context Transformer and Adaptive Method with Visual Transformer for Robust Facial Expression Recognition. Applied Sciences 14 (4), 1535.
DOI: 10.3390/APP14041535