Revista ELECTRO, Vol. 46, 2024, pp. 351 - 356
https://itchihuahua.mx/revista_electro
ISSN Electrónico: 3061-774X

RECONOCIMIENTO DE EMOCIONES MEDIANTE IMÁGENES DE EXPRESIONES FACIALES BASADO EN ARQUITECTURA DE RED SIAMESA

Luis Xavier Nevárez Ochoa, Mario I. Chacón-Murguía, Juan Alberto Ramírez Quintana
Laboratorio de Percepción Visual, Tecnológico Nacional de Mexioc/I T Chihuahua,
Ave. Tecnológico No. 2909, 31310, Chihuahua, Chih., México
+52(614)2012000, Fax +52(614)4135187
luisxavier21@hotmail.com, mchacon@ieee.org, juan.rq@chihuahua.tecnm.mx

RESUMEN.

El interés por estado emocional de una persona se ha incrementado en gran medida en el sector salud debido a las consecuencias que puede tener una situación de un estado emocional deteriorado no atendido. Por lo tanto, en este trabajo se propone un modelo de red profunda con arquitectura siamesa para el reconocimiento de emociones humanas por medio de imágenes, y que sea capaz de funcionar sobre un sistema embebido de manera rápida, portable y sin complicaciones de incompatibilidad de paquetes o una pobre optimización. Para desarrollar el modelo final de red siamesa se realizó una etapa previa llevando a cabo entrenamientos con pares de emociones que siguieran estas características: que fueran similares entre sí, que fueran medianamente similares, y que fueran distintas. El número de emociones se fue incrementando hasta que el modelo mantuviera una precisión general por arriba del 70%. Esta restricción se logró alcanzar con 5 emociones. El modelo final resultó tener un desempeño de 83.83% de precisión.

Palabras Clave: Clasificación de emociones, red siamesa, red convolucional.

ABSTRACT.

The interest in a person's emotional state has significantly increased in the health sector due to the consequences that can arise from an unattended deteriorated emotional state. Then, this paper proposes a deep network model with a Siamese architecture for recognizing human emotions through face images, which can operate on an embedded system quickly, portably, and without complications related to package incompatibility or poor optimization. The development of the Siamese network model is based on a preliminary stage conducted by training with pairs of emotions that follow these characteristics: being similar to each other, being moderately similar, and being distinct. The number of emotions was gradually increased until the model maintained an overall accuracy above 70%. This threshold was achieved with five emotions. The final model resulted in a performance of 83.83% accuracy.

Keywords: Emotion classification, Siamese network, convolutional network.

1. INTRODUCCIÓN

El campo del reconocimiento de emociones humanas ha sido de interés para distintas áreas de investigación como por ejemplo el de medicina o psicología, ya que entrega información sobre el estado de las personas y cómo podrían reaccionar ante distintos escenarios. Debido al avance de tecnología de los últimos años, se han logrado diseñar y crear metodologías que permitan reconocer a cierto grado, las emociones que presenta un individuo en un instante de tiempo. Entre los modelos de reconocimiento de emociones más destacados se puede encontrar aquellos que hacen uso de imágenes para reconocer las expresiones faciales, las que hacen uso de interfaces cerebro computadora (BCI), el análisis de texto y el reconocimiento por medio de la voz del individuo. Cada uno de los métodos descritos anteriormente tienen sus fortalezas y debilidades: BCI se considera un poco molesto por el usuario, ya que requiere la colocación de electrodos en la cabeza de este, sin embargo, su desempeño es difícil de vulnerar y tiene buen rendimiento. El análisis de texto puede fallar mucho si la persona no es consistente en su escritura o contiene faltas ortográficas, pero tiene mucho potencial al usarse en redes sociales. El análisis de la voz puede contener ruido de fondo y depende de la duración de la grabación, pero es un método no intrusivo y no requiere dispositivos especializados para recoger la información. Finalmente, el reconocimiento por medio de expresiones faciales depende de la resolución de las imágenes y es sensible a la iluminación, con la ventaja de que no es intrusivo y es el método más cercano al utilizado por los humanos, puesto que se depende en gran medida de la información que obtenida por medio del sistema de visión humano. Es este último método el que se utilizará en el presente trabajo.

Entre los trabajos analizados, se puede encontrar diversos algoritmos para la detección de emociones. Por ejemplo, en [1] se utiliza una detección multimodal utilizando la transformada de Hilbert Huang para características visuales y de audio. Otros, en cambio, utilizan técnicas de aprendizaje profundo como en [2] [3], donde se utiliza una red neuronal convolucional (CNN por sus siglas en inglés). Otras técnicas encontradas en la literatura que optan por modelos que no incluyen redes neuronales son los expuestos en [4], [5], los cuales logran buenos resultados en el análisis de distintas emociones en imágenes o secuencias de video. En otros artículos se utilizan diferentes métodos para llevar a cabo el reconocimiento, como en [6], donde se analizan las expresiones faciales, tono de voz, electrocardiograma, pulso sanguíneo, electroencefalograma, respiración, temperatura corporal, electromiografía y actividad electrodérmica. En unos casos se utiliza únicamente un tipo de red o algoritmo, pero como se vio en la literatura, existen métodos diferentes para la extracción de rasgos, como por ejemplo la utilización de modelos híbridos que combinan una o más redes para mejorar la

precisión del reconocimiento. Tal tarea es realizada en [7], que combina un clasificador tipo K vecinos más cercanos (k-NN) con una red neuronal tipo perceptrón multicapa (MLP) para el reconocimiento en imágenes 3D. En [8] se utiliza una CNN en conjunto con una red neuronal bidireccional recurrente (BRNN), donde la imagen de entrada se introduce en la primera red para la extracción de rasgos de alto nivel. Luego, en la BRNN, se aprenden los cambios de cada una de las imágenes de entrada. En otro trabajo se propone una arquitectura 3D CNN en conjunto con celdas de memoria de corto y largo plazo (LSTM) para llevar a cabo la tarea del reconocimiento de emociones, así como la combinación CNN y redes neuronales recurrentes (RNN) [9]. En [4], la identificación se lleva a cabo utilizando múltiples fuentes tales como expresiones faciales, poses, movimiento del cuerpo y voz. Estas características espacio temporales se introducen en una 3D CNN y redes bayesianas profundas (DBN), donde se modela la información presente en audio y video para realizar un reconocimiento óptimo. La última combinación mencionada, CNN-RNN, es también utilizada en [10].

A pesar de la investigación para reconocer emociones humanas con modelo profundos es amplia, no siempre se contempla en su diseño aspectos de la complejidad del modelo y requerimiento computacionales. Es por ello que en este trabajo se presenta el diseño de un modelo de red de baja complejidad que puede funcionar sobre un sistema embebido de manera rápida, portable y sin complicaciones de incompatibilidad de paquetes o un pobre desempeño. En las siguientes secciones se describen las bases de datos utilizadas para realizar el diseño, así como el diseño del modelo desarrollado.

2. BASE DE DATOS

Las bases de datos utilizadas para el entrenamiento y evaluación de la metodología empleada, así como en varias de las encontradas en el estado del arte, son las bases de datos Extended Cohn-Kanade Dataset (CK+), FER-2013 y KDEF. La base de datos de Cohn-Kanade (CK) fue introducida en el año 2000 con el propósito de promover la investigación en el campo de la detección de expresiones faciales en secuencias de video. Una actualización de la base CK fue liberada en el año 2010 con el fin de incrementar la cantidad de muestras y el número de sujetos de prueba, así como una revisión y validación de la información contenida en la base de datos. Dicha actualización fue nombrada Extended Cohn-Kanade Dataset (CK+). Para este trabajo se utilizó una versión reducida de CK+, limitada a 7 emociones. Esta reducción se consiguió al extraer los últimos 3 cuadros de cada secuencia de video de la base de datos original. Las características principales de la base de datos son las siguientes:

- 123 sujetos de prueba de edades comprendidas entre 18 a 50 años, de los cuales:
  - 69% son mujeres.
  - 81% Euroamericanos.
  - 13% Afroamericanos.
  - 6% de diversas etnias.
- 981 imágenes de expresiones faciales distribuidas de la siguiente manera:
  - Enojo: 135 imágenes.
  - Sorpresa: 249 imágenes.
  - Tristeza: 84 imágenes.
  - Desagrado: 54 imágenes.
  - Alegría: 207 imágenes.
  - Miedo: 75 imágenes.
  - Disgusto: 177 imágenes.
- Disgusto se refiere a un sentimiento de pesadumbre causado por contrariedad o accidente, mientas que desagrado se refiere a una sensación de rechazo por algo o alguien.
- Esta base de datos se enfoca a emociones positivas y negativas, por lo que no se usa la clase neutral.
- Todas las imágenes son a escala de grises y con dimensiones de 48x48 píxeles.

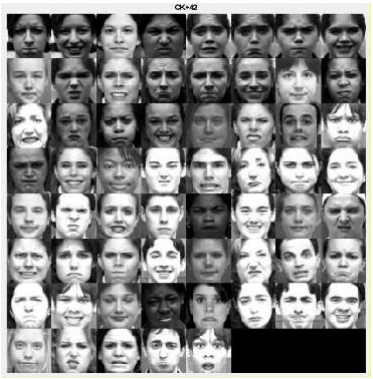

Figura 1. Muestra de la base de datos CK+.

La base de datos Facial Emotion Recognition Challenge, comúnmente abreviado como FER-2013, fue creada por Pierre-Luc Carrier y Aaron Courville como parte de un proyecto de investigación [11]. En este trabajo se utiliza una versión preparada por Goodfellow et al. para que concordara con 7 emociones: enojo, disgusto, miedo, alegría, tristeza, sorpresa y neutral. Su información es la siguiente:

- 35,887 imágenes de expresiones faciales distribuidas de la siguiente manera:
  - Enojo: 4953 imágenes.
  - Sorpresa: 4002 imágenes.
  - Tristeza: 6077 imágenes.
  - Alegría: 8989 imágenes.
  - Miedo: 5121 imágenes.
  - Disgusto: 547 imágenes.
  - Neutral: 6198 imágenes.
- Todas las imágenes son a escala de grises y con dimensiones de 48x48 pixeles.
- Toda la base de datos se concentra en un archivo con formato Comma Separated Values (CSV).

Revista ELECTRO, Vol. 46, 2024, pp. 351 - 356
https://itchihuahua.mx/revista_electro
ISSN Electrónico: 3061-774X

La base de datos KDEF, del inglés Karolinska Directed Emotional Faces, es un conjunto de 4,900 imágenes de expresiones faciales de 70 distintos individuos. Fue creada por el departamento de Neurociencia clínica, sección de Psicología del Instituto Karolinska, en Suecia [12]. Las emociones presentes en la base de datos son las siguientes: Miedo (afraid), Enojo (angry), Disgusto (disgusted), Alegre (happy), Neutral (neutral), Triste (sad), Sorprendido (surprised). Algunos ejemplos de imágenes de las bases de datos mencionadas se muestran en la Figura 2.

a) b) c)

Figura 2. Muestra de la base de datos a) CK+. b) FER-2013. c) KDEFpara 5 emociones.

3. PREPROCESAMIENTO

Las bases de datos contienen imágenes con resoluciones variables, algunas de 48x48 pixeles a escala de grises, como lo son CK+ y FER-2013. Debido a que las redes neuronales requieren que los datos de entrada compartan la misma estructura, se realizó una normalización de las imágenes, asegurándonos que el espacio de color sea a escala de grises y contengan la misma resolución. Esto último se logra al redimensionar las imágenes utilizando técnicas como la interpolación bicubica.

4. RED SIAMESA

Una red siamesa contiene dos o más instancias de una subred, los pesos y parámetros son los mismos en todas ellas. Este tipo de arreglo permite al sistema medir la distancia entre clases y utilizar esta información para aprender si las entradas dadas a cada subred son similares o diferentes [13], [14]. Su objetivo es encontrar una función de similitud entre clases. En el caso que las entradas sean imágenes, es posible utilizar CNNs para la etapa de extracción de rasgos. Una manera gráfica de representar la arquitectura de una red siamesa se muestra en la Figura 3. Una red siamesa realiza un mapeo no lineal ϕ de su dominio de entrada X, dado por una imagen I, a un espacio euclídeo $\mathbb{R}$, es decir, extrae características del espacio X y lo convierte a otro espacio real,

$$\phi: X \rightarrow \mathbb{R} \quad (1)$$

Este mapeo se logra gracias a la etapa de extracción de rasgos realizada por las subredes y el aprendizaje de parámetros discriminativos. Dichas subredes serán definidas como N_k, donde k es el número de instancias de red existentes y todas compartirán los parámetros de los pesos W. A cada subred se le asigna una entrada X_k y producirá un vector de características $f_w(X_k)$. En este trabajo, la incrustación no lineal ϕ varía con respecto el método de extracción de rasgos.

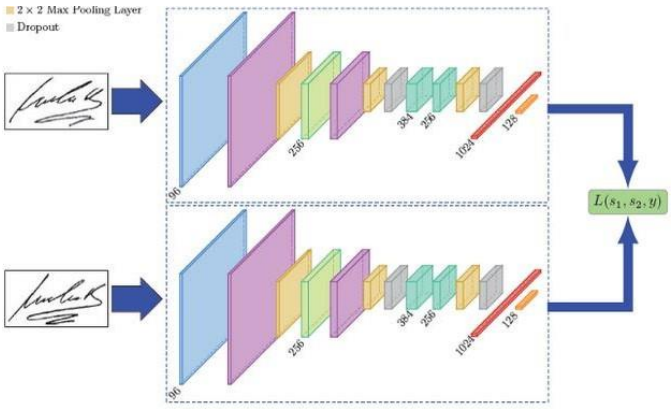

Figura 3. Ejemplo de una red siamesa para verificación de firmas [15].

Para obtener el nivel de similitud entre las clases, se utiliza la función de distancia D_W, es decir, las entradas X que pertenezcan a la misma clase tendrán un valor bajo, mientras que las que sean de clases diferentes tendrán una alta separación. La manera en la que se obtiene dicho valor de proximidad es utilizando la distancia Euclidiana

$$D_W(X_1, X_2) = \|f_w(X_1) - f_w(X_2)\| \quad (2)$$

El resultado obtenido es utilizado por el método de clasificación seleccionado para predecir si las imágenes comparadas son similares o diferentes.

El entrenamiento de las redes siamesas es distinto al de arquitecturas de redes tradicionales debido a que se requieren los vectores de características de las subredes y la comparación entre ellas con la función de distancia D_W. Por lo que, la función de pérdida se encuentra limitada y generalmente se utilizan tres modelos para llevar a cabo el entrenamiento: pérdida contrastiva, por triplete y por cuatrillo. En este trabajo se utilizó la función contrastiva

$$L(X_1, X_2, y) = (1-y)\frac{1}{2}D_W^2 + y\frac{1}{2}max(0, m - D_W)^2 \quad (3)$$

donde X_1 y X_2 son las entradas al sistema, y es una etiqueta binaria que indica si pertenecen a la misma clase o no (0 y 1 respectivamente), m es un margen que anula la pérdida de dos clases distintas que presenten una distancia muy grande. De manera gráfica, la función de pérdida contrastiva opera como se muestra en la Figura 4.

Revista ELECTRO, Vol. 46, 2024, pp. 351 - 356
https://itchihuahua.mx/revista_electro
ISSN Electrónico: 3061-774X

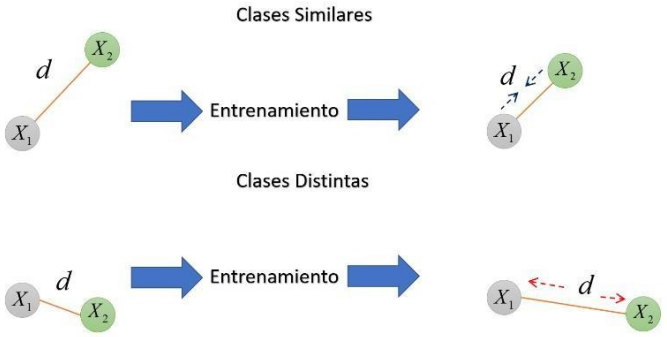

Figura 4. Funcionamiento de función de pérdida contrastiva.

5. MODELO DISEÑADO

En este trabajo se decidió utilizar una red tipo CNN para la extracción automática de rasgos. A continuación, se detalla su arquitectura.

El modelo de la red siamesa diseñado se compone de una a primera etapa de extracción de rasgos que se muestra en la Figura 5, seguida de la etapa de clasificación mediante una red completamente conectada. El modelo completo se ilustra en la Figura 6 y su descripción en la Tabla 1. En este modelo las características se obtienen automáticamente por medio de una red CNN cuya salida es un vector de 2048 dimensiones, el cual será utilizado para obtener las distancias entre las imágenes. La estructura de la CNN consta de 1 capa de entrada, 3 capas convolucionales con función de activación ReLu y cada una conectada a una capa de Max Pooling. Finalmente, las características se introducen a una capa completamente conectada cuya salida.

Tabla 1. Estructura de la red siamesa.

Capa	Características
Entrada	48x48x1
Conv1	10x10 (128 filtros)
Conv2	7x7 (256 filtros)
Conv3	4x4 (256 filtros)
Max-Pooling	2x2 (stride = 2)
Fully-Connected	2048 (Narrow-normal)
Número de parámetros	61986

6. RESULTADOS

Debido a la forma en la que funcionan las redes siamesas se obtuvieron resultados para detección de similitud y para clasificación. Para ambos casos, la imagen ancla se obtiene al extraer una imagen de la base de datos. Posteriormente, se obtienen los vectores de características de la emoción ancla y de todas las otras imágenes de la base de datos. La manera en la que las redes siamesas miden el grado de similitud entre una clase u otra es por medio de la distancia euclídea entre sus vectores de características, por lo que se procede a obtener la diferencia entre la imagen ancla y todas las otras emociones del conjunto de datos. Finalmente, se ordenan de menor a mayor, siendo las distancias más pequeñas las que más posibilidades tienen de ser de la misma clase que la imagen ancla. Con este vector contenedor de las menores distancias es posible catalogar que imágenes pueden ser consideradas similares por medio de una umbralización. Por el contrario, si se desea clasificar la imagen, se consideran las clases de las ***N*** imágenes más cercanas a las características de la imagen ancla. La clase que se asignará será aquella etiqueta con mayor número de apariciones.

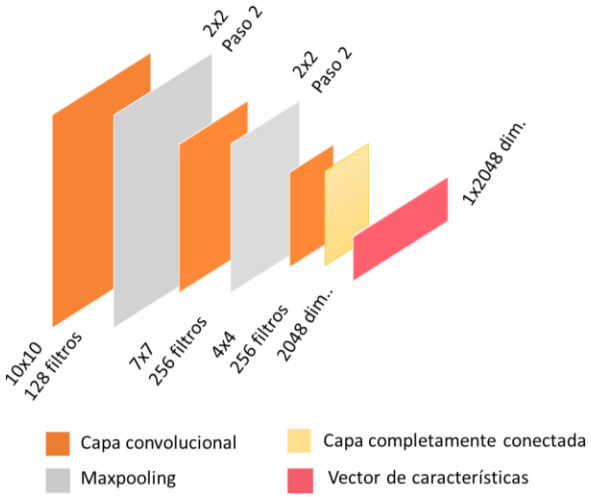

Figura 5. Arquitectura CNN para extracción de rasgos.

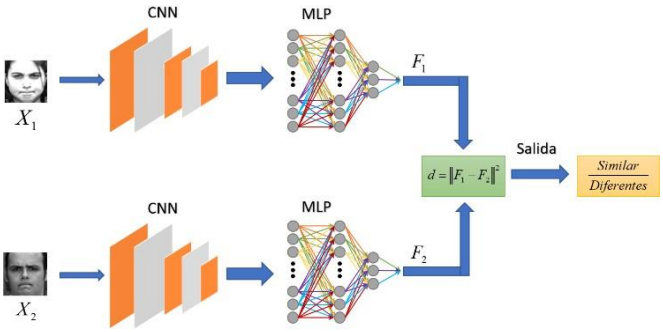

Figura 6. Modelo de la red siamesa.

La red siamesa se entrenó con las bases de datos de CK+, FER-2013 y KDEF por separado. Se dividieron en 70% de muestras de cada base de datos para entrenamiento, 15% de validación y 15% para prueba. El entrenamiento se realizó con el optimizador ADAM, taza de aprendizaje de 0.0001, 100 épocas y un tamaño de lote de 180. La implementación de las redes se realizó con 5 emociones, las cuales son cuatro que comparten las bases de datos y neutral.

Un primer modelo fue entrenado con la base de datos CK+. Posteriormente, se realizó un rentrenamiento del modelo. Para el rentrenamiento se cambió a la base de datos de KDEF, publicada por el Instituto Karolinska con la finalidad de proporcionar una base de datos para la identificación de emociones, la cual cuenta con más imágenes, 4,900 dividido entre 70 individuos, y se añaden emociones, siendo en total las emociones de: miedo, enojo, disgusto, alegre, neutral, triste y sorprendido, con 700 fotos cada una. Además, la resolución de las imágenes también es diferente, aumentando de los 42x42 pixeles a 100x100. Para esto se realizó un ajuste del tamaño de entrada a las nuevas dimensiones. Debido a que la base de datos y el número de emociones fue diferente el reentrenamiento se realizó siguiendo el siguiente esquema:

- Se entrena el modelo con 2 emociones siguiendo el esquema: emociones similares, medianamente similares y diferentes.
- Se anotan los resultados y se analizan cuales emociones funcionan bien juntas y cuáles no.
- Se incrementa el número de emociones en 1 y se repite el proceso.
- Lo anterior es repetido hasta que el desempeño de la red disminuya a menos del 70% de precisión para tres o más emociones.

En la prueba con 5 emociones arrojó resultados positivos, con precisiones arriba del 70% en prueba. Algunos de los resultados obtenidos se muestran en la Tabla 2, donde se utiliza el conjunto de pruebas para la tarea de encontrar la similitud entre imágenes.

Tabla 2. Resultados de ambas redes de la red siamesa para 5 emociones.

Emociones	TP	TN	FP	FN	Precisión
Enojo	10	158	6	35	80.38%
Alegría	36	169	0	4	98.09%
Sorpresa	27	156	12	14	87.56%
Neutral	24	165	5	15	90.43%
Miedo	12	141	28	28	73.21%
Enojo	18	118	46	27	65.07
Alegría	31	169	0	9	95.69
Sorpresa	35	167	1	6	96.65
Neutral	15	138	32	24	73.21
Tristeza	19	144	25	21	77.99

El modelo para 5 emociones se implementó en una computadorapersonal y el sistema embebido Nvidia Jetson Xavier. La tarjeta Jetson Xavier fue seleccionada porque es una plataforma adecuada para sistemas embebidos de computación afectiva implementada en robots y camaras. Los resultados obtenidos se analizaron y compararon para determinar la viabilidad de implementar este tipo de arquitectura de red en dispositivos con bajos recursos computacionales.

Las especificaciones de los equipos se encuentran contenidos en la Tabla 3, donde se muestran los elementos que cada uno contiene. Tanto la computadora como la Jetson Javier tardaron un tiempo de un segundo aproximadamente para la clasificación.

Tabla 3. Especificaciones principales de los equipos.

	PC	Nvidia Jetson Xavier
CPU	Intel i5 10400f 6-cores	ARM v8.2 8-core
RAM	16 Gb LPDDR4	32Gb LPDDR4*
GPU	Nvidia RTX 3070	Nvidia Volta 512core
VRAM	8 Gb	32Gb LPDDR4*
Tensor Cores	184 Tensor Cores	64 Tensor Cores

*Memoria compartida.

7. CONCLUSIONES

Este trabajo presentó el diseño de una red siamesa para el reconocimiento de emociones. El modelo se compone de dos etapas, la etapa de extracción de rasgos mediane un modelo CNN, y la etapa de clasificación usando una red completamente conectada. El mejor desempeño del modelo al clasificar múltiples emociones se obtuvo al clasificar 5 emociones con un desempeño promedio en precisión de 83.83%. El modelo diseñado fue implementado en una computadorapersonal y el sistema embebido Nvidia Jetson Xavier. En ambos casos la complejidad del modelo fue lo bastante reducida para implementarse con los recursos disponibles y el tiempo de clasificación fue de 1 segundo.

8. AGRADECIMIENTOS

Los autores agradecen al tecnológico Nacional de México/I.T. Chihuahua 19182.24-P por el financiamiento de esta investigación.

9. REFERENCIAS

[1] S. Mo, J. Niu, S. Y. y S. Das, A novel feature set for video emotion recognition. Neurocomputing,., vol. 291, pp. 11-20, 2018. https://doi.org/10.1016/j.neucom.2018.02.052

[2] J. Yan, Z. W. Z. Cui, T. C. Z. T. y Y. Zong, Multi-cue fusion for emotion recognition in the wild. Neurocomputing, , vol. 309, pp. .27-3, 2018. https://doi.org/10.1016/j.neucom.2018.03.068

[3] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali y M. Zareapoor, Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters, vol. 115, pp. 101-106, 2018. https://doi.org/10.1016/j.patrec.2018.04.010

[4] R. Favarett, P. K. S. Musse, F. Vilanova y Â. Costa, Detecting personality and emotion traits in crowds from video sequences. Machine Vision and Applications, vol. 30, no 5 , pp. 999-1012, 2018. https://doi.org/10.1007/s00138-018-0979-y

[5] S. Kumar, M. Bhuyan, B. Lovell y Y. Iwahori, Hierarchical uncorrelated multiview discriminant locality preserving projection for multiview facial expression recognition. Journal of Visual Communication and Image Representation, vol. 54 , pp. 171-181, 2018. https://doi.org/10.1016/j.jvcir.2018.04.013

[6] M. Al, A. Mosa, F. Machot y K. Kyamakya, Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review. Studies in Systems, Decision and Control, vol. 109, pp. 287-302, 2017. https://doi.org/10.1007/978-3-319-58996-1_13

[7] P. Tarnowski, M. Kołodziej, M. A. y R. Rak, Emotion recognition using facial expressions. Procedia Computer Science, vol. 108, pp. 1175-1184, 2017, vol. 108, pp. 1175- 1184, 2017. https://doi.org/10.1016/j.procs.2017.05.025

[8] J. Yan, Z. W. Z. Cui, C. Tang, T. Zhang y Y. Zong, Multi-cue fusion for emotion recognition in the wild. Neurocomputing, vol. 309, pp. 27-35, 2018. https://doi.org/10.1016/j.neucom.2018.03.068

[9] Z. Yu, G. Liu, Q. Liu y J. Deng, Spatio-temporal convolutional features with nested LSTM for facial expression recognition. Neurocomputing , vol. 317, pp. 50-57, 2018. https://doi.org/10.1016/j.neucom.2018.07.028

[10] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali y M. Zareapoor, Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters, vol. 115, pp. 101-106, 2018. https://doi.org/10.1016/j.patrec.2018.04.010

[11] G. I. et al, Challenges in Representation Learning: A Report on Three Machine Learning Contests. de Lecture Notes in Computer Science, Berlin, Heidelber, Springer, 2013., pp. 117-124. https://doi.org/10.1007/978-3-642-42051-1_16

[12] E. Lundqvist, F. D. y A. Öhman, The Karolinska Directed Emotional Faces – KDEF. Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, Suecia, 1998.

[13] Y. Gao y et al, Calculating Color Differences of Images via Siamese Neural Network. de 2024 IEEE International Symposium on Circuits and Systems (ISCAS), Singapore, Singapore, 2024.

[14] R. Vachhani, S. Mandal y B. Gohe, Low-Resolution FaceRecognition Using Multi-Stream CNN in Siamese Framewor. de *2023 Seventh International Conference onImage Information Processing (ICIIP)*, Solan, India, 2023. 10.1109/ICIIP61524.2023.10537740

[15] S. Ravichandiran, Hands-On Deep Learning Algorithmswith Python. 1st ed. Birmingham: Packt, 2019, p. 437-456