Deep Neural Network Architecture: Application for Facial Expression Recognition

M. García, and S. Ramírez

***Abstract*—There are great challenges to build a model or architecture in Deep Learning and integrate it into a real-time application. Two of them are the construction or acquisition of large quality datasets (thousands or millions of objects) as well as the computing power needed for the learning process. Finally, efficient architectures are needed for the design of deep neural networks, which requires expertise, human experience and practical work. This work presents a deep neural network architecture to classify two feelings of facial expression: happy and sad. A dataset is also created presenting variations in: image environments, facial expression, pose, age, ethnicity and others. The evidence presented shows a competitive architecture and indicates an accuracy greater than 90% with noisy data. Finally, the implementation of a real-time application for facial expression recognition is shown.**

***Index Terms*—CNN architecture, Deep learning, Facial expression recognition, Visual classification of sentiment.**

I. Introducción

El crecimiento exponencial del contenido visual en medios sociales ha motivado a investigar y analizar los medios sociales multimedia a gran escala. Dentro de estos esfuerzos de investigación, se encuentra el análisis de las emociones y sentimientos en el contenido de medios visuales, que en los últimos años ha incrementado la atención en la investigación y las aplicaciones prácticas. Las imágenes y los videos que muestran fuertes sentimientos pueden fortalecer la opinión transmitida en el contenido e influir efectivamente en la audiencia. Comprender el sentimiento expresado en el contenido visual beneficiará enormemente la comunicación de los medios sociales y permitirá amplias aplicaciones en educación, publicidad y entretenimiento [1].

El reconocimiento de las emociones a través de la detección de expresiones faciales es uno de los campos de estudio importantes en la interacción humano computadora. La expresión y sentimiento de una persona difiere bajo diferentes situaciones con diferentes contextos. Uno de los retos más importantes para el reconocimiento de rostro es descubrir de una forma eficiente y discriminativa descriptores faciales que sean resistentes a grandes cambios. Por ejemplo, ante la existencia de diferencias en la iluminación, diversidad en la expresión, pose, edad, oclusiones parciales y otros cambios [2]. La mayoría de técnicas existentes se enfocan en clasificar 7 expresiones básicas que se han considerado universales en las diferentes culturas y subgrupos, las cuales son: neutral, feliz, sorprendido, con miedo, enojado, triste y disgustado [3], observándose que entre algunas de las categorías existe una gran similitud.

El avance en visión por computadora, un campo de la ciencia que crece rápidamente, se debe principalmente a tres factores: 1) en la actualidad se encuentran cámaras más económicas y con mejores capacidades; 2) el poder de procesamiento disponible; y 3) los algoritmos en está área se están consolidando constantemente. Todo ello ha permitido la integración y desarrollo de sistemas de aplicación de una manera más rápida. La librería de visión por computadora OpenCV, ha hecho posible en gran medida el crecimiento de esta área por permitir a miles de personas hacer más productivo el desarrollo de aplicaciones en múltiples tareas de visión. Con el enfoque de visión en tiempo real, OpenCV ayuda a estudiantes y profesionales a implementar proyectos eficientemente, proporcionándoles la herramienta para continuar con sus investigaciones. La visión por computadora es la puerta de entrada para el desarrollo de aplicaciones que involucran el reconocimiento facial y en conjunto con técnicas de aprendizaje profundo se cuenta con un sistema de identificación de expresiones del rostro para determinar el sentimiento del o los individuos en una escena.

Este trabajo tiene como aporte una arquitectura de una red convolucional profunda que emplea un conjunto de datos en el que se incluyen rostros de diferentes edades, etnias, posiciones y otras condiciones de los rostros en sus elementos. Se ha logrado una precisión de más del 95% en conjuntos de datos en donde las imágenes mantienen las mismas condiciones y mayor a un 90% en el repositorio de información que contiene imágenes heterogéneas. Se presenta la integración de un sistema de identificación del sentimiento de dos expresiones básicas (feliz o triste) en tiempo real, en el cual se utilizan herramientas de visión por computadora y aprendizaje profundo. Además se hace disponible un conjunto de datos de rostros para dos expresiones faciales, el cual cuenta con 3278 imágenes, incluyendo personas de diferentes edades, etnias y ambientes diversos de fondo, posición, oclusión y diferente calidad en las imágenes.

La organización de este documento es el siguiente: En la sección I se describe al lector el problema a resolver, los conceptos principales y los elementos de solución; en la sección II se proporciona un panorama de los trabajos en el reconocimiento de objetos, expresiones faciales y detección de rostro; en la sección III se presenta la arquitectura de la red neuronal profunda; en la sección IV se describe el conjunto de datos construido para el entrenamiento de la red neuronal; la sección V presenta las pruebas, los resultados obtenidos y la implementación del sistema de aplicación, finalmente en la sección VI se presentan algunas conclusiones.

Ambos autores están en la Facultad de Ingeniería Eléctrica de la Universidad Michoacana de San Nicolás de Hidalgo.

M. García Villanueva (email:moigarciav@gmail.com).

S. Ramírez Zavala (email:szavalaram@gmail.com

II. Reconocimiento de Objetos

Enfoques actuales para el reconocimiento de objetos, hacen uso esencialmente de técnicas de aprendizaje de máquina como en [4], [5]. Para mejorar su desempeño, es necesario construir conjuntos de datos muy grandes, aprender modelos mucho más robustos y utilizar mejores técnicas para prevenir el sobreentrenamiento de los sistemas. Las tareas de reconocimiento simples se pueden resolver bastante bien con conjuntos de datos pequeños, aquellos que se componen de decenas de miles de ejemplos, especialmente si se complementan con transformaciones que conservan las etiquetas de las categorías a las que pertenecen, algunos ejemplos como en [6] muestran resultados comparables al desempeño de reconocimiento de un humano. Para el aprendizaje de miles de objetos a partir de millones de imágenes, es necesario un modelo con una gran capacidad de aprendizaje. En los últimos años, el área de aprendizaje profundo, un subcampo de aprendizaje de máquina, ha desarrollado nuevas arquitecturas o modelos de aprendizaje, basados en los principios de redes neuronales. Las Redes Neuronales Convolucionales (CNN, del inglés Convolutional Neural Networks), constituyen una clase de estos modelos de gran poder de aprendizaje [6], [7], [8]. Su capacidad se puede controlar variando su profundidad y amplitud, además de que hacen suposiciones sólidas y en su mayoría correctas sobre la naturaleza de las imágenes (es decir, la estacionariedad de las estadísticas y la ubicación de las dependencias de píxeles). Así, en comparación con las redes neuronales estándar avanzadas con capas de tamaño similar, las CNN tienen muchas menos conexiones y parámetros, por lo que son más fáciles de entrenar. Cuando se cuenta con un conjunto de datos pequeño, los modelos complejos de aprendizaje, como los CNN, es muy fácil que sobre entrenen los datos. Para solucionar este problema de entrenamiento de un clasificador de alta capacidad en conjuntos de datos pequeños, los trabajos relacionados en esta área han recurrido al uso de tareas de aprendizaje por transferencia, donde los pesos de la CNN se inicializan con aquellos de una red previamente entrenada, utilizando el conjunto de datos objetivo. Este enfoque ha logrado consistentemente mejores resultados, en comparación con el entrenamiento directo de la red en el conjunto de datos pequeño.

A. Reconocimiento de Expresiones Faciales (REF)

Existen identificadas principalmente dos metodologías para el reconocimiento de expresiones faciales. La primera se refiere a los métodos convencionales, en los cuales se extraen manualmente las características físicas que representan el rostro y entonces se utilizan algoritmos de decisión, principalmente SVM [9], [10]. La segunda son los métodos basados en aprendizaje profundo, en donde principalmente se incluyen CNN y RNN (redes neuronales recurrentes, del inglés Recurrent Neural Network). La principal ventaja de una CNN se refiere a que permite el aprendizaje a partir de las imágenes de entrada, es decir, no existe ningún preprocesamiento y elimina o reduce en gran parte la extracción manual de las características físicas en que se basan los métodos convencionales [11], [12]. Muchos de los trabajos que emplean aprendizaje profundo han adaptado una CNN directamente para la detección de los movimientos específicos de los músculos faciales, denominados unidades de acción [13], [14], [15], [16]. Debido a que los métodos basados en CNN no pueden reflejar variaciones temporales en las componentes faciales, un enfoque híbrido reciente presenta una combinación de una CNN para las características espaciales en los fotogramas en una secuencia de video, con una red LSTM (del inglés long short-term memory) para las características temporales [17], [18], [19]. Las ventajas de una red LSTM toman importancia cuando estas son empleadas en secuencias de imágenes (video). Algunos de los estudios representativos que utilizan una combinación de CNN y LSTM (RNN) son: [18], [19], [20], [14], [15], [16], [21], [22].

En los trabajos realizados para el reconocimiento de expresiones faciales, son consideradas secuencias de videos y conjuntos de individuos muy pequeños, cantidades de no más de 100 o 200 personas diferentes, además no son considerados los bebes, niños y ancianos en el modeloado de las expresiones faciales.

En la presente propuesta se emplea una CNN entrenada para diferentes tamaños de imágenes de entrada empleando un conjunto de rostros heterogéneos. Se emplean los diferentes modelos en la predicción de una secuencia de imágenes, la cual se obtiene por mayoría de la predicciones que proporcionan los modelos CNN en forma individual. Algunas ventajas y desventajas de las arquitecturas de aprendizaje profundo en el problema de REF se ilustran en la Tabla I.

TABLA I
Ventajas y desventajas de arquitecturas de aprendizaje profundo en REF

	CNN	RNN
Ventajas	Sin preprocesamiento en los datos de entrada; Simples de implementar; menos conexiones y parámetros.	Considera secuencias de video. Intenta descubrir implicitamente variaciones de los fotogramas en el tiempo. Conformación de información temporal en forma automática.
Desventajas	Falla en capturar dependencias contextuales, es decir, la relación que guardan diferentes regiones de la imagen [23]. No considera la información espacial en secuencias de video, las variaciones entre los fotogramas.	Requieren de mayores recursos para el entrenamiento. En ocasiones requieren de información generada por otras arquitecturas.

B. Métodos de Detección de Rostro

El rostro humano es uno de los objetos más importantes en una imagen o video. Durante los últimos años, el problema de detección de rostros ha recibido un enfoque importante debido a la variedad de sus aplicaciones en el comercio y la aplicación de la ley. Además, en los últimos años se han propuesto muchos métodos de reconocimiento de patrones y heurísticas para detectar el rostro humano en imágenes y videos [24]. La detección o reconocimiento de expresiones faciales en imágenes o video es un área de investigación activa. El análisis de expresiones faciales por una máquina es una

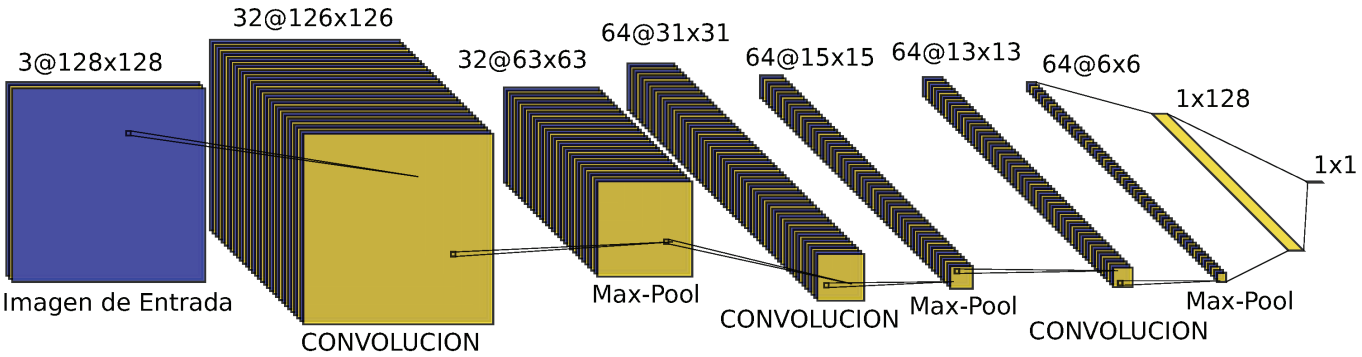

Fig. 1. Arquitectura de la CNN para la clasificación binaria.

tarea que enfrenta importantes retos pero con una amplia gama de aplicaciones. La detección de un rostro humano significa que para una imagen o video dado, determinar cuando en uno de estos se incluye una región de un rostro, en caso de si existir, determinar el número, la posición exacta y el tamaño de los rostros [25]. El desempeño de varias aplicaciones basadas en la identificación de rostros, recae en la exactitud y eficiencia de la técnica utilizada para la detección de estos [26]. Los rostros humanos son difíciles de modelar, debido a que es muy importante capturar todas las características considerando todas las variaciones de apariencia probables atribuibles a los cambios en la escala, ubicación, orientación, expresión facial, condiciones de iluminación y oclusiones parciales, por mencionar algunas. Existen dos enfoques principales para la detección de rostro: a) Basado en Características y b) basado en Imagen. En las técnicas basadas en características, los objetos regularmente son reconocidos por sus características únicas. Existen algunas características estructurales del rostro humano, las cuales pueden ser reconocidas entre un rostro y algunos otros objetos, tales como: los ojos, la nariz, la boca, las cejas, y algunos más, que identifican las técnicas basadas en características para la identificación del rostro humano. El problema con estos algoritmos es que las características son dañadas por situaciones de iluminación, oclusión o ruido en la imagen. El otro conjunto de técnicas basada en imagen considera técnicas de redes neuronales, enfoques estadísticos y subespacios lineales. Se basan principalmente en intentar obtener la mejor coincidencia entre las imágenes de entrenamiento y las imágenes de prueba. La librería de OpenCV para la detección de un rostro humano, tiene implementado el método publicado por [4], [5], el cual es considerado uno de los más utilizados por su velocidad y eficiencia, gracias a la discriminación de regiones que realiza, identificando aquellas en donde existe o no un rostro y procesando solamente las áreas que considera pertinentes.

III. Arquitectura de la Red Neuronal Profunda

Una forma común de redes neuronales profundas son las redes neuronales convolucionales, las cuales están compuestas de múltiples capas de convolución. En tales redes, cada una de las capas genera en forma sucesiva un nivel de abstracción superior de los datos de entrada, denominado mapa de características, en el cual se preserva información única y esencial de las imágenes [27].

En la actualidad, el diseño de arquitecturas de redes neuronales convolucionales requiere tanto pericia humana como trabajo práctico. Las nuevas arquitecturas se elaboran manualmente mediante una cuidadosa experiencia o se modifican a partir de un conjunto de redes existentes [28].

Para tareas de diseño, existen plataformas que facilitan la implementación de una red neuronal profunda. En el presente trabajo se utilizó el software libre Keras [29] para implementar el modelo que se propone, la arquitectura se muestra en la Fig. 1. Cuenta con 3 capas convolucionales en 2D y una capa completamente conectada de 128 neuronas, la salida es una sola neurona para la clasificación binaria, con una función de activación sigmoidea. Todas las capas convolucionales y la capa densa tienen como función de activación ReLU (Rectified Linear Units). El Max-Pool se fijó en todas las capas de convolución de dimensión 2×2. El tamaño de los filtros en todas las capas 2D fue de 3×3. Finalmente el valor dropout establecido respectivamente en cada una de las capas fue de: 0.20, 0.10, 0.10 y 0.5.

Para llegar a la arquitectura propuesta, se estimó el rendimiento del modelo empleando la métrica de precisión (E), definida por (1).

$$E = \frac{P_c}{N_{TP}} \tag{1}$$

En donde P_c es la cantidad de predicciones correctas del modelo y N_{PT} es el número total de predicciones. También fue considerado el comportamiento de la función de pérdida binaria entropía cruzada ($H_{y'}(y)$) [30], que para el caso de dos clases se establece por (2).

$$H_{y'}(y) = -\frac{1}{N}\sum_{i=1}^{N} y_i^{'} \cdot log(y_i) + (1 - y_i^{'}) \cdot log(1 - y_i) \tag{2}$$

En donde $y_i^{'}$ es el valor de la etiqueta para la clase a la que pertenece el elemento i, mientras que y_i corresponde a la predicción hecha por el modelo, el valor de probabilidad que se encuentra en la capa de salida en la arquitectura CNN, y N la cantidad total de elementos en el conjunto de datos.

En las pruebas que se realizaron, primero se estableció el tamaño de 64×64 píxeles para la capa de entrada. Se utilizaron 2, 3 y 4 capas de convolución sin el elemento de descarte (dropout o DO OFF en las Figuras 2 a 4) para investigar su efecto en la métrica de evaluación y la función de pérdida. El comportamiento de la precisión después de

las 50 épocas tiende a disminuir e incrementar su distancia ante el rendimiento que presentan los datos de entrenamiento, lo que nos lleva a inferir un sobre entrenamiento en las arquitecturas. En el aspecto de la función de pérdida se tiene un comportamiento contrario a partir del rango ente 25 a 50 épocas en el proceso de entrenamiento y lo esperado es un rendimiento semejante con los datos de entrenamiento y validación. Con estos resultados previos, se procedió a incluir la operación de descarte en las diferentes capas. La combinación de valores que se asignó entre las capas osciló en el rango de 0.1 a 0.5. Los mejores rendimientos permitieron elegir el modelo propuesto. Las gráficas en las Figuras 2 y 3 muestran ejemplos del comportamiento de la precisión para 3 y 4 capas convolucionales (indicado como 3C y 4C respectivamente en las gráficas). Se señala además la inclusión (DO ON) o no (DO OFF) de la operación de descarte. El rendimiento observado es incremental en las arquitecturas para el caso en que se asigna la operación de descarte. En la Fig. 4 también se incluye un ejemplo del comportamiento registrado por la función de pérdida por un modelo que incluye la operación de descarte, cuya tendencia es más próxima al rendimiento con los datos de entrenamiento. Para justificar la selección de la cantidad de filtros en las capas, se procedió de forma similar al proceso de elección del número de capas. Se realizaron pruebas con las cantidades 8, 16, 32 y 64, después de observar los mejores rendimientos en las evaluaciones se establecieron los valores para los filtros en cada capa. La Fig. 5 es un ejemplo del resultado obtenido para el valor establecido en la primer capa, se observa que el rendimiento es creciente hasta la cantidad de 32 filtros, mientras que en el entrenamiento del modelo con 64 filtros se disminuyó.

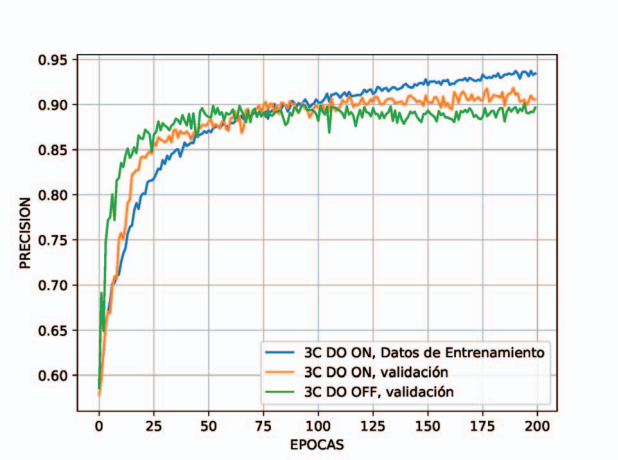

Fig. 2. Comportamiento de la precisión en el proceso de entrenamiento con 3 capas.

IV. Conjuntos de Datos

Un análisis basado en imágenes o videos en 2D tiene dificultades para manejar grandes variaciones de pose, sutiles comportamientos faciales, edad, género, dimensión y calidad en las imágenes. Para lograr un sistema de clasificación competitivo en ambientes reales, el conjunto de datos deberá ser demasiado hetereogeneo y un repositorio que nos proporciona una gran

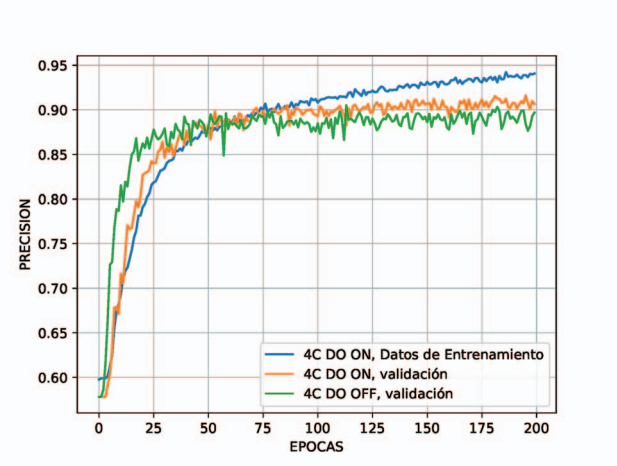

Fig. 3. Comportamiento de la precisión en el proceso de entrenamiento con 4 capas.

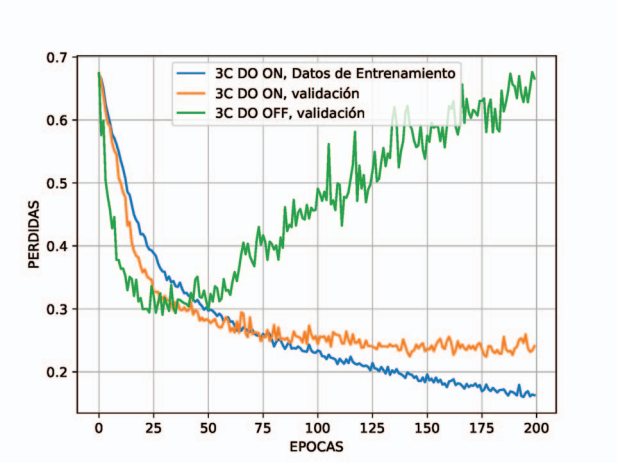

Fig. 4. Comportamiento de la función de pérdidas en el proceso de entrenamiento.

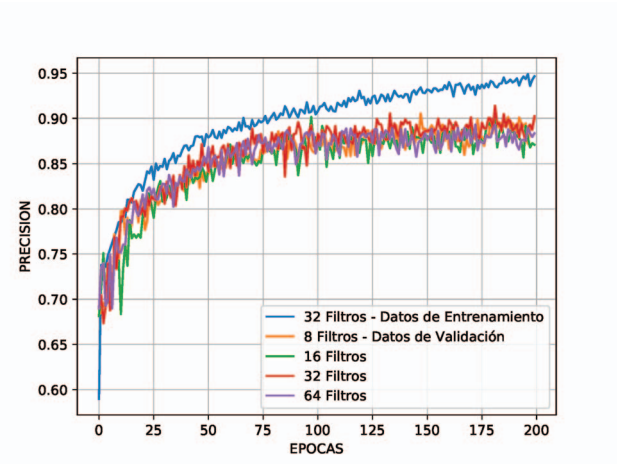

Fig. 5. Rendimiento del modelo para diferentes cantidades de filtros en la capa 1.

diversidad de imágenes es Flickr, que cuenta además con una interfaz para obtener grandes volúmenes de información, es un sitio web que permite almacenar, ordenar, buscar, vender

y compartir fotografías o videos en línea, a través de Internet. Los usuarios de esta plataforma comparten fotografías y videos creados por ellos mismos. Se procedió a crear el conjunto de datos utilizando el API de Flickr, obteniendo las imágenes que el API logró recuperar a las siguientes consultas: *happy person, happy friend, laugh kids, sad people, melancholy, cry younger, depressed person*. A continuación se obtuvieron las regiones identificadas como rostros, utilizando las librerías de OpenCV, finalmente se etiquetó cada uno de los rostros con alguno de los dos sentimientos: Positivo (Feliz) o Negativo (Triste). La Tabla II muestra la cantidad de imágenes que compone el conjunto de datos y la partición de los mismos que se utilizó para el entrenamiento de la arquitectura neuronal profunda: entrenamiento (80%) y validación (20%). El conjunto de datos considera una gran cantidad de poses, calidad de las imágenes, diversas edades de los rostros (bebés, niños, jóvenes y adultos), género, oclusiones, diversos fondos o ambientes en las imágenes, dimensión y otros. En la Fig. 6 se muestran algunos de los ejemplos de las imágenes con los rostros o regiones de interés que fueron obtenidas, y posterior al preprocesamiento de cada una de las imágenes obtenidas con el API de Flickr fueron etiquetadas en las clases Positivo y Negativo. El preprocesamiento consistió en identificar el rostro y solamente mantener la región de interés que corresponde al mismo.

TABLA II
TOTAL DE IMÁGENES EN EL CONJUNTO DE DATOS Y LAS CANTIDADES UTILIZADAS PARA EL ENTRENAMIENTO AL PARTICIONAR LOS DATOS

Sentimiento	Cantidad de Imágenes		
	Entrenamiento (80%)	Validación (20%)	Total
Positivo	1635	409	2044
Negativo	1750	437	2187

(a) Positivo

(b) Negativo

Fig. 6. Regiones de interés de las imágenes obtenidas por las consultas en Flickr y etiquetadas como: (a) Positivo y (b) Negativo.

FER 2013

Este conjunto de datos ha sido ampliamente utilizado en los diferentes trabajos de REF [31], [32], [33]. Fue creado utilizando el API de búsqueda de imágenes de Google, contiene 35887 imágenes en escala de grises, de dimensiones 48×48 píxeles, distribuidas en 7 categorías o expresiones. Se caracteriza además porque son imágenes heterogéneas, tomadas en condiciones naturales, lo que significa una problemática mayor al enfrentar oclusiones y variaciones de pose. Sólo se consideraron las imágenes de las categorías feliz y triste, para el entrenamiento se contó con 12045 imágenes, para la validación con 1548 y para prueba 1473 imágenes.

Para validar la eficiencia de la arquitectura propuesta y verificar que tanto fueron generalizados los datos, es necesario hacer pruebas con un conjunto de información que nunca se haya utilizado en el proceso de entrenamiento y validación del modelo. Para lograrlo, se obtuvieron adicionalmente 2 conjuntos de imágenes para prueba. El primero fue construido manualmente a partir de los resultados obtenidos en el motor de búsqueda de google.com, 110 rostros para el sentimiento Feliz y 133 para Triste, aplicando el mismo preprocesamiento de datos. El segundo conjunto se obtuvo de imágenes contenidas en el conjunto de datos Karolinska Directed Emotional Faces (KDEF) [34] para los sentimientos aquí evaluados, considerando únicamente los rostros de frente, que consiste en 280 imágenes, 140 para cada sentimiento.

V. PRUEBAS Y RESULTADOS

Esta sección reporta el proceso de entrenamiento de la arquitectura neuronal profunda que se ha propuesto y las pruebas realizadas para la identificación de sentimiento facial en los conjuntos de datos descritos previamente.

A. Entrenamiento de la Arquitectura

Se entrenó la red neuronal profunda para 4 diferentes tamaños de imagen de entrada, ver la Tabla III, en donde se muestra la mejor precisión obtenida en 200 épocas para los datos de validación, se procedió a almacenar el mejor modelo en cada caso durante este periodo de entrenamiento, adicionalmente se indican los tiempos de entrenamiento por cada época del modelo de red neuronal profunda que se obtuvieron. Las características del hardware utilizado tanto para el entrenamiento de los modelos como la implementación del sistema de la Fig. 9 son: Procesador Intel core i5 de 6th generación a 2.3 GHz, memoria RAM de 8GB, que se encuentran en equipos comunes y comerciales. Esto último sirve para enfatizar el poder de cómputo y tiempo que se emplea en el entrenamiento de una arquitectura neuronal en una prueba. A manera de ejemplo para el tamaño de imagen de 128×128 píxeles en 200 épocas se requirieron 83955 segundos en promedio para entrenar la red propuesta.

TABLA III
PRECISIÓN Y TIEMPO DE ENTRENAMIENTO DE LA ARQUITECTURA PROPUESTA PARA DIFERENTES DIMENSIONES DE LA CAPA DE ENTRADA

Tamaño de Imagen de entrada (ancho x alto)	Precisión en el conjunto de validación (%)	Tiempo de entrenamiento por época (segundos)
64×64	93.2	6
128×128	94.8	34
150×150	93.1	42
200×200	91.4	62

En la Fig. 7 se observa el comportamiento de la precisión durante el proceso de entrenamiento de la red neuronal durante una de las pruebas, se indican los resultados tanto para el

conjunto de entrenamiento como para el conjunto de validación, el tamaño de las imágenes de entrada es de 128×128 píxeles en este caso. Con esta gráfica es posible señalar que la generalización de los datos durante el proceso de aprendizaje es aceptable y no existe un sobre entrenamiento, es decir, el modelo obtendrá un desempeño al menos en el rango de los resultados de validación, esta aseveración se establece porque las curvas tienen una tendencia incremental a lo largo del tiempo de entrenamiento y se encuentran alineadas en forma paralela. La gráfica de la función de pérdida del modelo (mostrada en la Fig. 8), es una herramienta de apoyo que nos indica si el desempeño del modelo cuando se encuentre en producción será semejante al mostrado durante el proceso de entrenamiento. Para que ello se cumpla, es deseable que el comportamiento obtenido de las curvas de perdidas sea un comportamiento paralelo, en nuestra gráfica dicho comportamiento paralelo se mantiene hasta un poco antes de las 100 primeras épocas, posterior a ello los datos de validación presentan una tendencia constante.

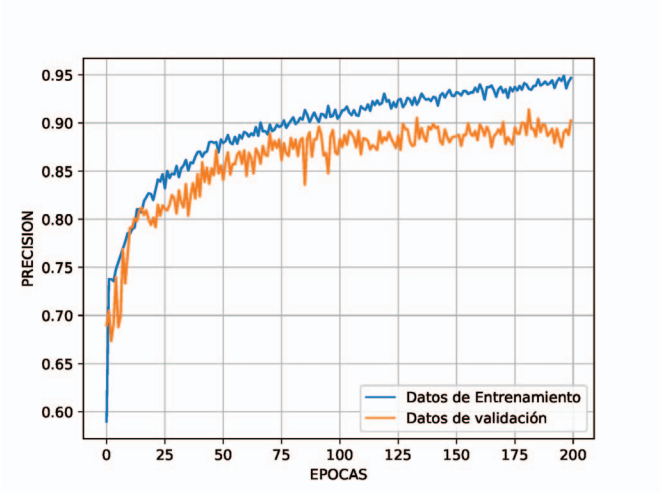

Fig. 7. Comportamiento de la precisión durante el proceso de entrenamiento del modelo de Red Neuronal Profunda propuesta.

B. Prueba de Desempeño

Para probar el desempeño de la arquitectura y realizar una comparación de los resultados, se utilizaron los dos conjuntos de datos que se crearon para prueba, datos nunca vistos por el clasificador. En la Tabla IV se indica la matriz de confusión y precisión obtenida a diferentes tamaños de la capa de entrada en la arquitectura o clasificador utilizando el conjunto de datos de prueba que se construyo manualmente de imágenes de google.com. El mejor resultado se obtuvo con las dimensiones de imagen para la capa de entrada de la arquitectura de 64×64 píxeles, alcanzando una precisión del 91.13%. Las diferencias de precisión respecto a las obtenidas con el conjunto de validación son en promedio del 3%. Se observa además que independientemente de las dimensiones de entrada se logra una precisión aproximada al 90% para este conjunto de datos.

Para fines de comparación de los resultados, se consideró el conjunto de datos de [34] para las expresiones Feliz y Triste.

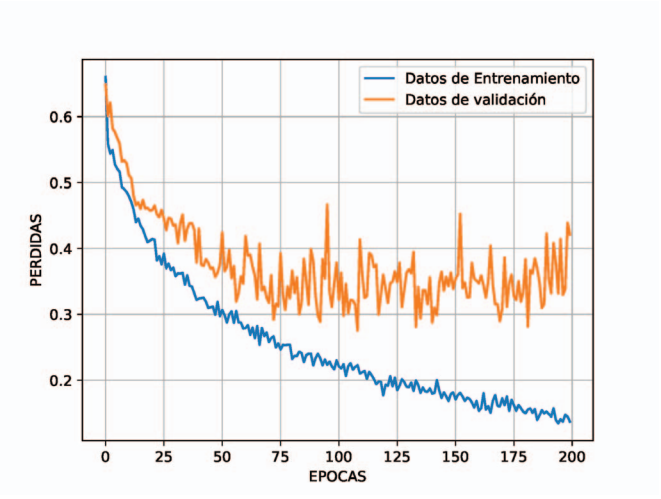

Fig. 8. Comportamiento de la función de pérdida del modelo durante el proceso de entrenamiento.

TABLA IV
RESULTADOS CON EL CONJUNTO DE DATOS CONSTRUIDO MANUALMENTE DE GOOGLE

Dimensión de Entrada	Matriz de Confusión				Precisión (%)
	Clase	Positiva	Negativa	Total	
64×64	Positiva	94	16	110	**91.13**
	Negativa	5	122	127	
128×128	Positiva	100	10	110	90.29
	Negativa	13	114	127	
150×150	Positiva	94	16	110	89.45
	Negativa	9	118	127	
200×200	Positiva	102	8	110	89.45
	Negativa	17	110	127	

En el trabajo de [35], el cual utiliza un clasificador SVM para 7 expresiones faciales, reporta los resultados obtenidos para las expresiones Feliz y Triste de 88.9% y 100% respectivamente (el promedio es del 94.45%), para el mismo conjunto de datos. En la Tabla V se señala la matriz de confusión obtenida para el conjunto de datos KDEF a diferentes dimensiones de la capa de entrada en la arquitectura.

TABLA V
RESULTADOS CON EL CONJUNTO DE DATOS KDEF

Dimensión de Entrada	Matriz de Confusión				Precisión (%)
	Clase	Positiva	Negativa	Total	
64×64	Positiva	133	7	140	96.07
	Negativa	4	136	140	
128×128	Positiva	135	5	140	96.07
	Negativa	6	134	140	
150×150	Positiva	135	5	140	95.71
	Negativa	7	133	140	
200×200	Positiva	130	10	140	**96.42**
	Negativa	0	140	140	

Los resultados aquí presentados son competitivos con el conjunto de datos KDEF, independientemente de la dimensión de entrada en la arquitectura. Es de señalarse que los datos KDEF tienen una complejidad menor, al no tener lentes, barba, bigote o aretes en los rostros, y además el rango de edad de

los participantes esta entre los 20 y 30 años, lo que justifica el desempeño mayor de nuestra arquitectura con este conjunto de datos respecto del otro.

C. Comparación de Resultados

En [9] se cita el trabajo de [36] en la categoría de técnicas para el REF basadas en Aprendizaje Profundo, y se reporta el desempeño más alto (77.90% de exactitud) en el reconocimiento de expresiones faciales utilizando solamente una imagen en el proceso de REF en 7 categorías. Para mostrar el desempeño de la red neuronal profunda propuesta y realizar la comparación de resultados bajo las mismas condiciones, se implementó la arquitectura de [36] para el caso de 2 categorías y se entrenaron ambos modelos para uno de los conjuntos de datos reportados en el mismo trabajo (FER 2013). Se consideraron las mismas condiciones de entrenamiento, estableciendo las dimensiones de la capa de entrada en 48×48 y 64×64 píxeles. El parámetro de tasa de aprendizaje va decreciendo en forma polinomial de la forma: $base\_lr(1-iter/max\_iter)^{0.5}$, señalando una tasa de aprendizaje base con valor $base\_lr = 0.01$, $iter$ es la iteración actual y max_iter es le máximo de iteraciones permitidas, se emplearon 200 épocas. Los resultados de desempeño en los distintos conjuntos de datos que se tienen para probar los modelos, se muestran en la Tabla VI, en donde se observa que la arquitectura propuesta es superior en dos de los tres conjuntos de datos (FER 2013 y Google). Para el caso del conjunto KDEF, nuestro modelo es competitivo respecto al resultado que presenta [36]. El incremento de desempeño de la arquitectura propuesta se atribuye principalmente al incremento en la cantidad de datos para el entrenamiento. Se observa además que el cambio en las dimensiones de la capa de entrada no afecta en el desempeño de la arquitectura.

TABLA VI
PRECISIÓN DE LAS ARQUITECTURAS ENTRENADAS CON LOS CONJUNTOS DE DATOS DE PRUEBA

Conjunto de Datos	Arquitectura	Dimensión de Entrada	
		48×48	64×64
FER 2013	Deeper CNN [36]	89.20	89.47
	Propuesta CNN	91.51	**91.92**
KDEF	Deeper CNN [36]	**99.64**	**99.64**
	Propuesta CNN	99.28	99.28
Google	Deeper CNN [36]	85.23	91.56
	Propuesta CNN	**92.40**	**92.40**

D. Sistema para Identificar la Expresión Facial

La implementación del sistema para determinar el sentimiento visual de una expresión facial que se propone, está constituida por los siguientes elementos: a) Captura de la imagen (camara); b) Detección y obtención del rostro (Visión por Computadora); c) Modelo de una Red Neuronal Profunda (Aprendizaje Profundo), véase la Fig. 9, que presenta el diagrama de bloques del sistema propuesto.

El bloque para la captura de la imagen se realiza a través de la interfaz con una cámara web embebida en el equipo. Para el bloque de identificación de rostro se utilizó el algoritmo de

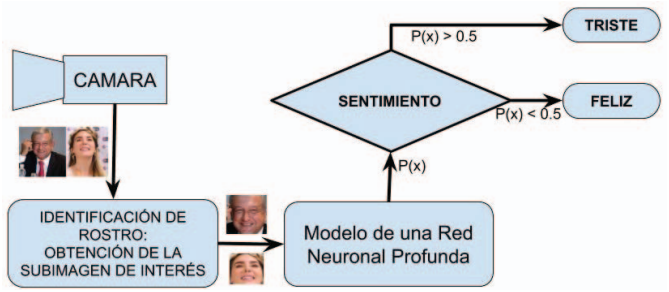

Fig. 9. Diagrama de flujo del sistema de análisis de sentimiento visual de rostros.

[4] implementado en OpenCV. Finalmente el bloque que lee el modelo de aprendizaje profundo fue implementado por las librerías Keras y Tensorflow.

En la integración de los elementos de visión por computadora y aprendizaje profundo para realizar la clasificación de sentimiento visual del rostro, se hicieron pruebas en forma independiente con cada uno de los modelos, ver la Tabla IV y VI, obteniendo la mejor respuesta los modelos que se entrenaron con el conjunto de datos FER 2013. Para un sistema más robusto, se propone introducir un clasificador por votación, es decir, introduciendo 3 de los modelos que fueron entrenados, de tal forma que la decisión final de clasificación se determina en base a la mayoría de las predicciones hechas por los modelos.

Para evaluar el sistema se utilizó el video "Faces from around the world" a la resolución de 1280×720 píxeles, cuenta con rostros heterogéneos que se han planteado en el presente trabajo. La cantidad de fotogramas analizados fueron 3312. El algoritmo de identificación de rostro consideró 1587 imágenes con información relevante, de las cuales 12 no corresponden a un rostro o este se encuentra incompleto, resultando en un total de 1575. Una persona manualmente etiquetó las imágenes que contienen un rostro en cada una de las clases y con esta información se obtuvo la matriz de confusión que se muestra en la Tabla VII. El valor de precisión que se obtiene es de 87.49%.

TABLA VII
MATRIZ DE CONFUSIÓN DE LAS PREDICCIONES REALIZADAS POR EL SISTEMA QUE IDENTIFICA LA EXPRESIÓN FACIAL

	Feliz	Triste	Total
Feliz	894	66	960
Triste	131	484	615
Total	1025	550	1575

VI. CONCLUSIONES

En el aprendizaje profundo se plantea una arquitectura neuronal profunda que con un conjunto de datos pequeño, considerando las cantidades de datos requeridas para el entrenamiento de las arquitecturas de la literatura (de decenas de miles de imágenes), ha presentado una buena respuesta, lo que nos ha permitido obtener una eficiencia del 91% en los datos de prueba, siendo competitivo respecto a trabajos previos citados. Se ha creado un conjunto de datos para el análisis de

sentimiento de rostros, el cual se encuentra disponible para futuros desarrollos, aplicaciones o trabajos de investigación. Se concluye además, que los nuevos desarrollos en software y su disponibilidad de código abierto, permite realizar en forma rápida y sencilla aplicaciones de IA. Se ha desarrollado un sistema que permite identificar el sentimiento expresado por una persona a través de medios visuales, clasificándolo entre feliz o triste. Al integrar la presente aplicación con otros sistemas, se podrá identificar la existencia de una persona en una escena y el estado de ánimo que guarda.

Como trabajo futuro se propone realizar el entrenamiento de un modelo de aprendizaje profundo que categorice una mayor cantidad de sentimientos, tales como enojado, disgustado, neutro, etcétera. Integrar en un robot de telepresencia o en sistemas de videovigilancia la identificación facial de sentimiento, ello ayudaría a determinar la forma en cómo tratar a las personas o tomar algunas reservas o decisiones de lo que se plantee en el actuar de las personas involucradas.

AGRADECIMIENTOS

Los autores agradecemos el apoyo de la Universidad Michoacana de San Nicolás de Hidalgo en el desarrollo de este trabajo.

REFERENCIAS

[1] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.

[2] X. Tan and B. Triggs, "Fusing gabor and lbp feature sets for kernel-based face recognition," in *International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 235–249, Springer, 2007.

[3] K. Yu, Z. Wang, L. Zhuo, J. Wang, Z. Chi, and D. Feng, "Learning realistic facial expressions from web images," *Pattern Recognition*, vol. 46, no. 8, pp. 2144–2155, 2013.

[4] P. Viola, M. Jones, *et al.*, "Robust real-time object detection," *International journal of computer vision*, vol. 4, no. 34-47, p. 4, 2001.

[5] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *Joint Pattern Recognition Symposium*, pp. 297–304, Springer, 2003.

[6] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," *arXiv preprint arXiv:1202.2745*, 2012.

[7] K. Jarrett, K. Kavukcuoglu, Y. LeCun, *et al.*, "What is the best multi-stage architecture for object recognition?," in *2009 IEEE 12th international conference on computer vision*, pp. 2146–2153, IEEE, 2009.

[8] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*, pp. 609–616, ACM, 2009.

[9] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, 2018.

[10] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 53–79, 2020.

[11] R. Walecki, O. Rudovic, V. Pavlovic, B. W. Schuller, and M. Pantic, "Deep structured learning for facial action unit intensity estimation," *CoRR*, vol. abs/1704.04481, 2017.

[12] I. Talegaonkar, K. Joshi, S. Valunj, R. Kohok, and A. Kulkarni, "Real time facial expression recognition using deep learning," *Available at SSRN 3421486*, 2019.

[13] B. Ko, E. Lee, and J. Nam, "Genetic algorithm based filter bank design for light convolutional neural network," *Advanced Science Letters*, vol. 22, no. 9, pp. 2310–2313, 2016.

[14] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[15] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 2983–2991, 2015.

[16] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3391–3399, 2016.

[17] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 443–449, ACM, 2015.

[18] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 467–474, ACM, 2015.

[19] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 223–236, 2017.

[20] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 25–32, IEEE, 2017.

[21] M. Sivaram, V. Porkodi, A. S. Mohammed, and V. Manikandan, "Detection of accurate facial detection using hybrid deep convolutional recurrent neural network.," *ICTACT Journal on Soft Computing*, vol. 9, no. 2, 2019.

[22] J. Yan, W. Zheng, Z. Cui, and P. Song, "A joint convolutional bidirectional lstm framework for facial expression recognition," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 4, pp. 1217–1220, 2018.

[23] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 18–26, 2015.

[24] R. Tsang, J. P. Magalhaes, and G. D. Cavalcanti, "Combined adaboost and gradientfaces for face detection under illumination problems," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2354–2358, IEEE, 2012.

[25] H. Hatem, Z. Beiji, and R. Majeed, "A survey of feature base methods for human face detection," *International Journal of Control and Automation*, vol. 8, no. 5, pp. 61–78, 2015.

[26] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790–799, 2014.

[27] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[28] B. Baker, O. Gupta, N. Naik, and R. Raskar, "Designing neural network architectures using reinforcement learning," *arXiv preprint arXiv:1611.02167*, 2016.

[29] F. Chollet, *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG, 2018.

[30] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, pp. 8778–8788, 2018.

[31] S. Li and W. Deng, "Deep emotion transfer network for cross-database facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3092–3099, IEEE, 2018.

[32] W. Wang, Q. Sun, T. Chen, C. Cao, Z. Zheng, G. Xu, H. Qiu, and Y. Fu, "A fine-grained facial expression database for end-to-end multi-pose facial expression recognition," *arXiv preprint arXiv:1907.10838*, 2019.

[33] S. Ramalingam and F. Garzia, "Facial expression recognition using transfer learning," in *2018 International Carnahan Conference on Security Technology (ICCST)*, pp. 1–5, IEEE, 2018.

[34] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces (kdef)," *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, vol. 91, p. 630, 1998.

[35] H. Alshamsi, H. Meng, and M. Li, "Real time facial expression recognition app development on mobile phones," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 1750–1755, IEEE, 2016.

[36] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE*

Winter conference on applications of computer vision (WACV), pp. 1–10, IEEE, 2016.

Moisés García Villanueva nació en Pátzcuaro Michoacán México, recibió el grado de Ingeniero Electricista y Maestro en Ingeniería Eléctrica con Opción en Sistemas Computacionales en la Facultad de Ingeniería Eléctrica de la Universidad Michoacana de San Nicolás de Hidalgo en 1999 y 2001 respectivamente. Actualmente es Profesor e Investigador de tiempo completo en la misma Facultad. Sus áreas de interés son el reconocimiento de patrones, visión por computadora, robótica móvil, minería de datos y clasificación de objetos a través de aprendizaje profundo.

Salvador Ramírez Zavala nació en Morelia, Mich. Recibió el grado de Ingeniero Electricista en la Universidad Michoacana de San Nicolás de Hidalgo, el grado de Maestro en Ingeniería Eléctrica en la misma Institución en 1990 y 1998 respectivamente. Actualmente es Profesor Investigador de tiempo completo de la misma Facultad. Sus áreas de interés son Electrónica de Potencia, Robótica, Control e Instrumentación.