Universidad Autónoma del Estado de Morelos

Instituto de Investigación en Ciencias Básicas y Aplicadas

Centro de Investigación en Ciencias

Reconocimiento de expresiones faciales usando redes de aprendizaje profundo

T E S I S
PARA OBTENER EL TÍTULO DE

LICENCIADO EN CIENCIAS

ÁREA TERMINAL CIENCIAS COMPUTACIONALES Y COMPUTACIÓN CIENTÍFICA

PRESENTA:

Fredy Marín Flores

DIRECTOR DE TESIS
Dr. Juan Manuel Rendón Mancha

CUERNAVACA, MORELOS 3 de abril de 2024

VOTOS DE APROBATORIOS

SECRETARIA EJECUTIVA
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS
UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

P R E S E N T E

Por medio del presente le informamos que después de revisar la versión escrita de la tesis que realizó el **C. MARIN FLORES FREDY** con número de matrícula **20164007730** cuyo título es:

"Reconocimiento de expresiones faciales usando redes de aprendizaje profundo"

Consideramos que **SI** reúne los méritos que son necesarios para continuar los trámites para obtener el título de **LICENCIADO EN CIENCIAS ÁREA TERMINAL CIENCIAS COMPUTACIONALES Y COMPUTACIÓN CIENTÍFICA.**

Cuernavaca, Mor a 26 de abril de 2024

Atentamente
Por una humanidad culta

Se adiciona página con la e-firma UAEM de los siguientes:

DRA. LORENA DÍAZ GONZÁLEZ	**(PRESIDENTE)**
DR. JORGE ALBERTO FUENTES PACHECO	**(SECRETARIO)**
DR. JUAN MANUEL RENDÓN MANCHA	**(VOCAL)**
DR. JORGE HERMOSILLO VALADEZ	**(SUPLENTE)**
DR. MAURICIO ROSALES RIVERA	**(SUPLENTE) NO PARTICIPA**

MIE/VRRC/eae

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

JUAN MANUEL RENDON MANCHA | Fecha:2024-04-26 16:22:17 | FIRMANTE

v8Hhfzl7OU8dkN8GRGxuqrAPWKCgRUT+zHxome+El6oAJGTlVb36sAgzjXZvNCj62oZ2rd73BLdgW1lw3vISrGM4awHQAi92/2RaIGPCUKinZowqQxmPIB2foajXCVPlg8APN35qk2d1osJcJdoy51YMVVLrFgmGbdKZ6flt2NwLN27jlfCo7S+j1nvtg4p9qpDd90zs2QFZRnWmOy8YFcuVaBfINU2nnsmmD5W5vILJZVU+wbKSPv/U3pl7OdzO9mhVRh+zLvQZGnd45MPZs/CnLOzd9cQDOQpOyxbmRiFheHp2tUpAgtQqsCS8J6nlpvEtqf/aDopYqGaJLE0xoA==

JORGE HERMOSILLO VALADEZ | Fecha:2024-04-27 11:10:30 | FIRMANTE

Vpv/y4Yd+TKFlbmnw9D/OQ6gGYPMk3de8gzgvnjVqy8OSniyNT6c+Iklf0w31+t3ce34DnJWi3HxMFrJZ9xBCGK/f0PLA7TlmA+bfKDLaO/L4g4E1bX0eKqAzIEfBMsyW87I8OUsopn8zshxG89r7xvsfvKMfX0wztaY8mp0HSU1TgDnn6VIQJ7yoxSg7/8gHYuCQ21SPSJpcStOS/rj3r7eLTlzGmL/i1aBY4s50+cbbqJH91ZmMwrYc1my7OtVoC2taeus5KvW7b0SuFqJv+VylWq7s+yW4lwMG3h2ifZsQtyH9NNx1SKw2qsf7nQHP+7zbESRhMLZVjsdb33SVQ==

JORGE ALBERTO FUENTES PACHECO | Fecha:2024-04-27 11:27:12 | FIRMANTE

YDlZpRmBhq0ESNdo6jSaE9d97b2FzPu9ZeKJdhX57Bmg0ze1Ko9iIEOeQDo+oOMM1oIaI+Gl4l+sjn1oDDlWQp249M87GsGmAZum9m0qAgNhtrMVdz9I/bAT1WrzQwGItN1QCi9yV9y8laSkvVjbf05KO+z1d2qumcpNkrI1B81slqzm2y5jOcJGuMNBd/Keaz7qFqAycXuSQojgyeeEWGm9fZzQwiu4e2lprlXlRKMN9Bx2gfK3EupWnWwaBFRXo3mqBboU4j+0rjAMYwpwJFOQby7QDUVen0vxbhjk/0GUlS6GCDUD1wXKUK73cR6tppzbJ+O/TX4MhFjZtHahVA==

LORENA DIAZ GONZALEZ | Fecha:2024-04-29 10:25:38 | FIRMANTE

oUcpOZr1+5lyP8cvpr5bjjpSbbSLBhgFPvHLEnJ8VRfKT/bI3FD7Ay62MOu4zH7xudyo+xwWSpbvGefWceumIDfdC+xubU8PeTcU3i5vXiBxegtSbBs/0ydtRc4bC0MFNTxSxxawf/ly1s09HWc22+jHYAeoeWthxjpPsxGg6/frRiJ9EHJZnJAbZL7+FBKY2tWAwzko7tFf7mziCf21X927HCw1ZwNjH2UM0rG/3tjXP+vrqmV8A8mh5jDHdKnA8ifpgDMDTMOJTLAYxwfYyPaV9uB4GcGnXBU3XFpiO9tyrMZ1IUU/zQAuezpV1LVti8IU2XpI+ThaQELng5dqHg==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

64fizO90A

https://efirma.uaem.mx/noRepudio/EaTLd33Oh77TWVj73pvk82wiDdSlfIue

RESUMEN

Reconocimiento de expresiones faciales usando redes de aprendizaje profundo

Fredy Marín Flores

El reconocimiento de expresiones faciales se enfoca en la detección de las emociones que son expresadas a través del rostro tales como enojo, felicidad, miedo, entre otras. Para su correcta detección se han desarrollado múltiples modelos de redes, basadas en aprendizaje profundo que brindan soluciones aceptables en términos de tiempo de entrenamiento y precisión. En este trabajo de tesis tomaremos como base el trabajo realizado por [45]. Él demostró que, aplicando un paso de pre-procesamiendo adicional de recorte a las imágenes la red fue capaz de alcanzar resultados aceptables en términos de precisión y tiempo de entrenamiento.

En el área del procesamiento de imágenes para el reconocimiento facial se disponen de múltiples conjuntos de datos que contienen variaciones de luz en las imágenese águlos del rosto, en este trabajo se basa en en el conjunto de datos de CK+[59] debido a su fácil acceso y al número de expresiones que se añaden en este conjunto de datos además de ser el conjunto utilizado por [45].

La precisión alcanzada de los modelos de red para el reconocimiento facial han logrado resultados muy superiores, sin embargo, estos resultados se han obtenido al entrenar los modelos con imágenes de laboratorio mientras que el deterioro en los resultados se da en conjuntos con imágenes del mundo real. Durante los experimentos en este trabajo los resultados de precisión obtenidos del modelo muestran que al incrementar la región facial puede permitir a la red un aprendizaje superior a travéz de la selección de caractéristicas de las expresiones faciales que brinda este incremento.

Índice general

CAPÍTULO

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Las expresiones faciales son de las características más importantes del reconocimiento de las emociones humanas que fueron introducidas por Darwin en su libro titulado "The Expresion of the Emotions in Man and Animals" [12]. De acuerdo a [74] las expresiones pueden ser definidas como los cambios faciales en respuesta al estado emocional, las intenciones o la comunicación social de una o más personas. Actualmente, los sistemas de detección de expresiones faciales han tenido una amplia variedad de aplicaciones. En años recientes, con el surgimiento de las redes neuronales (NN), la detección de expresiones ha sido utilizada en múltiples tareas de Machine Learning (ML), sus aplicaciones son numerosas tanto en el tratamiento de imágenes como en la visión por computadora, en tareas de clasificación y segmentación, por mencionar algunas.

Los trabajos realizados sobre el reconocimiento de expresiones se han intensificado en los últimos años. Los modelos tradicionales eran desarrollados para tareas específicas usando imágenes estáticas o secuencias de video usando técnicas tradicionales de procesamiento de imágenes como el suavizado, las transformaciones geométricas, la extracción manual de características y una extensa lista.

Por otro lado, las tareas de procesamiento de imágenes han logrando avances significativos en múltiples áreas de aplicación como la astronomía, róbotica, medicina y seguridad además de otros sistemas de detección. Woody Bledsoe, Helen Chan Wolf, and Charles Bisson en la década de los 60's realizaron parte de los primeros trabajos en este campo. Recientemente han surgido numerosas propuestas para desarrollar modelos de Inteligencia Artifical (AI) eficientes que brinden los mejores resultados, así como la aplicación de distintas técnicas las cuales involucran el reconocimiento de patrones, descripción de contornos, de características, de región, mapeo, por mencionar algunas.

Dado el surgimiento de las Redes Neuronales de Convolución (CNN), un campo nuevo en ML que ha brindado resultados favorables en aquellas

tareas en las que no se tenían buenos resultados a través de los modelos tradicionales. Las CNNs surgen como una mejora a las redes convencionales, las diferencias entre ellas son considerables en distintos aspectos, pero mantienen usos comunes. El estudio de las expresiones faciales bajo el enfoque de las redes de aprendizaje profundo (Deep Leaning, DL) ha mostrado tener la capacidad de poder modelar y caracterizar en gran medida las partes que definen un objeto. En cuanto a la detección de rostros y expresiones faciales, ofrecen una ventaja frente a los enfoques tradicionales de AI, tales como HOG (2.1.4) , LBP (2.1.1), LGC (2.1.2) [42].

Durante los primeros años de desarrollo del ML, los métodos desarrollados ofrecieron resultados aceptables en distintas áreas, sus aplicaciones lograron mejorar multitud de sistemas que operaban de forma ineficiente. Sin embargo, en comparación con los modelos actuales de DL, las redes convencionales solo eran capaces de resolver tareas sencillas mientras que los enfoques de DL son capaces de lidiar con problemas complejos, además, brindan un resultado excepcional que puede llegar a superar en precisión a un experto humano. Una particularidad de los modelos tradicionales es la extracción manual de características, además, estos modelos tienden a consumir demasiado tiempo en procesamiento, entrenamiento y un alto uso de memoria (recursos). Es frecuente el uso de modelos pre-entrenados así como una correcta selección de hiperparamétros que es también una parte vital en casi todos los modelos.

1.1. Problemática

El problema del reconocimiento de expresiones faciales (FER) consiste en determinar en un rostro (imagen) una de las 6 emociones universales de acuerdo a [17]: angustia, disgusto, miedo, felicidad, tristeza y sorpresa. Algunos sistemas también reconocen la expresión neutral, es decir, donde no hay una emoción reflejada. Estas emociones son expresadas a través de componentes conductuales como las expresiones faciales, así mismo, pueden ser categorizadas en expresiones intencionales y espontáneas, siendo estas últimas complicadas de detectar debido a que suelen estar o presentarse en una fracción de segundo.

Las expresiones faciales son catalogadas mediante Unidades de Acción utilizadas por primera vez en [17] y definidas por [28] que describe una UA como una contracción o la relajación de uno o más músculos basado en la Codificación Facial que es un sistema para denominar movimientos faciales humanos por su apariencia en la cara, basado en un sistema desarrollado originalmente por un anatomista [28].

Los sistemas de reconocimiento facial constan de tres pasos esenciales de acuerdo a [74]: detección facial, extracción y representación de características faciales y reconocimiento de la expresión. Estos sistemas se dividen además en dos principales enfoques: métodos basados en imágenes estáticas y aquellos que utilizan secuencias de imágenes. Los enfoques basados en imágenes estáticas no utilizan información temporal, por otro lado, los métodos basados en secuencias utilizan la información temporal de las imágenes (o el video) para reconocer la expresión capturada de uno o más cuadros.

En el reconocimiento de expresiones con DL, se debe seleccionar un conjunto de datos capaz de aportar la información necesaria para una correcta predicción. Generalmente estos conjuntos son etiquetados y el número de ejemplos es bajo. Las condiciones bajo las cuales son obtenidos estos conjuntos es un factor importante, usualmente cada una de las expresiones son tomadas de una secuencia corta de video que inicia con una expresión neutra y avanza hasta una expresión pico (peak), como suele llamarse; los últimos frames expresan en su totalidad la emoción.

Existen algunos otros elementos que afectan los resultados de predicción en los modelos como accesorios en el rostro, lentes y barba por mencionar

algunos. Por otro lado, una mala elección de características así como algunos otros factores como el ángulo del rostro en la imagen, luz, oclusión o algún efecto como la distorsión o baja calidad en la imagen causan un bajo rendimiento en los modelos.

En este trabajo se aborda el uso de las CNNs para el tratamiento de imágenes y está enfocado en la detección de expresiones faciales. Aquí, la principal tarea será detectar el tipo de emoción que se está expresando en un rostro.

1.2. Justificación

Decidimos implementar en este trabajo de tesis el método de [45] porque es un método reciente, con buenos resultados y fácil de implementar ya que la red neuronal utilizada es relativamente simple.

1.3. Objetivo General

1. Implementar el método de reconocimiento de expresiones de [45] y reproducir los resultados reportados.

1.4. Objetivos particulares

1. Implementar el método de [45] y comprobar si se obtienen los resultados reportados en su artículo.
2. Proponer una modificación al método de recorte de [45] para mejorar el desempeño del método de reconocimiento de expresiones.

1.5. Metodología básica

La metodología usada para este trabajo consiste en la implementación del método propuesto por [45], el cual propone un paso de pre-procesamiento necesario antes de realizar el entrenamiento de la red.

Capítulo 2

Antecedentes

A mediados del siglo XX surgió la teoría de *feedback facial* [80], en la que se postula la vinculación entre los movimientos musculares de la cara y la generación de las emociones, básicamente relaciona de manera directa las emociones y las expresiones faciales, además se determina que los estados de ánimo pueden ser transmitidos en forma directa o indirecta a otros individuos mediante estas expresiones. Los movimientos musculares que ocurren al expresar una emoción son los responsables de iniciar una experiencia emocial, o sea que según esta hipótesis los movimientos y gestos de la cara pueden evocar reacciones afectivas o no en el sujeto que las realiza, en otras palabras, las emociones están ligadas a las expresiones faciales y pueden ser transmitidas: un estado de felicidad provocará una sonrisa, un estado de tristeza mostrará un rosto desanimado.

Más tarde en 1976 [17] formula el planteamiendo sobre la universalidad de las emociones básicas: sorpresa, tristeza, desprecio, miedo, ira, alegría y asco. También, explica que otras emociones o sensaciones llamadas expresiones secundarias (vergüenza, empatía, confianza, entre otras), pueden ser generadas en combinación con estas primeras expresiones.

Por otro lado, con el surgimiento de uno los primeros modelos de red neuronal artificial propuesto por Warren McCulloch y Walter Pitts (1944) [19] a finales del siglo XX, comienzan a surgir modelos predictivos basados en este primer acercamiento a las redes, mismos que comienzan a dar los primeros resultados poco atractivos debido a las limitaciones computacionales en esta temprana etapa. Sin embargo, en 2003 [21] proponen un modelo de red jerárquica para el reconocimiento visual de patrones. Este modelo propone un método para el reconocimiento de números escritos a mano como una versión mejorada del neocognitrón [22] donde el porcentaje de reconocimiento fue del 98.5 %.

En 1999 [61], usaron representaciones 2D de la Wavelet de Gabor para el reconocimiento de expresiones faciales, además, el modelo podía predecir

el género y raza con excelentes resultados.

En el año 2004 [70], usaron imágenes de personas con vista frontal y lateral con un modelo de red neuronal. Los resultados fueron los esperados de acuerdo al trabajo, sin embargo, el modelo presentaba deficiencias al trabajar con imágenes de sujetos usando accesorios en el rostro, esto provocó que los resultados en la evaluación fueran de tan solo 86.3 %.

Años más tarde, en 2009 [83], proponen usar más elementos e información del contexto e involucran otras ramas como la psicología y la lingüística. En otras palabras, se aborda la idea de añadir información del contexto como el tema de conversación y así aumentar la capacidad del modelo para predecir la expresión con ayuda de esta información extra. La idea principal de este trabajo es tomar el reconocimiento de expresiones desde una perspectiva psicológica.

En 2011 [86], proponen un modelo usando secuencias de imágenes con operadores espacio-temporales para la descripción de expresiones faciales además de LBP-TOP (Local binary patterns from three orthogonal planes) como descriptor de características. Los rasgos faciales basados en componentes se presentan para combinar información geométrica y de apariencia proporcionando una forma efectiva de representar las expresiones faciales. Los resultados exprerimentales demuestran que su uso en conjunto con Máquinas de Vectores de Soporte (SVM por su siglas en inglés) brindan excelentes resultados y manejan adecuadamente las variaciones de iluminación.

Según [55], los sistemas de reconocimiento facial constan de un proceso de entrenamiento en tres etapas: aprendizaje de características, selección de características y construcción de un clasificador, en este orden. La etapa de aprendizaje es responsable de la extracción de todas las características relacionadas a la expresión facial. La selección de características obtiene las mejores características para representar la expresión. Finalmente, un clasificador o más son usados para inferir la expresión facial dada esta selección anterior de características.

Recientemente, con la llegada de las Redes Neuronales (NN) y el Aprendizaje Profundo (DL), el uso de estas últimas redes en combinación con otros modelos de red como el propuesto por [37] en 2015, muestran un enfoque híbrido RNN-CNN (RNN: redes neuronales recurrentes) para propagar información sobre una secuencia de imágenes utilizando una respresentación

de capa oculta de valor continuo. Con este modelo los autores proveen una arquitectura para las tareas de reconocimiento facial que puede superar a un enfoque de solo redes neuronales de convolución (CNN).

En 2015 [35] usaron dos tipos de CNN, el primer modelo de red extrae características de apariencia temporal de las secuencias de imágenes, el segundo modelo extrae características de geometría temporal de puntos de referencia temporales. Mediante esta combinación de modelos lograron mejorar el rendimiento del reconocimiento de expresiones faciales.

[88] propusieron el aprendizaje de múltiples etiquetas y región profunda (Deep Region and Multi-label Learning, DMRL) para la detección de unidades de acción (UAs) [79] siendo un modelo de red pronfunda unificada permitiendo que dos tareas aparentemente no relacionadas puedan contribuir entre ambas, el aprendizaje de región (RL) [15] y el aprendizaje de múltiples etiquetas (ML) y métodos como AdaBoost [50], GentleBoost [33] o SVMs [59], puedan interactuar directamente. Compararon su enfoque frente a uno similar: Joint Patch and Multi-label Learning (JPML) [89], el modelo propuesto superó al otro por encima de 5 % en sus resultados.

2.1. Enfoques convencionales de FER

Los modelos para el reconocimiento de expresiones desarrollados a finales y principios del siglo XX fueron enfoques convencionales de visión por computadora que comparten una notable cualidad, y es que son altamente dependientes de la extracción de características de manera manual. Algunas de las principales técnicas de extracción de características para la categorización de expresiones faciales son los modelos basados en la geometría y apariencia, unidades de acción individual/grupo de músculos y no basados en UAs, así como los descriptores de características que proporcionan la información que define a un objeto.

Algunos de los principales métodos para la descripción de características que han sido utilizados ampliamente y que aún están presentes en trabajos recientes son:

2.1.1. Local binary pattern (LBP)

Originalmente fue propuesto para el análisis de texturas en 1990 por [14], y usado más tarde por [67]; además, se ha determinado que cuando se combina con el descriptor histograma de gradientes orientados mejora considerablemente el resultado de detección. Es una técnica descriptiva simple pero altamente efectiva para la clasificación de objetos filtrando píxeles adyacentes. Esta técnica codifica la relación de la intensidad del píxel central con la intensidad de los píxeles adyacentes a este asignando una etiqueta a cada píxel en una P-vecidad de píxeles (P valor del píxel igualmente espaciado con un radio R y denotado por G_p) al establecer el umbral de sus valores con el valor central (g_c) y convertir estos valores de umbral en un número decimal dado por la ecuación:

$$LBP_{p,R}(X_C, Y_C) = \sum_{p=0}^{p-1} s(g_p - g_c) \qquad donde,\ s(x) = \begin{cases} 1, & x \geqslant 0 \\ 0 & e.o.c \end{cases}$$

Debido a su elevada capacidad discriminatoria, constituye una aproximación usual para la solución de multitud de problemas además de ser un método altamente robusto.

2.1.2. Scale invariant feature transform (SIFT)

Este descriptor fue propuesto por [58], es utilizado para una gran cantidad de propósitos en visión por computadora. Este descriptor fue propuesto con la finalidad del análisis de imágenes con percepción 3D y reconocimiento de objetos basado en la vista. Sin embargo, también ha sido aplicado a imágenes en escala de grises y de color. El descriptor SIFT es invariable a las traslaciones, rotaciones y transformaciones de escala en el dominio de la imagen. En la práctica, se ha demostrado que es muy útil para la comparación de imágenes y el reconocimiento de objetos en situaciones del mundo real.

2.1.3. Local directional pattern (LDP)

Este método propuesto por [31], es un descriptor facial robusto para la representación de expresiones faciales.

En este método, se convolucionan ocho máscaras de Kirsh [40] de tamaño 3×3 con regiones de la imagen de la misma dimensión para obtener un conjunto de 8 valores de máscara. Estos valores de máscara se clasifican

y los tres primeros se asignarán con uno en el código binario de 8 bits y los otros con cero, el valor decimal correspondiente a este código binario se asignará como el valor del píxel central de la región de 3×3 tomada. LDP genera una imagen la cual es dividida en bloques definidos que se concatenan para formar el descriptor final.

2.1.4. Histogram of gradient orientations (HOG)

El histograma de gradiente orientado es un descriptor de características invariante a cambios de iluminación. HOG se encuentra mediante la consideración de magnitud/píxel para cada píxel en una imagen. Primero, se calculan los gradientes X y Y de la imagen aplicando filtros ($Gx = [-1, 0, 1]$, $Gy = [-1, 0, 1]^T$) para las direcciones horizontal y vertical. La orientación de los ángulos van en un rango de $0 - 180$ grados o $0 - 360$ grados, con signo y sin signo respectivamente. Este proceso nos dará dos nuevas matrices: una almacena los gradientes en la dirección x y la otra en la dirección y. El segundo paso es calcular la magnitud y orientación usando estos valores y crear un histograma de celdas, el valor de la orientación calculada se coloca en el rango del ángulo que le corresponde para denotar la frecuencia que tiene cada píxel en la imagen, a este proceso se conoce como conteo por votos, otras opciones para el peso de los votos podrían incluir la raíz cuadrada o el cuadrado de la magnitud del gradiente. El tercer paso es tener en cuenta el cambio en la iluminación y el contraste, las intensidades del gradiente deben normalizarse localmente, lo que lleva a agrupar celdas en bloques más grandes.

En [4] propusieron un modelo híbrido de red neuronal convolucional con agregador de transformación invariante a escala densa (SIFT). El método involucra una red neuronal convolucional y SIFT para la extracción de características para el reconocimiento de expresiones faciales, este modelo logró un porcentaje de reconocimiento del 99.1 % usando el conjunto de CK+[59] mientras que el porcentaje decreció hasta 73.4 % en precisión al usar otros conjuntos de datos.

2.2. Enfoques basados en Deep Learning para tareas de FER

En los últimos años, los enfoques de DL han sido utilizados para la extracción automática de características (dejando de lado los descriptores antes mencionados), tareas de clasificación y tareas de reconocimiento. Debido a esto, su uso se ha extendido en múltiples campos incluyendo la detección de objetos, el reconocimiento facial, reconocimiento de patrones y FER. Existen distintas arquitecturas de red profunda que han sido ampliamente utilizadas y algunas otras resultantes de la modificación de estos modelos base, algunas de estas son:

2.2.1. CNN

En los trabajos realizados por [6] usaron técnicas de visualización de CNN para obtener un modelo aprendiendo conjuntos de datos de FER y demostraron la capacidad de las redes para aprender tareas de detección de emociones y otras tareas relacionadas con FER. [35] usaron un modelo de CNN de dos etapas, la primera etapa extrae características de apariencia temporal de puntos faciales de referencia y la segunda etapa extrae características geométricas de puntos de referencia faciales temporales. El uso combinado de estas redes ofrecen una mejora de rendimiento en el reconocimiento de expresiones faciales.
Otro enfoque destacable es el propuesto por [90], un tipo de red unificada que usa una capa con funciones de feed-forward para indicar regiones faciales importantes en la imagen. Esto obliga a los pesos aprendidos a capturar información sobre la estructura del rostro. Esto ayuda a obtener representaciones más robustas sobre variaciones dentro de una región local como lo es boca, ojos, nariz, por mencionar algunas de estas regiones.

Estos modelos de red basados en CNN presentan excelentes resultados, sin embargo, aún quedan limitantes, por ejemplo, estos métodos no pueden manejar variaciones temporales en los componentes faciales de una imagen, no logran capturar las caraterísticas espaciales en una secuencia de imágenes (frames). Para hacer frente a estos desafíos, se desarrolló un modelo llamado Long short-term memory (LSTM), este modelo contiene conexiones de retroalimentación que no solo son capaces de procesar puntos de datos individuales (imágenes), sino también sencuencias completas como voz y video lo que las hace excelentes para la clasificación de videos así como de música.

Principalmente es un modelo de red que funciona bastante bien con datos que se mueven o están expresados en el tiempo.

En 2017 [10] propusieron un modelo híbrido para la detección multinivel de unidades de acción (AU) [59] combinando características espaciotemporales, este modelo consiste en dos etapas, la primera, extraía representaciones espaciales usando un modelo de CNN, la segunda, consistía en un modelo de LSTM para extraer las características espaciales contenidas.

En [57] propusieron un modelo de CNN para el reconocimiento de expresiones faciales, los resultados de experimentación llegaron a 96.76 % con seis clases, usaron el conjunto de CK+[59], JAFFE [60] y BU-3DFE [49] siendo este último uno de los conjuntos de datos más desafiantes para tareas de FER debido a la multitud de poses en las que se encuentran los rostros, incluyendo expresiones espontáneas y variaciones de luz por mencionar algunos. El método aplica distintas técnicas de pre-procesamiento así como recorte de imágenes, aumento de datos y rotaciones, este paso demuestra que es una parte importante en el modelo, así mismo, el modelo resulta ser fácil de implementar. Además, durante las pruebas de entrenamiento resulta ser bastante rápido. El modelo fue probado en un entorno real con excelentes resultados.

En [82] utilizaron un modelo de CNN con codificadores automáticos profundos y dispersos (deep sparse autoencoders, DSAE) logrando un porcentaje de reconocimiento de 95.79 % usando el conjunto de CK+[59] y utilizando características geométricas y de apariencia espacial. El enfoque demuestra que la información de apariencia espacial contiene la informacioń precisa y completa de las emociones. El DSAE es utilizado para reconocer las expresiones faciales con alta precisión mediante el aprendizaje de características sólidas y discriminatorias de los datos.

En [72] utilizaron un modelo de CNN con un clasificador de bosque aleatorio para tareas de reconocimiento, alcanzando un porcentaje de reconocimiento del 96.38 %. Este enfoque además utiliza unidades de acción (UAs [17]) para el reconocimiento de las expresiones mediante validación cruzada (twofold). Una de las partes importantes es que realizan una medición de las expresiones, esto lo hacen mediante el seguimiento de puntos de características faciales del modelo de apariencia activa (AAM [25]), además utilizan un rastreador óptico Lucas-Kanade (LK [2]) y mediante la estimación de los movimientos de los puntos de características (vectores) logran

estimar la expresión que está marcada en el rostro. Posteriormente, estas características se transforman a un bosque aleatorio para determinar las UAs, tras cada nivel en el bosque aleatorio se logra la clasificación de la expresión. El método propuesto logra un rendimiento mayor en comparación con otros enfoques basados en UAs.

2.2.2. RNN

Red Neruronal Recurrente, esta es una red que captura información temporal por lo que la vuelve un modelo más adecuado para la predicción de datos secuenciales, esto permite trabajar con secuencias de videos donde las expresiones son mostradas continuamente.

En el trabajo realizado por [26] usaron un modelo de red recurrente (RNN) que considera las dependencias temporales que están presentes en una imagen. Durante la fase de pruebas usaron dos tipos de LSTM (LSTM bidireccional y LSTM unidireccional), el estudio comprobó que la red bidireccional proveía un rendimiento mejor que una LSTM unidireccional.

Otros trabajos realizados consideran secuencias de video para la detección de expresiones como en [18] que proponen un modelo híbrido de CNN-RNN que usa convoluciones 3D. La combinación de estos modelos codifican la información de movimiento y apariencia de distintas formas. En concreto, la RNN toma características de apariencia extraídas por la CNN sobre cuadros de video individuales como entrada, mientras que la CNN 3D modela la apariencia y el movimiento del video simultáneamente.

En [66] propusieron un modelo de red recurrente basándose en pequeñas secuencias de videos, el modelo parte de tres componentes principales: 1) Detección facial donde se enfocan principalmente en las regiones de ojos y boca. 2) Extracción de características basadas en geometría. Durante este paso utilizan los puntos de referencia (landmarks) como discriminante de cada una de las emociones. Las secuencias fueron tomadas de la base de datos de BioVid Emo. Por último, 3) para el reconocimiento de expresiones utilizan un modelo híbrido de LSTM-CNN para clasificación alcanzando un porcentaje de reconocimiento del 81 % siendo un resultado favorable debido a la dificultad de capturar las expresiones provenientes de secuencias de videos.

2.2.3. CNN-LSTM

La unión de estos dos modelos proporciona un modelo de red más robusto así como una mejora sustancial frente a un único enfoque. Como resultado de esto, existen modelos híbridos que combinan o trabajan en conjunto con múltiples redes especializadas en una única tarea.

Las LSTM son un caso particular de las CNN, este enfoque fue usado por [73], el trabajo realizado se basa en la detección de micro expresiones faciales. Este método supone la detección de expresiones faciales espontáneas, las cuales, solo duran alguna fracción de segundo en el rostro, lo que las hace una tarea díficil de llevar a cabo ya que estas expresiones pueden no aportar la información necesaria sobre la emoción que se está expresando. El enfoque propuesto se basa en dos modelos de CNN 3D para la extracción de información espacial y temporal, simultáneamente se aplica una operación de convolución 3D sobre la secuencia de video.

En [38] usaron un modelo LSTM para la representación de características aprendidas para el reconocimiento de expresiones faciales. Estas características espaciales en las imágenes son aprendidas a través de una CNN, mientras que las características temporales de la representación de las características aprendidas por la CNN se aprenden mediante una LSTM. Así mismo, el enfoque fue usado para el reconocimiento de micro expresiones y expresiones deliberadas con resultados superiores a los métodos más avanzados, los resultados de reconocimiento fueron de 78.61 % usando el conjunto de datos MMI y 60.98 % usando CASME II.

2.3. Conjuntos de datos para tareas de FER

El tema de elegir un conjunto de datos adecuado ha sido discutido ampliamente debido a que las personas de diferentes rangos de edades, culturas y genéros muestran e interpretan una expresión facial en distinta forma, por ello, es importante la elección correcta de un conjunto de datos con ejemplos abundantes y que sean capaces de brindar la información necesaria para que los modelos puedan generalizar adecuadamente las distintas expresiones con base en las diferencias faciales. Existen dos tipos principales de conjuntos, 2D corresponde a las imágenes estáticas, un análisis basado en este enfoque presenta dificultades para manejar las variaciones de pose (ángulo) y comportamientos faciales sutiles, es decir, asociados a la geometría facial.

Los trabajos relacionados al análisis 3D proponen facilitar el análisis de los sutiles cambios estructurales en las expresiones faciales permitiendo, en algunos casos, llenar información faltante. La oclusión es un tema importante en este aspecto. Se presentan algunos de los conjuntos de datos más populares para tareas de FER:

- The Extended Cohn-Kanade Dataset CK+[59]: Es un conjunto de datos de imágenes en resolución de 640×480 y 640×490 de 123 sujetos de entre 18 y 30 años de edad, además contiene alrededor de 593 secuencias de video tanto de expresiones espóntaneas como provocadas. Asimismo, proporciona protocolos y resultados de referencia (landmarks) para el seguimiento de rasgos faciales, UAs y reconocimiento de emociones mediante etiquetado.
- Japanese Female Facial Expressions JAFFE [60]: Contiene 213 imágenes de siete tipos de emociones tomadas de diez modelos de origen japonés. Cada una de las imágenes tiene un tamaño de 256×256 píxeles.
- MMI [71]: Consiste en 2900 secuencias de video e imágenes de alta resolución de 75 participantes. El tamaño de las imágenes es de 720×576 píxeles, el conjunto está completamente etiquetado.
- The Karolinska Directed Emotional Face (KDEF) [1]: Consiste en 4900 imágenes de 70 participantes, presenta 7 emociones en cinco distintos ángulos. El tamaño de las imágenes es de 562×762 píxeles.
- Extended Yale B face (B+) [68]: Este conjunto consta de 16,128 imágenes faciales tomadas. El tamaño de cada imagen es de 320×243 píxeles.
- FER2013 [47]: Es un conjunto de datos usado durante ICML en 2013 *Challenges in Representation Learning*. FER2013 corresponde a imágenes recopiladas por la API de búsqueda de imágenes de Google, el tamaño de las imágenes es de 48×48 píxeles, contiene 28,709 imágenes de entrenamiento, 3,589 imágenes para prueba y 3,589 imágenes para validación.

Para el presente trabajo se tomará el conjunto de CK+[59] debido a las condiciones en las que se encuntran las imágenes y la facilidad de obtener tal conjunto.

2.4. Estructura general de una red de convolución

Las redes neuronales de convolución (CNN) utilizan una forma de vinculación de paramétros que reduce enormemente el número total de paramétros libres, lo que las hace útiles para el tratamiento de imágenes. El uso de las CNN propuestas en 1998 por [43] ha mostrado una efectividad en el aprendizaje de características con un alto nivel de abstracción cuando se utilizan arquitecturas más profundas, es decir, cuando se hace un amplio uso de capas de convolución. En general, este tipo de red jerárquica tiene tres tipos alternativos de capas, incluidas capas convolucionales, capas de submuestreo y capas completamente conectadas.

- Capa de convolución, contiene un conjunto de filtros que son convolucionados en toda la imagen, cada uno de estos filtros corresponde a ciertas características de la imagen que serán la entrada para la siguiente capa en la red. Las operaciones de convolución están asociadas a tres beneficios: 1) conectividad local que aprende correlación entre los píxeles vecinos en la imagen; 2) pesos compartidos en el mismo mapa de características, esta ventaja reduce en gran medida el número de paramétros a ser aprendidos por la red e, 3) invarianza de desplazamiento a la ubicación del objeto.

- Capa de submuestreo, esta capa es usada para reducir el tamaño espacial de los mapas de características [11] (resultantes de la capa anterior), además de reducir el costo computacional de la red, de lo contrario el tamaño y recursos necesarios para procesar se elevaría considerablemente. Existen dos tipos de sub-muestreo para este tipo de capa conocidos como maximum-pooling y average-pooling [11].

- Capa completamente conectada, esta capa es colocada al final de la red para garantizar que todas las neuronas de la capa estén conectadas en la capa anterior y para permitir que los mapas de características se conviertan a un mapa de características 1D, esto para una mayor representación y clasificación de las características aprendidas.

Una de las principales ventajas de las CNN es que la entrada del modelo puede ser una imagen sin procesar en lugar de un vector o conjunto de características seleccionadas a mano. La imagen 2.1 muestra la estructura

general una red de convolución (CNN), el número de capas es variable, algunos modelos de red híbridos han surgido tomando la base de este modelo.

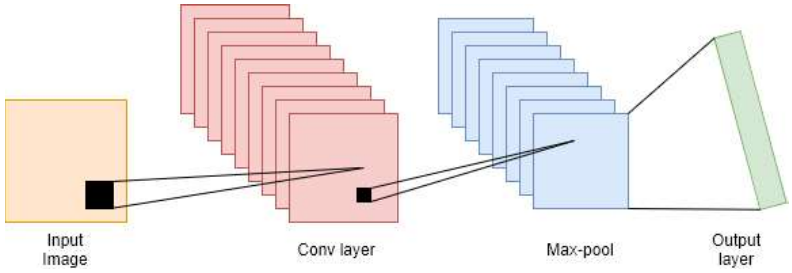


Figura 2.1: Estructura CNN.

2.4.1. Modelos populares de CNN

A medida que el hardware fue incrementando su capacidad, las CNNs comenzaron a popularizarse como un enfoque de aprendizaje eficiente en tareas de visión por computadora, reconocimiento de lenguaje, además de otros campos de ML. Tras este esquema general de CNN, han surgido modelos híbridos de los que se ha presentado un poco en el capítulo anterior, además de los trabajos realizados con estos modelos y sus resultados en FER. Algunos de los principales modelos usados en el campo de la visión por computadora, así como su uso para tareas de FER, son:

- LeNet (1998), originalmente este modelo fue propuesto en 1999, pero debido a las limitaciones computacionales no fue implementado sino hasta 2010 por [43]. Este modelo fue propuesto con el algoritmo de back-propagation y se experimentó con digítos escritos a mano. La arquitectura básica de la red consta de 2 capas de convolución, 2 capas de sub muestreo y 2 capas completamente conectadas más una capa de salida con conexión gaussiana.
- AlexNet (2012), quizá uno de los modelos de red profunda mayormente conocidos, este modelo propuesto por [41] durante *ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012*. Su desempeño logró superar a los últimos métodos propuestos. Fue un punto de separación entre el uso de modelos tradicionales y los nuevos modelos basados en aprendizaje profundo. La estructura básica de la red se basa en 4 capas de convolución junto con una capa de max-pooling tras cada capa de convolución, 2 capas completamente conectadas (fully connected) y una capa de salida con una función softmax. Dos nuevos conceptos son introducidos en este modelo, Local Response Normalization (LRN) y Dropout [77]. Así mismo, dos redes con una estrctura similar y el mismo número de mapas de características son entrenadas de forma paralela para este modelo. Un punto a considerar en este modelo es que el número de paramétros de la red es considerablemente alto.
- Visual Geometry Group (VGGNET, 2014), este modelo fue presentado en la competencia ILSVRC [75]. Una de las principales contribuciones de este modelo es mostrar que la profundidad de las redes es un componente crítico para lograr un mejor reconocimiento o precisión en el modelo. Esta arquitectura está constituida por dos capas de convolución junto con una función de activación ReLU, una capa de max

pooling y múltiples capas completamente conectadas con una función ReLU, se añade una capa final para clasificación con una función softmax. Otras variantes de red VGG fueron propuestas siguiendo este modelo inicial.

- GoogLeNet (2014), este modelo también fue presentado en la competencia de ILSVRC por [78] de Google con el objetivo de reducir la complejidad computacional de las CNNs. El modelo GoogLeNet mejoró los resultados de precisión de múltiples modelos presentados al incorporar 22 capas en total, siendo un modelo mucho mayor en cuanto a capas a cualquier otro modelo anterior. Sin embargo, el número de paramétros de la red de GoogLeNet es de apenas 5M frente a los modelos de AlexNet o VGG con 60M y 138M respectivamente.

- Residual Network (ResNet, 2015), fue el modelo ganador en la competencia de ILSVRC de 2015 desarrollado por [27], este enfoque fue propuesto con la idea de obtener un modelo de red ultra-profundo que no sufra de desvanecimiento de gradiente que presentaban muchos modelos anteriores [[5], [23]]. Esta arquitectura se basa en un amplio número de capas que va incrementando, algunos de estos son: 50, 101, 152 y 1202 capas. El modelo más popular es ResNet50 el cual consiste en 49 capas de convolución y una capa de clasificación completamente conectada.

- Densely Connected network (DenseNet, 2017) fue desarrollado por Gao Huang en 2017 [29]. La estructura general de esta red es que la salida de cada capa está conectada con todas las capas sucesivas en un denso bloque. La idea de esta arquitectura es reutilizar los mapas de características que son obtenidos tras cada capa, este novedoso enfoque permite la reducción de paramétros de entrenamiento en la red. DenseNet consiste en múltiples bloques densos y bloques de transición que se colocan entre cada par de bloques densos.

En el presente trabajo se usará un modelo de CNN que sigue este esquema general a fin de implementar un modelo de red que sea sencillo de generar, entrenar y realizar predicciones. Para la evaluación de la red, se usarán algunas métricas comunes en tareas de FER las cuales se abordarán en el capítulo 4.

Capítulo 3
Estado del Arte

El trabajo realizado por Kuan Li se enfoca en el uso de imágenes estáticas y del reconocimiento de seis expresiones proponiendo un método de recorte facial a las imágenes de entrenamiento. Como parte del pre-procesamiento se incorpora la normalización z-score y la equalización de histograma al conjunto de entrenamiento. Para probar el método se usaron los conjuntos de CK+[59] y JAFFE [60], los resultados mostraron que el método es competitivo en términos de tiempo de entrenamiento, pruebas y precisión.

El presente trabajo se enfocará en la aplicación de un modelo de CNN debido al emergente uso de este enfoque, además de utilizar algunas de las técnicas de pre-procesamiento así como de descripción de características. Se expone una comparación de dos tareas muy importantes de la visión por computadora: la deteccion de expresiones faciales y la detección de rostros. Se pretende hacer una comparación entre los enfoques tradicionales de ML, es decir, enfoques dependientes de la extracción manual de características y los enfoques de Deep Learning que han mostrado mejoras significativas. Estos modelos comparten un conjunto de prueba en común, MMI [71].

Pensamos que la implementación de un método de CNN de vanguardia nos permitirá adentrarnos a la parte teórica y experimental no solo de la visión por computadora sino en las tareas de recocimiento de expresiones faciales y el análisis detallado de los conjuntos de datos que son usuales en tareas de FER. Además, hacemos mención de algunos trabajos realizados así como sus resultados con el fin de contrastar los resultados obtenidos al usar distintos conjuntos de datos así como los modelos de red utilizados.

En [8] propusieron un método basado en redes profundas (CNN) sin algún tipo de pre-procesamiento manual. La arquitectura de red consiste en 7 capas de convolución, 5 de max-poooling, 2 capas de concatenación y una capa de normalización, este modelo sigue cuatro pasos. La primera parte es responsable del pre-procesamiento automático de datos mientras que lo restante se ocupa de la extracción de características. La evaluación del modelo fue usando los conjuntos de CK+[59] y MMI alcanzando un porcentaje de

reconocimiento del 99.6 % y 98.63 % respectivamente. No se garantiza que los sujetos (ejemplos) usados para entrenamiento no se hayan usado para probar el modelo.

En [76] desarrollaron un sistema de reconocimiento facial utilizando redes de convolución. Este modelo consta de 5 capas. De acuerdo a los autores, es común obtener un sobre-ajuste en la red si se dispone de pocos ejemplos para el entrenamiento. Debido a esto, los autores proponen el aumento de datos a fin de incrementar el número de ejemplos usados. Además se hace uso de drop-out [77] durante la fase de entrenamiento. Los experimentos se realizaron usando el conjunto de CK+[59] y otros tres conjuntos de datos creados por los autores, el uso de validación cruzada se hizo en cada experimento alcanzando un porcentaje del 99.2 %, sin embargo cada experimento contó con solo cinco expresiones (enojo, felicidad, tristeza, sorpresa y neutral).

Otros trabajos realizados se enfocan en el uso de Unidades de Acción (UAs) [[3], [48], [46]]. Estos trabajos presentan el uso de las unidades de acción como un factor importante a considerar en el estudio y reconocimiento de las expresiones faciales, la selección adecuada de estas unidades así como la selección misma de las regiones de intéres. Otros puntos de importancia como el recorte basado en puntos de referencia faciales (landmarks) que eliminan por completo el fondo en las imágenes o la selección de regiones faciales muestran resultados favorables. Estos enfoques se basan en el cambio que ocurre en el rostro, es decir, las UAs que están presentes antes y después en una secuencia de imágenes, así como las UAs que están activas en una imagen estática determinan la emoción que se está expresando. Sin embargo, mientras que el uso de estas UAs resultan una vía favorable para atacar el problema de la detección facial, presentan una desventaja al usar imágenes donde el rostro no es frontal o donde los sujetos no expresan las unidades de acción correspondientes, es decir, las emociones son las mismas en distintas personas pero estas son expresadas de forma distinta.

3.1. Método de Kuan Li

Se realizaron experimentos a fin de probar la versatilidad del método de recorte por lo que los autores realizaron dos comparaciones usando los conjuntos de CK+[59] para entrenar la red y JAFFE [60] para probar. Poste-

riormente repitieron el proceso usando JAFFE [60] para entrenar y CK+[59] para probar. Durante este paso, solo se usan 6 expresiones dado que el conjunto de JAFFE [60] no incluye las mismas expresiones de CK+[59] por lo que se utilizan aquellas que comparten ambos conjuntos, miedo, disgusto, angustia, felicidad, tristeza y sorpresa.

De acuerdo a los resultados reportados por los autores, la expresión Neutral fue la responsable de la baja tasa de clasificación en el experimento, resultando por debajo del 50 % en promedio de clasificación al entrenar el modelo usando CK+[59] y probar los resultados usando JAFFE [60].

Otros trabajos como en [62], dónde los autores obtuvieron un porcentaje de reconocimiento del 98 % usando el conjunto de JAFFE [60] y siete clases con un modelo de redes de convolución profunda (DCNN, por sus siglas en inglés) y máquinas de vectores de soporte (SVM) para clasificación. Sin embargo, este método usó validación cruzada dejando uno fuera (leave-one-subject-out, LOSO) el cual implicó un alto coste computacional. Otro trabajo similar al que usó Kuan Li para comparar su recorte fue realizado por [55] y [56] donde se usó validación cruzada (eightfold) usando 7 clases con un porcentaje de reconocimiento del 96.7 % y 96.76 % respectivamente.

Partiendo de estos trabajos realizados, las tablas 3.1 y 3.2 muestran el porcentaje alcanzado en tareas específicas de FER utilizando el conjunto de MMI. Los resultados que brindan los enfoques de DL superan en gran medida a los métodos tradicionales. Aunque los resultados experimentales con CNN son exitosos, aún hay algunas cuestiones que vale la pena mencionar que pueden mantenerse en mente y que servirán de guía en el resto de este proyecto.

- Se requiere de un gran poder cómputo a medida que se desea profundizar aumentando el número de capas en un modelo de CNN.
- Se requiere de conjuntos de datos etiquetados, usualmente a mano.
- Se requiere un modelo de red que en la práctica sea sencillo de implementar.
- El tiempo de entrenamiento y pruebas es una parte esencial. En aplicaciones del mundo real es un factor importante que un modelo esté bien balanceado.

Tipo	Enfoque propuesto	Precisión
Enfoques convencionales de ML en tareas de FER	Sparse representation classifier with LBP features [30]	59.18
	Sparse representation classifier with local phase quantization features [81]	62.72
	SVM with Gabor wavelet features [85]	61.89
	Sparse representation classifier with LBP from three orthogonal planes [87]	61.19
	Sparse representation classifier with local phase quantization feature from three orthogonal planes [34]	64.11
	Collaborative expression representation CER [44]	70.12
	Promedio	63.20

Cuadro 3.1: Comparativa de modelos tradicionales usando únicamente el conjunto de MMI [69].

Por otro lado, la tabla 3.3 muestra una comparativa de los modelos más representativos de FER en conjuntos de datos como CK+[59] y JAFFE [60].

Tipo	Enfoque propuesto	Precisión
Enfoques de DL en tareas de FER	Deep learning of deformable facial action parts [53]	63.40
	Joint fine-tuning in deep neural networks [36]	70.24
	AU-aware deep networks [63]	69.88
	AU-inspired deep networks [51]	75.85
	Deeper CNN [65]	77.90
	CNN + LSTM with spatio-temporal feature representation [84]	78.61
	Promedio	72.65

Cuadro 3.2: Comparativa de modelos de DL usando el conjunto de MMI.

Tipo	Enfoque propuesto	Precisión
Enfoques de DL en tareas de FER	Facial Expression Recognition Method Based on Sparse Batch Normalization CNN [9]	95.24
	Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy [45]	97.38
	Facial expression recognition with Convolutional Neural Networks Coping with few data and the training sample order [56]	96.76
	Extended deep neural network for facial emotion recognition [32]	95.23
	A deep learning perspective on the origin of facial expressions [7], solo CK+[59]	98.62
	Promedio	96.64

Cuadro 3.3: Comparativa de modelos de DL usando los conjuntos de CK+[59] y JAFFE [60].

Capítulo 4

Metodología

4.1. Implementación del Método de Kuan Li

4.1.1. Elección del conjunto de datos

- El conjunto de CK+[59] cuenta con un total de 10,726 imágenes de ocho categorias de las cuales solo 5,754 imágenes están etiquetadas. Cada una de las expresiones es mostrada entre 10 y 22 imágenes por 123 sujetos. Así mismo, la secuencia de cada emoción inicia con una expresión neutral y evoluciona hasta una expresión pico. Para este trabajo se considera la primera y última imagen de todas estas secuencias dando un total de 439 imágenes únicas para las ocho expresiones. El conteo de cada expresión es el siguiente:

  Enojo: 44 imágenes

  Desprecio: 17 imágenes

  Disgusto: 59 imágenes

  Miedo: 24 imágenes

  Felicidad: 69 imágenes

  Neutral: 116 imágenes

  Tristeza: 27 imágenes

  Sorpresa: 83 imágenes

4.1.2. Pre-procesamiento

Durante este paso se propone aumentar el número de ejemplos de entrenamiento (imágenes) al conjunto de datos base de CK+[59] a fin de mejorar la precisión del modelo, la generalización, y controlar el sobre-ajuste (overfitting). Para el trabajo de Kuan Li se consideran solo seis expresiones: sorpresa, tristeza, felicidad, miedo, disgusto y enojo. Se realizó una división

del conjunto de datos 90/10 para entrenamiento y pruebas, 279 y 27 imágenes respectivamente para cada conjunto.

- Rotaciones (face alingment), durante este paso se realiza una corrección a las imágenes del conjunto de entrenamiento usando Dlib Toolkit [39] para la localización de puntos de referencia. Para cada imagen de entrenamiento se obtuvieron 68 puntos de referencia faciales (facial landmarks) y se usaron solo 12 de estos para identificar los ojos en la imagen. Estos corresponden a los puntos del 36 al 47, los primeros seis definen el ojo izquierdo y el resto define al ojo derecho. Para realizar la rotación de la imagen, el ángulo de rotación es calculado por la siguiente fórmula:

$$angle = tan^{-1} \frac{\sum_{n=42}^{47} y_n - \sum_{n=36}^{41} y_n}{\sum_{n=42}^{47} x_n - \sum_{n=36}^{41} x_n} \tag{4.1}$$

  donde (x_n, y_n) es la coordenada x del n-ésimo punto.

  La rotación hace que el ángulo formado entre el centro de un ojo y el otro y el eje horizontal sea cero. La corrección puede no resultar perfectamente alineada con el eje horizontal después de la rotación.

- Recorte (image cropping), se elimina el fondo de cada imagen permitiendo a la red centrarse solo en la parte del rostro, es decir, removiendo información que no contribuye a la expresión (orejas, parte de la frente). Para el recorte se usan los puntos 1, 9 y 17 de los puntos faciales (landmarks) considerados en el paso anterior y un cuarto punto. Para la localización de este último, se calcula el centroide de cada ojo con la siguiente fórmula:

$$centroid_x = \frac{\sum_{i=0}^{n} x_i}{n} \tag{4.2}$$

$$centroid_y = \frac{\sum_{i=0}^{n} y_i}{n} \tag{4.3}$$

  donde (x_i, y_i) corresponde a la coordenada del i-ésimo punto, y n al total de puntos.

Se traza una línea horizontal tomando los centroides calculados y se localiza el punto medio de dicho segmento, el cuarto punto se coloca a una distancia de 60 píxeles de forma vertical respecto al punto medio localizado.

Para la obtención final de las imágenes que serán usadas para el entrenamiento y pruebas se siguió en orden esta serie de pasos como parte del pre-procesamiento.

- Ecualización de histograma (histogram equalization), se aplica al conjunto de datos para entrenamiento sin alguna modificación u operación adicional.
- Normalización z-score, se calcula la media y desviación estándar de todas las imágenes, posteriormente se aplica la normalización a cada imagen usando los valores encontrados.
- Finalmente, se reduce el tamaño de la imagen (downsampling) a 32×32 píxeles. Este reescalado puede garantizar que las distintas partes en el rostro estén en la misma posición para distintas imágenes (personas).

(a) Imagen original (b) Imagen alineada (c) Imagen recortada

Figura 4.1: Alineación y método de recorte a las imágenes de CK+[59].

Para confirmar el método de recorte usado por [45], los autores compararon con otros dos métodos de recorte conocidos utilizando el modelo de red: LeNet5 [43] y AlexNet [41]. El tamaño de las imágenes son de 32×32 y 224×224 píxeles respectivamente, los autores reportan que su modelo de red y recorte usado logra una ligera mejora en los resultados de predicción frente a estos modelos de red.

4.1.3. Aumento de datos

Para este paso se realizó una rotación aleatoria entre -2° y 2° a fin de incrementar el número de ejemplos y evitar un sobre-ajuste del modelo, también se usa volteo horizonal (horizontal flipping). Este paso se hace con ayuda de las bibliotecas de Keras y Tensorflow disponibles en Python. Tras este proceso se logró aumentar el número ejemplos de entrenamiento. Este incremento se realiza únicamente durante la etapa de entrenamiento.

4.1.4. Elección de un modelo de CNN

- La arquitectura de la red convolucional (CNN) se muestra en la figura 4.2. La tabla 3.1 muestra los paramétros utilizados para la red.

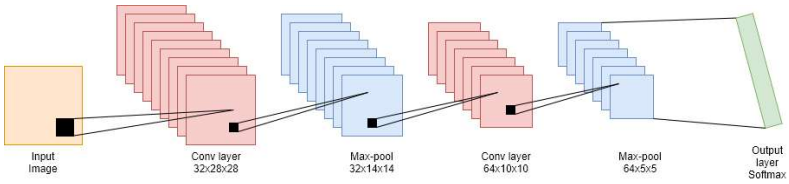


Figura 4.2: Estructura CNN, tomado de [45] pág. 6.

Como puede verse en la imagen, en la última capa de la red se conecta directo a una capa de salida que corresponde a un vector 1600-dimensional.

Paramétros	Valor
Random rotation	-2° a 2°
Stride	2
Kernel size	2×2
Dropout	0.5 (segunda capa de conv)
Initializer	Xavier
Optimizer	Momentum (lr = 0.001)
Batch size	16
Loss funtion	Cross-entropy
Epochs	120

Cuadro 4.1: Paramétros de la red.

4.1.5. Entrenamiento y evaluación

- En las tareas de clasificación existen múltiples criterios para evaluar los resultados de los estudios, en esta sección consideraremos algunas de las métricas más utilizadas en reconocimiento facial. La tasa de error y exactitud del modelo se obtiene mediante la aplicación de las fórmulas 4.4 y 4.5. Otros criterios de evaluación son la precisión 4.6 y recall 4.7, mientras que para tareas de multiclasificación usualmente se utiliza la métrica de F1 score 4.8.

$$err = \frac{1}{m}\sum_{i=1}^{m} g(f(x_i) \neq y_i) \tag{4.4}$$

$$acc = \frac{1}{m}\sum_{i=1}^{m} g(f(x_i) = y_i) \tag{4.5}$$

$$pre = \frac{TP}{TP + FP} \tag{4.6}$$

$$rec = \frac{TP}{TP + FN} \tag{4.7}$$

$$F1 = \frac{2 \times pre \times rec}{pre + rec} \tag{4.8}$$

Donde TP, FP, TN, FN, son verdadero positivo, falso positivo, verdadero negativo y falso negativo respectivamente y m el número de ejemplos.

Durante el entrenamiento, el modelo de red recibe un conjunto de datos de entrenamiento que comprende imágenes en escala de grises con su respectiva etiqueta. Para asegurar que el rendimiento en el entrenamiento no sea afectado se separaron algunas pocas imágenes para validación mientras que el resto fueron usadas para entrenamiento.

Figura 4.3: Imagen ilustrativa sobre la rotación aleatoria de entre -2° y 2° durante el entrenamiento.

4.2. Propuesta de mejora al Método de Kuan Li

Durante esta sección mostraremos los cambios realizados en este trabajo de tesis como una propuesta de mejora al método implementado por [45]. Los cambios realizados fueron sobre el método de recorte y cambiando el modelo de red profunda. Para el método de recorte se hace un ligero incremento tanto en las imágenes de entrenamiento como en las de prueba a fin de obtener una región o área de interés mayor, las figuras 4.4 y 4.5 muestran este cambio. El incremento realizado es sobre el valor de distancia de 60 píxeles ([45]) a 90 píxeles, que está asociado al proceso de Recorte mostrado en 4.1.2.

4.2.1. Experimentos de reconocimiento sobre el recorte propuesto

(a) Enojo (b) Disgusto (c) Miedo (d) Felicidad (e) Tristeza (f) Sorpresa

Figura 4.4: Método de recorte propuesto por [45]

(a) Enojo (b) Disgusto (c) Miedo (d) Felicidad (e) Tristeza (f) Sorpresa

Figura 4.5: Nuevo método de recorte.

La elección de este recorte en la figura 4.5 fue determinado arbitrariamente y considerando las Unidades de Acción (UAs, por sus siglas en inglés) [59], aunque el modelo propuesto por [45] originalmente no se enfoca en el uso o revisión de UAs. Un método similar fue propuesto por [56] donde los autores usaron un factor α para el recorte, lo que le permite manejar situaciones en las que la geometría facial difiere de otras imágenes (rostros). En la figura 4.5.f se muestran ligeros cambios en la unidad de acción conocidad como AU1 (Inner Brown Raiser) figura 4.7 la cual está involucrada en las emociones de tristeza, sorpresa y miedo. El uso completo de estas unidades de acción fue realizado por [13]. Esta nueva región facial añadida puede perderse como se muestra en la figura 4.4.f, por el contrario, puede observarse en 4.5.e y 4.5.c este incremento de región. En los experimentos realizados no se observó pérdida de información al utilizar este método de recorte propuesto y que impacte directamente en los resultados de predicción del modelo usado por Kuan Li, por el contrario, al conducir varios experimentos los resultados fueron ligeramente mejores a los mostrados por [45] esto presupone que al tener una región facial adecuada puede resultar en un desempeño favorable y una mejora en las predicciones del modelo.

En este trabajo de tesis, se propone un nuevo modelo de red basado en la conocida red de LeNet [43] para mejorar la exactitud mostrada por Kuan Li, dicho modelo se muestra en la figura 4.6, mientras que los paramétros de entrenamiento se visualizan en la tabla 4.2. Estos mismos paramétros se usaron en todos los experimentos realizados.

Paramétros	Valor
Random rotation	-2° a 2°
Stride	2
Kernel size	5×5
Dropout	0.5 (segunda capa de conv)
Initializer	Xavier
Optimizer	Momentum (lr = 0.001)
Batch size	16
Loss funtion	Cross-entropy
Epochs	120

Cuadro 4.2: Paramétros usados para entrenar el modelo de red propuesto.

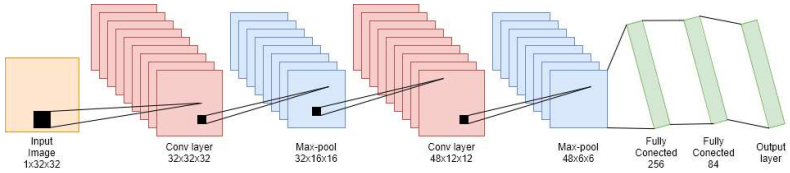


Figura 4.6: Modelo de red propuesto, basado en LeNet.

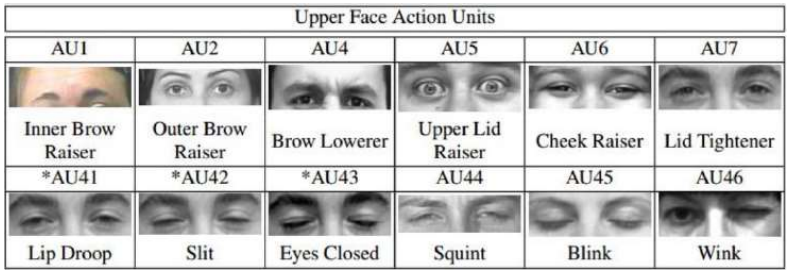


Figura 4.7: Unidades de Acción (UAs). Tomado de IntraFace [13] pág. 5.

Capítulo 5
Resultados

En este capítulo abordaremos los resultados obtenidos de acuerdo al modelo de red y método de recorte usados por [45]. También, se abordarán los resultados obtenidos del modelo de red y recorte propuestos en este trabajo de tesis. Además, indicaremos los paramétros de red usados para la fase de entrenamiento. Se realizaron varios experimentos a fin de determinar una correcta evaluación de ambos modelos, también se muestran algunos gráficos obtenidos durante el entrenamiento de los modelos de red.

5.1. Resultados usando el modelo de Kuan Li

Durante la fase de pruebas, [45] realizan múltiples pruebas a fin de determinar un modelo de CNN adecuado a través de cuatro experimentos usando validación cruzada para la selección del ángulo de rotación y número de neuronas en la capa final. Tras cada experimento se utilizaron los mismos paramétros de red, ver tabla 4.1. En los experimentos usaron un conjunto de imágenes de CK+[59] utilizando el método de recorte mencionado e imágenes de entrenamiento con fondo. El porcentaje de exactitud de la red se incrementó cuando se eliminaba una capa oculta, la red alcanzó un porcentaje del 92.42 % conectando la última capa de submuestreo con una capa final usando un función de activación softmax.

Sin embargo, este resultado no pudo ser reproducido durante la fase de entrenamiento y pruebas en este trabajo de tesis. Se obtuvieron resultados con una menor exactitud. Los bajos resultados obtenidos se deben al número de imágenes en el conjunto de prueba que corresponde a un total de 27 imágenes (ver 4.1.2) mientras que Kuan Li obtiene un número de 415. Además, no encontramos en el artículo de Kuan Li cómo se obtuvo este incremento de imágenes en el conjunto de prueba lo que representa una inquietud ya que el alto valor de exactitud se pudo alcanzar usando una misma imagen o sujeto que este presente tanto en el conjunto de entrenamiento como en el de prueba. El valor máximo alcanzado fue de 90 % mientras que

al reentrenar se logra un porcentaje de hasta 89 % en precisión. Las gráficas **??** ilustran la precisión y pérdida obtenida durante el entrenamiento. El promedio de precisión usando el modelo de red de Kuan Li fue de 85.8 %, usando el conjunto de prueba para evaluar el modelo. Nuevamente, este bajo resultado puede deberse al escaso número de ejemplos en el conjunto de prueba.

5.2. Resultados de reconocimiento sobre el nuevo recorte propuesto

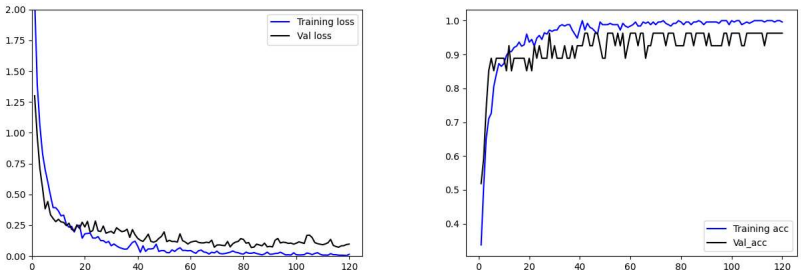
Para la etapa de pruebas, se utilizó el mismo paso de pre-procesamiento con el cambio en la modificación del recorte y conservando el número de ejemplos en ambos conjuntos inicialmente, es decir, el entrenamiento se realizó usando 279 imágenes el cual se divide en entrenamiento y validación, además, el aumento de datos se realiza únicamente sobre este último conjunto de entrenamiento, mientras que para el conjunto de pruebas se obtuvieron 27 imágenes. No se realiza aumento sobre este conjunto. La evaluación del modelo se realizó al final de cada experimento usando el conjunto de prueba, posteriormente se pasa al siguiente fold. Las gráficas **??** corresponden a los experimentos realizados y tomando el k-fold más alto durante cada uno de los experimentos.

Los experimentos se realizaron mediante cinco entrenamientos y verificados por validación cruzada (tenfold) con un promedio del 88 % usando el enfoque propuesto, mientras que al replicar los experimentos usando el modelo usado por Kuan Li se alcanzó un promedio del 85.8 % mostrando que el nuevo enfoque tiene una ligera mejora en los resultados al proponer este incremento de región facial.

Los resultados de cada uno de los experimentos fueron los esperados tras cada entrenamiento, es decir, los resultados de precisión mostrados en el conjunto de entrenamiento son mejores en comparación al conjunto de validación, por otro lado, el error es menor usando el conjunto de entrenamiento y ligeramente superior en el conjunto de validación. Esta pequeña diferencia entre el error y precisión en ambos conjuntos resulta aceptable, además es posible ver que no existe un sobreajuste en los datos.

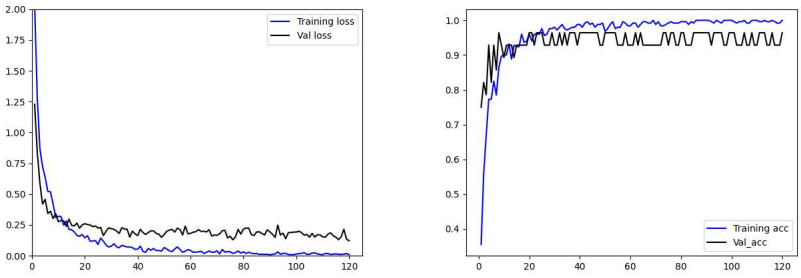
Tras estos experimentos es posible ver una mejora en los resultados de predicción del modelo como resultado del recorte propuesto permitiendo

tener una región facial más amplia de extracción de características de las expresiones faciales. No obstante, estos resultados no logran igualar los resultados reportados por Kuan Li, donde la red alcanzó una exactitud del 92.42 %, este resultado no fue reproducido en nuestras pruebas posiblemente por el limitado número de ejemplos en el conjunto de prueba. Como consecuencia de esto los gráficos siguientes muestran escalones en casi todos los experimentos realizados.



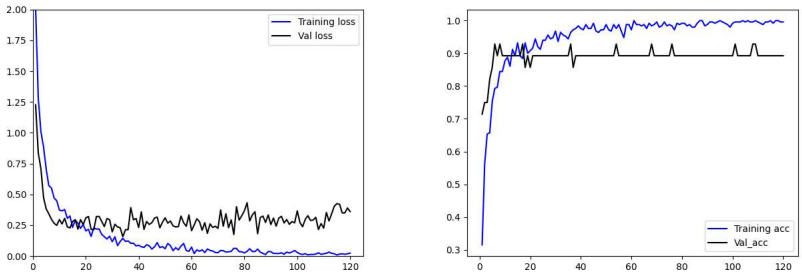
(a) Valor de pérdida (b) Valor de precisión

Figura 5.1: Resultados de experimento 1. Modelo de Kuan Li



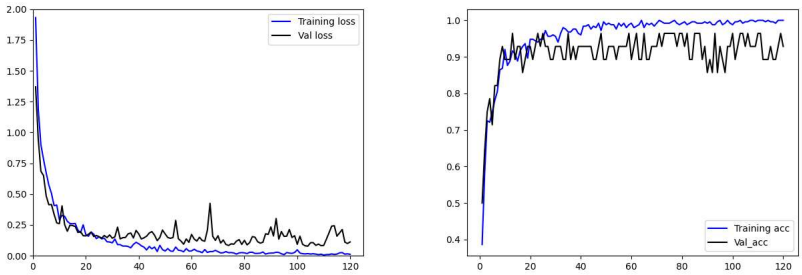
(a) Valor de pérdida (b) Valor de precisión

Figura 5.2: Resultados de experimento 2. Modelo de Kuan Li



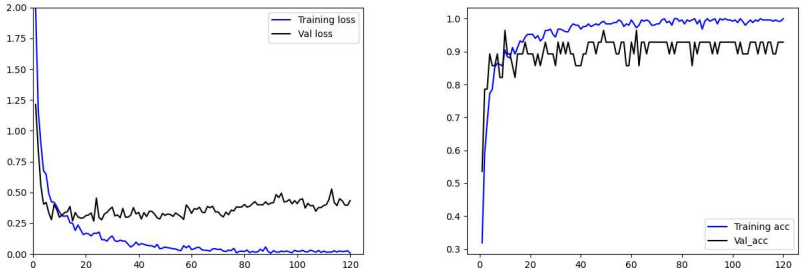
(a) Valor de pérdida (b) Valor de precisión

Figura 5.3: Resultados de experimento 3. Modelo de Kuan Li



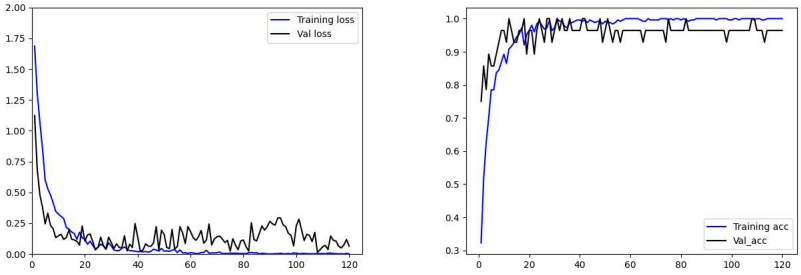
(a) Valor de pérdida (b) Valor de precisión

Figura 5.4: Resultados de experimento 4. Modelo de Kuan Li



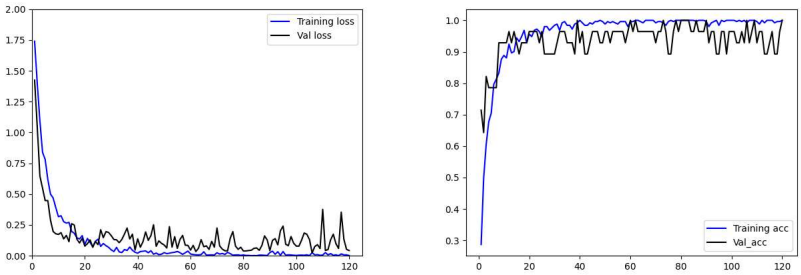
(a) Valor de pérdida (b) Valor de precisión

Figura 5.5: Resultados de experimento 5. Modelo de Kuan Li



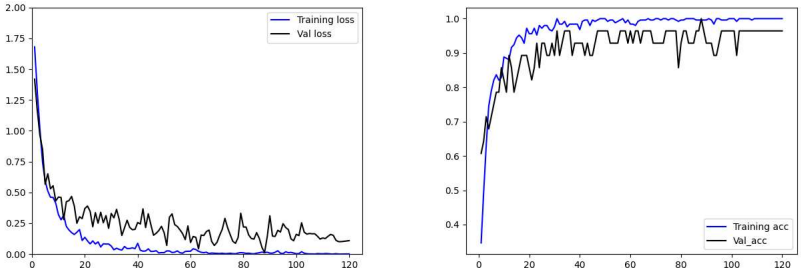
(a) Valor de pérdida (b) Valor de precisión

Figura 5.6: Resultados de experimento 1. Modelo de red propuesto



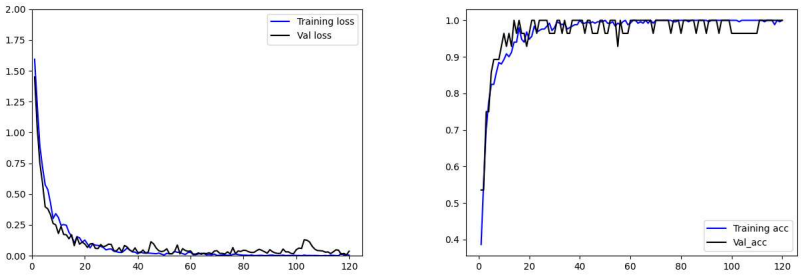
(a) Valor de pérdida (b) Valor de precisión

Figura 5.7: Resultados de experimento 2. Modelo de red propuesto



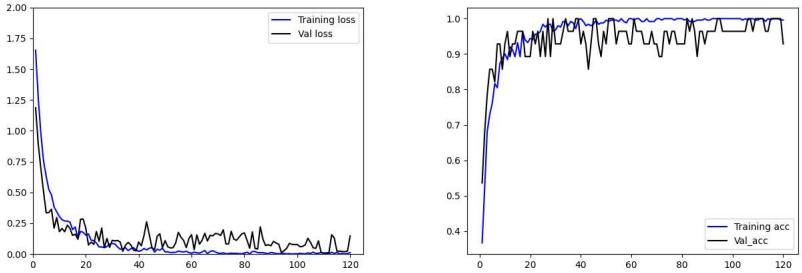
(a) Valor de pérdida (b) Valor de precisión

Figura 5.8: Resultados de experimento 3. Modelo de red propuesto



(a) Valor de pérdida (b) Valor de precisión

Figura 5.9: Resultados de experimento 4. Modelo de red propuesto



(a) Valor de pérdida (b) Valor de precisión

Figura 5.10: Resultados de experimento 5. Modelo de red propuesto

Capítulo 6
Conclusiones

En este trabajo se abordó el uso de los puntos de referencia faciales para la detección de expresiones faciales tomando como base el trabajo realizado por [45] Tomando como base este artículo se decidió proponer un cambio en la metodología y proponer una modificación en el método de recorte facial en las imágenes que consitió en incrementar la región facial usando los puntos de referencia faciales 1, 9, 17 y un cuarto punto como resultado de calcular el centroide de los puntos que corresponden a los ojos y uniendo estos centroides mediante un segmento para obtener el punto medio, finalmente este cuarto punto se coloca a una distancia de 90 píxeles vertical hacia arriba partiendo del punto medio. Esta modificación permitió tener una región facial más amplia permitiendo que la red pueda tomar las caracteríscas adecuadas que definen a cada una de las expresiones faciales. También se propuso cambiar el modelo de red a LeNet. Las pruebas realizadas siguieron la misma metodología, así, con esta propuesta y tras cada experimento realizado los resultados mostrados fueron según los esperados de acuerdo al modelo y número de ejemplos. Sin embargo, no conseguimos reproducir los mismos resultados reportados por [45] que muestran resultados superiores a los obtenidos y frente a enfoques similares mediante el uso de puntos de referencia faciales (landmarcks), debido al bajo número de ejemplos en la etapa de entrenamiento y prueba.

Podemos concluir que tomando como punto de partida el trabajo de Kuan Li y proponiendo una mejora en el método de recorte así como el uso del modelo de red de LeNet, logramos ver un incremento en los resultados de entrenamiento aún con el bajo número de ejemplos en el conjunto de entrenamiento. Finalmente, con esta propuesta y los resultados obtenidos se logró llevar acabo los objetivos planteados en este trabajo de tesis.

Capítulo 7
Perspectivas

Un trabajo futuro sobre este proyecto involucraría el uso de conjuntos de datos que contengan imágenes de entornos no controlados [[20], [16], [1], [24]], aquí el rol de la extracción de características basada en apariencia puede resultar desafiante o el uso de métodos basados en información espacial resulta de especial interés. Una exhaustiva revisión a estos trabajos [[64], [54], [52]] así como sus métodos empleados resulta vital antes de abordar esta problemática. Los trabajos realizados con conjuntos de datos públicos y bajo condiciones controladas parecen brindar excelentes resultados sin complicación, sin embargo, cuando son sometidos a prueba con imágenes de entornos reales su desempeño decae al 50 % o menos. Parece ser que la información extraída no es suficiente para representar las expresiones. El uso de las CNN logra evitar la dependencia de la extracción manual de características, esto sucede gracias a que la red es capaz de aprender de forma automática el conjunto de características que mejor modela la clasificación deseada por lo que para mejorar en la clasificación se requiere de una amplia cantidad de datos.

Estos trabajos brindan un panorama en el que presuponen que la extracción de características profundas resulta importante. Por otro lado, el número de ejemplos de entrenamiento parece ser un factor importante siendo quizá una deficiencia en los modelos orientados a imágenes del mundo real. Para abordar este problema algunos pasos de pre-procesamiento a las imágenes permiten reducir esta dependencia.

Como trabajo futuro, se plantea la posibilidad de investigar otros métodos de aprendizaje para aumentar la precisión, además de que utilizando un conjunto de datos mixto entre los mencionados puede brindar información adicional que ayude a los sistemas de reconocimiento a mejorar en sus resultados, y por último, un modelo de red robusto capaz de aprender las características adecuadas de cada una de las expresiones faciales. Esta parte resulta clave para la correcta clasificación de las emociones.

Bibliografía

[1] The karolinska directed emotional faces (kdef). available online: http://www.emotionlab.se/resources/kdef (accessed on 2021-01-13), 2018.

[2] F. Abdat, C. Maaoui, and A. Pruski. Real time facial feature points tracking with pyramidal lucas-kanade algorithm. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pages 71–76, 2008.

[3] S. Al-Darraji, K. Berns, and A. Rodić. Action unit based facial expression recognition using deep learning. volume 540, pages 413–420, 11 2017.

[4] M. Al-Shabi, W. Cheah, and T. Connie. Facial expression recognition using a hybrid cnn-sift aggregator. 08 2016.

[5] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[6] R. Breuer and R. Kimmel. A deep learning perspective on the origin of facial expressions. *CoRR*, abs/1705.01842, 2017.

[7] R. Breuer and R. Kimmel. A deep learning perspective on the origin of facial expressions. *ArXiv*, abs/1705.01842, 2017.

[8] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki. Dexpression: Deep convolutional neural network for expression recognition. *CoRR*, abs/1509.05371, 2015.

[9] J. Cai, O. Chang, X.-L. Tang, C. Xue, and C. Wei. Facial expression recognition method based on sparse batch normalization cnn. In *2018 37th Chinese Control Conference (CCC)*, pages 9608–9613, 2018.

[10] W. 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)*, pages 25–32, 2017.

[11] D. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. Flexible, high performance convolutional neural networks for image classification. pages 1237–1242, 07 2011.

[12] C. Darwin. *The Expression of the Emotions in Man and Animals*. CreateSpaceIndependent Publishing Platform, 1872.

[13] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn. Intraface. 05 2015.

[14] Dong-chen He and Li Wang. Texture unit, texture spectrum, and texture analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 28(4):509–512, 1990.

[15] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[16] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[17] P. Ekman and W. Friesen. Facial action coding system: a technique for the measurement of facial movement. 1978.

[18] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. 11 2016.

[19] F. B. Fitch. Warren s. mcculloch and walter pitts. a logical calculus of the ideas immanent in nervous activity. bulletin of mathematical biophysics, vol. 5 (1943), pp. 115–133. *Journal of Symbolic Logic*, 9(2):49–50, 1944.

[20] Y. Fu, T. M. Hospedales, T. Xiang, Y. Yao, and S. Gong. Interestingness prediction by robust learning to rank. In *ECCV*, 2014.

[21] K. Fukushima. Neocognitron for handwritten digit recognition. *Neurocomputing*, 51:161–180, 04 2003.

[22] K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In S.-i. Amari and M. A. Arbib, editors, *Competition and Cooperation in Neural Nets*, pages 267–285, Berlin, Heidelberg, 1982. Springer Berlin Heidelberg.

[23] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

[24] D. große MPI Gesichtsausdruckdatenbank. Available online: https://www.b-tu.de/en/graphic-systems/databases/the-large-mpi-facial-expression-database (accessed on 2021-01-13). Mpi facial expression database.

[25] J. GyeongSic, Kim., and Y. Guk. *Journal of Information Processing Systems*, 6(2):261–268, 06 2010.

[26] B. Hassani and M. H. Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. *CoRR*, abs/1705.07871, 2017.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[28] C.-H. Hjortsjo. Man's face and mimic language / carl-herman hjortsjo; [translation, w. f. salisbury]. 1970.

[29] G. Huang, Z. Liu, and K. Weinberger. Densely connected convolutional networks. page 12, 08 2016.

[30] M. Huang, Z. Wang, and Z. Ying. A new method for facial expression recognition based on sparse representation plus lbp. In *2010 3rd International Congress on Image and Signal Processing*, volume 4, pages 1750–1754, 2010.

[31] T. Jabid, M. H. Kabir, and O. Chae. Facial expression recognition using local directional pattern (ldp). In *2010 IEEE International Conference on Image Processing*, pages 1605–1608, 2010.

[32] D. Jain, P. Shamsolmoali, and P. Sehdev. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120, 04 2019.

[33] S. Jaiswal, B. Martinez, and M. Valstar. Learning to combine local models for facial action unit detection. pages 1–6, 05 2015.

[34] B. Jiang, M. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. pages 314–321, 03 2011.

[35] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2983–2991, 2015.

[36] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2983–2991, 2015.

[37] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. 11 2015.

[38] 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*, 10(2):223–236, 2019.

[39] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60):1755–1758, 2009.

[40] R. A. Kirsch. Computer determination of the constituent structure of biological images. *Computers and Biomedical Research*, pages 315–328, 1970.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, page 2012.

[42] J. Kumari, R. Rajesh, and K. Pooja. Facial expression recognition: A survey. *Procedia Computer Science*, 58:486 – 491, 2015. Second International Symposium on Computer Vision and the Internet (VisionNet'15).

[43] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[44] S. H. Lee, W. J. Baddar, and Y. M. Ro. Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos. *Pattern Recognition*, 54:52 – 67, 2016.

[45] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen. Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. *The Visual Computer*, 36:391–404, 2019.

[46] L. Li, T. Baltrusaitis, B. Sun, and L. Morency. Edge convolutional network for facial action intensity estimation. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 171–178, 2018.

[47] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, pages 1–1, 2020.

[48] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 103–110, 2017.

[49] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 211–216, 2006.

[50] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image Vision Comput.*, 24(6):615–625, June 2006.

[51] M. Liu, S. Li, S. Shan, and X. Chen. Au-inspired deep networks for facial expression feature learning. *Neurocomput.*, 159(C):126–136, July 2015.

[52] M. Liu, S. Li, S. Shan, and X. Chen. Au-inspired deep networks for facial expression feature learning. *Neurocomputing*, 159:126 – 136, 2015.

[53] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. pages 143–157, 11 2014.

[54] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. pages 143–157, 11 2014.

[55] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[56] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 61:610 – 628, 2017.

[57] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 2016.

[58] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999.

[59] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[60] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. volume 1998, pages 200 – 205, 05 1998.

[61] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[62] V. Mayya, R. M. Pai, and M. Manohara Pai. Automatic facial expression recognition using dcnn. *Procedia Computer Science*, 93:453 – 461, 2016. Proceedings of the 6th International Conference on Advances in Computing and Communications.

[63] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013.

[64] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013.

[65] 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)*, pages 1–10, 2016.

[66] A. Mostafa, M. I. Khalil, and H. Abbas. Emotion recognition by facial features using recurrent neural networks. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, pages 417–422, 2018.

[67] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585 vol.1, 1994.

[68] B. D. A. online: https://computervisiononline.com/dataset/1105138686 (accessed on 2021-01-19). B+ database.

[69] M. F. E. D. A. online: https://mmifacedb.eu/ (accessed on 2021-01-13). Mmi facial expression database.

[70] M. Pantic and L. J. M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(3):1449–1461, 2004.

[71] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 5 pp.–, 2005.

[72] X. Pu, K. Fan, X. Chen, L. Ji, and Z. Zhou. Facial expression recognition from image sequences using twofold random forest classifier. *Neurocomputing*, 168:1173 – 1180, 2015.

[73] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee. Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks. *CoRR*, abs/1904.01390, 2019.

[74] A. K. J. S. Z. Li. *Handbook of Face Recognition, Springer Science Business Media*. Springer, 2011.

[75] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[76] I. Song, H. Kim, and P. B. Jeon. Deep learning for real-time robust facial expression recognition on a smartphone. In *2014 IEEE International Conference on Consumer Electronics (ICCE)*, pages 564–567, 2014.

[77] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 06 2014.

[78] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[79] Y. . Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[80] S. S. Tomkins and R. McCarter. What and where are the primary affects? some evidence for a theory. *Perceptual and Motor Skills*, 18(1):119–158, 1964. PMID: 14116322.

[81] Z. Wang and Z. Ying. Facial expression recognition based on local phase quantization and sparse representation. In *2012 8th International Conference on Natural Computation*, pages 222–225, 2012.

[82] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273:643 – 649, 2018.

[83] Z. Zeng, M. Pantic, G. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31:39–58, 02 2009.

[84] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu. Recent advances in convolutional neural network acceleration. *Neurocomputing*, 323:37 – 51, 2019.

[85] S. Zhang, X. Zhao, and B. Lei. Robust facial expression recognition via compressive sensing. *Sensors (Basel, Switzerland)*, 12:3747 – 3761, 2012.

[86] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607 – 619, 2011.

[87] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29:915–28, 07 2007.

[88] K. Zhao, W. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3391–3399, 2016.

[89] K. Zhao, W.-S. Chu, F. De la Torre, J. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015:2207–2216, 06 2015.

[90] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. pages 3391–3399, 06 2016.