**UNIVERSIDAD POLITÉCNICA DE MADRID**

**ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INDUSTRIALES**

**Aplicación de técnicas de visión artificial y Machine Learning en el desarrollo de algoritmos para el reconocimiento en vivo de expresiones faciales humanas e interpretación del lenguaje de señas de sordomudo americano**

**Autor**: Pedro José Zampella

**Tutor**: Sergio Domínguez

Madrid, Julio 2019

**Máster Universitario en Automática y Robótica**

## RESUMEN

El presente trabajo final de máster expone el desarrollo de dos programas: uno capaz de interpretar expresiones faciales y otro capaz de interpretar señas del lenguaje de sordomudos (en versión americana), ambos utilizando técnicas de visión artificial y Machine Learning.

El alcance del proyecto incluye el diseño de los algoritmos capaces de capturar las imágenes obtenidas mediante la cámara integrada de un ordenador portátil, su tratamiento y adecuación para interpretación, así como el entrenamiento e implementación de las redes neuronales que realizan la clasificación de las imágenes, todo esto implementado en lenguaje Python.

Además, se creó una interfaz ejecutable tanto en sistema operativo Windows como Linux, que permite al usuario ejecutar cualquiera de los dos programas, indicar el número máximo de objetivos a interpretar en cada cuadro de vídeo, y que como resultado muestra en vivo la imagen captada por la cámara, donde se identifica y rastrea el objeto de interés, junto a la clasificación del mismo, según sea el caso.

El hecho de que se utilice el lenguaje Python durante su programación permitirá a estos programas ser utilizados como servicios en la nube que permitan a un robot humanoide utilizarlos libremente para demostraciones, o para permitir mayor eficacia durante su comunicación con seres humanos.

**Palabras clave:** Inteligencia Artificial, Perceptrón Multicapa, Red Neuronal Convolucional, Python, Visión artificial.

## ÍNDICE DE CONTENIDO

[RESUMEN ii](#_Toc5458668)

[ÍNDICE DE CONTENIDO iii](#_Toc5458669)

[ÍNDICE DE FIGURAS v](#_Toc5458670)

[ABREVIATURAS Y ACRÓNIMOS vi](#_Toc5458671)

[1. INTRODUCCIÓN 7](#_Toc5458672)

[1.1. Contexto 7](#_Toc5458673)

[1.2. Estructura del documento 8](#_Toc5458674)

[2. ESTADO DEL ARTE 10](#_Toc5458675)

[2.1. Machine Learning y redes neuronales artificiales 10](#_Toc5458676)

[2.1.1. Perceptrón multicapa 11](#_Toc5458677)

[2.1.2. Mapa auto-organizado 12](#_Toc5458678)

[2.1.3. Red neuronal convolucional 12](#_Toc5458679)

[2.2. Visión artificial 14](#_Toc5458680)

[2.3. Python 15](#_Toc5458681)

[2.3.1. Tensorflow y Keras 15](#_Toc5458682)

[2.3.2. OpenCV 15](#_Toc5458683)

[2.4. Proyectos similares 16](#_Toc5458684)

[2.4.1. Detecting Skin in Images & Video Using Python and OpenCV 16](#_Toc5458685)

[2.4.2. Clasificación de número escritos a mano con redes neuronales 17](#_Toc5458686)

[2.4.3. Sign Language Recognition with Unsupervised Feature Learning 18](#_Toc5458687)

[3. MOTIVACIÓN Y OBJETIVO 20](#_Toc5458688)

[3.1. Motivación 20](#_Toc5458689)

[3.2. Objetivo general 20](#_Toc5458690)

[3.3. Objetivos específicos 21](#_Toc5458691)

[3.3.1. Reconocimiento de expresiones faciales humanas 21](#_Toc5458692)

[3.3.2. Interpretación de lenguaje de sordomudo versión americana 21](#_Toc5458693)

[4. METODOLOGÍA Y DESARROLLO 23](#_Toc5458694)

[4.1. Elección del lenguaje de programación 23](#_Toc5458695)

[4.1.1. Librerías utilizadas 23](#_Toc5458696)

[4.2. Obtención de bases de datos para entrenamiento de redes neuronales 24](#_Toc5458697)

[4.2.1. Base de datos de rostros expresando emociones 24](#_Toc5458698)

[4.2.2. Base de datos de manos haciendo señas del lenguaje de sordomudos 25](#_Toc5458699)

[4.3. Pre-procesamiento del dataset 26](#_Toc5458700)

[4.3.1. Pre-procesamiento de imágenes de rostros 26](#_Toc5458701)

[4.3.2. Pre-procesamiento de imágenes de señas 27](#_Toc5458702)

[4.4. Entrenamiento de redes neuronales artificiales 27](#_Toc5458703)

[4.5. Clasificación en vivo 28](#_Toc5458704)

[4.5.1. Identificación de objetos en el rango de color de la piel humana 28](#_Toc5458705)

[4.5.2. Seguimiento de objetos 29](#_Toc5458706)

[4.5.3. Identificación de rostros humanos 29](#_Toc5458707)

[4.5.4. Clasificación mediante redes neuronales artificiales 30](#_Toc5458708)

[5. RESULTADOS 31](#_Toc5458709)

[5.1. Pruebas post-entrenamiento 31](#_Toc5458710)

[5.2. Pruebas en vivo 32](#_Toc5458711)

[5.2.1. Clasificación de expresiones faciales humanas 32](#_Toc5458712)

[5.2.2. Clasificación de señas del lenguaje de sordomudo americano 33](#_Toc5458713)

[6. POSIBLES APLICACIONES 35](#_Toc5458714)

[7. MEJORAS Y FUTUROS DESARROLLOS 36](#_Toc5458715)

[8. BIBLIOGRAFÍA 37](#_Toc5458716)

[9. ANEXOS 38](#_Toc5458717)

## ÍNDICE DE FIGURAS

[Figura 1: Arquitectura del perceptrón multicapa 10](#_Toc5458562)

[Figura 2: Mapa auto-organizado durante proceso de entrenamiento 11](#_Toc5458563)

[Figura 3: Arquitectura de una red neuronal convolucional 12](#_Toc5458564)

[Figura 4: Funcionamiento de una neurona convolucional 13](#_Toc5458565)

[Figura 5: Detección de piel y filtrado en imagen 15](#_Toc5458566)

[Figura 6: Ejemplo de dígitos de la base de datos MNIST 16](#_Toc5458567)

[Figura 7: Métodos de procesamiento de imagen 17](#_Toc5458568)

[Figura 8: Alfabeto del lenguaje de señas americano 21](#_Toc5458569)

[Figura 9: Dataset utilizado para identificación de emociones 24](#_Toc5458570)

[Figura 10: Alfabeto contenido en el dataset 25](#_Toc5458571)

[Figura 11: Secuencia de imágenes mostrando tratamiento seguido al identificar piel humana 28](#_Toc5458572)

[Figura 12: Tablas de resultados de pruebas después del entrenamiento 30](#_Toc5458573)

[Figura 13: Sujeto uno mostrando expresiones de “ira” y “disgusto” 31](#_Toc5458574)

[Figura 14: Sujeto dos mostrando expresiones de “tristeza” y “sorpresa” 31](#_Toc5458575)

[Figura 15: Sujeto tres mostrando expresiones de “alegría” y “miedo” 32](#_Toc5458576)

[Figura 16: Los tres sujetos de prueba mostrando la expresión “neutral” 32](#_Toc5458577)

[Figura 17: Sujeto mostrando las señas para las letras “A”, “F”, “I”, “L”, “O”, “P”, “Q”, “Y” 33](#_Toc5458578)

## ABREVIATURAS Y ACRÓNIMOS

CNN: Red neuronal convolucional (del inglés *Convolutional Neural Network*)

KNN: K vecinos más próximos (del inglés *K-Nearest Neighbor*)

ML: Machine Learning (Aprendizaje automático, en español)

MLP: Perceptrón multicapa (del inglés *Multilayer Perceptron*)

MNIST: (Base de datos) modificada del Instituto Nacional de Estándares y Tecnología (del inglés *Modified National Institute of Standards and Technology Database*)

PCA: Análisis de componentes principales (del inglés *Principal Component Analysis*)

SOM: Mapa auto-organizado (del inglés *Self-Organized Map*)

# INTRODUCCIÓN

### Contexto

Durante las últimas dos décadas, ha habido un crecimiento exponencial de la robótica en el ámbito industrial, comercial y social. Es cada vez más común encontrar robots haciendo trabajos repetitivos que antes eran realizados por seres humanos, quienes ahora se encuentran con el tiempo necesario para encargarse de tareas más creativas y que agregan un mayor valor a la sociedad.

Estos robots, a su vez, se van haciendo cada vez mejores al realizar sus tareas, y hay industrias que no son concebibles en la actualidad sin robots: automotriz, aeroespacial, distribución, etc. Sin embargo, la introducción de los robots en la parte comercial y social ha sido un poco más lenta, principalmente debido a dos razones: primero, las habilidades necesarias en un robot en este ámbito son más complejas y requieren un tacto que hasta hace pocos años se consideraba como únicamente humano; y segundo, los seres humanos sentimos miedo a ser sustituidos.

En la actualidad, se pretende desarrollar herramientas que permitan a los robots debilitar ambas barreras a través de la comunicación directa con los seres humanos. En un principio, la comunicación entre un ser humano y un robot se hacía totalmente a través del código con el que se le programaba, de forma tal que sólo los técnicos podían enviar órdenes e interpretar respuestas. Este método de comunicación es suficiente en el caso de robots industriales, puesto que su interacción con el resto del personal en el área de trabajo debe ser mínimo, y en el mejor de los casos, nulo. Sin embargo, en el caso de robots enfocados en el ámbito comercial y social la situación cambia radicalmente, ya que su principal función es la de interactuar con seres humanos, por lo que se espera que la comunicación sea a través de un lenguaje más natural.

Ya existen muchos trabajos que tratan la interpretación del lenguaje natural para robots, ya sea escrito o hablado. En 2014 se creó el primer programa no solo capaz de interpretar el lenguaje natural escrito, sino que fue capaz de pasar el test de Turing (aunque algunos no lo consideran cierto). Por otra parte, en la actualidad nuestra sociedad se ha inundado de programas capaces de interpretar nuestro lenguaje y ofrecernos respuestas, desde las contestadoras automáticas de las grandes compañías hasta chatbots. Es por ello que este trabajo final de máster trata el desarrollo de dos herramientas que permitirían la interacción de los robots de una forma diferente.

Primero, una herramienta capaz de ayudar al robot a interpretar las expresiones faciales del interlocutor, permitiendo así reducir la brecha entre las meras palabras expresadas por este, y un significado que puede variar dependiendo la intención que va por detrás de ellas.

Segundo, otra herramienta capaz de interpretar el lenguaje de señas de sordomudos en su versión americana, de forma tal que un robot pueda ser capaz de interpretar lo que una persona con déficit de audición intente comunicarle sin la utilización de un teclado. Esta herramienta está más enfocada a robots humanoides que tienen interacción directa con el público, en el que la inclusión de personas con dificultades de habla puede marcar una diferencia positiva para la percepción de los robots sociales.

A lo largo de este trabajo, se explicarán las tecnologías y técnicas utilizadas durante su realización, así como sus posibles aplicaciones en casos de la vida real donde se le pueda sacar el máximo provecho posible.

### Estructura del documento

Este documento consta de una estructura donde se analiza y explica el desarrollo de este trabajo.

Luego de la introducción que se está realizando en el primer capítulo, el segundo realizará una revisión del estado del arte de la tecnología actual sobre la que se está apoyando el presente desarrollo. Aquí también se mostrarán trabajos previos que han inspirado las estrategias seguidas.

El tercer capítulo explicará la motivación para su realización con mayor detalle, así como una exposición clara de los objetivos que se buscan cumplir con su realización.

El capítulo cuatro busca mostrar los detalles de la metodología seguida durante el desarrollo del proyecto, la selección de técnicas y métodos utilizados durante el mismo, así como su integración para obtener cada uno de los programas finales.

El capítulo cinco expone los resultados obtenidos en la actual implementación, tanto durante el desarrollo de los programas como ejemplos de su aplicación real.

El sexto capítulo y séptimos capítulos incluyen las posibles aplicaciones y mejoras que se le pueden dar al proyecto, y en conjunto dan una idea sobre el futuro que pueda tener este proyecto.

Finalmente, todos los documentos y referencias utilizadas a lo largo de este trabajo serán anexadas al final.

# MOTIVACIÓN Y OBJETIVO

Las principales motivaciones para la elección de este proyecto se explicarán a continuación, y de igual manera, tanto los objetivos generales como los objetivos específicos de este trabajo serán desglosados según su índole.

### Motivación

La consultora multinacional Ernst & Young, en su sede en Madrid, cuenta desde hace varios años con un Centro de Excelencia especializado en Inteligencia Artificial, conocido como wavespace, que es único en la empresa a nivel mundial, y en donde se trabajan principalmente proyectos de innovación especializados en automatización de procesos, desarrollo web, visión artificial, Big Data, visualización de datos, realidad virtual, realidad aumentada y robótica de humanoides, para clientes internos y externos, nacionales, europeos e internacionales.

Este centro tiene a su disposición dos robots en su auditorio de exposiciones, modelos Pepper y Nao, los cuales han sido modificados y reconfigurados con avanzados programas de visión artificial, reconocimiento de imágenes, identificación de personas, interpretación de lenguaje natural, identificación de emociones mediante el contenido del mensaje, traducción en vivo entre varios idiomas y mapeo para desplazamiento dentro del propio auditorio.



Figura 1: Robots Pepper y Nao presentes en el wavespace de EY

Estos robots son el centro de atención de clientes y demás empleados de la organización, por lo que se quiere seguir aumentando el abanico de posibilidades que puedan ofrecer al público.

Es por ello que la propuesta de equiparlos con servicios capaces de mejorar la comunicación visual con seres humanos ha interesado a la junta directiva, razón por la cual ha puesto a disposición los recursos que permitan su desarrollo en cooperación con la Universidad Politécnica de Madrid.

### Objetivo general

El objetivo principal de este proyecto es el desarrollo de programas en Python basados en inteligencia artificial, aprendizaje automático y visión artificial que puedan ser utilizados por robots para mejorar la comunicación visual con seres humanos en vivo, y que puedan funcionar como servicios en la nube de forma tal que sean extensibles a diferentes locaciones geográficas.

### Objetivos específicos

De manera más específica, este proyecto busca enfocarse en la solución de dos problemas bien diferenciados, pero que comparten muchas características en común desde el punto de vista técnico, como lo son el reconocimiento de las expresiones faciales de seres humanos y la interpretación del lenguaje de sordomudos, en su versión americana, utilizando técnicas avanzadas de visión artificial y aprendizaje automático, más específicamente, redes neuronales artificiales.

A pesar de sus puntos en común, cada uno de estos programas será un módulo independiente al que se podrá acceder y ejecutar independientemente, siempre y cuando el hardware utilizada cumpla con unos requerimientos mínimos (principalmente, un dispositivo de vídeo de entrada).

A continuación se especifican los objetivos de cada uno de los dos programas:

#### Reconocimiento de expresiones faciales humanas

Para el programa encargado del reconocimiento de las expresiones faciales humanas, se espera alcanzar los siguientes objetivos.

* Filtrar dentro de cada cuadro de vídeo los objetos cuyo color se encuentre dentro de los rangos de la piel humana.
* Identificar de entre los objetos filtrados como “piel humana”, aquellos que sean rostros humanos.
* Realizar el seguimiento de los objetos identificados por la cámara cuadro a cuadro, mientras se encuentren dentro del campo de visión de la misma.
* Entrenar una red neuronal artificial capaz de clasificar los rostros humanos, de acuerdo con las siguientes expresiones faciales mostradas:
  + Neutral
  + Alegría
  + Tristeza
  + Ira
  + Miedo
  + Desagrado
  + Sorpresa
* Clasificar los rostros observados (limitando su número pero no disminuir desempeño) en el campo de visión de la cámara utilizando la red neuronal artificial mencionada anteriormente.
* Mostrar en vivo y sobre el vídeo la región de interés identificada y la clasificación realizada.

A pesar de que para este proyecto se desea mostrar en resultado en la pantalla, para su aplicación sólo se requeriría transmitir al robot el resultado de la clasificación.



Figura 2: Ejemplo de sujetos mostrando diferentes expresiones faciales

La figura anterior muestra cuatro sujetos demostrando diferentes expresiones faciales de forma actuada, con un fondo monocromático y sin objetos.

#### Interpretación de lenguaje de sordomudo versión americana

Se esperan alcanzar los siguientes objetivos para la interpretación del lenguaje de sordomudos, en su versión americana:

* Filtrar dentro de cada cuadro de vídeo los objetos cuyo color se encuentre dentro de los rangos de la piel humana.
* Identificar de entre los objetos filtrados como “piel humana”, aquellos que no correspondan con rostros humanos, los cuales se asumirán como manos.
* Realizar el seguimiento de los objetos identificados mientras se encuentren dentro del campo de visión de la cámara.
* Entrenar una red neuronal artificial capaz de clasificar las posiciones de las manos correspondientes a cada una de las letras del lenguaje de sordomudos (versión americana), con la excepción de la “J” y la “Z”, por no ser estáticas.
* Clasificar la posición las manos observadas en el campo de visión de la cámara utilizando la red neuronal descrita anteriormente.
* Mostrar en vivo y sobre el vídeo la región de interés identificada y la clasificación realizada.

En este caso también se desea mostrar en resultado en la pantalla, a pesar de que para su aplicación sólo se requeriría transmitir al robot el resultado de la clasificación.



Figura 3: Alfabeto del lenguaje de señas americano

La figura anterior muestra el lenguaje de señas utilizado para este proyecto, excluyendo aquellas que implican movimientos de la mano (“J” y “Z”) al requerir una metodología diferente a la que se está estudiando, y que por lo tanto escapan del alcance del mismo.

De igual manera, en este caso también se está utilizando un fondo monocromático y sin objetos, lo cual deberá ser replicado durante su uso.

# ESTADO DEL ARTE

Como se ha mencionado anteriormente, el presente proyecto está basado en dos programas diferenciables. Sin embargo, ambos programas tienen en común una estructura con dos módulos diferenciables: el desarrollo de modelos de aprendizaje automático o Machine Learning (ML), y el procesamiento de imágenes usando técnicas de visión artificial. Ambos módulos fueron desarrollados en lenguaje Python.

### Machine Learning y redes neuronales artificiales

El Machine Learning (ML) es una rama de las ciencias de la computación y de la inteligencia artificial, que consiste en desarrollar técnicas que permitan a programas informáticos aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos[1].

Una de las técnicas de ML más utilizadas, y en la que se basó este proyecto, son las redes neuronales artificiales. Estas se definen como un modelo computacional inspirado en el comportamiento observado en su homólogo biológico. Consisten en un conjunto de unidades, denominadas neuronas artificiales, las cuales se encuentran conectadas entre sí transmitiendo señales basado en ciertas condiciones. La información de entrada atraviesa la red neuronal (donde se somete a diversas operaciones) produciendo unos valores de salida[2].

En forma general, estos modelos se utilizan principalmente con dos objetivos: predicción y clasificación. Para ambos casos, la metodología utilizada es similar: se debe construir un dataset con una gran variedad de casos que se deben procesar. Este dataset se divide en una parte que será utilizada para el entrenamiento del algoritmo, y otra que será utilizada para validar los resultados del mismo.

Hay dos tipos de redes neuronales: supervisadas y no supervisadas. La principal diferencia entre ambas es que, en el caso de las redes neuronales supervisadas, el dataset incluye el valor deseado para cada muestra, mientras que en el caso no supervisado, dicho valor no se conoce.

Durante la etapa de entrenamiento, la conexión entre cada una de las neuronas artificiales varía su peso (el umbral a partir del cual dejará pasar la señal o no, es decir, se “activará”) dependiendo de si la salida se acercó al valor esperado o no, y este proceso se repetirá iterativamente hasta que el programador considere que el aprendizaje ha culminado. Luego, el modelo se evalúa utilizando la parte del dataset que no se usó para el aprendizaje, de forma tal que sea posible evaluar la precisión del mismo.

Es importante tomar en consideración que la decisión sobre si la etapa de aprendizaje ha culminado o no depende enteramente del programador, y la efectividad de la red neuronal puede depender en gran parte de su experticia: si se detiene el aprendizaje de la red neuronal antes de tiempo, sus resultados no serán confiables ya que no logró analizar suficientes casos; si se alarga la etapa de aprendizaje demasiado tiempo, la red neuronal podría caer en el estado de “sobreentrenamiento”, y no ser confiable tampoco.

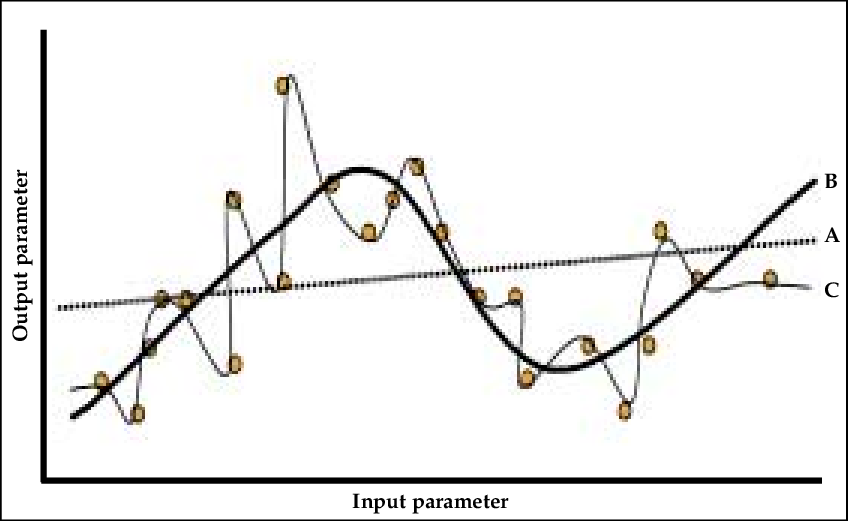


Figura 4: Ejemplo de sobre-entrenamiento y bajo entrenamiento

En la figura anterior se pueden observar los tres posibles casos que se mencionaron anteriormente.

Para una función coseno con un ruido introducido, se tomaron 23 muestras, y se entrenaron tres redes neuronales para que encontraran la función original sin ruido:

* Caso A: esta red neuronal no fue entrenada por suficiente tiempo, por lo que, si bien muestra una línea de tendencia similar a la esperada, no es capaz de seguir la función original.
* Caso B: esta red neuronal fue entrenada por un tiempo suficiente, y se puede observar cómo sigue el valor de la función a pesar de casi no tocar ninguna muestra. Es decir, está logrando el objetivo de obtener la función original sin el ruido.
* Caso C: esta red neuronal ha sido sobreentrenada, razón por la cual se ha visto afectada considerablemente por el ruido y se ha alejado de la forma de la función original. Se podría decir que en lugar de “aprender” de la muestra, lo que ha hecho ha sido “memorizarla”.

#### Perceptrón multicapa

Existen diversas morfologías para las redes neuronales, siendo el perceptrón multicapa (MLP) el más extendido en la actualidad.

Este consiste en una red neuronal artificial conformada por múltiples capas, lo cual le otorga la capacidad de resolver problemas que no son linealmente separables. Cada una de estas capas está conectada con la capa anterior a través de sus neuronas. Al igual que en su homóloga biológica, para que el estímulo pase de una neurona a otra, este debe alcanzar el umbral de activación. Este umbral de activación puede variar dependiendo del programador, pero las dos más utilizadas son las siguientes funciones sigmoides:

En el caso en el que las capas están totalmente interconectadas, cada salida de una neurona de la capa “i” es entrada de todas las neuronas de la capa “i+1”, aunque dependiendo del caso, este diseño puede variar ligeramente[3].

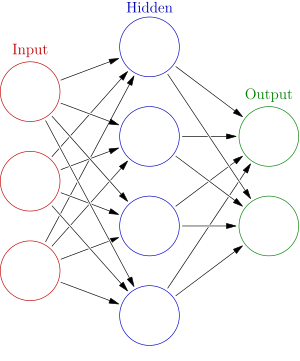


Figura 5: Arquitectura del perceptrón multicapa

Entre los parámetros que se suelen ajustar durante el entrenamiento, se encuentran:

* La función de activación (usualmente sigmoide).
* El número de capas ocultas.
* El número de neuronas en cada capa oculta.
* El tamaño del lote que se utiliza en cada iteración.
* El número de ciclos (un ciclo es el número total de iteraciones que debe ocurrir hasta que todos los lotes que conforman el dataset hayan sido procesados).

Adicionalmente, el número de neuronas en la capa de entrada es definido por el número de dimensiones de cada una de las muestras, y el número de neuronas en la capa de salida, depende del número de dimensiones de la respuesta esperada.

#### Mapa auto-organizado

Los mapas autoorganizados (SOM por sus siglas en inglés) son un tipo de red neuronal estrictamente no supervisada, que utiliza una función de vecindad para preservar las características topológicas del espacio de entrada, permitiendo visualizar vistas de baja resolución en datos de alta dimensión[4].

Asociado a cada neurona hay un vector de pesos, de la misma dimensión que los vectores de entrada, y con una posición en el mapa. Usualmente, las neuronas se disponen en una red de dos dimensiones, ya sea hexagonal o rectangular. Los mapas auto-organizados describen el mapeo de un espacio de mayor dimensión a uno de menor dimensión.

El procedimiento para ubicar un vector del espacio de los datos en el mapa es encontrar la neurona con el vector de pesos más cercano (menor distancia métrica) al vector del espacio de los datos. Esto puede traer como consecuencia que dos puntos que son vecinos en el espacio de mayor dimensión, puedan encontrarse lejos en el espacio dimensional inferior.

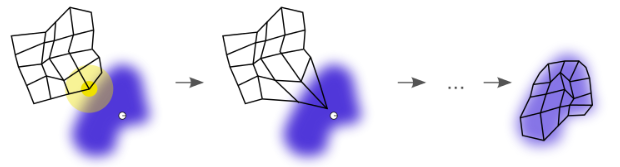


Figura 6: Mapa auto-organizado durante proceso de entrenamiento

El entrenamiento utiliza el denominado aprendizaje competitivo. Cuando una matriz de datos es procesada, la distancia euclidiana a todos los vectores de pesos de las neuronas es calculada. La neurona cuyo vector de pesos sea más similar a la entrada es llamada unidad de mejor correspondencia, y tanto sus pesos como los de las neuronas cercanas a ella en la cuadrícula del SOM son ajustados hacia el vector de entrada. La magnitud de los cambios se decrece con el tiempo (en iteraciones) y con la distancia desde la unidad de mejor correspondencia.

Este proceso es repetido para cada vector de entrada durante un número de ciclos usualmente grande. La red va asociando las neuronas de salida con grupos o patrones en el conjunto de entrenamiento.

#### Red neuronal convolucional

Otra morfología muy utilizada en la actualidad es la red neuronal convolucional (CNN en inglés). Este es un tipo de red neuronal donde las neuronas se corresponden a los campos receptivos imitando a sus equivalentes biológicos en la corteza visual del cerebro[5].

La función de activación es típicamente una capa de rectificación (RELU en inglés), la cual no más que una función de activación simple y centrada en cero:

La morfología de esta red es una variación del perceptrón multicapa explicado anteriormente, y al ser aplicado a matrices bidimensionales, se muestran particularmente útiles en problemas de visión artificial, especialmente pero no limitado a casos de clasificación y segmentación de imágenes.

Su arquitectura consiste en múltiples capas de filtros convolucionales que pueden tener una o más dimensiones. Entre cada una de las capas, se puede añadir una función para realizar un mapeo causal no-lineal. Las primeras capas están compuestas de neuronas convolucionales y de reducción de muestreo, con el fin de extraer las características de la imagen en general, y no de cada uno de sus puntos. En las últimas capas se encuentran neuronas de perceptrón sencillas para realizar la clasificación final sobre las características extraídas.

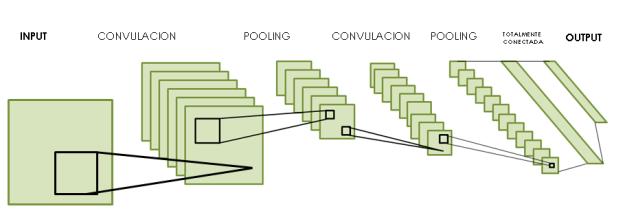


Figura 7: Arquitectura de una red neuronal convolucional

La fase de extracción de características es equivalente al proceso estimulante en las células de la corteza visual, y está compuesta de capas alternas de neuronas convolucionales y neuronas de reducción de muestreo. En la medida que se avanza a través de las capas, se reduce el número de neuronas, siendo las neuronas en capas finales menos sensibles a perturbaciones en los datos de entrada, y activándose cada vez por características más complejas[6].

Las neuronas convolucionales se diferencian de los perceptrones ya que son procesadores de matriz cuadrada que realizan una operación de convolución sobre un campo en 2D, en lugar de trabajar sobre un solo píxel. La salida de cada neurona se calcula de la siguiente manera:

Donde la salida **Yj** de una neurona **j** es una matriz que se calcula por medio de la combinación lineal de las neuronas **Yi** en la capa anterior **i**, cada una de ellas, operadas con el núcleo convolucional **K** correspondiente a esa conexión. Esta cantidad es sumada a una influencia **b** y luego se pasa por una función de activación  **(·)** no lineal.

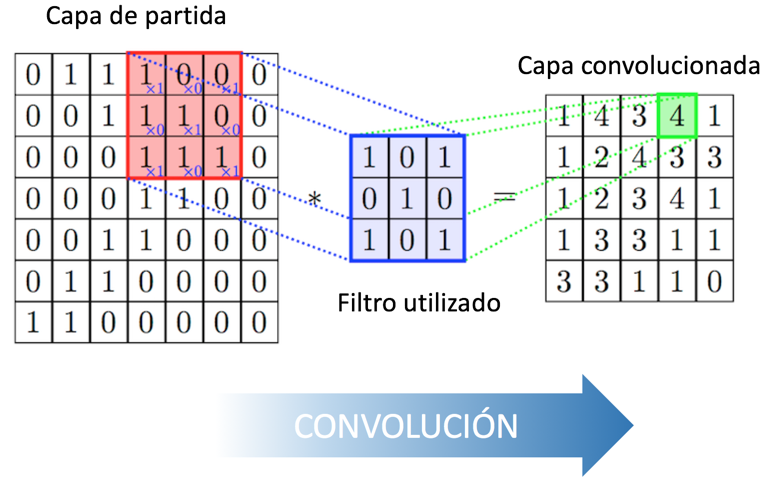


Figura 8: Funcionamiento de una neurona convolucional

Las neuronas de reducción de muestreo permiten aumentar la tolerancia de la red neuronal antes pequeños cambios entre dos imágenes (por ejemplo, una misma imagen pero trasladada unos pocos píxeles). Para ello, estas neuronas utilizan operaciones de subsampling, principalmente max-pooling, que permiten obtener tomar un valor resumen de las características de una cierta área, reduciendo de esta manera el tamaño de los datos.

* Subsampling: es una operación constante que permite la reducción de la muestra evitando perder sus características principales. En CNN, usualmente la operación utilizada es el max-pooling.
* Max-Pooling: se utiliza para reducir el tamaño de la muestra, tomando una matriz cuadrada de un tamaño definido (2x2, 3x3…) a lo largo de toda la imagen, y sustituyendo el subconjunto de píxeles dentro de cada matriz por un único nuevo píxel con el valor máximo encontrado.

Las neuronas de clasificación son la última etapa de la red, y su morfología es la de un MLP clásico, puesto que los datos que recibe ya han sido depurados y la clasificación se puede realizar de forma genérica.

Entre los parámetros que se suelen ajustar durante el entrenamiento, se encuentran:

* La función de activación (usualmente una capa RELU, que puede ser suavizada o no).
* El número de capas ocultas.
* El número de neuronas en cada capa oculta.
* El tamaño del lote que se utiliza en cada iteración.
* El número de ciclos (un ciclo es el número total de iteraciones que debe ocurrir hasta que todos los lotes que conforman el dataset hayan sido procesados).
* El tamaño del filtro de convolución.
* El tamaño de la matriz de max-pooling.

De igual manera, el número de neuronas en la capa de entrada y de la capa de salida, depende respectivamente del número de dimensiones de la imagen de entrada y del número posible de respuestas esperadas.

### Visión artificial

La visión artificial es una rama de las ciencias del computador que incluye métodos y técnicas para adquirir, procesar, analizar y comprender imágenes en el mundo real, y transformarlos en información numérica que pueda ser interpretada por un computador. Esta interpretación se logra a través de distintos campos como la geometría y la estadística, principalmente. La adquisición de los datos se consigue por varios métodos, que van desde archivos de imágenes simples, secuencias de vídeo grabado o en vivo, o hasta datos multidimensionales obtenidos por un escáner médico.

Entre las principales tecnologías que utilizan visión artificial, se encuentran:

* Limpieza de imágenes
* Reconocimiento de objetos
* Seguimiento de objetos (tracking)
* Reconstrucción de escenas
* Segmentación de imágenes

Durante el desarrollo de un algoritmo de visión artificial, se debe tomar en cuenta que, si bien la calidad y resolución de imagen ayuda a obtener mejores resultados, el procesamiento de estas imágenes va a consumir un mayor procesamiento, por lo que se debe buscar un punto intermedio sobre el cual trabajar.

Una manera de mitigar este problema es realizar transformaciones sobre las imágenes para reducir la carga sobre el procesamiento. Algunas de estas transformaciones son:

* Reducción de la resolución de la imagen (resize)
* Recorte del área de interés (crop)
* Transformación de la imagen a escala de grises
* Análisis de componente principal (PCA por sus siglas en inglés)

### Python

El lenguaje de programación Python cuenta con un importante número de librerías tanto para trabajar con redes neuronales artificiales como para aplicar técnicas de visión artificial. Es además un lenguaje ampliamente utilizado y con documentación completa y accesible.

#### Tensorflow y Keras

Las librerías Tensorflow y Keras son las más utilizadas en el campo de la investigación para trabajar con ML en general, por contar con funciones y métodos especializados para trabajar con redes neuronales artificiales.

Tensorflow fue desarrollada por Google específicamente para ML. Su unidad de procesamiento es el Tensor, el cual es una construcción específica similar a un vector, y es la base arquitectónica de este lenguaje.

Keras por su parte, es una librería basada en Tensorflow y que permite programas y entrenar redes neuronales de varios tipos en pocas líneas de código. De esta forma, una vez decidida su arquitectura, no es necesaria su construcción capa por capa, sino que se debe llamar a la función correspondiente, introducir los parámetros necesarios, y utilizar métodos para entrenar la redes o probar sus resultados.

Este librería cuenta con todos los métodos presentes en Tensorflow para el desarrollo de redes neuronales, entre los que destacan reductores de medias, optimizadores, convolucionadores, funciones de max-pooling, entrenadores de redes, calculadores de eficacia de la red, entre otros.

#### OpenCV

La librería OpenCV es la librería más amplia y versátil que se puede encontrar en este lenguaje para el procesamiento de imágenes. Cuenta con funciones y métodos de fácil uso que permiten realizar una amplia serie de acciones sobre archivos de imágenes: recortes, transformaciones de color, erosión, dilatación, seguimiento de objetos, tratamiento de ruido, entre otros.

Para lograr esto, Python aprovecha la interpretación de una imagen como una matriz que hace el computador, para realizar distintas operaciones matriciales y realizar el procesamiento.

OpenCV cuenta con todas las operaciones necesarias para tratar cualquier imagen de forma secuencial y extraer la información que sea necesaria de esta, de forma tal que puede ser utilizada para el tratamiento en vivo de imágenes en vídeo.

### Proyectos similares

En el pasado reciente, ha habido un interés cada vez mayor en el uso de tecnologías como las mencionadas anteriormente para la clasificación de imágenes de todo tipo. Las técnicas utilizadas en cada uno de ellos varían dependiendo de las necesidades del problema, pero tienen en común el procesamiento previo de imágenes y su posterior clasificación mediante redes neuronales.

#### Detecting Skin in Images & Video Using Python and OpenCV

Este trabajo, publicado[7] abiertamente por Adrian Rosebrock, consiste en la elaboración de un algoritmo capaz de detectar y filtrar la piel humana en un vídeo en vivo mediante el uso de la librería OpenCV para visión artificial, en Python.

Utilizando la librería OpenCV, el autor transforma la imagen a escala HSV y luego define un rango de colores dentro del cual considera que está la piel humana. El rango de colores utilizados en HSV varían entre:

* [0,48,80]:
* [20,255,255]:

A partir de este punto, todos los píxeles de la imagen que estén dentro de estos valores se consideran piel, por lo que se pasa a transformar la imagen a binario (blanco y negro), utilizando como filtro una operación que sólo permite píxeles dentro de dicho umbral.

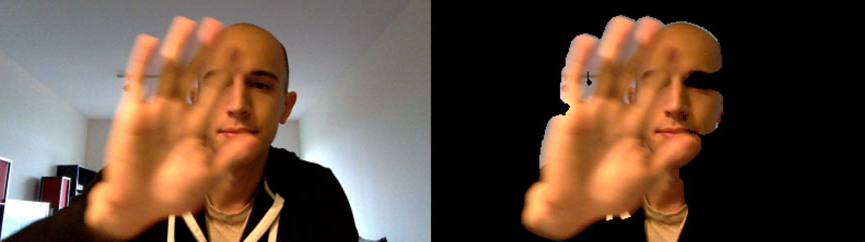


Figura 9: Detección de piel y filtrado en imagen

Finalmente, utilizando esta transformación como una máscara, es posible que sólo la piel sea observada en la imagen final.

#### Clasificación de número escritos a mano con redes neuronales

Desde hace varios años, este se ha vuelto un problema clásico en el estudio de las redes neuronales. Fue planteado utilizando la base de datos modificada Instituto Nacional de Estándares y Tecnología (en inglés MNIST), que consiste en una base de datos indexada de decenas de miles de muestras de dígitos escritos a mano provenientes de archivos civiles de EEUU. Está normalizada como imágenes en escala de grises y resolución de 28x28 píxeles.

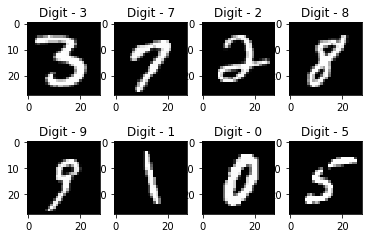


Figura 10: Ejemplo de dígitos de la base de datos MNIST

En un trabajo anterior, se procedió a analizar la efectividad de diferentes clasificadores sobre sobre esta base de datos para lograr identificar los dígitos escritos a mano. Se utilizó 10.000 muestras para esto, de las cuales el 80% se utilizó para entrenamiento y el 20% restante para pruebas.

Para reducir el tiempo de procesamiento y optimizar el proceso de clasificación, se procedió primero a identificar píxeles que no estuvieran aportando información (principalmente en las esquinas). Para ello se revisó el valor de cada uno de los píxeles para las 10.000 imágenes, y aquellos que tuvieran exactamente el mismo valor en todas ellas, se eliminó. Este proceso redujo la dimensionalidad en un 10%.

Posteriormente se aplicó un análisis de componente principal (en inglés, PCA) para reducir aún más la dimensionales perdiendo la menor información posible. El número de dimensiones utilizadas dependió del clasificador que se estudió en cada caso.

Los resultados obtenidos fueron los siguientes:

* Clasificador KNN: es un método de clasificación supervisado, que estima la probabilidad de un elemento de pertenecer a una clase dependiendo de la clase a la que pertenezca el K número de vecinos. El porcentaje de acierto para este clasificador fue del 94,55%.
* Clasificador bayesiano: es un clasificador probabilístico que aplica el Teorema de Bayer y los supuestos de independencia fuertes (ingenuos) entre las clases para determinar a dónde pertenece cada muestra. En lugar de crear un límite en función de la posición de cada muestra individual, este clasificador crea el límite en función de la distribución probabilística de cada conjunto (clase) de muestras. El porcentaje de acierto para este método fue del 85,65%.
* Clasificador MPL: este tipo de red neuronal se entrena y ajusta sus ponderaciones para reproducir el mismo resultado con entradas similares. Se decidió utilizar un PCA para reducir los datos hasta 100 dimensiones, y el resultado final fue del 93,05%.
* Clasificador SOM: esta red neuronal se comporta como un mapa que trata de cubrir la mayoría de las muestras de datos, de modo que a cada muestra se le asigna a una neurona. Se utilizó una topología de 60x7 neuronas y un PCA para reducir la dimensionalidad a 200. El porcentaje de acierto para este clasificador fue del 73,75%.

#### Sign Language Recognition with Unsupervised Feature Learning

Este trabajo fue publicado[8] por Justin Chen en la Universidad de Stanford (EEUU) en 2014, y trata el reconocimiento del lenguaje de señas americano para sordomudos utilizando redes neuronales no supervisadas y procesamiento de imágenes. Está inspirado en el trabajo sobre clasificación de números escritos a manos, y utiliza una dataset de 1200 muestras para cada una de las señas de las 10 primeras letras del alfabeto latino, omitiendo la ”J” al requerir movimiento.

Para su procesamiento, se experimentó utilizando diferentes aproximaciones:

* Segmentación de bordes: esta estrategia pretendía utilizar la función “Canny” de OpenCV, la cual consiste en un detector de bordes en la imagen. Para ello, realiza derivaciones tanto vertical como horizontalmente, de manera que identifique los cambios repentinos en la imagen como bordes. Una vez teniendo esta imagen de bordes, se procede a la clasificación.
* Identificación de puntas de los dedos: la estrategia a seguir en este caso es utilizar técnicas geométricas para definir la posición aproximada de las puntas de los dedos. Para ello, se identifica la mano como una curva geométrica cerrada, y se procede a identificar los vértices de la misma mediante el método “Contour” de OpenCV. Utilizando el número de dedos identificados, así como su posición relativa al centro de masa de la imagen, se procede a la clasificación.
* Segmentación por color de piel: esta última estrategia hace una conversión de la imagen a blanco y negro, utilizando como umbral para cada píxel el hecho de si pertenece a el rango definido como piel humana. Posteriormente, se aplica esta imagen en blanco y negro como una máscara en la imagen principal, se forma tal que sea posible segmentar la mano y aplicación la clasificación directamente.

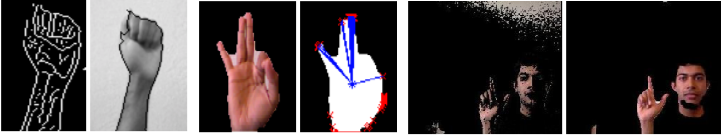


Figura 11: Métodos de procesamiento de imagen (de izquierda a derecha: segmentación de bordes, identificación de puntas de los dedos y finalmente, segmentación por color de piel)

Una vez detectada la mano, se procedió a recortarla y pasarla a una resolución de 32x32 píxeles, en escala de grises.

Para el entrenamiento, se utilizó un dataset de 1.200 imágenes por letra, para un total de 12.000 muestras. Cada muestra tiene una resolución de 32x32 píxeles. Por esta razón la topología utilizada fue de 32x32 neuronas (una por cada píxel) en la capa de entrada, 100 neuronas en la única capa oculta, y 10 neuronas en la capa de salida (una por cada letra). El número de iteraciones para el entrenamiento se definió en 400.

De las tres metodologías utilizadas, la que demostró un mejor rendimiento fue el uso de la segmentación por color de piel. Los resultados finales lograron superar el 95% de aciertos, lo cual es comparable con los resultados del clasificador de dígitos escritos a mano.

# METODOLOGÍA Y DESARROLLO

En esta sección se explicará con detalle la metodología utilizada para resolver la problemática planteada y los retos que surgieron durante la elaboración de este proyecto, así como el procedimiento que se ha seguido durante el desarrollo.

### Elección del lenguaje de programación

Como se mencionó en el segundo capítulo, el lenguaje de programación seleccionado para el desarrollo de este proyecto fue Python, más específicamente, en su versión 3.6.

Las razones de esta decisión se pueden resumir en dos puntos principales:

* Python es el lenguaje de programación más ampliamente difundido en la actualidad para proyectos basados en inteligencia artificial, más específicamente, en aprendizaje automático (ML, por sus siglas en inglés).
* Tanto el robot Pepper como el robot Nao tienen entre sus posibilidades el llamado de servicios en la nube que estén desarrollados en Python.

Si bien el primer punto tuvo un gran peso en la decisión final, fue el segundo el que descartó por completo a MatLab como una de las posibilidades para este proyecto, a pesar de ser la herramienta utilizada a lo largo del Máster y en la que se tenía una mayor experiencia.

Por otra parte, la elección de la versión 3.6 de Python por encima de la más difundida (v2.7) o la más reciente (v3.7) se debió al hecho de que todas las librerías que serían necesarias en este proyecto solo estaban disponible al mismo tiempo en esta versión al momento de iniciar el proyecto.

#### Librerías utilizadas

La siguiente lista enumera todas las librerías que son utilizadas durante la ejecución de los programas desarrollados en este proyecto:

* OpenCV (incluyendo OpenCV-Contrib)
* Numpy
* Os
* Shutil
* Math
* Scipy
* Skimage
* PIL
* Matplotlib
* Tensorflow
* Keras
* Sklearn
* Cmake
* Dlib

La instalación de las librerías mencionadas, luego de la instalación de la versión 3.6 de Python, es necesaria para la ejecución de los programas. De igual manera, para poder instalar las librerías Cmake y Dlib, es necesario instalar el módulo “Visual C++ Tools for Cmake” de Visual Studio 2015 o superior.

### Obtención de bases de datos para entrenamiento de redes neuronales

Para el entrenamiento de las redes neuronales artificiales necesarias para la solución de ambos problemas, se consideró primeramente la generación de las bases de datos utilizando una colección de fotografías creadas específicamente para el proyecto. Sin embargo, para alcanzar resultados aceptables, este trabajo tomaría mucho tiempo, por lo que se consideró el uso de bases de datos ya existentes y disponibles de forma gratuita.

#### Base de datos de rostros expresando emociones

La Universidad de Columbia, en EEUU, cuenta con una base de datos de 672 fotografías de 81 modelos, con una resolución de 506x650 píxeles en formato PNG que fueron recopiladas por la Dra. Nim Tottenham. Estas fotografías muestran el rostro de diferentes modelos expresando una serie emociones, tanto con la boca abierta como con la boca cerrada. Las expresiones mostradas por cada persona son:

* Neutral o calmado
* Alegría
* Tristeza
* Ira
* Miedo
* Desagrado
* Sorpresa

La composición demográfica del conjunto de 81 modelos es la siguiente: 58 de los participantes tienen origen caucásico (71,6%), 5 tienen origen africano (6,2%), 4 tienen origen latino (4,9%) y 4 tienen origen asiático (4,9%). Además, 51 de los modelos son mujeres (62,9%) y 30 son hombres (37,1%), y la media de edad es 23,4 años.

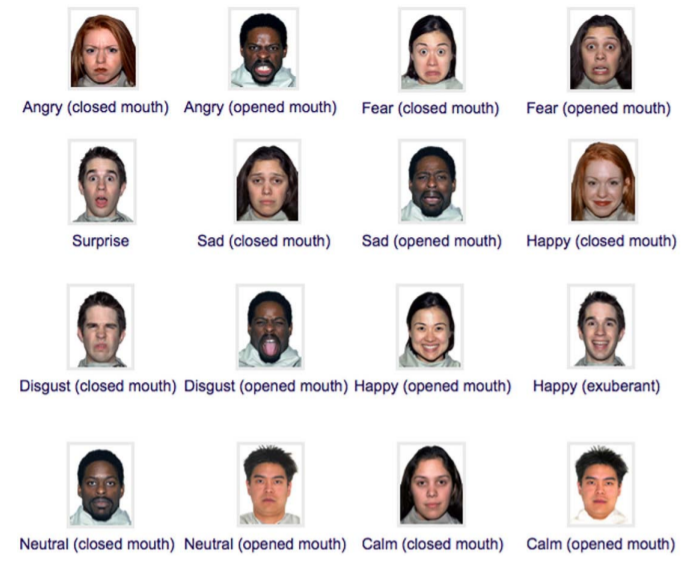


Figura 2: Dataset utilizado para identificación de emociones

Las expresiones faciales mostradas por los modelos son actuadas y exageradas, por lo que a pesar de ser claramente identificables, se nota que no son naturales. La etiqueta para la expresión mostrada en la fotografía forma parte del nombre del archivo de imagen.

El dataset se dividió aleatoriamente, de forma tal que 538 fotografías (80%) se utilizaron para el entrenamiento y 134 fotografías (20%) se utilizaron para pruebas.

#### Base de datos de manos haciendo señas del lenguaje de sordomudos

El mismo instituto mencionado en el segundo capítulo como propietario de la base de datos MNIST utilizada para el entrenamiento de programas para el reconocimiento de dígitos escritos a mano, también tiene a su disposición una base de datos con 34.627 fotografías de manos en las distintas posiciones del lenguaje de sordomudo americano, con la excepción de la “J” y la “Z”, al no ser estáticas.



Figura 3: Alfabeto contenido en el dataset

Cada una de estas imágenes se encuentra en resolución de 28x28 píxeles y en escala de grises de 256 bits. Además, se encuentran codificadas en un archivo CSV y en forma vectorial en lugar de matricial. Es decir, cada una de las imágenes tiene todos sus píxeles en una sola línea, y todas las líneas comienzan por un primer valor que es la etiqueta (donde “0” corresponde a “A”, “1” corresponde a “B”, “2” corresponde a “C”… “22” corresponde a “Y”). El total, incluyendo la etiqueta, es de 785 celdas por línea, y el total de líneas es de 34.627, divididas en 27.455 (80%) para entrenamiento, y 7.172 (20%) para pruebas.

### Pre-procesamiento del dataset

Con el objetivo de mejorar el rendimiento del clasificador, se procedió a realizar una serie de modificaciones a las imágenes originales del dataset utilizando las distintas herramientas que ofrece la librería OpenCV de Python.

#### Pre-procesamiento de imágenes de rostros

El primer paso que se realiza una vez cargadas las imágenes es transformarlas de RGB a escala de grises. Luego se utiliza el clasificador de Viola-Jones (incluido en la librería OpenCV) para identificar la zona de interés exacta donde se encuentra el rostro dentro de la imagen, y se recorta esta de manera tal que se minimice el número de píxeles que forman parte del fondo, cuello, ropa y cabello del modelo, dejando únicamente el rostro, y disminuyendo el potencial ruido.

Al tratarse de un dataset relativamente pequeño, se corre el riesgo de que la solución no responda correctamente ante mínimos cambios. Es por ello que se ha utilizado una función de expansión de datos, la cual consiste en replicar cada una de las imágenes haciendo ciertas modificaciones sobre ella, las cuales fueron:

* Rotación aleatoria de imágenes entre -15º y 15º.
* Introducción de ruido gaussiano.
* Inversión horizontal de la imagen.

Al hacer esto, el dataset original de 672 imágenes pasó a tener 10 veces más datos.

Por último, se procede a reducir la resolución de las imágenes a 32x32 píxeles. Esto se hace puesto que si se utilizara el dataset original, la capacidad de procesamiento necesaria para entrenar una red neuronal de tales dimensiones no está disponible en un ordenador convencional. Se realizaron varias pruebas y esta dimensión fue la que mejor proporción entre detalle y tamaño produjo.

#### Pre-procesamiento de imágenes de señas

De igual forma que en el caso anterior, es necesario hacer modificaciones a las imágenes originales del dataset para obtener los mejores resultados.

En este caso, las imágenes ya se encuentran en escala de grises de 256 bits y una resolución de 28x28, por lo que pueden ser utilizadas en un ordenador convencional. Sin embargo, se aplicó nuevamente la expansión de datos con los mismos parámetros que en el punto anterior para mejor la calidad de los mismos.

Cabe destacar que la inversión horizontal en este caso es muy importante, ya que todas las fotografías son de manos diestras, y puesto que se espera que el programa resultante sea capaz de interpretar personas diestras y zurdas por igual, es necesario tener también la versión de cada una de las señas en una mano zurda.

### Entrenamiento de redes neuronales artificiales

El proceso de entrenamiento de las redes neuronales artificiales para cada uno de los dos problemas fue muy similar, siendo la única diferencia entre ambas el número de neuronas de entrada y el número de neuronas de salida, debido a que tanto la resolución de las imágenes de entrada como el número de posibles resultados varía en cada caso.

La librería Keras de Python cuenta con el método “Train”, el cual permite entrenar una red neuronal definida previamente y únicamente necesitando como parámetros el dataset de entrada, las etiquetas de dicho dataset, la red neuronal propiamente dicha, el número de ciclos, el tamaño de los lotes, y la función de entropía.

* Dataset de entrada: ambos dataset, una vez procesados, son utilizados como una matriz 3D.
* Etiquetas del dataset: en ambos casos, las etiquetas se guardaron en una matriz separada y con codificación “One hot pixel”. Es decir, para cada etiqueta se genera un vector con tamaño igual al número de diferentes etiquetas posibles (7 para rostros, 24 para señas), en donde todos los valores son iguales a cero excepto por aquel que corresponda con el valor de la etiqueta, cuyo valor es uno.
* Morfología: al tratarse de imágenes con una resolución similar en orden de magnitud (28x28 y 32x32), se utilizó una morfología idéntica, cuyo filtro convolucional es una matriz cuadrada de tamaño 3.
* Número de ciclos: tras realizar varias pruebas, se observó que la precisión de la red neuronal se estanca alrededor del décimo ciclo, por lo que se decidió establecer este número para evitar sobre-entrenamiento.
* Tamaño del lote: se definió el tamaño del lote en 128 para evitar sobrecargar el procesamiento del ordenador.
* Función de entropía: en ambos casos se eligió “Binary Cross-Entropy”.

Una vez la red neurona artificial ha sido entrenada, se procede a probarla con el dataset que se dejó para validación. Una vez conocido el porcentaje de acierto del mismo, se procede a guardarla para su uso en datos reales.

### Clasificación en vivo

La obtención de imágenes en vivo se realiza utilizando la cámara web integrada del ordenador utilizando las herramientas ofrecidas por la librería OpenCV. Por cada uno de los cuadros de vídeo obtenidos, se debe identificar el objeto de interés utilizando como referencia el color de la piel humana.

#### Identificación de objetos en el rango de color de la piel humana

Para extraer de la imagen los objetos de interés, se utiliza el siguiente algoritmo para la identificación de piel humana:

* Transformar la imagen de RGB a HSV.
* Generar máscara con sólo los píxeles que se encuentren en el rango entre [0, 48, 80] y [20, 255, 255] (definido como piel humana).
* Erosionar la máscara para eliminar los objetos muy pequeños o píxeles aislados.
* Dilatar la máscara para no perder detalle en los bordes de los objetos identificados.
* Binarizar la máscara con sólo los píxeles más “brillantes” (aquellos con un valor menor de 60 sobre 255).
* Buscar los contornos más grandes en la imagen y filtrar aquellos que no tengan al menos un 50% de píxeles con el valor dentro del rango para piel humana. El número máximo de contornos que se utilizarán puede ser seleccionado por el usuario.



Figura 4: Secuencia de imágenes mostrando tratamiento seguido al identificar piel humana

Cada uno de estos contornos corresponde a un objeto cuyo color se encuentra dentro del rango de color de la piel humana.

#### Seguimiento de objetos

Una vez conocidos los contornos de los objetos de interés, se procede a hacer su seguimiento para cada cuadrado subsiguiente. En este caso, al existir la posibilidad de analizar más de un objeto a la vez, se utilizar la función “Multitracker”, y el tracker utilizado es del tipo MIL.

Se extrae a partir del contorno el rectángulo que envuelve el objeto analizado, y estas coordenadas son utilizadas para configurar el tracker. Una vez que el tracker está configurado, solo es necesario actualizarlo en cada nuevo cuadro de vídeo, y si se pierde de vista el objeto, el tracker se elimina. Adicionalmente, se imprime en la imagen original un rectángulo azul con las coordenadas mencionadas anteriormente mientras el tracker esté activo.

#### Identificación de rostros humanos

Para definir si el objeto encontrado es un rostro humano o no, se procedió a utilizar un clasificador cascada de patrones binarios locales (*LPE Cascade* en inglés), incluido en la librería OpenCV. Este clasificador identifica si el objeto es similar a un rostro visto frontalmente. En caso negativo, se asume que se trata de una mano.

#### Clasificación mediante redes neuronales artificiales

Una vez que se ha identificado el objeto de interés como rostro o como mano, se procese a realizar la clasificación que corresponda.

En ambos casos, la región de interés identificada se recorta de la imagen total, se convierte a escala de grises de 256 bits y se reduce su resolución (a 32x32 píxeles en caso de ser un rostro, y a 28x28 píxeles si no lo es, ya que se asume que es una mano).

Una vez que la imagen ya tiene las mismas características de las imágenes del dataset, se procede a utilizar la red neuronal que se entrenó previamente. El resultado de dicha clasificación va a ser un vector del mismo tamaño que el vector de etiquetas del dataset utilizado durante el entrenamiento, y el valor máximo en dicho vector corresponderá con la clasificación que se está arrojando.

Por último, se utiliza este valor para calcular el resultado real (una expresión o una letra, según sea el caso) y se imprime en la imagen original para su visualización en vivo.

# RESULTADOS

Como se mencionó en el capítulo anterior, ambos datasets están divididos en una proporción 80% - 20%, donde el grupo mayor se utilizó para el entrenamiento de las redes neuronales artificiales, y el restante 20% se utilizó para probar la correspondiente red.

### Pruebas post-entrenamiento

Inmediatamente después del entrenamiento cada una de las redes neuronales, se procedió a probar su desempeño sobre el dataset de pruebas. Para ello, se comparó la clasificación obtenida con la etiqueta de cada una de las imágenes.

Durante esta etapa del proyecto se probaron diferentes configuraciones para las redes neuronales, y los par5ámetros en el punto 4.4 corresponden con la configuración que mostró los mejores resultados, para cada uno de los casos.

Dichos resultados, para cada una de las redes, se muestran a continuación:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Porcentaje de acierto del clasificador de expresiones faciales | | | | | | | |
| Neutral | Alegría | Tristeza | Ira | Disgusto | Miedo | Sorpresa | **Total** |
| 77% | 77% | 80% | 76% | 80% | 79% | 74% | **79%** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Porcentaje de acierto del interpretador del lenguaje de sordomudos | | | | | | | | | | | | |
| A | B | C | D | E | F | G | H | I | K | L | M | N |
| 99% | 95% | 97% | 96% | 95% | 99% | 93% | 91% | 78% | 90% | 95% | 79% | 70% |
| O | P | Q | R | S | T | U | V | W | X | Y | **Total** | |
| 77% | 98% | 90% | 69% | 79% | 58% | 69% | 82% | 90% | 91% | 79% | **87%** | |

Figura 5: Tablas de resultados de pruebas después del entrenamiento

### Pruebas en vivo

Una vez que se seleccionaron las redes con el mejor desempeño, las mismas se aplicaron respectivamente en cada uno de los programas que utilizan el algoritmo explicado en el punto 4.5, con el objetivo de que realizaran la clasificación de expresiones faciales y señas, dependiendo del caso.

#### Clasificación de expresiones faciales humanas

Al ejecutar el programa totalmente integrado, primeramente se le pedirá al usuario que indique el número máximo de rostros que se analizarán por cada cuadro de vídeo. Se debe tomar en cuenta que entre mayor sea este número, el procesamiento de cada cuadro de vídeo será más exigente.

A continuación, para cada rostro identificado en cada cuadro de vídeo, se aplicará el clasificador de expresiones faciales. El desempeño general del mismo depende de varias variables, entre las que destacan la edad del sujeto, su género, etnia, iluminación del ambiente, distancia hasta la cámara, ángulo del rostro, calidad del vídeo, entre otras.

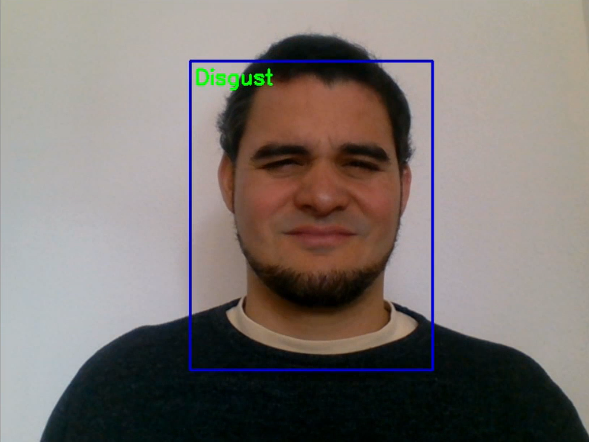


Figura 6: Sujeto uno mostrando expresiones de “ira” y “disgusto” (de izquierda a derecha)

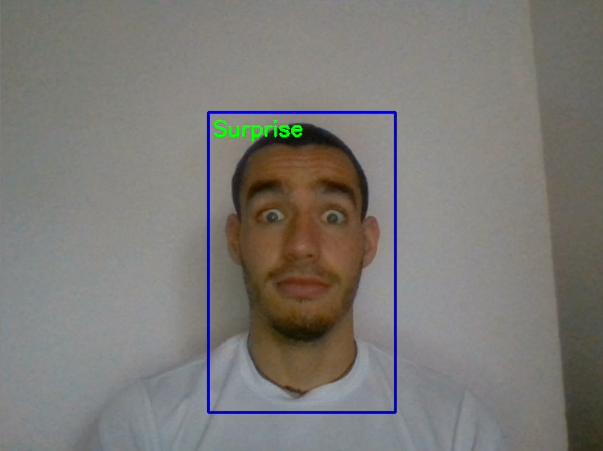


Figura 7: Sujeto dos mostrando expresiones de “tristeza” y “sorpresa” (de izquierda a derecha)

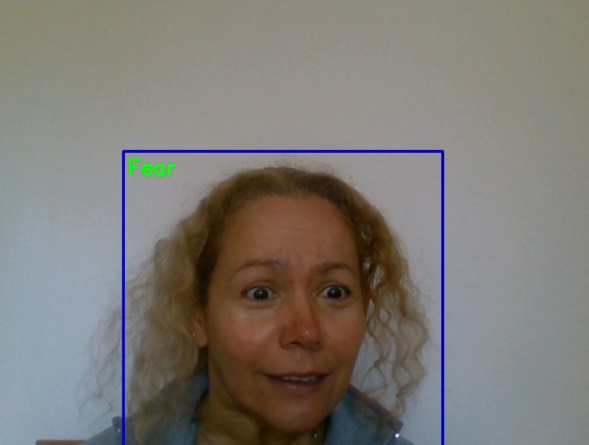


Figura 8: Sujeto tres mostrando expresiones de “alegría” y “miedo” (de izquierda a derecha)

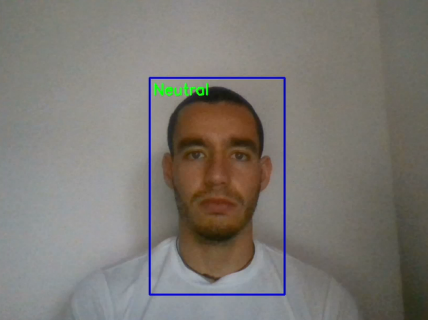
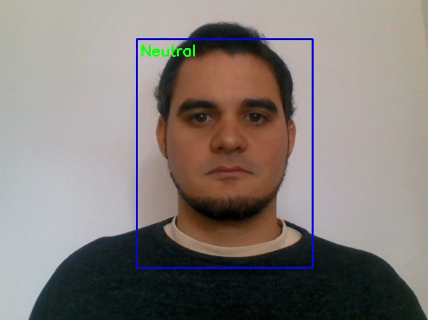


Figura 9: Los tres sujetos de prueba mostrando la expresión “neutral”

Al experimentar con distintos sujetos de estudio, se pudieron apreciar diferencias en el margen de acierto de algunas expresiones faciales sobre otras, dependiendo de la edad, género y etnia del sujeto. Esto se debió principalmente al hecho de que en el dataset de entrenamiento predominaban sujetos de entre 20 y 30 años, mayoritariamente caucásicos y con una predominancia femenina (ver punto 4.2.1).

#### Clasificación de señas del lenguaje de sordomudo americano

Al realizar la ejecución de este programa, se le pedirá usuario que indique el número máximo de manos que serán analizadas por cada cuadro de vídeo. Es indistinto el uso de la mano derecha o izquierda durante este experimento.

Se debe tomar en cuenta que entre mayor sea este número, el procesamiento de cada cuadro de vídeo será más exigente.

Luego, para cada mano identificada en cada cuadro de vídeo, se aplicará el clasificador de señales. El desempeño de dicho clasificador puede variar dependiendo principalmente de la iluminación ambiente, el ángulo de la mano, distancia hasta la cámara, la calidad del vídeo y qué tan diferente es la señal en cuestión con respecto a otra en el alfabeto. Además, al utilizarse el color de la piel para identificar la extremidad, es obligatorio el uso de indumentarias de un color diferente al de la piel y con mangas largas para el correcto funcionamiento del algoritmo.

A continuación, se muestran algunos ejemplos de detección para ocho diferentes letras:

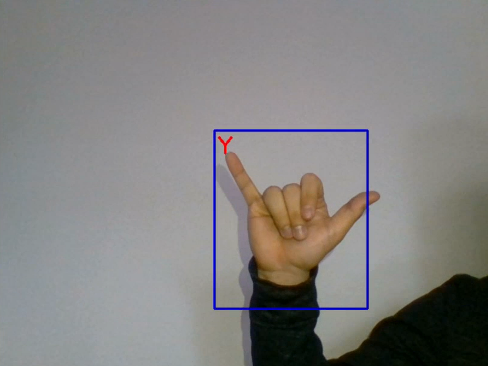
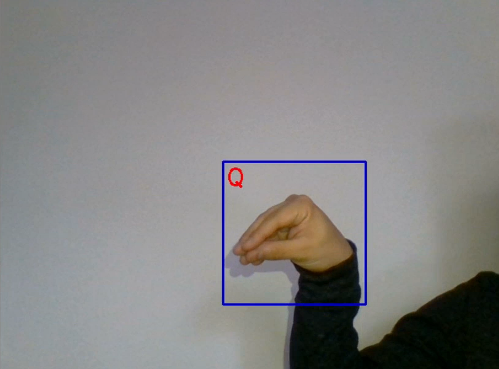
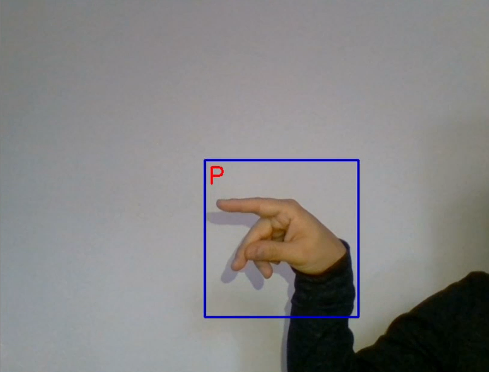
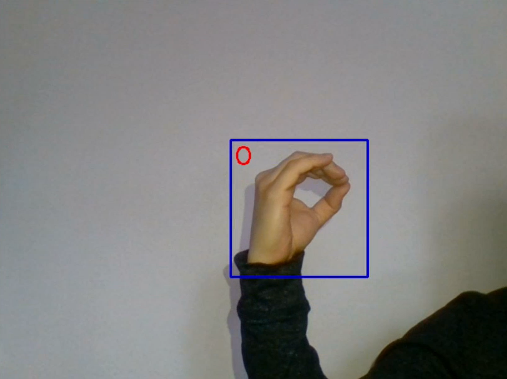
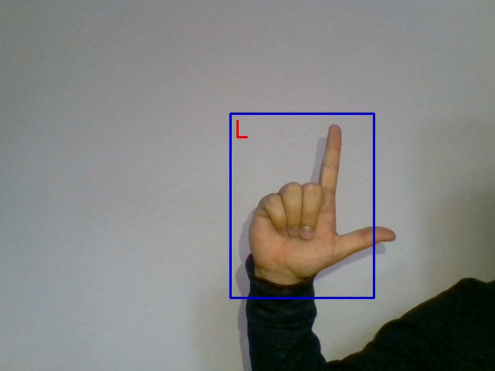
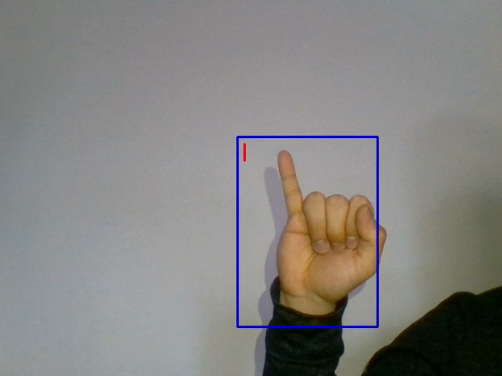
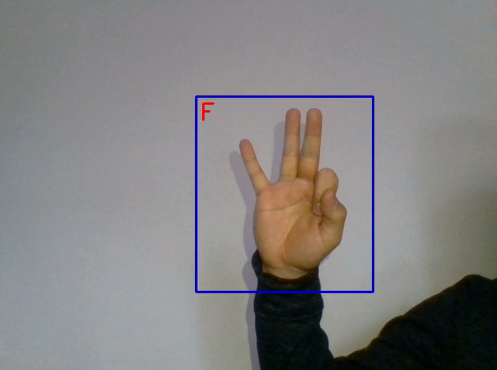
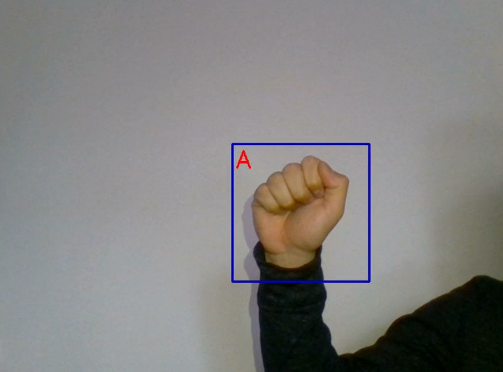


Figura 10: Sujeto mostrando las señas para las letras “A”, “F”, “I”, “L”, “O”, “P”, “Q”, “Y” (de izquierda a derecha t de arriba abajo)

Al experimentar con distintos sujetos y señas, se puede observar una marcada diferencia en el porcentaje de acierto entre diferentes señas. Los principales casos de desacierto se deben a la similitud entre dos o más señas. Entre ellas destacan las señas para las letras “A”, “M”, “N” y “S”; “F”, “V” y “W”; “P” y “Q”; “D”, “K” y “R”.

# POSIBLES APLICACIONES

La intención de este proyecto es dotar a los robots con nuevas herramientas que les permitan una mejor comunicación con seres humanos.

Durante el desarrollo del mismo, se mantuvo como principal objetivo a futuro su aplicación como un servicio en la nube al que pudieran acceder robots para la interpretación de las imágenes que estuviese recopilando en vivo durante su interacción con seres humanos. De esta manera, los robots que accediesen a dicho servicio serían capaces de mejorar el entendimiento de ciertas expresiones lingüísticas ambiguas, pero cuyo significado es fácilmente interpretable dependiendo de la expresión facial que muestra el interlocutor. De igual manera, dichos robots también serían capaces de comunicarse visualmente con cualquier persona que utilice el lenguaje de sordomudos americano.

El hecho de que ambos programas se hayan desarrollado utilizando el lenguaje Python, permite que su uso no esté limitado a los modelos Pepper y Nao que se tuvieron en cuenta durante su desarrollo, sino con la gran mayoría de robots que se están comercializando en la actualidad.

Adicionalmente, en el caso del interpretador del lenguaje de señas de sordomudo americano, existe la posibilidad de desarrollar una aplicación móvil basada en su algoritmo, y que permita a otros seres humanos ser capaces de comunicarse más efectivamente con personas que utilicen este lenguaje.

# MEJORAS Y FUTUROS DESARROLLOS

A pesar de que los resultados obtenidos fueron satisfactorios, una vez finalizado el proyecto se han identificado varios aspectos donde se pueden aplicar mejoras de cara a próximos desarrollos en esta área.

Entre ellos se destacan los siguientes:

* Utilizar un dataset más grande y más diverso para el entrenamiento de la red neuronal artificial que identifica expresiones faciales.
* Recopilar dentro del dataset mencionado anteriormente imágenes de rostros con expresiones faciales más naturales, y por lo tanto, menos exageradas.
* Incluir en la ejecución de ambos programas una modalidad que guarde ciertos cuadros para que el usuario evalúe si el clasificador acertó o no, y en base a ello reentrenarlo.
* Desarrollar un método que además de interpretar las señas del lenguaje de sordomudo, vaya generando palabras a partir de las letras identificadas.
* Incluir en las pruebas a sujetos que utilicen el lenguaje de señas de forma regular, de forma tal que se puedan evaluar casos más naturales.
* Desarrollar e incluir un clasificador de manos, de forma tal que no se tenga que asumir que un objeto es una mano sólo porque no fue identificado previamente como un rostro.
* Evaluar un rediseño general de los algoritmos para buscar reducir el procesamiento requerido durante las ejecuciones y mejorar la calidad final del mismo.

Se espera que algunas de las mejoras señaladas puedan ser desarrolladas antes de la aplicación de este proyecto a robots comerciales en los próximos meses.

# BIBLIOGRAFÍA

[1] Aprendizaje automático. https://es.wikipedia.org/wiki/Aprendizaje\_autom%C3%A1tico (consultado en marzo 2019)

[2] Artificial Neural Networks as Models of Neural Information Processing: Frontiers Research Topic. https://www.frontiersin.org/research-topics/4817/artificial-neural-networks-as-models-of-neural-information-processing (en inglés, consultado en febrero de 2018)

[3] Perceptrón multicapa: <https://es.wikipedia.org/wiki/Perceptr%C3%B3n_multicapa> (consultado en marzo 2019)

[4] Self-Organized Formation of Topologically Correct Feature Maps. Kohonen, Teuvo. Biological Cybernetics, página 43 (1982)

[5] Redes neuronales convolucionales. https://es.wikipedia.org/wiki/Redes\_neuronales\_convolucionales (consultado en marzo 2019)

[6] Receptive fields of single neurons in the cats striate cortex. Hubel, D. H.; Wiesel, T. N. “The Journal of physiology”, página 148 (1959).

[7] Detecting Skin in Images & Video Using Python and OpenCV. Rosebrock, Adrian. https://www.pyimagesearch.com/2014/08/18/skin-detection-step-step-example-using-python-opencv/ (2016)

[8] Sign Language Recognition with Unsupervised Feature Learning. Chen, Justin. http://cs229.stanford.edu/proj2011/ChenSenguptaSundaram-SignLanguageGestureRecognitionWithUnsupervisedFeatureLearning.pdf/ (2014)

# ANEXOS

Todo el código desarrollado durante este proyecto se encuentra publicado abiertamente en GitHub. Dentro de cada una de las carpetas, se puede encontrar el código que es necesario ejecutar para ver el algoritmo en acción, junto con una carpeta (CNN) donde se encuentra la red neuronal utilizada para la clasificación, y otra carpeta (Training) donde se encuentra el código utilizado para el entrenamiento de las redes neuronales artificiales:

* Detector de expresiones faciales:

<https://github.com/pzampella/Expression-and-Sign-Detector/tree/master/Emotion_Analytics>

* Interpretador de señas del lenguaje de sorde-mudos:

<https://github.com/pzampella/Expression-and-Sign-Detector/tree/master/Sign_Detection>