



UNIVERSIDAD DE MURCIA

TRABAJO FINAL DE GRADO

Supervisión visual de normas COVID

Autor:

Juan José MORELL

FERNÁNDEZ

juanjose.morellf@um.es

Tutor:

Alberto RUIZ GARCIA

a.ruiz@um.es

5 de mayo de 2021

Índice general

Resumen	3
Extended abstract	4
1. Introducción	5
1.1. Historia de la Visión Artificial	5
1.2. Reconocimiento Facial y COVID-19	7
2. Estado del arte	8
2.1. Python y OpenCV	9
2.2. Aplicaciones sin Deep Learning	9
2.3. Aplicaciones con Deep Learning	10
3. Análisis de objetivos y metodología	11
3.1. Prototipo	12
4. Diseño y resolución	13
4.1. Paul Viola and Michael Jones	13
4.2. Facial Landmark	19
4.3. Mediapipe	25
4.4. Tensorflow	28
4.5. Comparación	31
5. Conclusiones y vías futuras	32
Bibliografía	35

Índice de figuras

4.1. Haar-like Features	14
4.2. Funcionamiento de una <i>Imagen Integral</i>	15
4.3. Construcción del <i>Strong Classifier</i>	16
4.4. Construcción del <i>Multi-stage Classifier</i>	17
4.5. Fórmula <i>HOG</i>	20
4.6. 68 coordenadas del <i>facial landmark</i> con <i>iBUG 300-W dataset</i>	22
4.7. MediaPipe Graph utilizado en FaceMesh	26
4.8. Rendimiento de FaceMesh sobre dispositivos móviles	27
4.9. Arquitectura de una CNN con uso de SSD.	29

Resumen

En este proyecto se plantea el problema de...

Extended abstract

This project faces the problem of...

CAPÍTULO 1

Introducción

LA visión artificial es un ámbito de la informática que surgió hace 60 años, pensado en el estudio del procesamiento digital de las imágenes. En los últimos años ha tomado mucha importancia el reconocimiento facial, algoritmo capaz de identificar rostros humanos dentro de una imagen, gracias a la investigación realizada por *Viola y Jones* en 2001, y la reciente tendencia en *smartphones* con el desbloqueo facial.

1.1. Historia de la Visión Artificial

Uno de los primeros acontecimientos que propició la visión artificial fue la creación del cable Bartlane, capaz de transmitir una imagen a través del mar atlántico en los años 1920, con una duración de cerca a una semana. Pero, dentro del entorno del procesamiento de imágenes digitales, la investigación se centró en recuperar una estructura tridimensional del mundo real a través de una imagen para conseguir un entendimiento total de la escena que plasma la misma. Debido a esto aparecieron varios algoritmos de reconocimiento de líneas, donde uno de ellos fue creado por parte de Huffman en 1971.

Pero el avance de estas investigaciones pronto se ligarían con el del ordenador. Estos, se podrían resumir en varios acontecimientos importantes, tales como: la invención

del transistor por Bell Laboratories en 1948 y de los circuitos integrados en 1958 o la introducción por parte de IBM de los primeros ordenadores personales en 1981. [12]

Uno de los descubrimientos que inicio este movimiento no fue proveniente de la informática, sino de la psicología. Esta sería una de las principales fuentes sobre el entendimiento de como funciona la visión. Un par de psicólogos, David Hubel y Torsten Wiesel, describieron que el comportamiento de las neuronas encargadas de entender el entorno visual siempre empiezan con estructuras simples como vértices. Más tarde esta idea se convertirá en el principio central del *Deep Learning*.

Russell Kirsch, en 1959, es el primero en desarrollar un aparato que traducía las imágenes en datos que las máquinas pudiesen entender. Y, Lawrence Roberts en 1963 publica un estudio sobre como las máquinas perciben objetos sólidos de tres dimensiones, uno de los avances considerados precursores de la visión artificial moderna.

Durante los años 1980s, se desarrolló una red artificial capaz de reconocer patrones, mediante el uso de una red convolucional. Que propició la creación de un modelo llamado LeNet-5, la primera red convolucional moderna. Este modelo se caracteriza por usar la backpropagación. Mientras que en los años 1990s la visión artificial cambia totalmente de rumbo y los investigadores pasaron de intentar reconstruir objetos en 3D a intentar detectar objetos mediante sus características.

A partir del año 2000 se hacen muchos avances importantes y que actualmente son usados en aplicaciones reales. El primer detector facial llegaría en 2001 creado por Paul Viola y Michael Jones. Ambos consiguieron crear el primero que funcionase en tiempo real.

El problema de estos modelos es el uso de información para poder entrenarlos, y para esto se creo un proyecto llamado PASCAL VOC, que creo un dataset estándar para la clasificación de objetos. Posteriormente, apareció en 2010 ImageNet, el cual contiene más de un millón de imágenes para un total de mil objetos. Y junto a este apareció un modelo basado en una red convolucional llamado AlexNet. Desde entonces, el *Deep Learning* se ha convertido en eje central del avance en la visión artificial, conjuntamente con todos los avances matemáticos realizados anteriormente.

1.2. Reconocimiento Facial y COVID-19

En la actualidad, la visión artificial se utiliza en muchos proyectos y esta presente en investigaciones muy importantes para el futuro de la inteligencia artificial. Una de ellas se basa en el reconocimiento facial, y es usado en aplicaciones para reconocer personas, aspectos, contador de personas, etc.

La detección facial se inicia con una imagen arbitraria, con el objetivo de encontrar todas las caras que hay en una imagen, y posteriormente devolver otra con la localización exacta de cada una de las caras. Aunque esta tarea es natural para los humanos, es bastante complicada para los ordenadores. Ya que se encuentran muchos factores que lo dificultan, tales como: la escala, localización, punto de vista, iluminación, lentes, etc. Existen centenares de investigación/proyectos de detección facial, desde uno de los mas influyentes en los años 2000s, como *Viola and Jones face detection*, a proyectos basados en *Deep Learning*, con tecnologías como *Tensorflow* o *YOLO*. [23]

Tras el año 2020 y la aparición del COVID-19, el uso de mascarillas y el cumplimiento de las normas impuestas por la OMS (Organización Mundial de la Salud) están al orden del día. En este trabajo se tendrá como objetivo el estudio de todas estas tecnologías para poder controlar dichas normas, terminando con un prototipo capaz de detectar cuando una personas lleva mascarilla, ya sea al entrar a un comercio, evento u trabajo.

CAPÍTULO 2

Estado del arte

El reconocimiento facial es un ámbito de la visión artificial, que como su nombre indica, se centra en la búsqueda de rostros humanos dentro de imágenes digitales. Durante años se han desarrollado varias tecnologías capaces de realizar dicha acción, de las que se pueden distinguir dos grupos:

- Aplicaciones con sin el uso de *Deep Learning*
- Aplicaciones con uso de *Deep Learning*

Ambos se centran en las características de los rostros humanos para lograr identificarlos. Esto se conoce como *features*, que corresponden con puntos de la cara muy reconocibles, como: el mentón, ojos, cejas, nariz, etc. Esta técnica fue usada por primera vez en 2001, por Paul Viola y Michael Jones, y desde entonces se ha convertido en una de las técnicas principales en el reconocimiento facial.

Este ejercicio se complica cuando las imágenes presentan errores naturales en su toma (como baja luz, ruido, etc.) o los individuos visten complementos que tapen sus rasgos faciales. Este es el problema que se va a plantear en este trabajo, buscar una posible solución al reconocimiento facial con complementos faciales, en específico identificar si las personas llevan mascarillas.

2.1. Python y OpenCV

2.2. Aplicaciones sin Deep Learning

HAAR-like Features & ADABOOST

Facial Landmasking

2.3. Aplicaciones con Deep Learning

¿Que es el deep learning?

YOLO

TensorFlow

Keras

MediaPipe

IBM Watson

CAPÍTULO 3

Análisis de objetivos y metodología

El COVID-19 es un problema que ha provocado en la humanidad incontables problemas, y en el mundo de la visión artificial también, por eso me voy a centrar en la creación de un prototipo que sea capaz de revisar el cumplimiento de las normas COVID impuestas en España y en todo el mundo por la OMS. Concretamente, detectar cuando una persona lleva, de manera correcta, una mascarilla al entrar a un comercio, cine, restaurante, etc. Los objetivos que se plantean para llevar esto a cabo son los siguientes:

- Estudio de las tecnologías actuales, para comprobar su comportamiento con uso de mascarilla.
- Creación de un prototipo capaz de reconocer rostros y detectar si se lleva mascarilla.
- Estudiar la capacidad de que el prototipo pueda identificar si se lleva correctamente la mascarilla.
- Poder detectar la mascarilla, independientemente del tipo que se lleve.

3.1. Prototipo

Para cada uno de los apartados del desarrollo de este *TFG* se creará un prototipo con la finalidad de mostrar la tecnología expuesta en el mismo. Con el objetivo final de crear una aplicación que contenga todos los prototipos y se puedan ejecutar de forma sencilla.

Los prototipos serán probados en varios escenarios de prueba, todos ellos se realizarán en tiempo real. Se contará con una totalidad de tres escenarios:

- Detección de mascarillas a una distancia cercana.
- Detección de mascarillas a una distancia media.
- Detección de mascarillas desde una posición alejada, como la parte superior de una puerta.

Asimismo, las pruebas se repetirán en dos dispositivos diferentes. El primero de ellos sin GPU y el segundo con GPU (CUDA). A continuación se muestran las especificaciones de los dispositivos:

	PC 1	PC 2
CPU	Intel i7-1065G7 1.30GHz 8 núcleos	Intel i7 GHz núcleos
GPU	Intel Iris Plus Graphics	GTX 980M 2Gb
RAM	16 Gb	8 Gb
OS	Ubuntu 18.04	Ubuntu 18.04

Por último, en los anexos se mostrará una explicación de como se ha implementado más centrado en la programación, conjuntamente con un manual de usuario.

CAPÍTULO 4

Diseño y resolución

4.1. Paul Viola and Michael Jones

En 2001, el reconocimiento facial tuvo su primera aparición en el campo de la visión artificial como aplicación en tiempo real. Este avance fue de la mano de Paul Viola y Michael Jones. Análogamente, el punto de partida del estudio de este TFG. Durante este apartado, se estudiará el funcionamiento del algoritmo *Viola-Jones face detector*, ideado por estos dos investigadores y se realizará una implementación del mismo mediante *Python* y *OpenCV* para comprobar como se comporta en la situación actual.

Método de estudio

El trabajo de los expertos fue presentado por parte de la Universidad de Cambridge mediante un *paper* (ensayo de la investigación). Y se introduce como:

"This paper describes a machine learning approach for visual object detection which is capable of processing images extremely rapidly and achieving high detection rates"[27]

Para poder lograr esta afirmación se basan en un procedimiento de trabajo en dos

fases: entrenamiento y detección. Igualmente, Paul y Michael dividen el proyecto en tres ideas principales para poder lograr un detector que se pueda ejecutar en tiempo real. Y estas son: la imagen integral, Adaboost (algoritmo de Machine Learning) y un método llamado *attentional cascade structure*.

Con todos estos puntos combinados lograron ingeniar un prototipo capaz de detectar caras humanas con un *frame rate* de 15 fps. Fue diseñado para la detección de caras frontales, haciéndose difícil para posiciones laterales o inclinadas.

Las imágenes que se toman para realizar la detección pasan por una transformación del espacio de color a *grayscale*. Con el objeto de encontrar características en ellas, llamadas *haar-like features*. Nombradas así por su inventor Alfred Haar en el siglo XIX. En este trabajo se hacen uso de tres tipos de haar-like features, que son las siguientes:

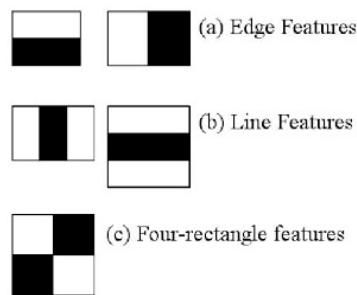


Figura 4.1: Haar-like Features

Las *Haar-like features*, o también conocidas como *Haar-wavelet* son una secuencia de funciones *rescaled square-shaped*, siendo similares a las funciones de Fourier y con un comportamiento parecido a los *Kernel* usados en las *Redes Convolucionales* (matrices que consiguen extraer ciertas *features* de la imagen de entrada). De manera que, las *Haar Features* serán las características de la detección facial.

En un estudio ideal, los píxeles que forma el *feature* tendrá una división clara entre píxeles de color blanco con los de color negro (Figura 4.1), pero en la realidad eso casi nunca se va a dar.

Más específicamente, las *Haar-like features* están compuestas por valores escalares que representan la media de intensidades entre dos regiones rectangulares de la imagen. Estas capturan la intensidad del gradiente, la frecuencia espacial y las direcciones, me-

dante el cambio del tamaño, posición y forma de las regiones rectangulares basándose en la resolución que se define en el detector. [21]

Estas características van a ayudar al ordenador a entender lo que es la imagen estudiada. Van a ser utilizadas mediante *Machine Learning* para detectar donde hay una cara o no, mediante un recorrido sobre toda la imagen. Esto conlleva una potencia de computación elevada. Para paliar este problema idearon el método de la *Imagen Integral*.

La *Imagen Integral* permite calcular sumatorios sobre subregiones de la imagen, de una forma casi instantánea. Además de ser muy útiles para las *HAAR-like features*, también lo son en muchas otras aplicaciones.

Si se supone una imagen con unas dimensiones de $\langle w, h \rangle$ (ancho y alto, respectivamente), la imagen integral que la representa tendrá unas dimensiones de $\langle w + 1, h + 1 \rangle$. La primera fila y columna de esta son ceros, mientras que el resto tendrán el valor de la suma de todos los píxeles que le preceden. [2] Ahora, para calcular la suma de los píxeles en una región específica de la imagen, se toma la correspondiente en la imagen integral y se suma según la siguiente fórmula (siguiendo la numeración de la Figura 4.2):

$$sum = L4 + L1 - (L2 + L3)$$

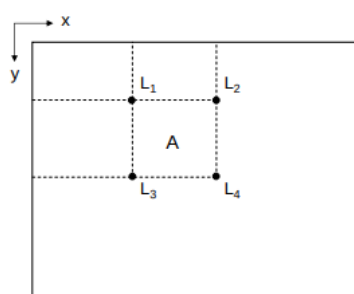


Figura 4.2: Funcionamiento de una *Imagen Integral*

Viola y Jones junta esta propuesta con los filtros *Haar-like features*, y consiguen computar dichas características de manera constante y eficaz. [7]

Una vez estudiada la obtención de características y con un set de entrenamiento, solo queda seleccionar un método de *machine learning* que permita crear una función de clasificación. Concretamente, se plantea el uso de una variante de **AdaBoost**, que permite seleccionar un pequeño conjunto de características y poder entrenar un clasificador.

Este algoritmo de aprendizaje esta basado en generar una predicción muy buena a partir de la combinación de predicciones peores y más débiles, donde cada uno de estas se corresponde con el *threshold* de una de las características *Haar-like*. La primera vez que aparece este algoritmo, de forma práctica, fue de la mano de *Freund y Schapire* [11]. Sin embargo, el usado por *Viola y Jones* es una modificación de este.

La salida que genera el algoritmo **AdaBoost** es un clasificador llamado *Strong Classifier*, como se ha mencionado anteriormente, compuesto por combinaciones lineales de *Weak Classifiers*.

El procedimiento para encontrar *Weak Classifiers* es ejecutar el algoritmo T iteraciones donde T es el número de clasificadores a encontrar. En cada iteración, el algoritmo busca el porcentaje de error entre todas las características y escoge la que menos porcentaje de error presente en dicha iteración. (Como se muestra en la *Figura 4.3*) [17]

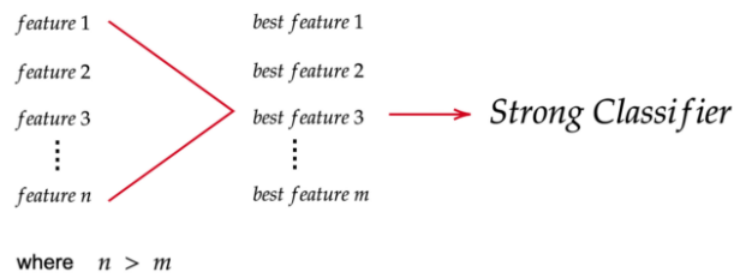


Figura 4.3: Construcción del *Strong Classifier*

Con estos clasificadores se procede a la construcción de una estructura en cascada para crear un *Multi-stage Classifier*, que podrá realizar una detección rápida y buena. Por tanto, la estructura de cascada esta compuesta por varios estados de *Strong Classifiers* generados por el algoritmo *AdaBoost*. Donde el trabajo de cada estado será identificar si, dada una región de la imagen, no hay una cara o si hay la posibilidad de que la haya. [11]

Si el resultado de uno de los estados es que no existe una cara en dicha región, esta

se descarta directamente. Mientras que, si hay la posibilidad de que exista una, pasa al siguiente estado de la estructura. De tal forma que, cuantos más estados atraviese una región de la imagen, con más seguridad se podrá afirmar que existe una cara en ella. La estructura completa se refleja en la *Figura 4.4*.

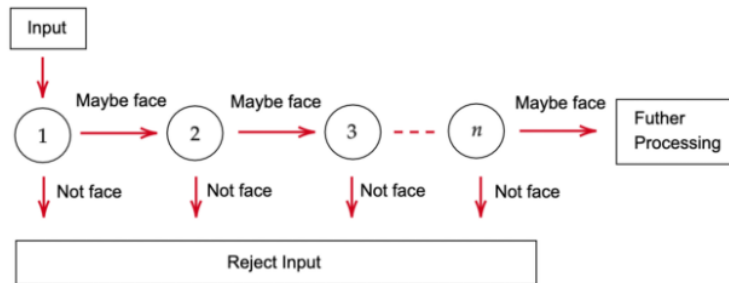


Figura 4.4: Construcción del *Multi-stage Classifier*

Implementación y Experimentación

El prototipo será implementado en *Python*, con el uso de *OpenCV*. Y, el objetivo es construir dos detectores de caras, donde el primero usará un modelo preentrenado de *OpenCV* de caras frontales. Mientras que en el segundo, se intentará modificar el programa, para que mediante el uso de varios modelos preentrenados se pueda detectar una cara con una mascarilla.

La **implementación básica** hace uso de un modelo preentrenado cargado mediante una clase de *OpenCV* llamada *Cascade Classifier*. Esta representa la base de *Machine Learning* explicado en el apartado anterior. Asimismo, *OpenCV* también proporciona una serie de archivos *xml* con diferentes modelos preentrenados. En concreto, para este prototipo se hace uso del modelo por defecto, detector de caras frontal, como se muestra en la investigación de *Viola y Jones*.

Finalmente, la detección se realiza, tras realizar una transformación del espacio de color a blanco y negro, mediante la función *detectMultiScale* de la clase, creada anteriormente, *Cascade Classifier*. Concretamente, su funcionalidad será encontrar caras dentro de las imágenes que vaya procesando.

[Explicación del prototipo custom]

OpenCV y Haar-like features + Machine Learning con PCA y SVM

Se implementa un identificador de caras conjunto a un modelo de *Machine Learning* que identifica cuando una persona lleva o no mascarilla, mediante una toma de muestras anterior. Gracias a los modelos PCA y SVM, se puede crear un modelo para su identificador. Lo malo: Solo funciona con los rostros/rostro que se toma como referencia para construir el modelo, igualmente pasa con el tipo de mascarilla (siendo la quirúrgica la que mejor funciona con este prototipo). Asimismo, su funcionamiento es de manera frontal y cercana, ya que si se coloca la cámara en la borde superior de una puerta o similar, el detector se pierde mucho y crea identificaciones falsa o no llega a reconocer nada.

Este procedimiento se podría llegar a usar con otra implementación, específicamente con HOG.

[Resultados] Bien, es un comienzo. Pero mal para un mercado amplio.

4.2. Facial Landmark

Con el objetivo de ampliar la idea anterior, se plantea el uso de Facial Landmark, una tecnología que nos permite el reconocimiento de puntos de interés en las caras que se han detectado en la imagen. Sus pasos de ejecución son: detectar cara dentro de la imagen (En este caso, se usará *Haar-like features*) y obtener dichos puntos de interés.

La implementación que se va a utilizar es la estudiada por *Kazemi y Sullivan* en 2014, con el paper *One Millisecond Face Alignment with an Ensemble of Regression Trees* [15], y usado en el *toolkit Dlib*. Este método se centra en localizar las siguientes zonas faciales: boca, cejas, ojos, nariz y mentón, gracias al uso de un conjunto de árboles de regresión. Estos son entrenados mediante un modelo formado por puntos de interés de un grupo de imágenes, etiquetados a mano y especificadas como coordenadas (x,y).

Dlib será el *toolkit (Open Source)* utilizado para la implementación de dicho método. Este contiene algoritmos de *Machine Learning* y herramientas capaces de crear software complejo en *C++* y *Python* para resolver problemas reales. Sobre todo centrado en robótica, dispositivos embebidos, móviles y ordenadores de gran capacidad. [9]

HOG - *Histogram of Oriented Gradients*

El primer paso en esta solución es encontrar una cara dentro de la imagen de entrada, y el encargado será el método HOG. El cual, sigue una idea similar al método de *Haar-like*, ya que se basa en la detección de *features* (características).

La idea teórica tras *HOG* es encontrar la apariencia y forma de un objeto mediante la distribución de la intensidad de gradientes locales, gracias a que estos obtienen una magnitud mayor en las cercanías de bordes o esquinas. Mientras que la implementación, divide la imagen en pequeñas regiones (llamadas celdas) y se calcula un histograma de gradientes de 1 dimensión para cada uno de los píxeles de cada celda. Para un mejor estudio se normaliza el contraste de la imagen de entrada. [5] Con este procedimiento se obtiene un *feature vector* a partir de la imagen de entrada, y la distribución resultante de los gradientes serán usados como las características. La implementación de HOG se

podría dividir en 5 pasos generales [20]:

1. *Pre-procesado*: Para una imagen de entrada de cualquier tamaño es tratada en regiones de ciertas escalas y analizadas en varias zonas de la imagen. La única restricción es que los tamaño de las regiones analizadas tienen una relación de aspecto fija.
2. *Cálculo de las imágenes de gradiente*: Para el calculo del histograma de gradientes es necesario realizar el calculo de los gradientes, tanto verticales como horizontales. Esto se puede obtener facilmente gracias al uso de *Kernels* (filtros). Posteriormente, se busca la magnitud de las direcciones de dichos gradientes con el uso de la siguiente fórmula:

$$g = \sqrt{g_x^2 + g_y^2}$$
$$\theta = \arctan \frac{g_y}{g_x}$$

Figura 4.5: Fórmula HOG

La fórmula esta implementada en *OpenCV* mediante la función *cartToPolar*. En este punto se puede obtener la imagen de los gradientes, eliminando toda la información no relevante de la imagen original, y se puede observar como en todos los pixeles el gradiente tomara una magnitud y una dirección. Si la imagen es RGB, los gradientes se evaluan sobre los tres canales de color, siendo la magnitud final el valor mayor entre los tres canales.

3. *Calculo de los histogramas de gradientes en celdas (8x8)*: La imagen se divide en celdas y se calculan los histogramas para cada una de ellas. Por ejemplo, si la celda es de tamaño 8x8, esta contendrá 192 píxeles (8x8x3), donde el gradiente tiene dos valores, descritos anteriormente (magnitud y dirección), por cada uno de los píxeles, lo que añade 128 valores (8x8x2). Esto facilita la representación de las regiones mediante el uso de un histograma, ademas mejora la influencia al ruido. El tamaño de la celda vendrá definida dependiendo de la escala de características (*features*) que se estén buscando.
4. *Normalización de bloques (16x16)*: Una vez creado el histograma, hay que tener en cuenta que la imagen es sensible al brillo que tenga. Por ejemplo, si el brillo de

la imagen se divide en dos, la magnitud del gradiente también lo hará. De igual forma, si el brillo se multiplica por dos, el gradiente ídem. Pero, se busca un descriptor que sea independiente a esto, por lo que se normaliza el histograma para que no se vea afectado por los cambios de luz/brillo. Disponiendo de celdas de 8x8, un bloque de 16x16 posee cuatro histogramas que pueden ser condensados en un unico vector normalizado.

5. *Calculo del vector de HOG*: El último paso es concatenar los vectores normalizados obtenidos en uno global.

Este procedimiento se repite varias veces sobre imagenes distintas y se introduce en un modelo *Machine Learning* del tipo SVM Linear, con el objetivo de obtener un detector de caras. En el caso de *Dlib*, posee un modelo preentrenado.

Sullivan Paper

El *paper* de Kazemi y Sullivan presenta la implementación usada en el *toolkit Dlib* del algoritmo que estima de forma precisa y eficiente los puntos de interés faciales. Esta basado en *gradient boosting* para el aprendizaje de un conjunto de arboles de regresión (*ensemble of regression trees*), que será el encargado de la predicción de los puntos de interés.[16]

Este método fue uno de los primeros que mejoro el rendimiento, a diferencia con los métodos anteriores, gracias a la detección de componentes esenciales para el *face alignment* y procesarlos para introducirlos en funciones de regresión en cascada. Cada una de estas funciones estima, de forma eficiente, la forma facial desde una estimación inicial y obtiene un conjunto de píxeles indexados a dicha estimación. En concreto, *Dlib* estimará un total de 68 píxeles indexados (Figura 4.6), ya que sigue las anotaciones del *dataset iBUG 300-W* [22], usado en los modelos pre-entrenados ofrecidos por *Dlib*.

Ademas, se introdujeron dos elementos clave a las funciones de aprendizaje de regresión. El primero gira en torno a la intensidad de los píxeles indexados con respecto a la estimación actual de la forma facial, ya que estos son muy influenciados por la deformación de la estimación de la forma y a los cambios de iluminación. El dilema que aparece

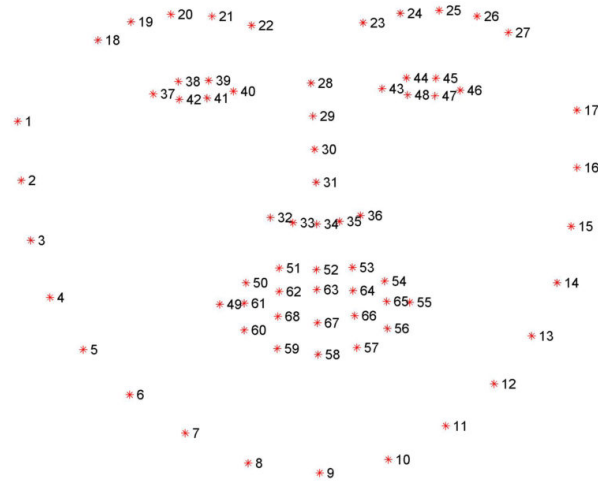


Figura 4.6: 68 coordenadas del *facial landmark* con *iBUG 300-W dataset*

es que necesitamos características confiables para obtener una predicción con precisión la forma y, por otro lado, necesitamos una estimación precisa de la forma para extraer características fiables. Para resolver este dilema se plantea el uso de un enfoque iterativo. La idea se basa en obtener una imagen transformada en un sistema de coordenadas normalizado basado en una estimación actual de la forma, y posteriormente extraer las características para predecir un vector de actualización para los parámetros de la forma. Este proceso suele repetirse varias veces.

El segundo elemento clave se trata de reducir la dificultad del proceso de inferencia/-predicción. El objetivo es obtener una función capaz de estimar la forma que concuerde con la información de la imagen y el modelo. Pero el problema es que esta función es no convexa en varios valores locales. Para resolver esto se plantearon dos soluciones a lo largo del tiempo. El primero afirma que las predicciones estimadas en zonas lineales del subespacio mienten, y efectivamente ayuda a evitar este problema. Pero, posteriormente se descubrió la segunda solución, donde se asume que la función mienten en zonas lineales del subespacio y no es necesario realizar trabajo adicional. En el caso de este *paper*, se implementa una solución donde se utilizan una combinación de ambas. Por lo que, cada regresor aprende mediante *gradient boosting* conjuntamente a una función de pérdida de error al cuadrado.

La entrada usada en el regresor, es seleccionada mediante *gradient boosting* sobre un conjunto de píxeles dispersos, y la prioridad de probabilidad sobre la distancia entre

pares de píxeles de entrada. Esto permite realizar un estudio sobre un mayor número de características relevante de forma eficiente. El resultado es una cascada de regresores que pueden localizar los puntos de referencia faciales cuando es inicializada con la media de la pose facial.

De forma más general, la investigación ofrece las siguientes contribuciones a las investigaciones de *Face Landmark* anteriores:

1. Un método de alineación que se basa en un conjunto de árboles de regresión que nos devuelve la forma facial mientras minimiza la función de error.
2. Método que maneja la predicción de puntos que faltan o no están presentes en la imagen. Por ejemplo, rostros que estén medio ocultos. (Esto solamente se realizará si el modelo creado con HOG detecta dicha cara oculta).
3. Resultados que demuestran que el método produce predicciones de alta calidad.
4. El efecto de la cantidad de datos de entrenamiento.

Prototipo

Mediante *OpenCV* y el *ToolKit Dlib* se obtiene la implementación de ambas ideas, mediante modelos pre-entrenados.

Próximos pasos

Con el objetivo de mejorar este prototipo se incorpora el *Deep Learning*. En las últimas décadas, se ha convertido en uno de los métodos más usados para la creación de aplicaciones de visión artificial. [23]. Está basado en las estructuras neuronales CNN (*Convolutional Neural Network*), principales responsables de que un ordenador pueda procesar de forma sencilla una imagen. CNN es una combinación entre capas neuronales y convolucionales, siendo estas últimas filtros con los que se obtendrán características de la imagen de entrada, los conocidos *features*. La principal función del Deep Learning es clasificar, ya sea una imagen, un objeto o incluso varios objetos a la vez. Esto es gracias

al uso de modelos, entrenados con miles de imágenes, que consiguen clasificar imágenes en varias categorías, por ejemplo *MobileNet*. [10]

En los siguientes apartados se probará implementar prototipos con el uso de Deep Learning con el objetivo de mejorar los resultado obtenidos hasta ahora. Se va a tratar con la API de Google llamada *MediaPipe* y la tecnología de *Transfer Learning* conjuntamente con *TensorFlow*, para poder reentrenar modelos para nuestro objetivo, detectar mascarillas.

4.3. Mediapipe

Mediapipe es una API *open-source* creada por Google, que ofrece servicios de *Machine Learning* para vídeos y fuentes multimedia. Entre ellas, se encuentra un servicio llamado *Face Mesh* que ofrece una solución que estima 468 puntos de interés de un rostro, que conforman una malla 3D en tiempo real. Este usa aceleración GPU conjuntamente con un modelo y el uso de una *pipeline*.

La *pipeline* que se utiliza en esta API consiste en dos modelos de *Deep Learning* que trabajan al mismo tiempo. Su funcionalidad es realizar una detección a partir de una imagen de los puntos de interés sobre una cara y construir un modelo *face landmark* 3D que aproxima la superficie de esta mediante regresión sobre dichos puntos. Esta tarea es facilitada si la cara, donde se tienen que detectar los puntos de interés, se encuentra recortada, haciendo así que el modelo se centre solamente en buscar los puntos, aumentando la precisión de la predicción. Asimismo, los recortes de las caras se puede generar a partir de las predicciones anteriores realizadas por el mismo modelo, y solamente es llamada la predicción nuevamente cuando no se consigue detectar la presencia de la cara.[13]

Todo esto es implementado gracias al framework *MediaPipe*, con la herramienta *MediaPipe graph*. Arquitectura caracterizada por estar formada por componentes llamados *Calculator*, nodos del grafo que tras la entrada de cero o más inputs generan cero o mas salidas. Todos estos nodos estan conectados mediante datos en forma de *Streams*, donde cada uno representa un conjunto de datos-tiempo en *Packets*. Por tanto, los *calculators* y *streams* definen el flujo de datos del *Graph*. [19]

El *pipeline* puede ser definido mediante la adición/modificación de *calculators* dentro del *graph*. Específicamente, el pipeline que se utiliza en esta solución (*FaceMesh*) esta formada por un *graph* compuesto por un *subgraph* de *face landmark* (proveniente del modulo, ya implementado, de face landmark de Mediapipe), donde a su vez usa otro *subgraph* proveniente de face detection module para la detección de caras, y un *face renderer subgraph* para mostrar el resultado. [13] En concreto el *graph* que se usa en esta implementación es el mostrado en la Figura 4.7

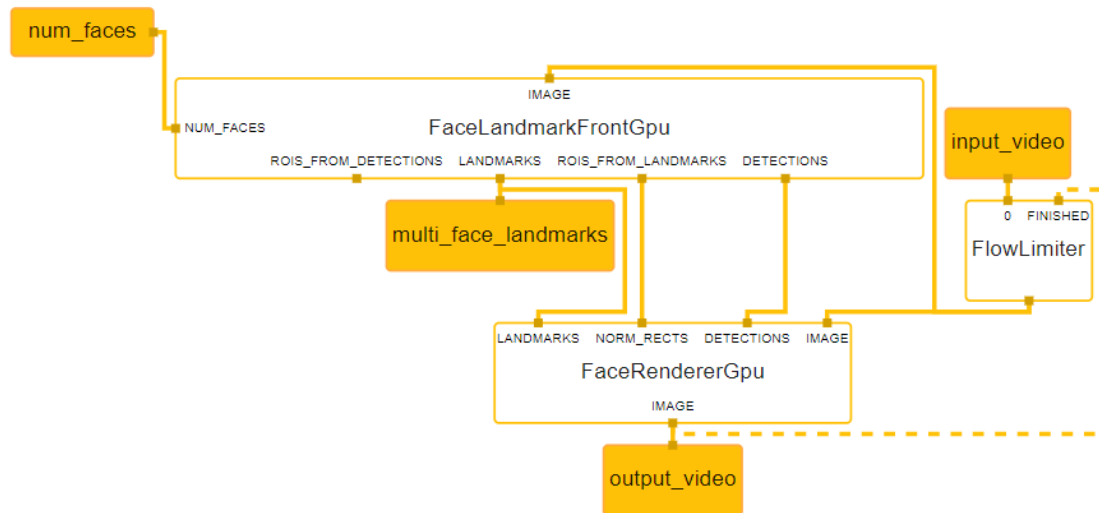


Figura 4.7: MediaPipe Graph utilizado en FaceMesh

Por tanto, el procesamiento de una imagen en este modelo sigue dos pasos. El primero (1) toma la imagen de entrada, capturada por la cámara, es procesada por un detector de caras *lightweight*, llamado *BlazeFace*, y produce unos rectángulos que definen el perímetro donde se encuentra la cara, conjuntamente con un par de puntos de interés superficiales (ojos, boca y nariz). Estos puntos se utilizarán para alinear la cara para el siguiente paso. Y el segundo (2), mediante el rectángulo obtenido en el paso anterior, se recorta la cara de la imagen inicial y es reescalado para utilizarse como entrada de la red neuronal que realiza la predicción de la malla. (El tamaño de reescalado será entre 256x256 en un modelo completo, hasta 128x128 en el modelo más pequeño). Tras la predicción, se obtiene como salida un vector de coordenadas *landmark 3D*, que serán mapeadas y dibujadas en la imagen original. [14]

Las coordenadas que se obtienen como salida están compuestas por unas coordenadas x e y provenientes de localizaciones del plano 2D propio de la imagen. Mientras que, la coordenada z es interpretada como una profundidad relativa a un centro de masa que compone la malla de la cara.

Se utilizan dos modelos para el funcionamiento de *FaceMesh*. El primero de ellos dedicado a la detección facial, llamado *BlazeFace* (mencionado anteriormente). Este es un modelo *lightweight* creado para GPU móviles, llegando a una velocidad de procesamiento de entre 200 a 1000 fps en dispositivos móviles punteros. Es inspirado en los

modelos *MobileNet*, tanto la primera version como la segunda, provenientes del *framework* SSD (*Single Shot Multibox Detector*). Este modelo produce una salida compuesta por un rectángulo perimetral y 6 puntos de interés faciales. [3]

El segundo modelo, *Face Landmark Model*, es generado mediante *Transfer Learning* buscando una serie de objetivos: crear coordenadas 3D (mencionadas anteriormente) y conseguir mostrarlas en la imagen de salida de forma correcta. [13] El *Transfer Learning* es una técnica de *Machine Learning* donde se puede hacer uso de un modelo preentrenado para personalizarlo y usarlo en una tarea determinada. Conviene destacar que un modelo entrenado es una red almacenada, entrenada previamente con un conjunto de datos con el objetivo de realizar una tarea de clasificación de imágenes a gran escala. [26]

FaceMesh propone una implementación mediante *TensorFlow Lite* que dispone de dos formatos: CPU y GPU, que presentan un rendimiento en dispositivos móviles (*Pixel3*, *Pixel2* y *iPhoneX*) muy fluido, impresionado sobre todo su funcionamiento sobre CPU. (Figura 4.8) [1]

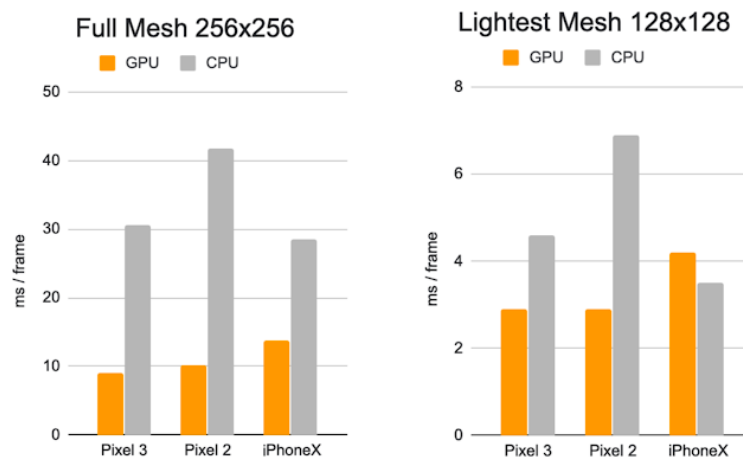


Figura 4.8: Rendimiento de FaceMesh sobre dispositivos móviles

Prototipo

Uso de *FaceMesh* del *framework* *MediaPipe* y un detector basado en *Haar-like features*, como el mencionado en el *paper* de Viola & Jones (4.1).

4.4. Tensorflow

TensorFlow es una plataforma de *Open Source* dedicada al aprendizaje automático. Permite compilar e implementar con facilidad aplicaciones con tecnología de AA. Esta se basa en tensores, matrices multidimensionales con un tipo uniforme, que si se está familiarizado con NumPy , los tensores son como *np.arrays*. Con la característica de que nunca se puede actualizar el contenido de un tensor, solo crear uno nuevo. [24]

Concretamente, se usarán los modelos y ejemplos de aprendizaje automático ofrecidos y entrenados mediante la API de alto nivel de Tensorflow para la implementación del último prototipo. Este recurso recibe el nombre de *TensorFlow Model Garden* [25], y se trata de un repositorio con un diferentes implementaciones de modelos y soluciones modeladas para Tensorflow. Los modelos estan pre-entrenados mediante un dataset llamado *COCO 2017* [4]. Siendo este un gran conjunto de datos a grande escala para *object detection*, segmentación y captación.

Model Garden dispone de un apartado dedicado a la visión artificial, donde se encuentran los ámbitos de clasificación de imágenes, como *MNIST*, *ResNet* o *EfficientNet*, y detección de objetos y segmentación, donde se destacan *RetinaNet*, *Mask R-CNN*, etc. Sin embargo, este prototipo se centrará en la implementación de uno de los modelos que ofrece Model Garden, llamado *SSD-MobileNet*, y aplicar *Transfer Learning* sobre el, para obtener un detector de lo que estamos buscando, comprobar si la mascarilla esta bien vestida.

Transfer Learning es una técnica en la que se reusa un model pre-entrenado para un nuevo problema. Últimamente su uso se esta popularizando, ya que permite el entrenamiento de una *deep neural network* con una pequeña cantidad de datos. Algo bastante revolucionario, puesto que todos los modelos hasta ahora necesitaban millones de datos, clasificados a mano y posteriormente necesitar una gran capacidad computacional para ser entrenados.[26] En nuestro caso, se busca utilizar un modelo general de detección de objetos y aplicando transfer learning, re-entrenarlo para una tarea más especifica pero sin desperdiciar sus conocimientos previos.

SSD-MobileNet

El prototipo de Tensorflow que se va a construir usará un modelo *SSD-MobileNet*, combinación de un modelo SSD, *Single Shot MultiBox Detector*, con otro llamado MobileNet. Este tipo de modelos se caracterizan por el uso de un método para detectar objetos en imágenes mediante una *deep neural network*, que genera valores de predicción sobre la presencia de cada objeto/categoría en cada una de las detecciones y devuelve el objeto con el más coincida.

El modelo SSD se caracteriza por estar formado por dos componentes principales: un modelo backbone y una cabeza SSD. El backbone es un modelo de clasificación de imágenes, que esta pre-entrenado para ser utilizado como un extractor de características. Normalmente, se implementa con modelos como *ResNet*. Mientras que, la cabeza SSD esta formada por uno o mas capas convolucionales que interpretan la salida del backbone como *bounding boxes* y clases de objetos. [8]

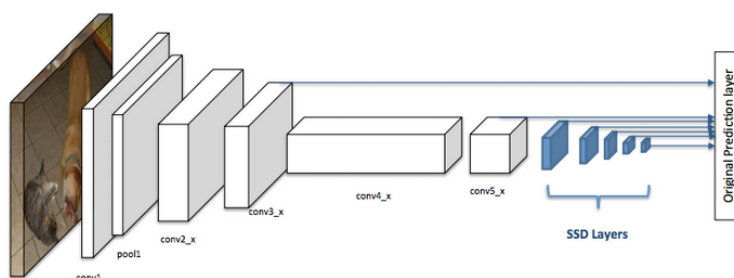


Figura 4.9: Arquitectura de una CNN con uso de SSD.

Otras característica destacables de SSD es la división de la imagen de entrada mediante un *grid* y realizar una predicción de la clase y la localización del objeto en cada una de las celdas de dicho grid. Si en dicho grid no existe ningún objeto, se considera como *background* y la localización es ignorado. A cada celda se le puede asignar varias *anchor boxes* (*Bounding boxes* con una altura y anchura predefinida) con un tamaño igual al de las celdas. Pero no todos los objetos que estan contenidos en la imagen tienen dicho tamaño, por eso se añade un parámetro (*ratio*) dedicado a especificar los diferentes valores que pueden tomar las *anchor boxes*. Asimismo, existe otro parámetro llamado zoom para especificar cuanto pueden escalar dichos *boxes*, tanto aumentando como reduciendo.[8]

El comportamiento y aportaciones específicas de este modelo, vienen planteadas en el paper con su mismo nombre, "*SSD: Single Shot MultiBox Detector*" [18]. Y son las siguientes:

1. Se introduce un modelo para la detección de múltiples categorías que es más eficaz y preciso que los modelos posteriores como YOLO.
2. El núcleo de SSD predice la categoría y localización de la *bounding box* mediante el uso de pequeños filtros convolucionales aplicados sobre mapas de características.
3. Para obtener una mayor precisión se realizan predicciones sobre mapas de características con diferente escalado.
4. Experimentación con análisis de tiempo y precisión sobre el modelo evaluado sobre *PASCAL VOC*, *COCO* y *ILSVRC*. Asimismo, es comparado con modelos recientes.

MobileNet es un modelo con arquitectura *CNN* para la clasificación de imágenes y visión artificial en móviles. Lo que hace especial este modelo es la potencia computacional necesaria para ejecutarlo o aplicar transfer learning sobre el. Esto lo convierte en un perfecto candidato para dispositivos móviles, sistemas integrados y ordenadores sin GPU o baja eficiencia computacional sin llegar a comprometer significativamente la precisión de los resultados.

Este modelo se basa en el uso de una arquitectura optimizada que usa convoluciones separables en profundidad para construir *deep neural networks* ligeros. Además, proporciona dos hiperparámetros globales que permiten ajustar de forma eficiente entre la latencia y precisión. Por consiguiente, con la combinación de ambos, se logra la creación de un modelo apto para dispositivos y ordenadores con poca potencia de GPU, capaz de obtener una detección de forma precisa y veloz.

Prototipo

4.5. Comparación

CAPÍTULO 5

Conclusiones y vías futuras

Finalizamos el trabajo estableciendo las conclusiones y vías futuras...

Bibliografía

- [1] Artsiom Ablavatski and Google AI Ivan Grishchenko, Research Engineers. Real-time ar self-expression with machine learning. 03 2019. Acceso: 25-04-2021.
- [2] AISHak. Integral images in opencv. <https://aishack.in/tutorials/integral-images-opencv/>, Jun 2010. Acceso: 07-04-2021.
- [3] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. 07 2019.
- [4] COCO. Coco, common objects in context, 2021. Acceso: 02-05-2021. URL: <https://cocodataset.org/#home>.
- [5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. doi:10.1109/CVPR.2005.177.
- [6] Rostyslav Demush. A brief history of computer vision (and convolutional neural networks). 02 2019. URL: <https://hackernoon.com/a-brief-history-of-computer-vision-and-convolutional-neural-networks-8fe8aacc79f3>.
- [7] Konstantinos Derpanis. Integral image-based representations. 1, 01 2007.
- [8] ArcGIS Developers. How single-shot detector (ssd) works?, 2021. Acceso: 05-05-2021. URL: <https://developers.arcgis.com/python/guide/how-ssd-works/>.
- [9] Dlib. Dlib library, 2021. Acceso: 25-04-2021. URL: <http://dlib.net>.

- [10] Govinda Dumane. Introduction to convolutional neural network (cnn) using tensorflow. 2020. Acceso: 28-04-2021. URL: <https://towardsdatascience.com/introduction-to-convolutional-neural-network-cnn-de73f69c5b83>.
- [11] Robert E. Schapire. Explaining adaboost. 2020. Acceso: 11-04-2021. URL: <https://www.math.arizona.edu/~hzhang/math574m/>.
- [12] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Pearson, 2018.
- [13] Google. Mediapipe face mesh. 2020. Acceso: 21-04-2021. URL: https://google.github.io/mediapipe/solutions/face_mesh.
- [14] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. 07 2019.
- [15] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. 06 2014. doi:10.13140/2.1.1212.2243.
- [16] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. doi:10.1109/CVPR.2014.241.
- [17] Soret Lee. Understanding face detection with the viola-jones object detection framework. 2020. Acceso: 11-04-2021. URL: <https://towardsdatascience.com/understanding-face-detection-with-the-viola-jones-object-detection-framework-c55c>.
- [18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016. URL: http://dx.doi.org/10.1007/978-3-319-46448-0_2, doi:10.1007/978-3-319-46448-0_2.
- [19] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. 06 2019.

- [20] Satya Mallick. Histogram of oriented gradients explained using opencv. 2016. Acceso: 02-05-2021. URL: <https://learnopencv.com/histogram-of-oriented-gradients/>.
- [21] Takeshi Mita, Toshimitsu Kaneko, and Osamu Hori. Joint haar-like features for face detection. *IEEE Int Conf Comp Vis*, 2:1619 – 1626 Vol. 2, 11 2005. doi:10.1109/ICCV.2005.129.
- [22] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. doi:10.1109/ICCVW.2013.59.
- [23] R. Szeliski. *COMPUTER VISION: Algorithms and aplications*. SPRINGER NATURE, 2021.
- [24] TensorFlow. Tensorflow core, 2021. Acceso: 02-05-2021. URL: https://www.tensorflow.org/guide?hl=es_419.
- [25] Tensorflow. Tensorflow official models, 2021. Acceso: 02-05-2021. URL: <https://github.com/tensorflow/models/tree/master/official>.
- [26] TensorFlow. Transferir el aprendizaje y la puesta a punto, 2021. Acceso: 25-04-2021. URL: https://www.tensorflow.org/tutorials/images/transfer_learning.
- [27] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Conf Comput Vis Pattern Recognit*, 1:I-511, 02 2001. doi:10.1109/CVPR.2001.990517.