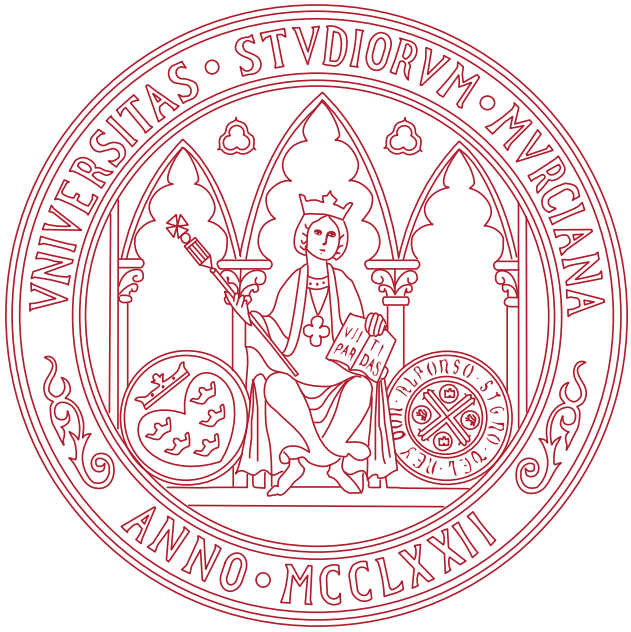

Universidad de Murcia

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

17 de abril de 2021

Me gustaría expresar mi agradecimiento a ...

Índice general

Índice de figuras

Resumen

En este proyecto se plantea el problema de...

Extended abstract

This project faces the problem of...

CAPÍTULO 1

Introducción

La vision artificial es un ámbito de la informática que surgió hace 60 años, pensado en el estudio del procesamiento digital de las imágenes. De hecho, uno de los primeros acontecimientos que propicio su aparición fue la creación del cable Bartlane, capaz de transmitir una imagen a traves del mar atlantico en los años 1920, con una duración de cerca a una semana.

1.1. Historia de la Visión Artificial

Dentro del entorno del procesamiento de imagenes digitales, se centró la investigacion en recuperar una estructura tridimensional del mundo real a traves de una imagen para conseguir un entendimiento total de la escena que plasma la imagen. A su vez aparecieron varios algoritmos de reconocimiento de lineas, donde uno de ellos fue creado por parte de Huffman en 1971.

Pero el avance de estas investigaciones pronto se ligarían al avance del ordenador. Estos, se podrian dividir en varios acontecimientos importantes, tales como: la invención del transistor por Bell Laboratories en 1948, la invencion de los circuitos integrados en 1958 o la introduccion por parte de IBM de los primeros ordenadores personales en 1981.

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 vision. Un par de psicologos, David Hubel y Torsten Wiesel, describieron que el comportamiento de las neuronas encargadas de entender el entorno visual siempre empiezan con estructuras simples como vertices. Más tarde esta idea se convertirá en el principio central del deep learning.

Este estudio se realiza antes de que los ordenadores pudiesen entender imagenes. Russell Kirsch, en 1959, es el primero en desarrollar un aparato que traducia las imagenes en datos que las maquinas pudiesen entender. Y, Lawrence Roberts en 1963 publica un estudio sobre como las maquinas perciben objetos solidos de tres dimensiones, uno de los avances considerados precursores de la vision artificial moderna.

En 1982, David Marr, le da un punto de vista diferente a la vision artificial. Idea un framework para vision que incluye un esquema con la representacion de las partes principales de la imagen, otro esquema con las superficies y la pronfundidad de la informacion, y un modelo 3D. Al mismo tiempo se desarrollo un red artificial capaz de reconocer patrones, mediante el uso de una red convolucional.

Tras estos descubrimientos se presento un modelo llamado LeNet-5, la primera red convolucional moderna. Este modelo se caracteriza por usar la backpropagación. En los años 1990s la vision artificial cambia totalmente de rumbo y los investigadores pasaron de intentar reconstruir objetos en 3D a intentar detectar objetos mediante sus caracteristicas.

Asimismo, esta epoca de 1980 se centro la investigacion y desarrollo de tecnicas matematicas para el entendimiento de imagenes y las escenas que estan representadas en ellas. Mientras que en 1990 se explotaron todas estas ideas conjuntamente con el desarrollo de la potencia de los ordenadores.

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, pero ademas que fuese en tiempo real. Se trata de un clasificador creado a partir de clasificadores mas debiles y busca las caras a partir de dividir la imagen de entrada en rectangulos y realiza un

estudio en cascada sobre los clasificadores debiles.

El problema de estos modelos es el uso de informacion para poder entrenarlos, y para esto se creo un proyecto llamado PASCAL VOC. Este creo un dataset estandar para la clasificacion de objetos. Posteriormente, aparecieron más dataset como este. Uno de ellos, en 2010, ImageNet contiene mas de un millon de imagenes para un total de mil objetos. Junto a este dataset aparecio un modelo basado en una red convolucional llamado AlexNet.

1.1.1. Reconocimiento Facial

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 en general. 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*.

Pero tras la aparición de COVID-19 y el uso de las mascarillas, la gran mayoría de las aplicaciones o investigaciones que hacían uso de estas tecnologías se han quedado obsoletas o funcionan de una forma mas pobre. Por eso, el objetivo de este trabajo será el estudio de posibilidades que solucionen este problema y terminar consiguiendo un prototipo con el que se puedan detectar rostros humanos que vistan una mascarilla, e incluso poder indicar cuando la viste mal o no la lleven.

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 caracteristicas 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 unas de las tecnicas principales en el reconocimiento facial.

Este ejercicio se complica cuando las imagenes 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 especifico mascarillas.

2.1. Aplicaciones sin Deep Learning

2.1.1. HAAR-like Features & ADABOOST

2.1.2. HOG & DLIB

2.1.3. Facial Landmasking

2.2. Aplicaciones con Deep Learning

¿Que es el deep learning?

2.2.1. YOLO

2.2.2. Tensorflow

Posiblemente meter modelos como Mobilenet, etc.

CAPÍTULO 3

Análisis de objetivos y metodología

Una vez tratado el estado del arte, realizamos en este capítulo el análisis de los objetivos del mismo, estableciendo los elementos concretos que nos van a permitir llevar a cabo dichos objetivos así como la metodología de desarrollo.

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"[9]

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, haciendose difícil para posiciones laterales o inclinadas.

Las imagenes que se toman para realizar la detección pasan por una transformación del espacio de color a *grayscale*. Con el objeto de encontrar caracteristicas 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:

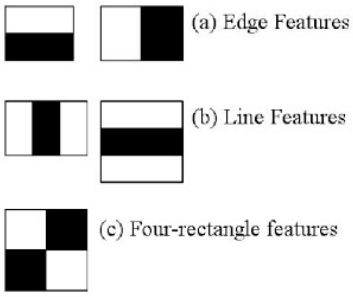

Figura 4.1: Haar-like Features

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 pixeles que forma el *feature* tendra una division clara entre pixeles de color blanco con los de color negro (Figura 4.1), pero en la realidad eso casi nunca se va a dar.

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

el cambio del tamaño, posición y forma de las regiones rectangulares basandose en la resolución que se define en el detector. [7]

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 instantanea. Además de ser muy útiles para las *HAAR-like features*, tambien lo son en muchas otras aplicaciones.

Si se supone una imagen con unas dimensiones de $< w, h >$ (ancho y alto, respectivamente), la imagen integral que la representa tendrá unas dimensiones de $< w+1, h+1 >$. La primera fila y columna de esta son ceros, mientras que el resto tendrán el valor de la suma de todos los pixeles que le preceden. [1] Ahora, para caluclar la suma de los pixeles en una region especifica 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):

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

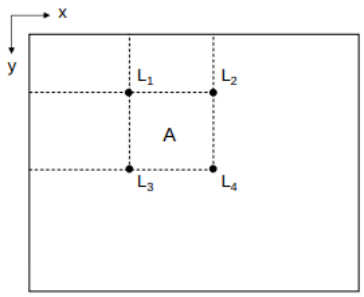

Figura 4.2: Funcionamiento de una *Imagen Integral*

Viola y Jones junta esta propuesta con los filtros *Haar-like features*, y consiguen computar dichas características de manera constante y eficaz. [3]

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* [4]. 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 escoje la que menos porcentaje de error presente en dicha iteración. (Como se muestra en la *Figura 4.3*) [6]

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. [4]

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*.

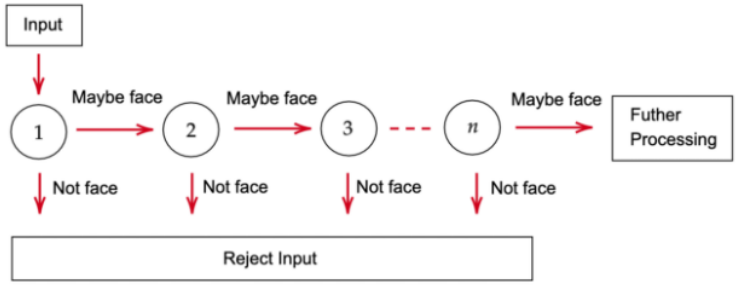

Figura 4.4: Construcción del *Multi-stage Classifier*

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 diferente 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.

[Explicacion del prototipo custom]

[Resultados]

4.2. Facial Landmark

Este apartado se podría adentrar en el ambito de deep learning, ya que sus pasos de ejecución serían: detectar una cara y posteriormente obtener los puntos de interes (Landmarks). El primer punto puede realizarse de dos maneras, tanto con la tecnica explicada anteriormente como con HOG + SVMLinear,

Haar y Facial Landmark

Implentacion de facial landmark con las ideas anteriores sobre bolsas de features.

Deep Learning

Implementacion de facial landmark para deteccion de mascarillas mediatne el uso de la API Mediapie o con el uso de otro tipo de modelo.

Implementación

4.3. YOLO

Mostrar la idea de lo que es una CNN y posteriormente explicar el funcionamiento de YOLO. Posible añadido de AlexNet si es util.

Implementación

YOLO Custom

4.4. Tensorflow

- Plantear idea
- Procedimiento
- Diferentes tipos de modelos
- Mostrar funcionamiento y aplicación al objetivo

CAPÍTULO 5

Conclusiones y vías futuras

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

APÉNDICE A

ANEXO I - Implementación Viola-Jones Face Detector

Bibliografía

[1] AIShak. Integral images in opencv. https://aishack.in/tutorials/integral-images-opencv/, Jun 2010. Acceso: 07-04-2021.

[2] 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.

[3] Konstantinos Derpanis. Integral image-based representations. 1, 01 2007.

[4] Robert E. Schapire. Explaining adaboost. 2020. Acceso: 11-04-2021. URL: https://www.math.arizona.edu/~hzhang/math574m/.

[5] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Pearson, 2018.

[6] Socret 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-c55cc2

[7] 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.

[8] R. Szeliski. *COMPUTER VISION: Algorithms and aplications.* SPRINGER NATURE, 2021.

[9] 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.