

Grado en Ingeniería en Tecnologías de
Telecomunicaciones
2022-2023

Trabajo Fin de Grado

“Análisis comparativo de sistemas de
computer vision.”

Javier Cuenca Gento

Tutor
Daniel Díaz Sánchez



Esta obra se encuentra sujeta a la licencia Creative Commons **Reconocimiento - No Comercial - Sin Obra Derivada**

RESUMEN

En este trabajo se ha desarrollado una comparación entre los frameworks de computer vision de las plataformas cloud de tres de las principales empresas a nivel mundial en tecnología; Google, Microsoft y Amazon. Para ello, habrá tres partes, una primera parte para el estudio y entendimiento del estado del arte de la visión por ordenador y de cada una de las plataformas, sus herramientas y características, una segunda parte para un estudio objetivo de estas plataformas con un script en Python y por último un estudio subjetivo que contará con una página web en la que los usuarios podrán elegir descripciones de cada API y también probarlas a tiempo real.

Palabras clave: Computer vision, Label Detection, Python, Google, Google Cloud, Amazon, Amazon Web Services, Rekognition, Microsoft, Microsoft Azure, API, scripts, web, Javascript, Node, Docker, DNS, survey, SQL, PostgreSQL.

ÍNDICE GENERAL

1. INTRODUCCIÓN	1
2. ESTADO DEL ARTE	2
2.1. Desarrollo de la visión por ordenador	3
2.1.1. Inicios y desarrollo.	3
2.1.2. Estado del arte - actualidad	4
2.1.3. Futuro próximo	5
2.2. Computer Vision dentro de la IA	7
2.3. Función del etiquetado de imágenes	9
3. FRAMEWORKS DE COMPUTER VISION RELEVANTES	11
3.1. Google Cloud Vision.	12
3.2. Amazon Web Services Rekognition	14
3.3. Microsoft Azure Computer Vision	16
3.4. Competidores y otras empresas	17
4. ANÁLISIS OBJETIVO DEL USO DE LOS FRAMEWORKS	18
4.1. Análisis del uso de las APIs	19
4.2. Definición y diseño del script	20
4.3. Comparación de precios	23
4.4. Resultados de tiempo y tamaño	24
5. ANÁLISIS SUBJETIVO DEL USO DE LOS FRAMEWORKS	26
5.1. Página web - sistema de obtención de datos	27
5.1.1. Proceso de creación y despliegue	27
5.1.2. Diseño y utilización de web	28
5.2. Resultados de las votaciones	32
6. CONCLUSIONES Y TRABAJO FUTURO	35
6.1. Conclusiones	35
6.2. Trabajo futuro	36
7. MARCO REGULATORIO	37
BIBLIOGRAFÍA	38

ÍNDICE DE FIGURAS

1.1	Ejemplo de computer vision. Reconocimiento de objetos.	1
2.1	Ejemplo de reconocimiento facial con markpoints.	2
2.2	Larry Roberts y sus primeros estudios.	3
2.3	ViT. <i>Vision Transformer</i>	4
2.4	Funcionalidades de YOLO. [4]	4
2.5	Estado del arte. Modelos de etiquetado o clasificación de imágenes.	5
2.6	Gráfica del ciclo del <i>hype</i> de la inteligencia artificial. Gardner.	6
2.7	Esquema de uso de SIFT.	7
2.8	Esquema de relación entre AI y Computer Vision. [8]	8
2.9	Ejemplo de etiquetado de imágenes para el control de tráfico.	9
2.10	Esquema de arquitectura de una red neuronal convolucional.	10
3.1	Logo de Google Cloud.	12
3.2	Gráfico utilización de Google Cloud.	12
3.3	Logo de Amazon Web Services.	14
3.4	Gráfico de utilización de Amazon Web Services.	14
3.5	Logo de Microsoft Azure.	16
3.6	Gráfico de utilización de Microsoft Azure.	16
3.7	Competidores en Computer Vision.	17
4.1	Esquema general y simplificado de las llamadas a APIs del script.	19
4.2	Ejemplo extraído del script.	20
4.3	Imagen resultado de la descripción de Google Cloud Vision.	21
4.4	Imagen resultado de la descripción de Amazon Web Services Rekognition.	21
4.5	Imagen resultado de la descripción de Microsoft Azure Computer Vision.	22
4.6	Gráfica resultado del script con la variable de tiempo.	24
4.7	Gráfica resultado del script con la variable de tamaño.	25
5.1	Esquema de utilización de herramientas en el proyecto web.	27

5.2	Página de Inicio. Escritorio y móvil.	28
5.3	Página de Encuesta. Escritorio y móvil.	29
5.4	Página de Términos y Condiciones.	29
5.5	Página de Pruébalo. Escritorio y móvil.	30
5.6	Página de Pruébalo una vez subida la imagen. Escritorio y móvil.	31
5.7	Página principal de Computer Vision.	31
5.8	Resultados de la encuesta clasificados por categorías.	32
5.9	Resultados de la sección <i>Pruébalo</i>	33
5.10	Resultados de la sección <i>Pruébalo</i> comparados con los de la encuesta. . .	34
5.11	Resultados totales del proyecto web.	34

1. INTRODUCCIÓN

La visión por ordenador o *computer vision*, es una herramienta, campo o técnica muy recurrente en los últimos tiempos en la ciencia y la tecnología, enfocado en dar a los ordenadores el poder de entender, identificar o catalogar imágenes y videos de una manera visible y entendible para el usuario. La variedad de casos de uso y de posibilidades de integraciones y despliegues ha hecho a este tipo de herramientas tener un gran peso en la tecnología y, en especial, en un campo de la informática que ha ganado globalidad y conocimiento público como es la inteligencia artificial.

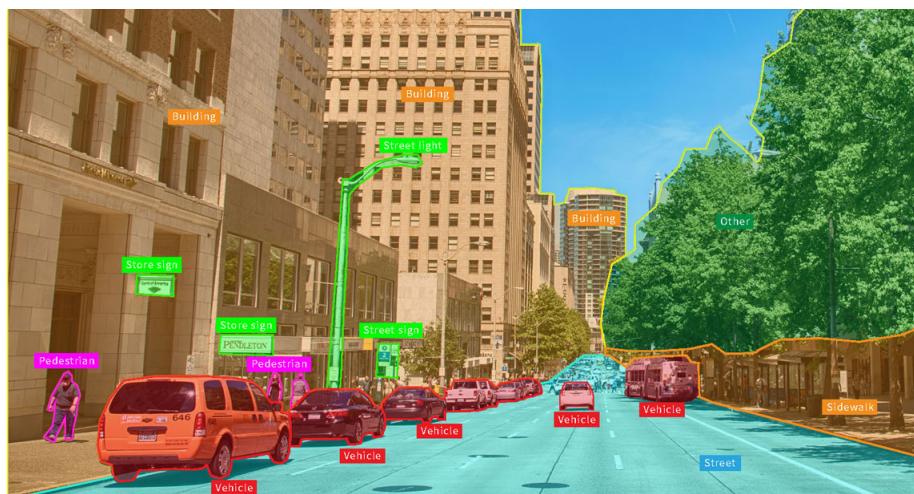


Fig. 1.1. Ejemplo de computer vision. Reconocimiento de objetos.

En este proyecto se describirán brevemente los conceptos de visión por ordenador, inteligencia artificial o etiquetado de imágenes, también se analizarán el desarrollo de la visión por ordenador y del etiquetado de imágenes, desde sus inicios hasta el futuro próximo, pasando por el estado del arte de este campo. Por último en cuanto a la investigación teórica, se definirán brevemente los fundamentos técnicos del etiquetado de imágenes utilizado en la mayor parte de este proyecto, siendo el objetivo de este la comparación entre tres frameworks punteros de manera objetiva y subjetiva.

2. ESTADO DEL ARTE

Profundizando en la introducción dada anteriormente, la visión por ordenador se puede definir como un conjunto de procesos, o entrenamientos, dados a un sistema para que este trate de entender un conjunto de píxeles relativamente en el mismo proceso que lo hace el cerebro humano, desde la unión de líneas y figuras a la comprensión de una imagen completa, o en su defecto de una serie cronológicas de imágenes, es decir, de un vídeo.

En este sentido, el campo de computer vision puede ser dividido en diferentes clases de resultados como el clásico reconocimiento de objetos, el análisis facial o el discutido etiquetado de imágenes o *image labeling*, el cual consiste en marcar diferentes imágenes o información en bruto con etiquetas a las cuales se les puede dar diferentes usos o valores.

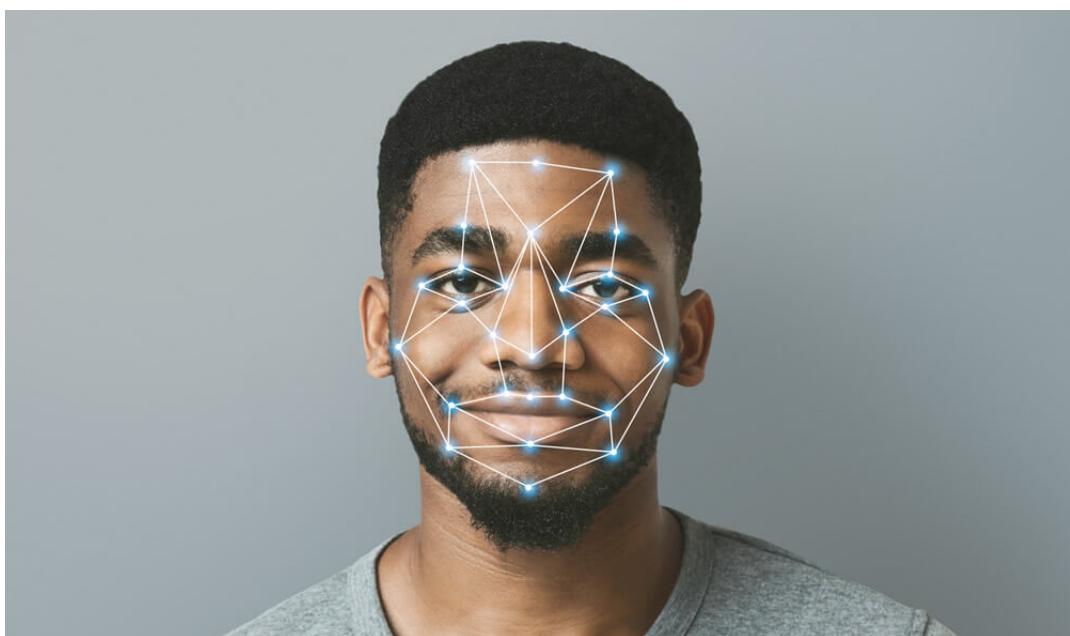


Fig. 2.1. Ejemplo de reconocimiento facial con markpoints.

Basando el recorrido de la visión por ordenador en una imagen global y, posteriormente, enfocado en el etiquetado de imágenes, a continuación se describirá el desarrollo, las nuevas integraciones con inteligencia artificial y el poder del etiquetado de imágenes en la visión por ordenador.

2.1. Desarrollo de la visión por ordenador

En esta sección se discutirá sobre el recorrido que han tenido las tecnologías que ahora son usadas y conocidas como computer vision, desde sus inicios unidos al desarrollo de los propios ordenadores hasta la actualidad con la variedad multidisciplinar en la que se encuentra, contando también el horizonte de futuro que le puede esperar a este tipo de herramienta.

Compartiendo recorrido con el desarrollo de la inteligencia artificial, la visión por ordenador tiene sus inicios en los años 60, un cierto estancamiento en las décadas posteriores y un fuerte crecimiento en los últimos años. A continuación se describirán estas etapas, además de lo que se puede esperar de este campo en el futuro próximo.

2.1.1. Inicios y desarrollo

Como se ha comentado anteriormente, los inicios de la visión por ordenador rondan el año 1960 y tienen a una importante figura de la tecnología detrás como es Larry Roberts, informático teórico, uno de los padres de internet y, también, de la visión artificial. Sus primeros estudios giraron alrededor de la posibilidad de captar información 3D de bloques 2D. A partir de ello investigadores, en especial del MIT, trataron de traspasar este hito a imágenes del mundo real. [1]

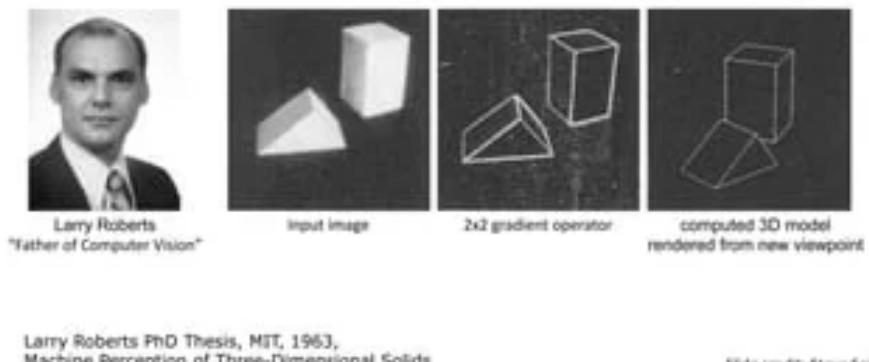


Fig. 2.2. Larry Roberts y sus primeros estudios.

En los años siguientes el progreso no fue el esperado, debido a los famosos períodos de *AI Winter*, en los cuales el desarrollo, la investigación o la producción de técnicas o herramientas tanto de inteligencia artificial como de visión por ordenador fue reducido a mínimos. Pese a ello hubo ciertos avances como el algoritmo Viola-Jones para la detección de objetos propuesto a principios de la década del 2000. Este, con una probabilidad de verdaderos positivos del 99,9 % y una probabilidad de falso positivos del 3,33 %, datos muy correctos para la época, combinado con una optimización como la propuesta por

Egorov podía aumentar entre 2 y 5 veces su velocidad reduciendo la tasa de precisión únicamente un 3 o 5 por ciento. [2]

A partir de esa época y hasta la actualidad se ha seguido avanzando en cada campo de la inteligencia artificial y de la visión por ordenador, con avances tan variados como la primera competición de *ImageNet* en 2010, esta plataforma cuenta ahora con casi 15 millones de imágenes en 20 mil categorías, siendo varias de ellas utilizadas en este proyecto. Como otros ejemplos del avance, en 2015 Google creó *FaceNet* como uno de los pioneros del reconocimiento facial o la creación de las VPU o *visual processing unit* como variación de las clásicas CPUs pero completamente enfocadas a la visión por ordenador entre otras técnicas de inteligencia artificial.

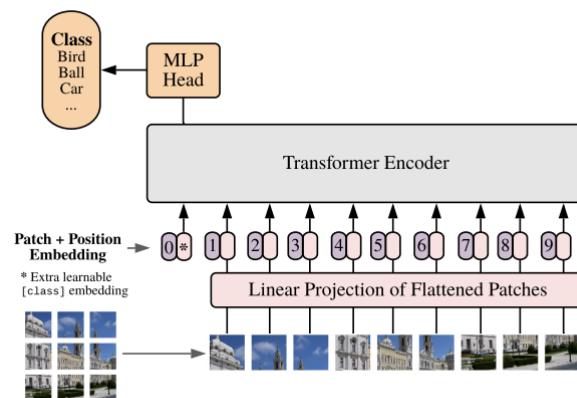


Fig. 2.3. ViT. *Vision Transformer*.

2.1.2. Estado del arte - actualidad

En los últimos años en el campo de *Computer Vision* han aparecido con fuerza los algoritmos YOLO o *You Only Look Once*, como sistemas de código libre para la detección y clasificación de imágenes, también se encuentran actualmente otros modelos como el *Segment Anything Model* de Meta, *GroundingDINO* como modelo de detección combinando un detector DINO basado en un *transformer* con un preentrenamiento, el *Detectron2*, *Mask RCNN* o *OpenAI Clip*. [3]

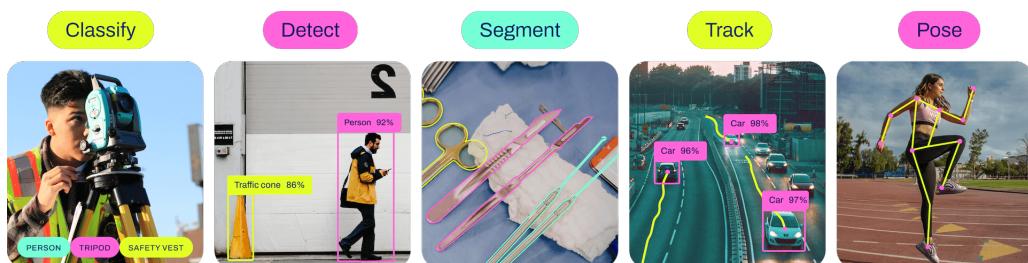


Fig. 2.4. Funcionalidades de YOLO. [4]

Estos modelos ofrecen detección, clasificación, segmentación o una combinación de todo ello, pero para este proyecto se buscará el estado del arte del etiquetado de imágenes. En este campo, los modelos actuales con mayor precisión son *CoCa*, *Model Soups*, *ViT-e* o *ViT-G* entre otros pero actualmente el modelo con el mayor valor de precisión es el *BASIC-L*, basado en *Lion* o *EvoLved Sign Momentum*, esta técnica entrena de manera muy eficiente redes neuronales profundas. Este tiene unos resultados de 88.3 % en *zero-shot*, un enfoque en el que se clasifican imágenes nunca vistas, y un 91.1 % en *fine-tuning*, el enfoque en el cual se preentrena el modelo, estos valores son un 2 % y un 0.1 % mayores que los recibidos con otros algoritmos como el *Adam*. [5]

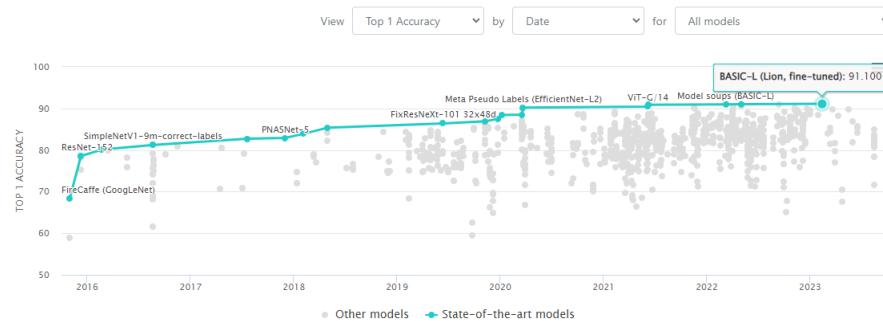


Fig. 2.5. Estado del arte. Modelos de etiquetado o clasificación de imágenes.

Hasta ahora se ha analizado el estado del arte en cuanto al campo de la visión por ordenador, ya sean algoritmos de un uso global o general o específicos para alguna técnica como puede ser la clasificación de imágenes. En cambio, para el objetivo de este proyecto, no se ha buscado el algoritmo ni el modelo con mayor precisión ni el más rápido, ya que la comparación se hará entre frameworks de grandes empresas, no únicamente entre los modelos en sí. En este sentido se pueden observar diferentes actualizaciones en Google Cloud Vision a finales de 2022 en cuanto a detección facial y de puntos de referencia, o del modelo OCR en mayo de 2022. En Amazon Web Services, en cambio a lo largo de 2023 han realizado muchas actualizaciones de características como detección de personas vivientes en vídeo o imagen, detección de rostros con mala visibilidad o parcialmente tapados, mientras que a finales de 2022 actualizaron tanto Rekognition Image como Rekognition Video. Por último Microsoft Azure sacó en mayo de 2023 Image Analysis 4.0, el cual mejora los detalles y el entendimiento semántico de las respuestas, mientras que van a retirar planean retirar la versión 3.0 en otoño.

2.1.3. Futuro próximo

Se estima que la visión por ordenador ya esté en su fase de establecimiento, atendiendo a las famosas gráficas del ciclo del *hype* de la inteligencia artificial, cerca de llegar a su pico de productividad, por ello se prevé que los cambios dejen de ser disruptivos y comiencen a estabilizar las técnicas y herramientas actuales, introduciéndolos en mayor medida en el mercado de la tecnología global. Pese a ello, la robustez de algunos modelos

sigue siendo un problema debido a sus complicaciones, como podría ser en la conducción autónoma de los vehículos, solución que necesitaría una precisión extremadamente alta y segura.

Por otro lado, no todos los expertos afirman que el enfoque del aprendizaje profundo sea siempre superior a un enfoque utilizando la física o a algún otro enfoque por descubrir, debido al continuo requerimiento de un número ingente de datos, sin afinar las técnicas dependiendo de lo requerido ya sea segmentación, seguimiento, clasificación o cualquier otro objetivo. Por ello, el futuro próximo de un campo tan amplio como la visión por ordenador es difícil de prever. [6]

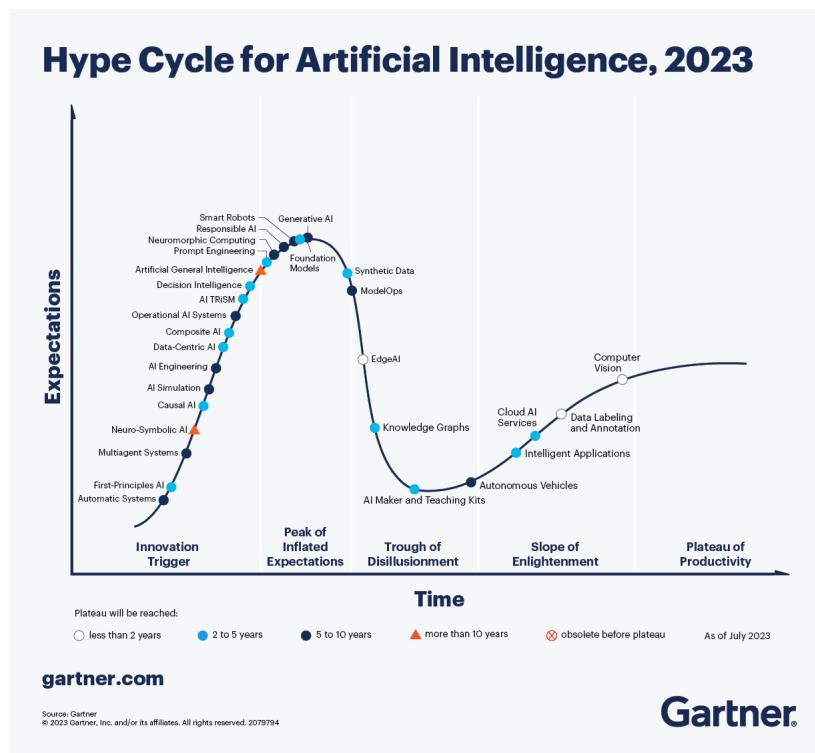


Fig. 2.6. Gráfica del ciclo del *hype* de la inteligencia artificial. Gardner.

2.2. Computer Vision dentro de la IA

En este apartado se hará una pequeña incursión en la inteligencia artificial, debido a la notoriedad que ha adquirido en los últimos tiempos no solo en el mundo tecnológico y científico si no también en el público general.

Es preciso diferenciar entre campos exclusivamente dentro de la inteligencia artificial, como podrían ser *Machine Learning* o *Deep Learning*, estas disciplinas son subconjuntos de la inteligencia artificial, la primera, también llamada aprendizaje automático consiste en entrenar conjuntos de datos masivos para así *aprender* de estos y sacar conclusiones de ellos para poder ser utilizados potencialmente con datos nuevos. En cambio el aprendizaje profundo, o *Deep Learning*, es otro subconjunto en sí del aprendizaje automático, el cual trata de aprender del propio proceso de entrenamiento para tener un cierto poder de elegir por sí mismo nuevas decisiones.

En este sentido, la visión por ordenador no es exclusivamente un subconjunto de la inteligencia artificial, pues puede consistir o no de varias disciplinas dependiendo de los casos de uso o herramientas utilizadas para cada objetivo, teniendo sobretodo en cuenta de que época es cada estudio o cada herramienta, siendo obviamente mucho más común encontrar uniones entre *Computer Vision* y *Machine Learning* en los últimos años.

Algunos de los modelos de computer vision fuera de la inteligencia artificial pueden ser *Scale-Invariant Feature Transform (SIFT)*, creado por David Lowe en 1999 y que se basa en detectar *keypoints*, figura a continuación, dentro de imágenes apoyándose en imágenes ya clasificadas guardadas en una base de datos, o *Features from Accelerated Segment Test (FAST)*, modelo creado básicamente y simplificando mucho para detectar esquinas y bordes, enfocándose rápidamente en distintos píxeles. [7]

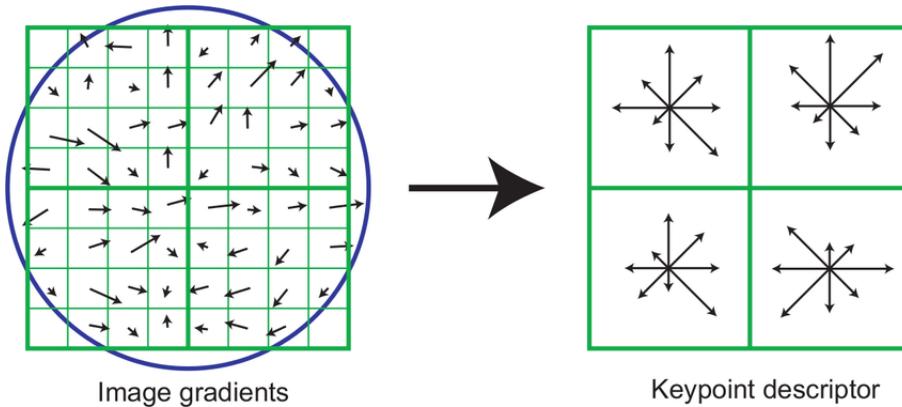


Fig. 2.7. Esquema de uso de SIFT.

Hasta la década del 2010, como se ha mencionado previamente, la visión por ordenador se basaba en herramientas para quitar ruido, encontrar bordes, detectar texturas y modificar diferentes formas para conseguir elementos deseados. Pero a partir de esta época, todo evoluciona en un mismo sentido, y la visión por ordenador también iba a ser influida por la inteligencia artificial. [8]

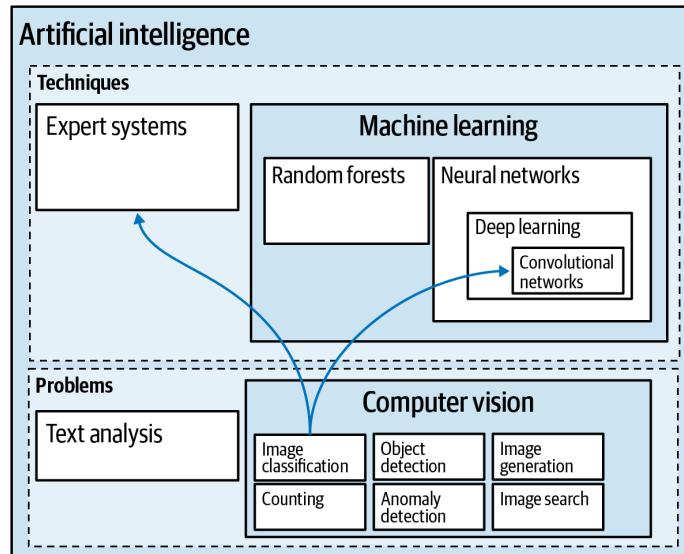


Fig. 2.8. Esquema de relación entre AI y Computer Vision. [8]

Un ejemplo claro de la nueva relación entre computer vision y *Machine Learning* o *Deep Learning* es la función clásica de reconocimiento de objetos. El típico modelo consta de tres partes, selección de la región con información, extracción de características y clasificación. La parte de extracción de características o *feature extraction* es también enormemente utilizado en el aprendizaje automático ya que se encarga de desechar aquella información que no será relevante a continuación. Por último, la clasificación es la parte principal que decide entre diferentes objetos o clases cual obtendrá mayor precisión, proceso realizado con los fundamentos del entrenamiento clásico de *Machine Learning*.

2.3. Función del etiquetado de imágenes

El etiquetado de imágenes o *image labeling* es el método de visión por ordenador que se utilizará y profundizará en este proyecto. Como se ha mencionado anteriormente, este método consiste en etiquetar o dar diferentes valores de etiquetas o clases a una imagen o a una sección de ella.

Una de las muchas características o funciones del etiquetado de imágenes es que podría denominarse como el método base de la visión por ordenador, es decir, el etiquetado de imágenes puede servir como base y mejora continua de muchos de los métodos más complejos de visión por ordenador como pueden ser el reconocimiento de objetos, tanto en imagen como en vídeo, que se podría denominar seguimiento de objetos, o la segmentación, también tanto en imagen como en vídeo. [9]

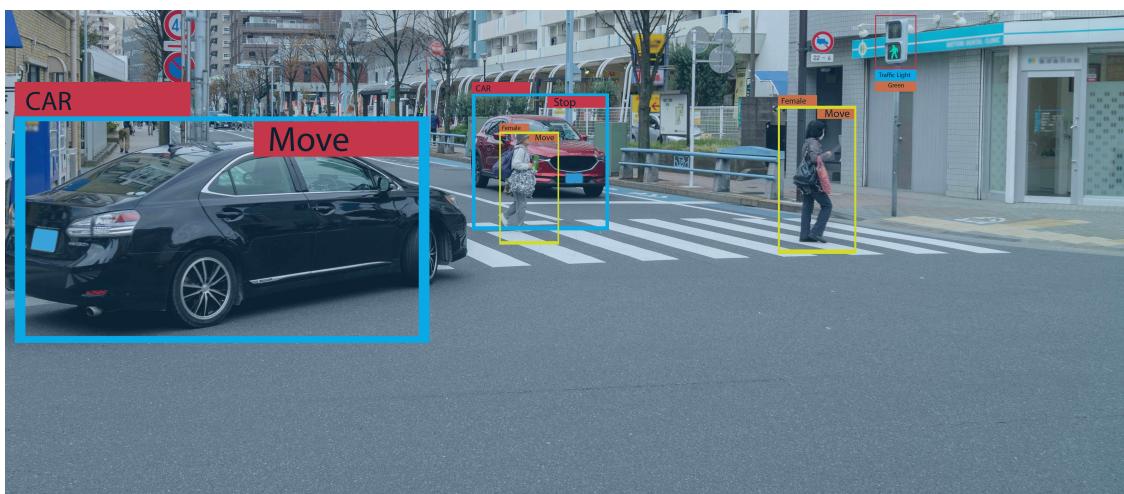


Fig. 2.9. Ejemplo de etiquetado de imágenes para el control de tráfico.

La arquitectura más usada en los últimos años en visión por ordenador, y especialmente en etiquetado de imágenes, es la de redes neuronales convolucionales o *CNN*. Esta arquitectura es el tipo de red neuronal más reconocible del campo de *Deep Learning*, al tener la ventaja de no necesitar supervisión humana en el entrenamiento. También cuenta con otras ventajas como la característica de *weight sharing* o reparto de pesos, característica que rebaja el número de parámetros de entrenamiento y por tanto contribuye a la generalización y reduce el sobreajuste.

Las redes neuronales convolucionales se basan en una estructura de capas, con distintas funciones dependiendo de donde se encuentren. El componente principal es la capa convolucional, la cual da nombre a la arquitectura y consiste de diferentes filtros convolucionales o *kernels*. La imagen es convolucionada con estos *kernels*, que son una cuadrícula de valores discretos que empiezan siendo aleatorios, o semialeatorios con ciertos métodos, y tras diferentes etapas del entrenamiento se convierten en los pesos que conseguirán extraer características. El resultado es un mapa de características que puede ser pasado por una capa de activación, la cual puede ser unitaria lineal rectilínea o *ReLU*, sigmoide o con una

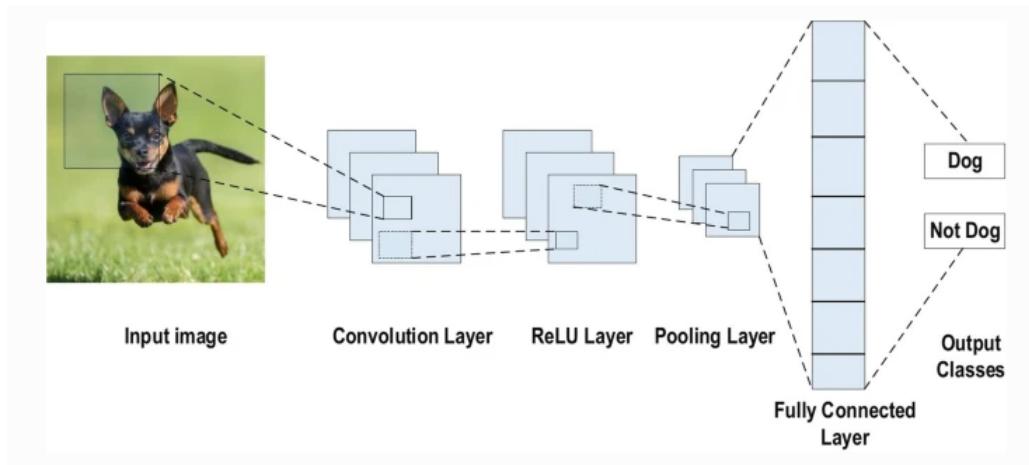


Fig. 2.10. Esquema de arquitectura de una red neuronal convolucional.

tangente hipérbolica entre otras. El resultado de nuevo se dirige a por la siguiente capa, la de agrupación o *pooling*, que puede ser de nuevo de distintas clases como por máximos, mínimos o por media global, esta capa crea diferentes agrupaciones de la información recibida sin variar demasiado la información. Por último se encuentra la capa totalmente conectada, que como bien dice su nombre conecta cada neurona anterior con cada una de las partes de esta capa, siendo la última parte de esta la que decide en qué clase o etiqueta se encuentra una imagen o una sección de ella. [10]

Con esta información se puede observar que el etiquetado de imágenes, pese a ser una técnica menos dirigida al caso de uso y a las integraciones directas con aplicaciones como sí se puede observar en el reconocimiento o segmentación para análisis y reacción de imagen y vídeo, es fundamental en el desarrollo de inteligencia artificial y visión por ordenador, siendo *image labeling* un pilar básico en este campo en continua evolución.

3. FRAMEWORKS DE COMPUTER VISION RELEVANTES

Para este proyecto se ha decidido centrarse en tres frameworks de computer vision, dentro de la gran variedad y competencia que hay en el mercado, estos son los de Google, Amazon y Microsoft, aprovechando también las plataformas de almacenamiento en cloud que ofrecen y el poder hacer una comparación de tres empresas que siempre compiten en cualquier campo de la tecnología y la información.

A continuación se describirá el proceso de inicialización en cada servicio cloud, el proceso de almacenamiento de imágenes y las correspondientes llamadas a las APIs de visión por ordenador, además de ciertos ajustes relativos a las llamadas y respuestas de estas herramientas para obtener resultados similares, por ejemplo como el uso de un traductor para las APIs que no disponen del idioma español. Para esto se dispondrá del módulo de Python del traductor de Google *googletrans* o la transformación de etiquetas de imágenes a una descripción construida también con inteligencia artificial, en este caso el creador de oraciones de *OpenAI*.

3.1. Google Cloud Vision

Se comenzará con Google Cloud y sus herramientas. Como herramienta clave para el objetivo de este proyecto se utilizará Cloud Vision API, que integra características como OCR o detección de caracteres, detección de contenido explícito o el etiquetado de imágenes, que será la función principal a usar para las llamadas principales. También como se ha comentado se usará una herramienta de almacenamiento como Cloud Storage para recoger las imágenes que serán usadas posteriormente.



Fig. 3.1. Logo de Google Cloud.

Para esto, es preciso registrarse en Google Cloud, y en primer lugar manejar los permisos para habilitar los servicios y las APIs desde la consola de Google Cloud. También es imprescindible para la mayoría de servicios y herramientas el crear una cuenta de facturación. Google Cloud también ofrece una herramienta de consola como es Google Cloud CLI, la cual es útil para instalar, inicializar y acreditarse a uno mismo para ser integrado donde se desee. Para ello es necesario realizar un auth login básico, seleccionar el proyecto en cuestión y entonces ejecutar el comando `gcloud auth application-default login`, con el que proporcionar las credenciales a ADC. En el caso de integrar con Windows también se debe habilitar el PowerShell correspondiente.

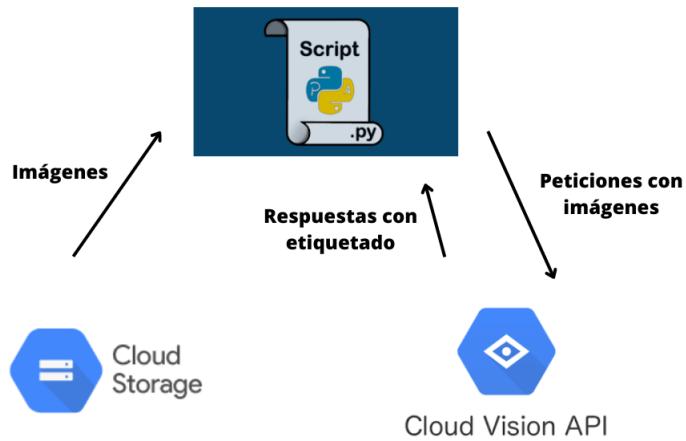


Fig. 3.2. Gráfico utilización de Google Cloud.

Las características principales y distintivas de Google Cloud Vision podrían ser la variedad de opciones que ofrece desde el momento de elección del producto. El servicio denominado Vision AI de Google engloba tres servicios diferenciados, por un lado

se encuentra Vertex AI Vision, un entorno de desarrollo que lleva dentro todas las características de visión por ordenador necesarias para gestionar aplicaciones directamente. Por otro lado tienen modelos personalizados de aprendizaje automático para entrenar directamente el conjunto de datos requerido para cada integración que desee el usuario. Y por último, está el producto básico utilizado en este proyecto, la API de Vision, esta API REST ofrece las facilidades de integración necesarias a un proyecto exterior como este, que requiere de diferentes plataformas cloud. [11]

3.2. Amazon Web Services Rekognition

A continuación es el turno de Amazon Web Services, con la herramienta principal AWS Rekognition, la cual cuenta también con distintas características como análisis facial, detección de texto en la imagen o, de nuevo la función utilizada, etiquetado de imágenes. Para el apartado de almacenamiento se utilizará la famosa herramienta de S3, en el que se utilizará un único bucket.



Fig. 3.3. Logo de Amazon Web Services.

La parte de inicialización y registro en Amazon Web Services es muy simple, en principio solo es necesario registrarse, en el caso de utilizar S3, crear un bucket de almacenamiento y descargar un simple archivo CSV que contendrá las credenciales del usuario, la región en la que se encuentra y diversos atributos más. Este archivo será uno de los pocos requerimientos que tiene AWS para integrar sus servicios con la aplicación objetivo.



Fig. 3.4. Gráfico de utilización de Amazon Web Services.

En este caso, las características y funciones más interesantes de AWS Rekognition o S3 podrían ser la facilidad que ofrecen para integrar estos dos servicios juntos (no se ha implementado para tener mayor paridad con las otras dos plataformas) y otros si fuera necesario, de manera que el producto es más compacto. Otra ventaja de Rekognition es la variedad de características diferenciales que ofrece, la mayoría de plataformas se ciñen al etiquetado de imágenes o al reconocimiento de caracteres y texto, en cambio AWS

dispone de análisis facial, comparación de rostros e incluso reconocimiento de famosos, como alguno de los ejemplos de técnicas disponibles. [12]

3.3. Microsoft Azure Computer Vision

Por último, ese el turno de Microsoft Azure, en este caso para el almacenamiento se utilizarán cuentas de almacenamiento clásico, creando un grupo de recursos común para todas las imágenes y con un bucket distinto para cada categoría. Para la parte de visión artificial se dispone del servicio Azure AI Computer Vision, el cual dispone directamente de la característica de definir con una frase lógica una imagen, aunque también cuenta con funciones de etiquetado o reconocimiento de texto.



Fig. 3.5. Logo de Microsoft Azure.

La inicialización en Microsoft Azure es sencilla de nuevo, es necesario registrarse y posteriormente crear una suscripción, la cual contará con las credenciales (en un archivo JSON de manera análoga al CSV de AWS) y con las características de facturación necesarias. Finalmente es necesario crear un grupo de recursos con los servicios que serán utilizados y la cuenta de Azure ya estaría lista para ser integrada.

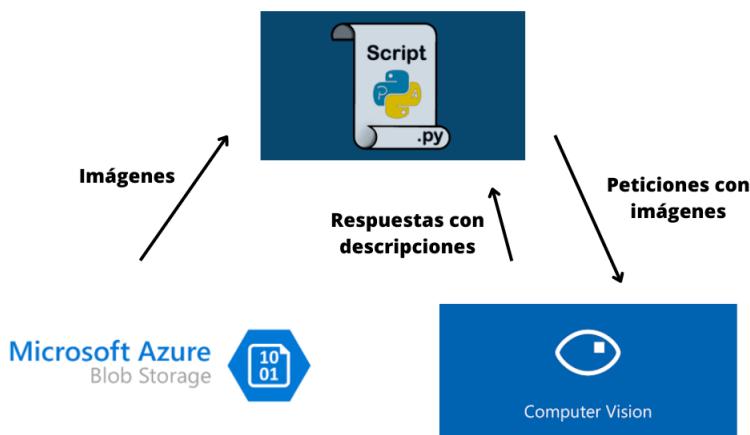


Fig. 3.6. Gráfico de utilización de Microsoft Azure.

Microsoft Azure cuenta con la mencionada característica, conveniente en este proyecto, de ofrecer directamente una descripción de la imagen, pero también cuenta con una herramienta diferencial que es el análisis espacial, este servicio analiza fuentes de vídeos para generar eventos sobre presencias o movimientos, lo cual es muy requerido últimamente para aplicaciones de seguridad o robótica. [13]

3.4. Competidores y otras empresas

Google, Amazon y Microsoft son, sin duda, tres superpotencias tecnológicas pero no son las únicas empresas interesadas en computer vision, y que por tanto ofrecen plataformas, herramientas o frameworks relevantes para la visión por ordenador. Algunas de estas empresas son OpenCV, IBM, Meta con Oculus o DataRobot.

Estas empresas, a diferencia de las tres estudiadas en este proyecto, suelen estar dedicadas a funciones o características más concretas. OpenCV es un framework *open source* enfocado completamente en visión por ordenador con inteligencia artificial para diferentes casos de uso con posibilidad de integración con C++, Python, Java soportado por Linux, Windows, iOS, MacOS o Android. Por otro lado, Meta con Oculus están enfocados directamente hacia la visión aumentada (VR) o realidad aumentada (AR) para la creación de gafas virtuales para experimentar estas AR/VR.



Fig. 3.7. Competidores en Computer Vision.

La principal ventaja de una herramienta como OpenCV es el ser *open source* de manera que tienen un uso libre, a diferencia de frameworks de empresas como las tres de este proyecto o Meta. Otra de las ventajas de OpenCV es la capacidad de aprendizaje sobre esta materia que aporta, ya que es necesario tener ciertos conocimientos, aunque una vez dominados, suele ser la opción elegida por los ingenieros.

Por otro lado, se pueden valorar estas ventajas como puntos negativos, ya que al ser *open source* y requerir un conocimiento previo, la curva de aprendizaje puede ser lenta y no todo el mundo podría acceder a estas herramientas. Google, Amazon y Microsoft en cambio tienen el poder necesario para ofrecer soluciones muy variadas, que por un cierto precio, pueden facilitar diversos acometidos para diferentes casos de uso.

4. ANÁLISIS OBJETIVO DEL USO DE LOS FRAMEWORKS

En este apartado se hará un análisis objetivo de las herramientas de computer vision utilizando diferentes scripts con Python con diferentes llamadas a las APIs de las tres empresas más reconocidas. El objetivo de este análisis es comprobar si hay diferencias objetivas entre las diferentes APIs, es decir, el tamaño de las respuestas que dan a cada imagen, el tiempo que tardan en responder y (de una manera más subjetiva) los tipos de precios que ofrecen para su uso y la dificultad que tienen de ser inicializados e integrados en diferentes despliegues, como podrían ser estos pequeños scripts con Python o como podría ser una API que los englobe o una página web que los requiera directamente.

Por tanto, y teniendo en cuenta que es complicado comparar de forma objetiva herramientas similares pero con distintas funciones y características, a continuación se detallarán los pasos para utilizar cada una de las plataformas, intentando destacar las ventajas e inconvenientes de cada una, poniendo el foco en las APIs de computer vision. También se harán breves descripciones de cada herramienta y por último se mostrará el script usado y los resultados obtenidos.

4.1. Análisis del uso de las APIs

Finalmente se analizarán las APIs de computer vision, combinadas con las herramientas de almacenamiento de cada plataforma, haciendo uso de un script que realizará las llamadas correspondientes inicializando las variables necesarias para después obtener unos resultados que puedan ser analizados correspondientemente.

Serán usadas 210 imágenes aleatoriamente seleccionadas de diferentes sets de internet, catalogadas en 7 diferentes categorías, (personas, animales, objetos, naturaleza, arquitectura, censurables y *memes*, o imágenes de internet) con 30 imágenes por categoría, para realizar las llamadas a las APIs. También es importante recalcar que para ahondar en cada una de estas plataformas cloud, se almacenarán las 210 imágenes en cada una de las herramientas de almacenamiento que tienen Google, Amazon y Microsoft para hacer las respectivas llamadas a sus APIs.

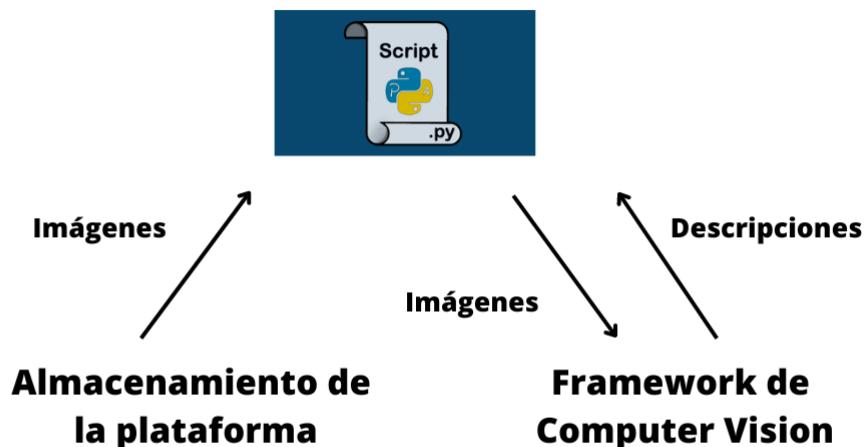


Fig. 4.1. Esquema general y simplificado de las llamadas a APIs del script.

4.2. Definición y diseño del script

El script, escrito con el lenguaje de programación Python, consta de varias partes. La primera parte es, por necesidad, la declaración e instalación de los módulos necesarios. Una vez instalados, se inicializará la conexión a la base de datos PostgreSQL con el módulo *psycopg2*. A continuación se inicializan los clientes de las plataformas de cada empresa, para esto se necesitan los módulos *vision* y *storage* de Google Cloud, unido a *service_account* de Google Auth, también es necesario el módulo *boto3* de Amazon Web Service y los módulos *CognitiveServiceCredentials*, *BlobServiceClient*, *ComputerVisionClient* y *VisionFeatureTypes* de Microsoft Azure. Cada uno de estos módulos serán combinados con los archivos de credenciales ya recogidos de su plataforma correspondiente y se podrá avanzar a la parte principal de las llamadas a las APIs.

La parte principal del script, como se ha mencionado, es la combinación de las herramientas de almacenamiento y computer vision de cada plataforma cloud. Por tanto la estructura será la de la figura 4.1, con leves modificaciones para adaptarse a cada plataforma, como se ve en las figuras 3.2, 3.4 y 3.6. Antes de ello, es importante definir las diferentes variables globales que servirán para comparar las distintas APIs, estas serán *start_time*, *elapsed_time* y *complete_time* para extraer la información temporal que requieren las llamadas a las APIs tras extraer las 210 imágenes de cada herramienta de almacenamiento. En el mismo sentido se utilizarán *response_size* y *complete_size* para tener la información sobre el peso en bytes de todas las respuestas.



Fig. 4.2. Ejemplo extraído del script.

Como ejemplo de las respuestas obtenidas se puede comprobar la figura 4.2, que corresponde a una de las 210 imágenes con las que se ha desarrollado el estudio. Además,

acompañado a ello, y gracias a tener ya disponible la API de OpenAI, se ha realizado el camino inverso, para poder ver si lo que entienden las distintas herramientas de etiquetado sobre una imagen luego se puede dar la vuelta y transformar de nuevo en una imagen reconocible. Siguiendo el mismo ejemplo anterior, se obtienen las figuras 4.3, 4.4 y 4.5 respectivamente para cada plataforma.



Fig. 4.3. Imagen resultado de la descripción de Google Cloud Vision.



Fig. 4.4. Imagen resultado de la descripción de Amazon Web Services Rekognition.



Fig. 4.5. Imagen resultado de la descripción de Microsoft Azure Computer Vision.

Atendiendo únicamente a esta imagen como ejemplo, se puede ver una mayor precisión en la descripción que ofrece Microsoft Azure Computer Vision aunque esta comparativa es relativamente subjetiva y se debería efectuar en un mayor número de imágenes para ofrecer un resultado o veredicto en cuanto a estos resultados al devolverle a una API la descripción para recuperar "la imagen en cuestión".

4.3. Comparación de precios

Para concluir la observación de las diferentes plataformas cloud con sus respectivos servicios, en especial los de computer vision, se analizarán las tarifas de precios ofrecidos por cada uno de ellos. Para este cometido, se tendrá en cuenta únicamente el precio de las APIs utilizadas, y en caso de hacer distinciones, de las herramientas de etiquetado de imágenes con Google y Amazon, y las descripciones con Microsoft. También hay que tener en cuenta las tarifas gratuitas que dispone cada herramienta para diferentes utilidades y que las plataformas suelen denominar distintos tipos de tarifas con "niveles".

Empresa	Gratis	Nivel 1	Nivel 2
Google	1000 cada mes	1.50\$ cada mil	1.00\$ cada mil
Amazon	5000 cada mes	1.00\$ cada mil	0.80\$ cada mil
Microsoft	5000 cada mes	1.50\$ cada mil	0.60\$ cada mil

Fuente: Documentación de Google, Amazon y Microsoft.

TABLA 4.1. COMPARATIVA DE TARIFAS DE PRECIOS.

Por tanto, como se ve en la tabla anterior primero se analizan las tarifas gratuitas, Google Cloud ofrece mil llamadas al mes a su API de etiquetado de imágenes de forma completamente gratuita, mientras que Amazon Web Services y Microsoft ofrecen 5000 llamadas con sus correspondientes imágenes al mes. A partir de aquí, las denominaciones varían levemente, como se verá a continuación.

El primer nivel en el caso de Google recoge desde la llamada mil a la cinco millones, para todas las llamadas dentro de este conjunto el precio es de 1.50 dólares por cada mil llamadas. Por otro lado, tanto Amazon como Microsoft recogen como primer nivel el primer millón de llamadas, en el cual el precio es de 1 dólar por cada mil imágenes en el caso de Amazon Rekognition y de 1.50 dólares cada mil en Microsoft Azure.

El segundo nivel de Google Cloud cuenta desde la llamada cinco millones en adelante, todo esto contabilizando mensualmente, y el precio es de 1 dólar por cada mil imágenes. En cambio, Amazon ofrece un nivel desde la llamada un millón hasta la cinco millones, cobrando 0.80 dólares por cada mil llamadas y además cuenta con niveles superiores para las siguientes 30 y 35 millones de imágenes siguientes. Microsoft Azure, de forma similar a Google, sólo recoge un nivel más, en este caso de la llamada un millón en adelante, con un precio 0.60 dólares por cada mil imágenes.

4.4. Resultados de tiempo y tamaño

Los resultados obtenidos principales son el espacio temporal que ocupa cada plataforma para recoger las imágenes y lanzar las llamadas a la API y el espacio en tamaño que ocupan las respectivas respuestas.

Los resultados temporales, teniendo en cuenta la variabilidad de las conexiones, localizaciones o cobertura, tienen un notable vencedor. Tanto el framework de Google como el de Amazon Web Services se van más allá de los diez minutos entre recoger cada imagen, mandarla a sus respectivas APIs recibiendo las etiquetas de cada una de ellas. Por otro lado, el framework de Microsoft Azure necesita poco más de tres minutos para hacer la misma estructura de pasos, con la ventaja de que ofrecen directamente una descripción de las imágenes sin pasar antes por el etiquetado.

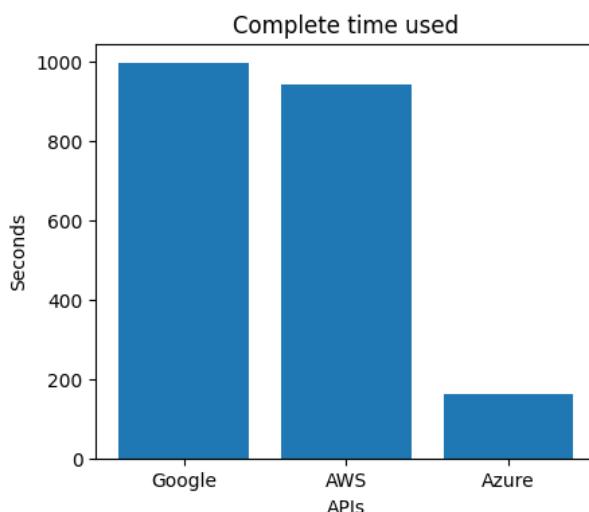


Fig. 4.6. Gráfica resultado del script con la variable de tiempo.

En contraste con los resultados temporales, se recibe la respuesta en tamaño y se puede confirmar que todos los frameworks mandan la misma cantidad de información, o más bien, se encapsulan de la misma manera en el script, por lo tanto en este caso no se podría sacar ventaja por la integración con ninguno de los tres frameworks.

Teniendo finalizado el análisis de las facilidades y complicaciones del acceso, uso e integración con los tres frameworks, del precio y los niveles que ofrece cada plataforma y de los resultados temporales y de tamaño, se pueden tomar ciertas consideraciones. Amazon Web Service destaca en cuanto a facilidad de acceso, uso e integración y en cuanto a precios para servicios o aplicaciones de pequeño a mediano tamaño. Por otro lado, Microsoft Azure toma la ventaja en cuanto a tiempo de respuesta y en cuanto a precio cuando el tamaño de la integración es grande. Google Cloud por último puede favorecerse cuando la integración necesita de diferentes pasos de autenticación y seguridad o en cuanto a variedad de características.

A continuación se realizará un análisis diferente en el que se recibirán resultados o

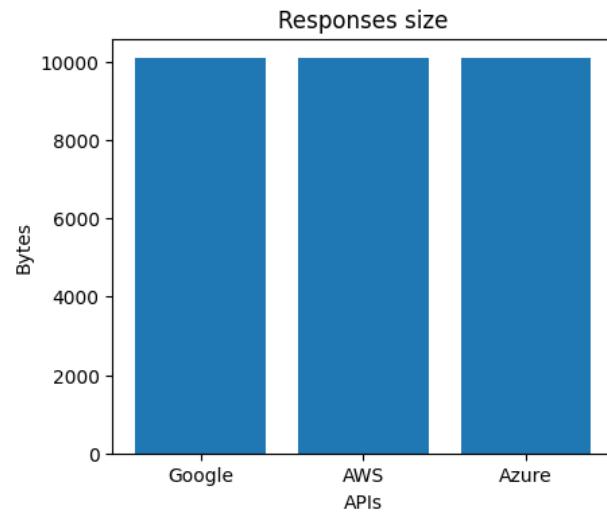


Fig. 4.7. Gráfica resultado del script con la variable de tamaño.

soluciones que podrán o no variar con las recibidas, pero se mantendrán los datos de esta sección para dar una solución u otra en la conclusión de la comparación.

5. ANÁLISIS SUBJETIVO DEL USO DE LOS FRAMEWORKS

En este caso, el análisis de las respuestas de computer vision para diferentes imágenes será realizado de manera subjetiva, apoyándose en una página web propia que será definida a continuación que dispondrá de una encuesta con el dataset mencionado en el análisis objetivo y con la posibilidad para los usuarios de la página web de probar las tres herramientas de visión por ordenador de manera simultánea con la opción de nuevo de elegir la respuesta que más cercana le parezca a la imagen subida.

Al tener dos opciones distintas para la comparación, se distinguirán los resultados tanto de manera conjunta, como distinguiendo entre ambas funciones de la web. Por un lado está la encuesta clásica que tiene las descripciones asociadas a cada imagen ya almacenadas en la base de datos, analizando los resultados también entre categorías en el último apartado y por otro lado se encuentra la opción en la que el usuario prueba las tres APIs de computer vision, recibiendo distintas descripciones dependiendo de la imagen que se suba a la web.

5.1. Página web - sistema de obtención de datos

Como se ha comentado anteriormente la página web cuenta con una encuesta prefabricada con 210 imágenes y tres respuestas por imagen, con un apartado en el que usuario puede probar las herramientas utilizadas en este proyecto y elegir entre ellas, y por último un apartado en el que se divulga sobre la visión por ordenador, con algunas definiciones, casos de uso, precios y ejemplos de varios tipos.

5.1.1. Proceso de creación y despliegue

Para este proyecto por tanto, se han utilizado muchas herramientas que se discutirán brevemente a continuación. Para comenzar, el lenguaje de programación utilizado principalmente es Javascript, contando con CSS para la parte visual de la web, teniendo en cuenta la responsividad para adaptarse de pantallas horizontales a verticales. La elección de Javascript fue prácticamente obligada teniendo en cuenta que el entorno de ejecución utilizado ha sido Node.js y Express.js, debido a las facilidades que estos entornos ofrecen para correr páginas web.

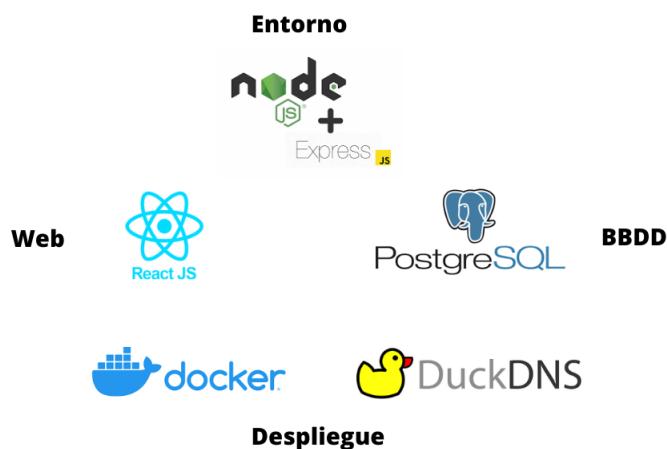


Fig. 5.1. Esquema de utilización de herramientas en el proyecto web.

Para el estructurado del proyecto se han utilizado estructuras similares a aquellas que se pueden encontrar en los proyectos que utilizan el framework Next.js al contar con directorios diferenciados para componentes, para las llamadas a APIs, para las páginas y para los ficheros públicos. Dentro de este contexto, se ha utilizado el framework de Javascript más reconocido, React.

Por otro lado, para las llamadas a las APIs, se han utilizado herramientas nativas de Javascript sumadas a los módulos de cada una de las plataformas comentadas duran-

te este proyecto, estos módulos son *@azure/cognitiveservices-computervision*, *@google-cloud/vision* y *aws-sdk*. En esta misma dirección, las llamadas a base de datos se realizan con herramientas nativas y el módulo *pg*, por tanto las respuestas del usuario a las encuestas se almacenan en una base de datos, utilizando para ello PostgreSQL.

Por último, el despliegue de la web se ha realizado encerrando el código del proyecto en un Docker, para manejar fácilmente cambios, por ejemplo, entre Linux y Windows. Después de esto y de un redireccionamiento de puertos del router utilizado, se ha desplegado y descubierto la página web usando el hosting gratuito de DNS dinámico de Amazon llamado DuckDNS.

5.1.2. Diseño y utilización de web

La primera parte de la web es la página inicio, esta cuenta con un menú superior (que va a ser idéntico para cada página), por el cual navegar entre las distintas páginas de la web. En la parte superior derecha cuenta con tres logos animados, que llevan a GitHub, a Linkedin y muestran un correo, respectivamente. En la parte inferior aparece un footer con el nombre del autor y de la institución en la que este proyecto se desarrolla, y por último aparecen tres cartas que llevan y describen cada una de las partes del proyecto.

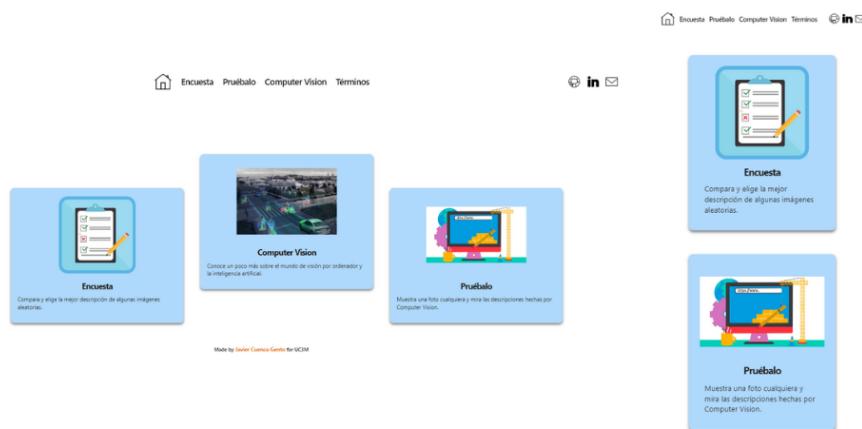


Fig. 5.2. Página de Inicio. Escritorio y móvil.

El primer apartado funcional de la web en ser analizado será la *Encuesta*. Esta página puede ser considerada la principal de la página web en lo que respecta a este proyecto, ya que de aquí se sacará la mayor parte de la información sobre como perciben los usuarios las descripciones de imágenes de cada plataforma. Estas imágenes son recogidas de

un directorio del proyecto mientras que las descripciones son todas almacenadas en un fichero generado previamente apoyado en el script de Python.

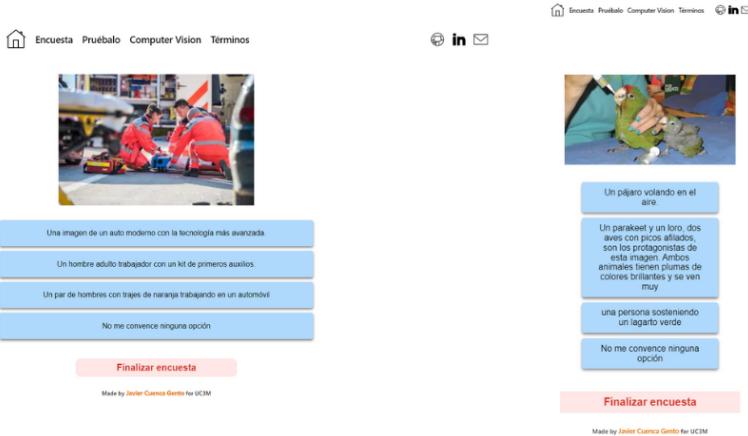


Fig. 5.3. Página de Encuesta. Escritorio y móvil.

La siguiente parte para analizar es *Pruébalo*, en este apartado el usuario puede subir una imagen de su galería o de la propia cámara y recibirá las descripciones de las tres APIs discutidas. Como se puede comprobar en el apartado *Términos* y *Condiciones* del menú, ni las imágenes ni ningún tipo de dato más allá de las elecciones en las encuestas, de las que solo se almacena la API seleccionada, es guardado en ningún lugar.

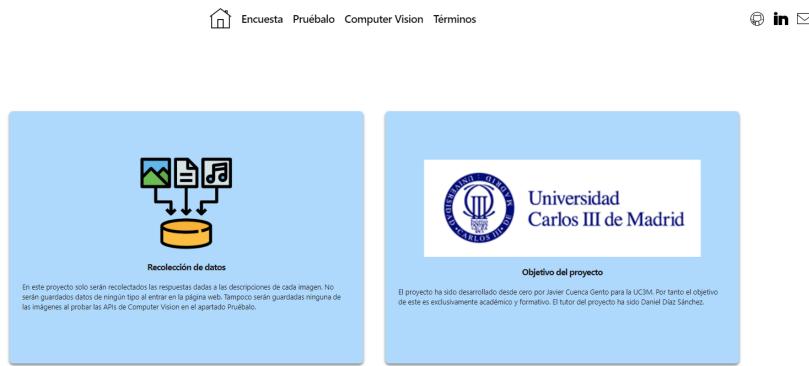


Fig. 5.4. Página de Términos y Condiciones.

Esta página está basada en el mismo formato que la encuesta principal, pero no recoge las imágenes y las descripciones de archivos previamente almacenados si no que la imagen es subida directamente por el usuario, esta es enviada a las tres APIs de manera similar a

como se hizo en el script de Python, pero esta vez usando los módulos de cada plataforma que ofrece Node.js y funciones nativas de Javascript como las *Promises*, para darle un funcionamiento asíncrono y esperar a tener todas las descripciones para mostrarlas y dar la opción de votar al usuario. Para el cometido de no recopilar ningún dato, se ha creado un pequeño script que borra cualquier imagen que haya sido subida a la página web, y este es ejecutado cada minuto gracias al módulo *cron*.

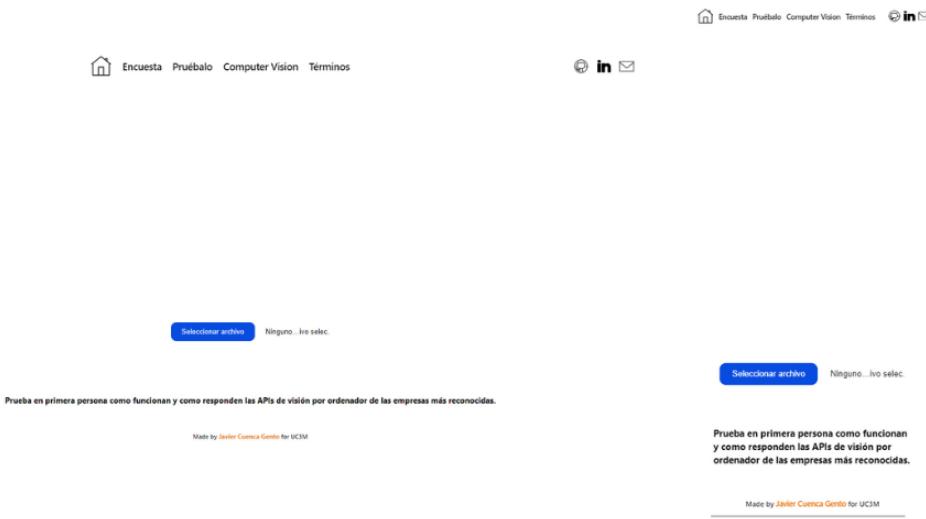


Fig. 5.5. Página de Pruébalo. Escritorio y móvil.

Por último aparece el apartado *Computer Vision*. Esta página está dedicada a explicar brevemente la visión por ordenador, las utilidades, la forma de usarlo o los precios, de manera divulgativa, de manera que se introducen elementos visuales y gráficas de la información que se pretenden hacer llegar, que resume de manera básica y breve algunos de los datos analizados en este proyecto.

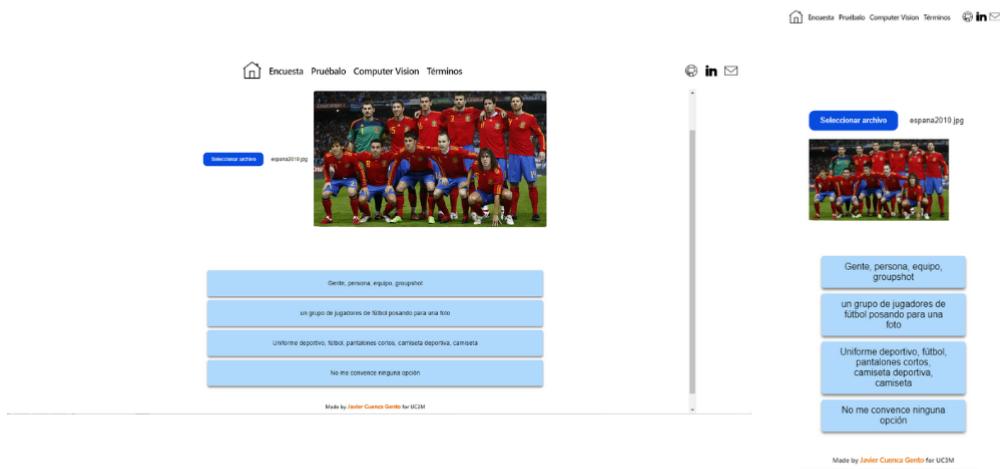


Fig. 5.6. Página de Pruébalo una vez subida la imagen. Escritorio y móvil.

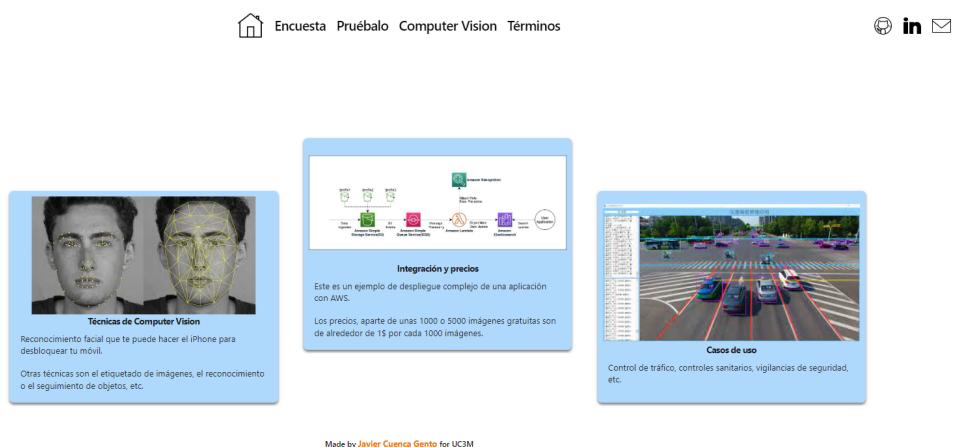


Fig. 5.7. Página principal de Computer Vision.

5.2. Resultados de las votaciones

Los resultados recolectados de la página durante varias semanas, recibiendo más de 500 respuestas con un número de usuarios alrededor de los cien, también divididos por diferentes secciones, que serán respuestas totales, respuestas separando la encuesta y la sección *Pruébalo* y por categoría. También se hará un pequeño análisis como se ha comentado del número de usuarios y de respuestas únicas en las dos secciones de la web.

Para comenzar se analizarán las selecciones de los usuarios en la encuesta con las imágenes del dataset, pero clasificando por categorías, para poder ver cuales son los puntos fuertes y débiles de cada framework. En la siguiente figura por tanto podemos ver la fuerza de Microsoft Azure en descripciones de animales y objetos, sin llegar a tener ningún punto débil. En cambio Google Cloud se ve favorecido en las categorías con mayores complicaciones para las APIs, como son la descripción de memes e imágenes censurables, pero no tiene tanta potencia en el resto de categorías. Por último Amazon Web Services obtiene unos resultados bastante favorables en todas las categorías excepto las mencionadas memes y censurables, ganando con cierto margen en las imágenes de arquitectura.

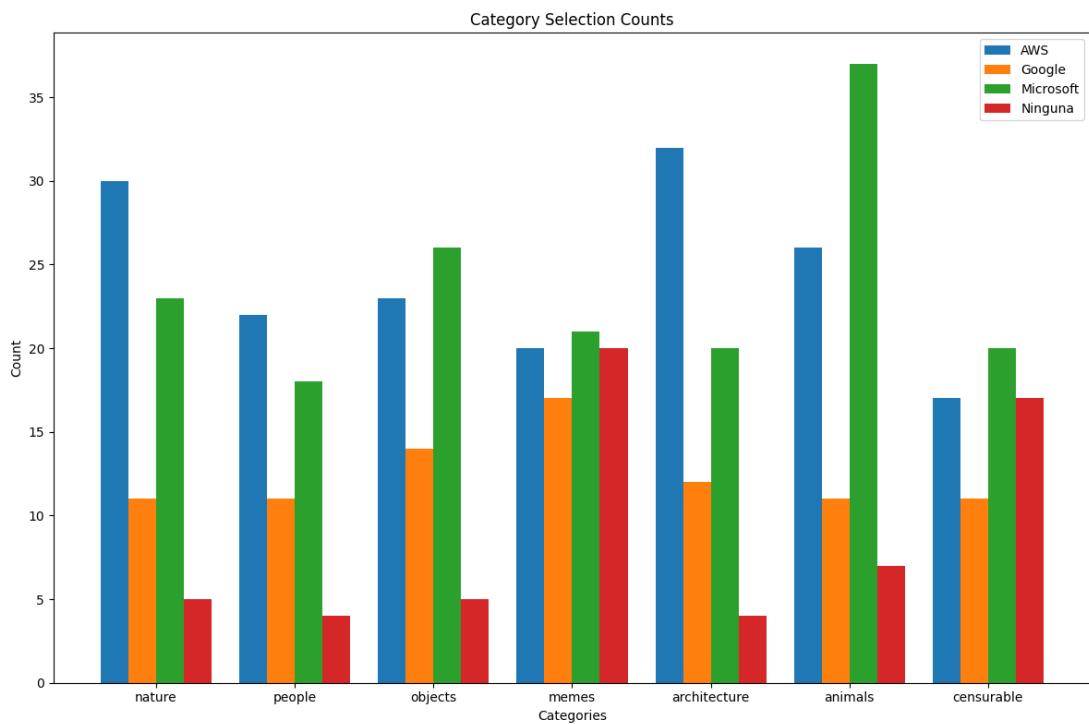


Fig. 5.8. Resultados de la encuesta clasificados por categorías.

A continuación se analizarán los datos obtenidos en la sección *Pruébalo* con las imágenes subidas y probadas por los propios usuarios en la página web. Esta sección ha sido menos usada, quizás por el hecho de tener que subir una imagen, pese a que se informaba en todo momento que la imagen no es almacenada en ningún caso. En este caso el framework que más selecciones ha recibido es Google Cloud Vision, algo por encima de Amazon Web Services Rekognition, mientras que Microsoft Azure ha recibido menos vo-

tos. Esto se puede deber a que en esta sección Amazon y Google mandaban las respuestas únicamente con etiquetas, mientras que Microsoft lanzaba una descripción que en algunos casos puede distanciarse de la imagen en cuestión.

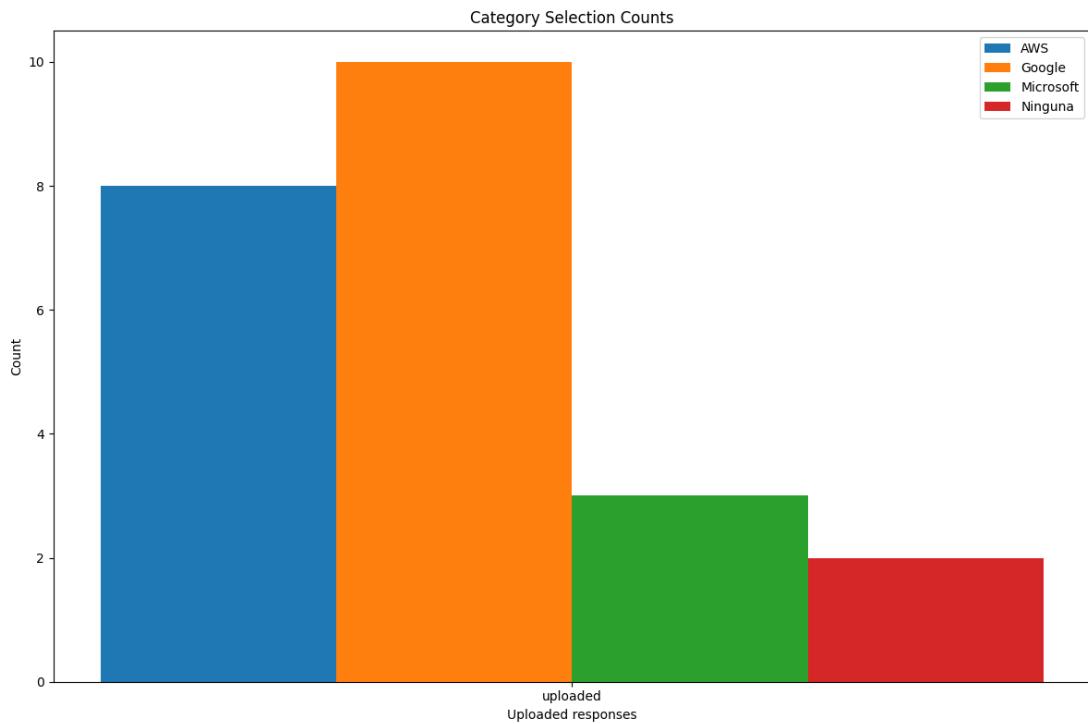


Fig. 5.9. Resultados de la sección *Pruébalo*.

Como se ha comentado anteriormente, la sección *Pruébalo* ha recibido un número menor de votos, dada la facilidad de responder a la encuesta y la complicación que puede suponer para algunas personas subir o hacerse una fotografía en el momento. Pese a ello, se puede hacer una estimación del resultado, atendiendo a varios datos. La comparación entre ambas secciones se encuentra en la figura 5.8 y, como se había visto, Google gana en votos en las imágenes de los usuarios, mientras que pierde con las imágenes de la encuesta respecto a sus dos competidores. También es destacable la notable disminución de votos en porcentaje de Microsoft Azure.

Para poner un último dato sobre la afluencia en el proyecto, se afirma de nuevo que el total de las respuestas recibidas es superior a las quinientas, con alrededor de cien usuarios, con unos veinte de ellos participando en ambas secciones. Estos datos pueden ser estudiados gracias a almacenar la marca de tiempo en que las respuestas han sido enviadas, no con ningún tipo de almacenamiento de información del usuario.

Para finalizar se realizará el análisis de los resultados totales, sumando ambas secciones del proyecto y clasificando únicamente por las diferentes selecciones posibles. En este análisis se pierden las anteriores diferenciaciones que ayudan al análisis de la comparativa dependiendo del caso de uso requerido pero también se gana una vista global sobre las respuestas generales del usuario.

Por tanto y teniendo en cuenta por este momento únicamente estas dos secciones,

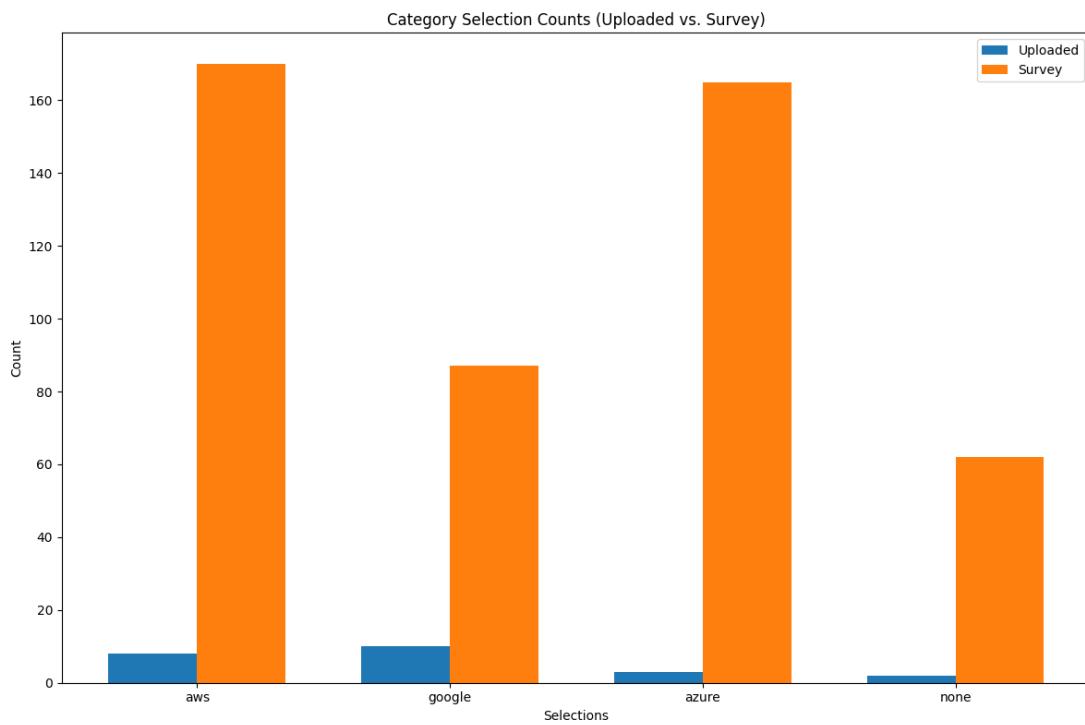


Fig. 5.10. Resultados de la sección *Pruébalo* comparados con los de la encuesta.

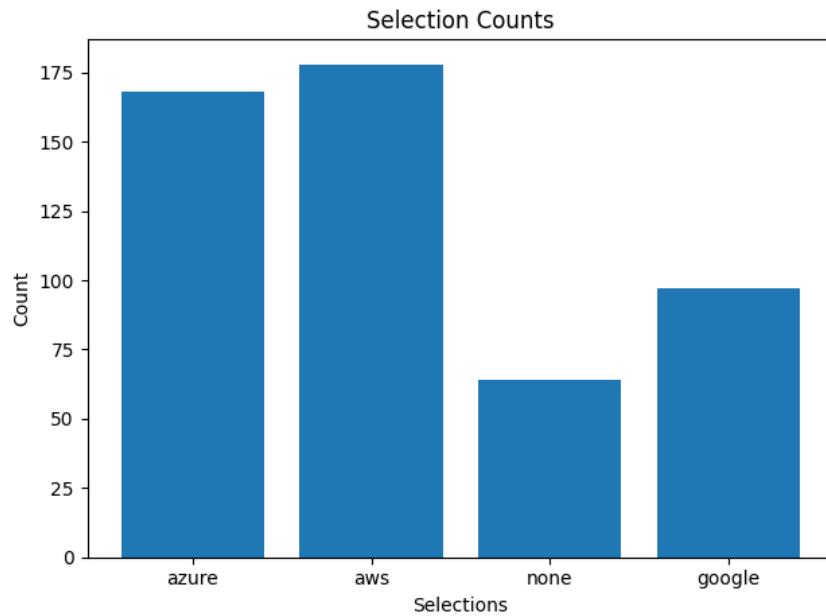


Fig. 5.11. Resultados totales del proyecto web.

se ven dos frameworks con cierta ventaja respecto al tercero y a la respuesta en la que ninguno es capaz de acertar. Estos son Amazon Web Services Rekognition y Microsoft Azure Computer Vision, con alrededor de 175 y 170 selecciones respectivamente. Algo más atrás se encuentra Google Cloud Vision API con alrededor de 100, dejando la opción de ninguno en menos de 70 selecciones.

6. CONCLUSIONES Y TRABAJO FUTURO

6.1. Conclusiones

En conclusión, este análisis comparativo ha intentado descubrir las ventajas y desventajas de cada framework de computer vision de las principales empresas tecnológicas, teniendo en cuenta también todas sus respectivas plataformas cloud. Una potencial decisión final siempre dependería del cliente en cuestión o de la integración necesaria pero en este proyecto se han encontrado ciertas soluciones.

Para aplicaciones de poca carga o de pequeño tamaño la posible solución podría ser *Amazon Web Services Rekognition* gracias a su facilidad de acceso y de uso sumado a su variedad de características y el potencial bajo importe por sus servicios. Por otro lado, para una aplicación de gran tamaño podría ser óptimo integrarse con *Microsoft Azure Vision* por su rapidez y precisión en las respuestas, especialmente creando descripciones de imágenes, sumado a su precio en el nivel de mayor uso. Por último, si se busca una mayor seguridad en autenticaciones con unas características más complejas la solución podría ser *Google Cloud Vision API*.

Por último y por tanto, el framework elegido para un proyecto como el de este trabajo hubiera sido **Amazon Web Services Rekognition** pero, en definitiva, estas comparaciones tienen distintas utilidades y por tanto soluciones para cada integración posible.

6.2. Trabajo futuro

Este tipo de proyectos del campo de la inteligencia artificial, y en este caso de computer vision (específicamente de etiquetado de imágenes), son inabarcables por la cantidad de posibilidades que ofrecen. En esta línea hay diferentes visiones de trabajo futuro, por un lado está la parte divulgativa y comparativa del proyecto y por otro la parte técnica y de investigación sobre las características de cada plataforma.

Por tanto no es descartable continuar con el desarrollo de la página web como proyecto personal, dándole una vista más profesional y ofreciendo más características de la visión por ordenador o añadiendo nuevas plataformas, al mismo tiempo que se va actualizando la sección divulgativa del proyecto.

Por otro lado, también sería interesante tratar de diseñar directamente una API que realice este etiquetado de imágenes o cualquier otra función de computer vision, entrenando directamente datos de distintos datasets, utilizando diferentes frameworks y herramientas de redes neuronales convolucionales, para con ello ver qué ventajas y desafíos ofrece entrenar tu propio modelo y no integrarte con el de una plataforma convencional.

7. MARCO REGULATORIO

Un proyecto de visión por ordenador debe tener en cuenta el aspecto legal, debiendo cumplir la *Ley Orgánica de Protección de Datos*. Para esto, se ha tenido en cuenta recolectar imágenes públicas de diferentes datasets de internet para el análisis objetivo y, con mayor importancia, se ha creado un script que elimina cada imagen subida a la página web. Por tanto, ninguna imagen es almacenada y no se debe cumplir de manera específica la mencionada *Ley Orgánica de Protección de Datos*.

BIBLIOGRAFÍA

- [1] T. S. Huang, “Computer Vision: Evolution and Promise,” 1996.
- [2] A. D. Egorov, “Algorithm for optimization of Viola–Jones object detection framework parameters,” *Journal of Physics*, 2018.
- [3] Roboflow. “Top Computer Vision Models.” (2023), [En línea]. Disponible en: <https://roboflow.com/models>.
- [4] Ultralytics. “Ultralytics YOLOv8.” (2023), [En línea]. Disponible en: <https://github.com/ultralytics/ultralytics>.
- [5] X. e. a. Chen, “Symbolic Discovery of Optimization Algorithms,” 2023.
- [6] J. e. a. Meng, “The Future of Computer Vision,” 2021, 2022.
- [7] Mwanikii, “A Non-Deep Learning Approach to Computer Vision,” 2023.
- [8] V. e. a. Lakshmanan, *Practical Machine Learning for Computer Vision*. O'Reilly Media, Inc., jul. de 2021.
- [9] J. Le, “The 5 Computer Vision Techniques That Will Change How You See The World,” abr. de 2018.
- [10] L. e. a. Alzubaidi, “Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, 2021.
- [11] Google Cloud. “Documentación Google Cloud,” Google Cloud. (2023), [En línea]. Disponible en: <https://cloud.google.com/docs?hl=es>.
- [12] Amazon Web Services. “Documentación Amazon Web Services,” Amazon. (2023), [En línea]. Disponible en: https://docs.aws.amazon.com/es_es/.
- [13] Microsoft Azure. “Documentación Microsoft Azure,” Microsoft. (2023), [En línea]. Disponible en: <https://learn.microsoft.com/es-es/azure/?product=popular>.
- [14] D. Jadhav, “The Use of Machine Learning and Deep Learning for Object Recognition in Computer Vision,” mar. de 2020. doi: [10.13140/RG.2.2.11858.84164](https://doi.org/10.13140/RG.2.2.11858.84164).