Universidad del País Vasco
Euskal Herriko Unibertsitatea

Máster Universitario en

Ingeniería Computacional y Sistemas Inteligentes

Departamento de Ciencias de la Computación e Inteligencia Artificial

Tesis de Máster

Estudio comparativo de diferentes aproximaciones a la clasificación de modelos 3D mediante Deep Learning

Aitor Ruiz de Samaniego López

Tutores

Basilio Sierra

Departamento de Ciencias de la Computación e Inteligencia Artificial

Facultad de Informática

Íñigo Mendialdua

Departamento de Ciencias de la Computación e Inteligencia Artificial

Facultad de Informática

KISA ICSI

KZAA /CCIA

Octubre 2020

Índice

1 INTRODUCCIÓN

En este trabajo se ha realizado una evaluación de diferentes métodos para la clasificación de modelos 3D. Los métodos utilizados han sido la aplicación de clasificación no supervisada sobre diferentes descriptores extraídos de mayas de modelos 3D.

1.1 MOTIVACIÓN

En los últimos años, la clasificación de modelos 3D se ha convertido en esencial para muchos aspectos de tecnologías que actualmente están más presentes en la sociedad, como la creación de nuevos modelos a partir de otros existentes para videojuegos o aplicaciones de realidad aumentada, o la identificación de objetos a partir de los nuevos sensores que se están incluyendo en todo tipo de dispositivos, como vehículos autónomos, o teléfonos móviles. En este trabajo, se pretenden evaluar distintos métodos de clasificación de modelos 3D basados en técnicas de machine learning, en concreto en métodos de clasificación supervisada, ya que son éstos los que más auge están teniendo, siendo las metodologías qué más aplicaciones están teniendo.

1.2 RESUMEN

En este trabajo se han comparado la clasificación de modelos 3D mediante la extracción de diferentes características de los modelos. Primero se ha realizado una clasificación basada en la voxelización de los modelos, similar a la realizada en el trabajo original de Modelnet realizado por la universidad de Priceton. En segundo lugar, se ha realizado extraído un histograma de ocupación de los modelos.

Para la implementación del primer caso se ha comparado una clasificación basada en una red neuronal densa y una red convolucional tridimensional, mientras que en el segundo caso se ha utilizado una red convolucional bidimensional. Además, en cada uno de los casos, se ha comparado el resultado con la utilización de la metodología pairwise con el mismo método de clasificación.

2 ESTADO DEL ARTE

Para establecer un contexto adecuado en el presente trabajo, se explica en éste apartado algunas de las aproximaciones actuales a la clasificación de modelos 3D. Mientras que para la clasificación de imágenes 2D se cuenta con multitud de software e incluso con servicios que permiten la extracción casi inmediata de información de la imagen, para 3D apenas se cuenta con trabajos que quedan puramente en el terreno académico, siendo recopilados algunos de ellos a través de la página Modelnet que sirve de conjunto de pruebas y ranking para establecer la bondad de dichas metodologías.

2.1 ESTADO DEL ARTE EN LA CLASIFICACIÓN DE IMÁGENES 3D

2.1.1 3D ShapeNets: A Deep Representation for Volumetric Shapes

Además de ser la base sobre la que se ha realizado el presente trabajo, sirve de punto de referencia a los demás investigadores para evaluar la validez de sus trabajos, estableciendo un ranking para la clasificación de 40 clases distintas de objetos 3D, y también de un subconjunto de 10 clases. Aunque los resultados de la clasificación obtenidos no son los mejores (un *accuracy* del 77% en la clasificación de 40 clases) la importancia de los resultados obtenidos para esta tesis hace necesario que se desarrolle la metodología que se llevó a cabo para la obtención del dataset utilizado.

Se realizó una búsqueda en 261 sites de modelos CAD. En ellos se consultaron las categorías más comunes y se eliminaron aquellas que no tenían más de 20 resultados, resultando un total de 660 categorías. Se eliminaron los resultados clasificados incorrectamente de forma manual, utilizando el servicio Amazon Mechanical Turk, que mostraba a un grupo de personas contratadas un modelo 3D y debían contestar si pertenecía o no a una categoría concreta. Después, se eliminaron los elementos irrelevantes del modelo, como el suelo, personas cerca del objeto etc., de tal forma que en modelo solo contuviese un objeto que perteneciese a la categoría etiquetada. Se eliminaron modelos poco realistas (objetos demasiado simplificados, aquellos que solo contenían imágenes del modelo) y modelos duplicados. Finalmente se obtuvieron 151.228 modelos CAD que pertenecían a 660 categorías únicas. De estas se

seleccionaron 40 para llevar a cabo la clasificación, y son las que actualmente sirven para establecer el mencionado ranking (Modelnet40).

2.1.2 Inductive Multi-Hypergraph Learning and Its Application on View-Based 3D Object Classification

Este ha sido de los métodos que mejores resultados ha obtenido en la clasificación de la base de datos Modelnet40, obteniendo una *accuracy* del 97,16%. Este método se basa en la aplicación de aprendizaje inductivo, es decir, la extracción de reglas generales a través de ejemplos, a hipergráfos generados a partir de los vértices de los modelos 3D. Los hipergráfos se contruyen relacionando los vértices del modelo con el resto de vértices a través de hiperaristas mediante kNN, conectando el vértice a los K vértices más cercanos. De este hipergráfo aplica el entrenamiento inductivo de forma recurrente para obtener el modelo buscado.

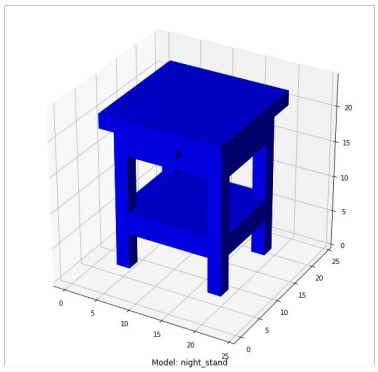
2.1.3 RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews from Unsupervised Viewpoints

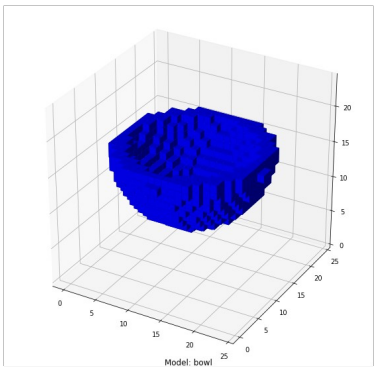
Otra de las apoximaciones con mejores resultados registradas en Modelnet con un 97,37% de *accuracy* es la llamada RotationNet, cuyo enfoque principal es realizar la clasificación no solo del modelo de entrada, si no de un conjunto de *vistas* obtenidas a partir de la colocación de la "cámara" que observa el objeto en distintos puntos alrededor del objeto a clasificar. Así pues, este método convierte la clasificación de objetos 3D, en una clasificación de imágenes 2D, lo que permite un rendimiento mucho mayor que otros como la voxelización, sin pérdida de capacidad predictiva.

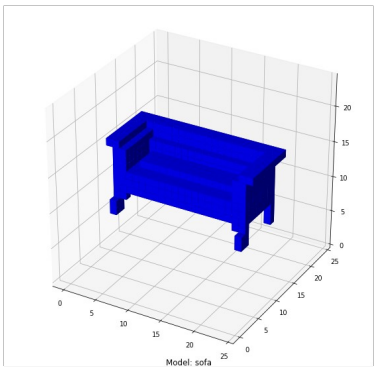
3 TÉCNICAS UTILIZADAS

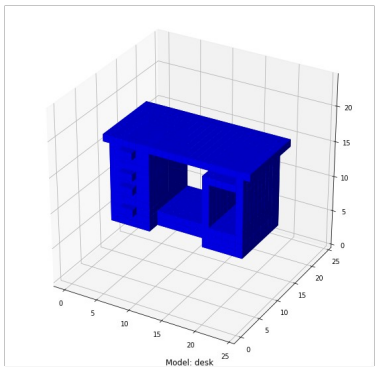
3.1 VOXELIZACIÓN

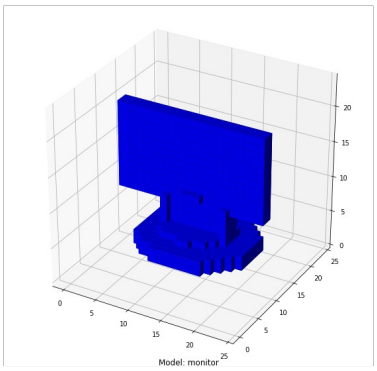
La voxelización es la técnica utilizada en el trabajo base en el que se creó la base de datos de evaluación utilizada. Consiste en la segmentación del espacio ocupado por un modelo 3D en cubos o *voxels* para determinar si el modelo se encuentra total o parcialmente en esa zona del espacio. Para el presente trabajo, no se ha tenido en cuenta el nivel de ocupación de cada *voxel*, indicando únicamente si el modelo 3D se encuentra en dicho *voxel*. El espacio se ha dividido en una matriz tridimensional de 24x24x24, habiendo hecho experimentos con distintas resoluciones (32x32x32 y 48x48x48) sin obtener diferencias altamente significativas.

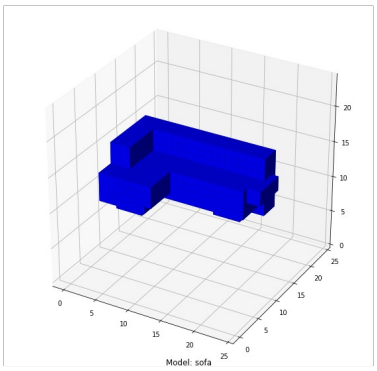






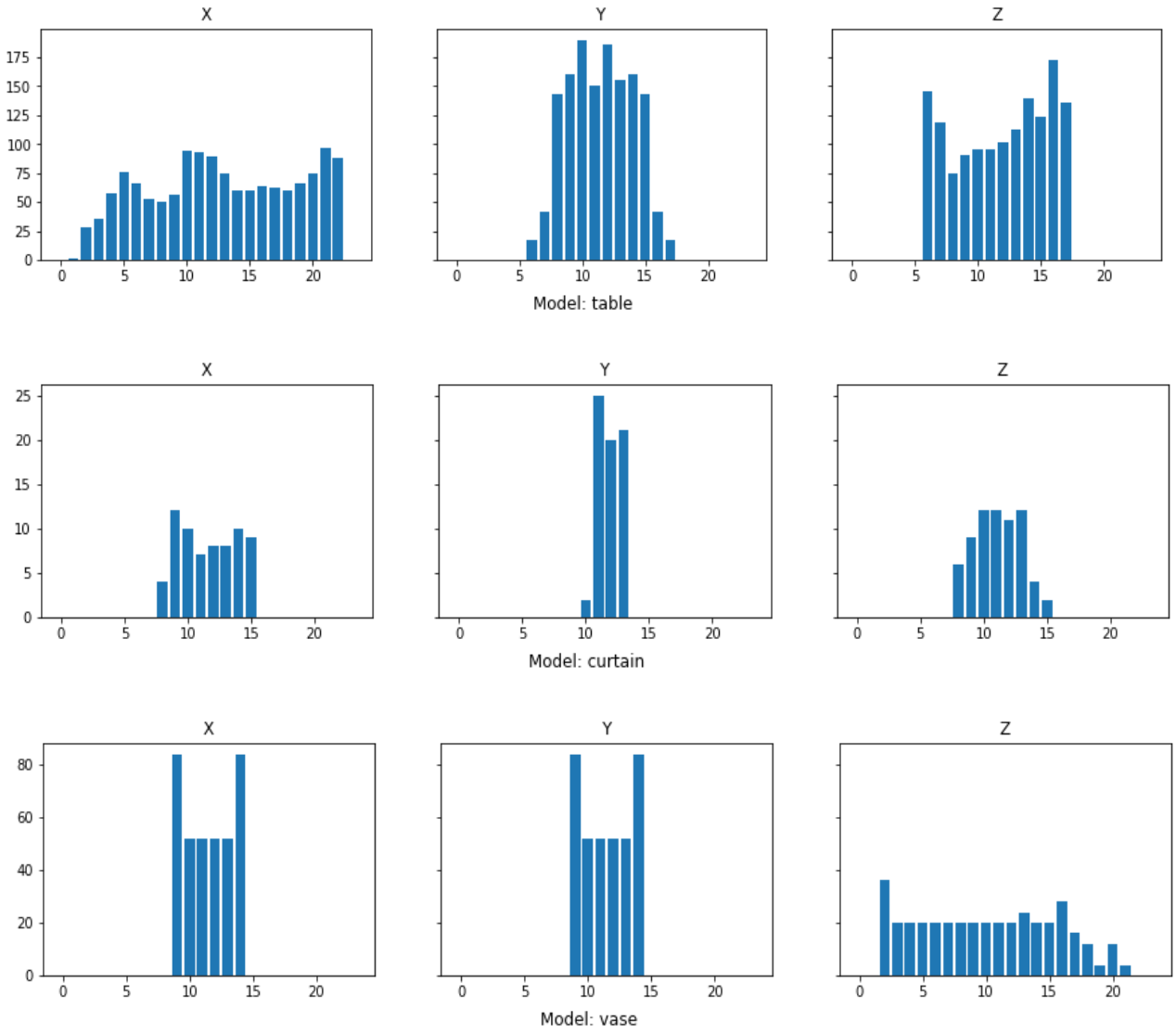






3.2 HISTOGRAMAS

Los histogramas utilizados representan la ocupación del modelo 3D del espacio que lo contiene. Para calcularlo, se han dividido cada una de las 3 dimensiones en 24 segmentos y se ha calculado cuantos de los segmentos de las otras 2 dimensiones están ocupados por el modelo.



3.3 PAIRWISE

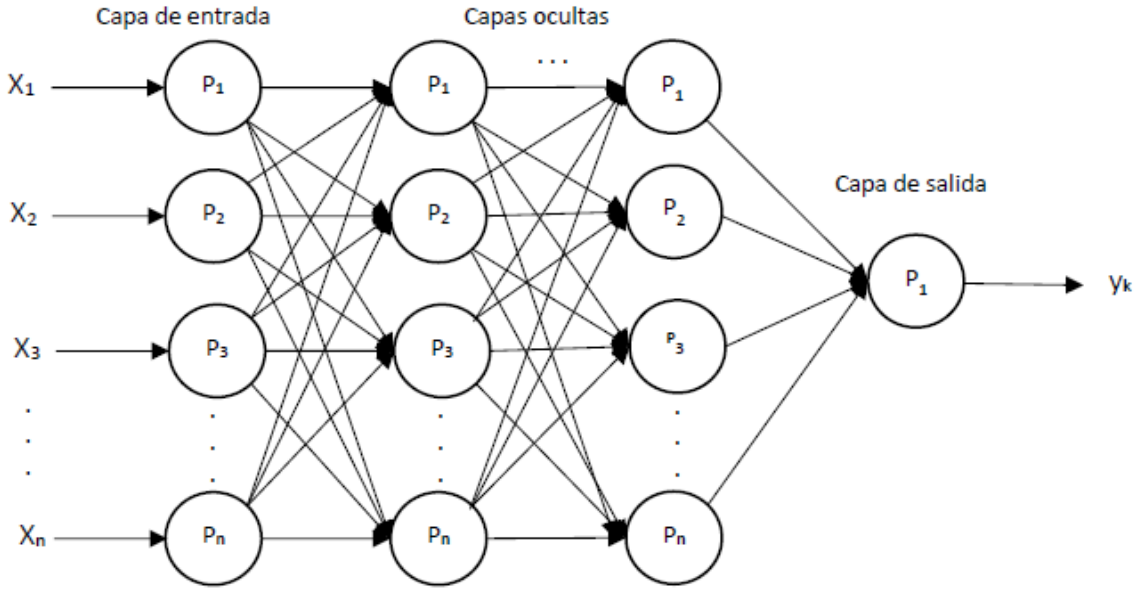
Es una técnica que se basa en dividir un problema de clasificación multiclase en una serie de problemas de clasificaciones binarias. Esta técnica se puede aplicar de distintas maneras, como por ejemplo, enfrentando cada clase con el resto, o enfrentando cada clase con cada una de las otras clases. Es este último sistema el que se ha aplicado en nuestros experimentos.

Para ello se han entrenado cada una de las clases a clasificar contra el resto de clases, obteniendo así una serie de $C=\frac{N+N}{2}-N$ clasificadores. Para el cálculo de la predicción se evaluará contra todos estos clasificadores. Como resultado se obtienen *C* valores para la clasificación de cada modelo y se seleccionará como resultado la clase que más resultados obtiene. Así, suponemos que al enfrentar la clase *A* a todos los clasificadores, obtendrá un valor *A* cuando se enfrenta en los clasificadores en los que está involucrada la clase *A* y valores aleatorios cuando se enfrentan cualquier otras dos clases.

3.4 RED NEURONAL DENSA

Las redes neuronales artificiales consisten en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre si para transmitirse señales y activarse entre si en aplicando lo que se llama función de activación. Producen un conjunto de salidas que se interpretan como una clasificación. Estos sistemas *aprenden* y se forman así mismos en lugar de ser programados de forma explicita modificando los valores (pesos) que se aplican a las funciones de activación.

[https://es.wikipedia.org/wiki/Red_neuronal_artificial]



3.5 RED NEURONAL CONVOLUCIONAL

Las redes neuronales convolucionales son un tipo de redes neuronales en las que se tienen en cuenta aspectos espaciales de las entradas de información para extraer características de los mismos y así reducir su dimensionalidad y poder realizar clasificaciones más eficientes. Por ejemplo, en 2 dimensiones, podemos aplicar una convolución de una imagen 8x8, con un *kernel* 3x3 para obtener una representación de dicha imagen con una dimensión de 4x4. El *kernel* es una matriz con la que se realizará el producto escalar de cada subconjunto de la imagen de entrada de igual dimensión, de izquierda a derecha y de arriba a abajo para obtener la nueva representación de la misma. Esta nueva *imagen* estará dibujando ciertas características de la imagen original que ayudará a poder distinguir los objetos contenidos en la imagen.

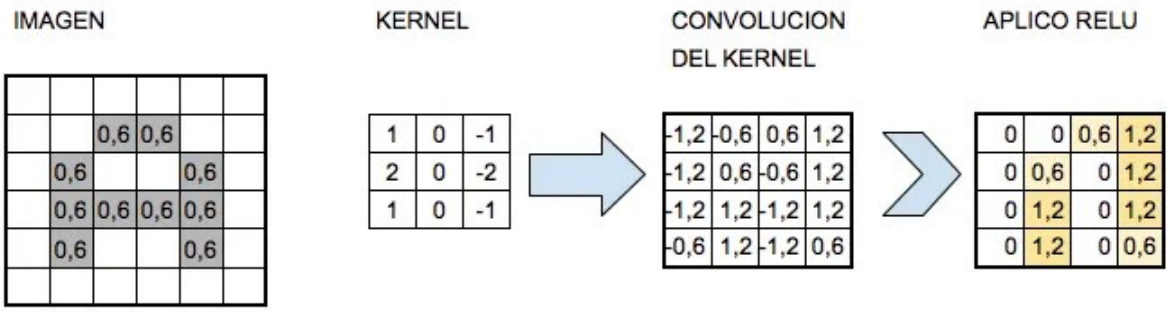


Figura 1: Extracción de características mediante convolución

4 EXPERIMENTOS

4.1 VOXELIZACIÓN

El primer paso para la realización de los experimentos ha consistido en la voxelización de todos los modelos contenidos en la base de datos Modelnet40. El total de modelos contenidos en esta base de datos es de más de 12.000, lo que implica un alto coste de tiempo de computación. La generación de esta voxelización se realizó con ayuda del software opensource *cuda voxelizer*, obteniendo como salida, una representación del espacio de 24x24x24. Esta voxelización será la entrada directa de los 4 primeros experimentos realizados (red neuronal densa, red convolucional, y sus correspondientes pairwise). También ha servido como entrada para el cálculo de los histogramas.

Para los experimentos se ha utilizado el software Jupiter. Para el aprendizaje se ha utilizado Tensorflow con la capa Keras para simplificar la generación y entrenamiento de los modelos.

Los experimentos se han ejecutado en el siguiente Hardware:

- Intel(R) Core(TM) i7-4510U CPU @ 2.00GHz
- 16G RAM
- GeForce 840M 2G
- Driver Version: 440.118.02 CUDA Version: 10.2
- Ubuntu 20.04

4.1.1 Red Neuronal Densa

Se ha creado mediante Keras una red neuronal con las siguientes capas:

- Flatten: Esta capa "aplana" la entrada tridimensional produciendo un único vector de 13824 posiciones.
- Dense (48): Esta capa conecta todas las entradas con la siguiente capa a través de 48 neuronas, que extraerán características con activación 'relu' (unidad linea rectificada).

- Dense(12): Una nueva capa que conecta todas las entradas con la siguientes con activación ‘sigmoid’ (que activa los valores altos y negativiza los bajos).
- Dropout(0.2): Una capa que negativiza aleatoriamente entradas para evitar el overfitting.
- Dense(40): Una última capa de salida para determinar la clasificación.

Los resultados obtenidos para cada clase se representan en la siguiente tabla:

Clase	accuracy
airplane	0.81
bathtub	0.82
bed	0.63
bench	0.38
bookshelf	0.53
bottle	0.65
bowl	0.60
car	0.85
chair	0.70
cone	0.53
cup	0.00
curtain	0.50
desk	0.53
door	0.60
dresser	0.66
flower_pot	0.00
glass_box	0.75
guitar	0.99
keyboard	0.46
lamp	0.50
laptop	0.58
mantel	0.80
monitor	0.79
night_stand	0.57
person	0.00
piano	0.73
plant	0.50

radio	0.00
range_hood	0.93
sink	0.29
sofa	0.76
stairs	0.00
stool	0.00
table	0.70
tent	0.24
toilet	0.92
tv_stand	0.42
vase	0.32
wardrobe	0.00
xbox	0.00

La exactitud total del experimento es de 0,6487, siendo el método que peores resultados ha obtenido. Es importante señalar que algunas clases ha obtenido resultados completamente negativos, lo que podría indicar unos modelos de calidad deficiente o muy similares a otros modelos.

4.1.2 Conv3D

Este modelo se ha compuesto de las siguientes capas:

- Conv3D, con un kernel de tamaño 3 y un salto entre voxels de 2. La función de activación utilizada ha sido de nuevo 'relu'. En esta capa se espera que el modelo sea capaz de extraer las características principales del modelo.
- Conv3D, con un kernel de tamaño 3 y un salto de 3. Con esta capa conseguimos reducir aún más la dimensionalidad del problema.
- Flaten: Con esta capa se aplana la salida para poder seguir utilizándola en la red
- Dense(24): Se crea una capa densa para mejorar el entrenamiento.
- Dense(40): Se añade una capa de salida.

Los resultados obtenidos con éste método se expresan en la siguiente tabla:

Clase	accuracy
airplane	0.93
bathtub	0.88
bed	0.82
bench	0.45
bookshelf	0.73
bottle	0.74
bowl	0.61
car	0.90
chair	0.84
cone	0.60
cup	0.38
curtain	0.48
desk	0.63
door	0.55
dresser	0.59
flower_pot	0.03
glass_box	0.86
guitar	0.96
keyboard	0.33
lamp	0.48
laptop	0.77
mantel	0.87
monitor	0.86
night_stand	0.66
person	0.38
piano	0.75
plant	0.57
radio	0.25
range_hood	0.95
sink	0.33
sofa	0.90
stairs	0.44
stool	0.62

table	0.81
tent	0.36
toilet	0.88
tv_stand	0.58
vase	0.60
wardrobe	0.27
xbox	0.50

La exactitud global del método es de 0.72 y no hay ninguna clase que obtenga una clasificación completamente errónea como con el método anterior, lo que parece indicar una mejoría (aunque sigue obteniendo puntuaciones bajas en las mismas clases, lo que también puede indicar un problema con los modelos).

4.1.3 Red Neuronal Densa + Pairwise

La red neuronal utilizada es similar es similar a la utilizada en el punto 4.1.1, con la excepción de la capa de salida que es binaria. Los resultados obtenido con ésta metodología se presentan en la siguiente tabla:

Clase	accuracy
airplane	0.80
bathtub	0.92
bed	0.76
bench	0.48
bookshelf	0.59
bottle	0.73
bowl	0.62
car	0.83
chair	0.60
cone	0.58
cup	0.50
curtain	0.45
desk	0.58
door	0.61
dresser	0.70
flower_pot	0.05
glass_box	0.83
guitar	0.96
keyboard	0.56
lamp	0.42
laptop	0.70
mantel	0.93
monitor	0.83
night_stand	0.66
person	0.58
piano	0.73
plant	0.60
radio	0.33

range_hood	0.99
sink	0.55
sofa	0.90
stairs	0.75
stool	0.58
table	0.68
tent	0.41
toilet	0.97
tv_stand	0.65
vase	0.39
wardrobe	0.80
xbox	1.00

El resultado de la exactitud global de este método ha sido de 0.71, significativamente por encima de la aplicación del mismo tipo de clasificación sin la metodología pairwise.

4.1.4 Conv3D + Pairwise

Al igual que en el caso anterior, se aplicó la misma red convolucional a cada pareja de clases, con la excepción del la capa de salida que era binaria. Los resultados se muestran en la siguiente tabla:

Clase	Accuracy
airplane	0.91
bathtub	0.97
bed	0.76
bench	0.54
bookshelf	0.66
bottle	0.81
bowl	0.75
car	0.91
chair	0.77
cone	0.79
cup	0.19
curtain	0.52
desk	0.62

door	0.54
dresser	0.74
flower_pot	0.04
glass_box	0.91
guitar	0.99
keyboard	0.53
lamp	0.50
laptop	0.58
mantel	0.92
monitor	0.84
night_stand	0.64
person	0.46
piano	0.73
plant	0.61
radio	0.14
range_hood	0.94
sink	0.55
sofa	0.92
stairs	0.78
stool	0.50
table	0.68
tent	0.50
toilet	0.96
tv_stand	0.72
vase	0.57
wardrobe	0.67
xbox	0.89

La exactitud general del experimento es de 0,75, siendo la más alta de todos los experimentos realizados.

4.1.5 Histogramas

Para este experimento se ha calculado el histograma de ocupación del espacio para cada uno de los modelos de los conjuntos de entrenamiento y test. Posteriormente, tomándolos como entrada, se ha entrenado una red neuronal con las siguientes capas:

- Conv2D: Una capa de convolución, en este caso 2D, para extraer las características de la representación del histograma, que tiene una dimensionalidad inicial de 24x3. Se le aplica un kernel de 2x3 sin salto.
- MaxPooling 2D: una capa de pooling para abstraer las características extraídas de su posición espacial.
- Flatter: una capa para conectar con el resto de la red.
- Dense (12): Una capa densa para mejorar el entrenamiento.
- Dense (40): Una capa de salida con cada clase de modelo a clasificar.

La exactitud por cada clase de este experimento se representa en la siguiente tabla:

Clase	accuracy
airplane	0.86
bathtub	0.82
bed	0.79
bench	0.47
bookshelf	0.78
bottle	0.72
bowl	0.75
car	0.93
chair	0.75
cone	0.65
cup	0.12
curtain	0.44
desk	0.68
door	0.59
dresser	0.65
flower_pot	0.12
glass_box	0.84
guitar	0.96

keyboard	0.59
lamp	0.56
laptop	0.57
mantel	0.90
monitor	0.88
night_stand	0.69
person	0.50
piano	0.73
plant	0.78
radio	0.27
range_hood	0.93
sink	0.16
sofa	0.90
stairs	0.38
stool	0.48
table	0.78
tent	0.37
toilet	0.94
tv_stand	0.64
vase	0.52
wardrobe	0.31
xbox	0.33

La exactitud general de este experimento fue de 0.73 siendo la mejor puntuación sin la aplicación de la metodología pairwise.

4.1.6 Histogramas + Pairwise

En este experimento se repitió el proceso de los anteriores: utilizando una red neuronal similar a la anterior, con la excepción de la capa de salida que en este caso era binaria, se enfrentó cada clase con el resto para crear un conjunto de clasificadores. Se realizó una evaluación con cada uno de ellos de los modelos del conjunto de test y se seleccionó como clase *ganadora* la clase que más resultados obtuvo.

El resultado del experimento por clases se representa en la siguiente tabla:

Clase	accuracy
airplane	0.54
bathtub	0.71
bed	0.47
bench	0.16
bookshelf	0.41
bottle	0.62
bowl	1.00
car	0.86
chair	0.27
cone	1.00
cup	0.00
curtain	0.29
desk	0.51
door	0.64
dresser	0.55
flower_pot	0.00
glass_box	0.91
guitar	0.61
keyboard	0.69
lamp	0.56
laptop	0.83
mantel	0.70
monitor	0.60
night_stand	0.52
person	0.33

piano	0.41
plant	0.61
radio	0.00
range_hood	1.00
sink	0.00
sofa	0.71
stairs	0.00
stool	1.00
table	0.66
tent	0.33
toilet	0.90
tv_stand	0.39
vase	0.38
wardrobe	0.50
xbox	0.67

La exactitud general del experimento fue de 0.51, siendo la peor clasificación de todos los experimentos.

5 COMPARACIÓN

En la siguiente tabla se comparan los distintos métodos con su correspondiente aplicación de la metodología pairwaise, indicando la mejoría o empeoramiento que produce dicha aplicación:

Tabla 1: Comparación metodologías

Método	Test Accuracy	+ Pairwise	% Mejora
Voxelización RN	0,6487	0,7092	9,33
Voxelización Conv3D	0,7220	0,75	3,87
Histograma	0,7331	0,5181	-29,32

Como se puede ver en la voxelización se produce una mejora significativa al aplicar la metodología pairwise, pero no cuando se trabaja con los histogramas. Tras analizar el problema de la metodología pairwise con los histogramas se pudo observar que, como en otros experimentos, clasifican muy mal (en este caso incluso enfrentándolas únicamente a otra clase).

6 CONCLUSIONES Y FUTUROS TRABAJOS

La identificación de objetos 3D es un área del aprendizaje automático donde aún queda mucho margen de mejora, sobre todo en lo referente al tratamiento de los modelos recibidos. La aplicación de nuevas técnicas de extracción de características como las convoluciones 3D, o enfoques de división del problema en problemas menos complejos, como el pairwise ofrecen una mejora en los resultados.

Se ha observado que hay una serie de clases que clasifican peor que otras, y se ha encontrado que se corresponden con que son clases que tienen menos modelos para realizar el entrenamiento. Sería interesante ampliar el número de modelos de estas clases para ver si con ello mejora su clasificación, por ejemplo, aplicando transformaciones espaciales a los modelos disponibles (rotaciones, inversiones…) o con la búsqueda de nuevos modelos a tratar.

También se ha observado que para esta clasificación no se ha tenido en cuenta las verdaderas dimensiones del modelo, ocupando el mismo espacio un avión que una botella. Este podía ser una entrada muy importante a la hora de realizar una clasificación más exacta.

Resultaría interesante en futuros trabajos realizar un análisis de combinación de los distintos clasificadores utilizados para comprobar en qué medida las carencias de uno de los métodos pueden ser suplidas por el resto.

7 BIBLIOGRAFÍA