PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

TÉCNICAS DE REPRESENTACIÓN Y RECONSTRUCCIÓN DE OBJETOS 3D EN EL COMPUTADOR: UNA REVISIÓN DE LITERATURA

TRABAJO DE INVESTIGACIÓN PARA LA OBTENCIÓN DEL GRADO DE BACHILLER EN CIENCIAS CON MENCIÓN EN INGENIERÍA INFORMÁTICA

AUTOR

SUMOSO VICUÑA, ERNIE LUDWICK

ASESOR:

SIPIRAN MENDOZA, IVÁN ANSELMO

Lima, Diciembre, 2020

Resumen

Actualmente en el mundo, las tecnologías de escaneo 3D se clasifican en dos grupos: de contacto y sin contacto. El primer grupo se caracteriza por la necesidad de reposar el escáner sobre el objeto (Sreenivasa K. 2003). Este tipo de escáneres representan un riesgo cuando los objetos en cuestión no pueden ser manipulados libremente debido a su fragilidad. Por otro lado, el segundo grupo de tecnologías son mayormente usadas en investigaciones y poseen una amplia variedad de aplicaciones en la industria medicinal y de entretenimiento. Este último grupo a su vez se divide en dos sub-grupos: activos y pasivos (Pears N. 2012). Las tecnologías de escaneo 3D activos se basan en el análisis y medición del tiempo de envío y retorno de una señal hacia el objeto para estimar la posición de la superficie. Por otro lado, las técnicas de escaneo sin contacto-pasivas no necesitan de la manipulación del objeto ni medición de señales ya que aprovechan la luz ambiental.

Dentro de las ciencias de la computación existe el problema de cómo sintetizar, procesar y analizar la información de una superficie obtenida mediante herramientas de escaneo 3D y guardarla en el computador con el fin de que este pueda ser visualizada y/o manipulada por otras herramientas informáticas. A lo largo de los años han surgido múltiples técnicas de representación de objetos en un espacio de tres dimensiones. Sin embargo, estas técnicas dependen fuertemente de las herramientas empleadas durante el proceso de escaneo. Es por ello que se han desarrollado también técnicas pasivas-sin contacto que permitan la obtención de superficies únicamente a partir de una colección de imágenes y haciendo uso de redes neuronales entrenadas en extensos conjuntos de datos. Para poder entender estas tecnologías emergentes es necesario investigar a profundidad cuales son los recientes métodos para generar superficies u objetos 3D, en qué casos se utilizan los distintos métodos y cuáles son los enfoques de los autores al emplear dichas técnicas.

Índice

Índice de Figuras

Índice de Tablas

1 Introducción

La reconstrucción de superficies 3D empleando técnicas de escaneo es un tema que se ha venido desarrollando a profundidad en la última década. El campo encargado de su estudio es la visión artificial dentro de las ciencias de la computación. Actualmente, en este campo existen múltiples técnicas para la generación de superficies a partir imágenes. Muchos de estos métodos han sido retomados recientemente a pesar de que los primeros trabajos datan de la década de 1990 (Blanz & Vetter, 1999). Con el reciente desarrollo de la tecnología se han impulsado este tipo de proyectos y consecuentemente se han descubierto nuevos métodos y tecnologías de reconstrucción de superficies.

Los trabajos de investigación en este campo pueden clasificarse dependiendo del propósito (o enfoque) de la reconstrucción. Existen métodos que facilitan el reconocimiento de un solo objeto en una imagen, otros que permiten detectar múltiples objetos a partir de una imagen panorámica, etc. Veremos que estos métodos serán empleados en casos específicos y para suplir necesidades específicas. También pueden clasificarse dependiendo de la técnica empleada para la representación del objeto 3D en el computador, ya que existen múltiples formas de adquirir, procesar y guardar la data. Para entender mejor estas tecnologías emergentes (así como la interpretación de resultados) ahondamos en el estado del arte empleando una revisión sistemática de la literatura.

2 Método

La reconstrucción de objetos en 3 dimensiones y su clasificación son áreas de investigación que se vienen desarrollando con mayor auge en los últimos años (2014-2019). Existen varias formas de clasificar este tipo de proyectos, sin embargo, nos centraremos en cómo se representan los objetos 3D y en el objetivo principal de los trabajos de investigación: clasificación, reconstrucción, texturas, etc.

Para la recaudación de información se realizó una revisión sistemática del estado del arte sobre la reconstrucción de superficies 3D mediante modelos de reconocimiento. Las tecnologías investigadas y empleadas en el área de interés nos permitirán evaluar e interpretar resultados para aplicarlo a nuestra problemática. Además, debido al gran número de trabajos de investigación involucrados se definirán las estrategias de búsqueda, así como los criterios de inclusión y exclusión.

2.1 Revisión Sistemática

2.1.1 Objetivos de revisión

El objetivo principal de esta revisión sistemática es identificar las diferentes técnicas de representación de superficies 3D en el computador, así como reconocer los métodos de reconstrucción de superficies que se vienen desarrollando recientemente. Esto nos brindará un contexto teórico adecuado y actualizado a las tecnologías emergentes en el campo de la visión artificial.

Se utilizarán los criterios PICOC para estructurar los elementos de los objetivos de la revisión sistemática.

Tabla 1. Criterios PICOC

Population	Escaneo de superficies 3D empleando cualquier técnica de representación de las superficies. Modelos entrenados para una visión artificial pero empleados para objetos de cualquier tipo.
Intervention	Aplicación de aprendizaje de máquina para el procesamiento de imágenes que permitan obtener una superficie altamente variable.
Comparison	Entre modelos de reconstrucción de superficies 3D a partir de una secuencia de imágenes o videos.

Outcome	Identificación de las recientes técnicas empleadas para representar superficies en el computador y el uso de modelos de machine learning para el escaneo 3D de objetos reales.
Context	Múltiples trabajos de investigación pertenecientes al área de ciencias de la computación que emplean deep learning para el escaneo 3D de objetos en el mundo real.

2.1.2 Preguntas de Investigación

Con el fin de poder entender las diferentes técnicas y métodos existentes en el estado del arte sobre las reconstrucciones 3D de objetos necesitamos clasificar los enfoques, así como las técnicas utilizadas para la representación de objetos 3D en el computador. Frente a estos métodos podemos plantear las siguientes preguntas de investigación:

- P1. ¿Qué enfoques se buscan con la generación de superficies u objetos reales a partir de imágenes?
- P2. ¿Cuáles son las técnicas que se utilizan para representar un objeto 3D en el computador?

2.1.3 Proceso de Búsqueda

Para llevar a cabo el proceso de búsqueda se usarán como herramientas múltiples motores de búsqueda reconocidos y respaldados por instituciones académicas a nivel mundial. Estos motores de búsqueda son:

- IEEE Xplore Digital Library: base de datos para investigaciones, como su nombre lo indica es una librería digital que contiene millones de artículos, revistas científicas, papers, etc. relacionados a temas de ciencias e ingeniería (IEEE 2020). Este motor de búsqueda está respaldado por ambas instituciones: IEEE (Institute of Electrical and Electronics Engineers) e IET (Institution of Engineering and Technology).
- ACM DL: por sus siglas en inglés Association for Computing Machinery Digital Library es una librería digital perteneciente a la asociación estadounidense ACM fundada en 1946 con fines académicos como una sociedad científica dedicada específicamente a la informática y ciencias de la computación, actualmente tiene presencia en más de 100 países.
- Springer: editorial alemana con publicaciones multinacionales cuyo motor de búsqueda contiene múltiples disciplinas relacionadas a la ciencia incluyendo ciencias de la computación e ingeniería.

- MIT libraries: versión oficial y digital de la librería del conocido MIT (Massachusetts Institute of Technology) cuyo contenido multimedia se encuentra distribuido entre las más de 400 bases de datos de la institución.

A continuación, se muestran los resultados de la búsqueda.

Tabla 3. Resultados de la Búsqueda

Motor de búsqueda	Cadena de búsqueda	Cantidad de resultados	Preguntas de revisión relacionadas
IEEE Xplore	"machine learning" AND "computer vision" AND ("3D object reconstruction" OR "3D shape recognition")	24	1
IEEE Xplore	("object classification" OR "object detection" OR "Object segmentation") AND "machine learning"	41	1
ACM	("computer stereo vision" OR "computer vision") AND "3D shape recognition"	3	1
ACM	"neural networks" AND ("generating shape" OR "volumetric shape")	2	2
Springer	"photometric mesh optimization"	1	2
MIT libraries	"convolutional neural networks" AND ("3D object classification" OR "3D shape recognition")	2	1
MIT libraries	"3D pose estimation" OR "3D geometry reconstruction"	2	1

2.1.4 Criterios de Inclusión y Exclusión

Se incluirán los estudios que cumplan con los siguientes criterios:

- Los resultados del estudio son replicables.
- El idioma del estudio es el inglés o español.
- El modelo involucrado en el estudio posee un score de precisión superior al 60%.

Se excluirán los estudios que cumplan con los siguientes criterios:

- El estudio no es de libre acceso.
- El estudio fue desarrollado hace más de 5 años (2015), debido a que los modelos de aprendizaje de máquina enfocados en el reconocimiento de superficies presentan mejores métricas en los últimos años.

- El estudio no realiza pruebas con videos (grabados del mundo real), puesto que el propósito es también investigar casos reales aplicables.

2.1.5 Datos Extraídos

El siguiente formulario de extracción de datos es aplicable para ambas preguntas de revisión planteadas previamente. De manera general, los datos extraídos de cada artículo son: autores, título, año de publicación, tipo de bibliografía, fecha de extracción, técnicas y enfoques presentados, imágenes y referencias.

Tabla 2. Formulación de extracción de datos

Campo	Descripción	Ejemplo	Pregunta
ID	-	P01	General
Autor/es	-	Pears N.	General
Título	-	Multi-view Convolutional Neural Networks for 3D Shape Recognition	General
Año de publicación	-	2015	General
Tipo de bibliografía	-	Paper, revista de investigación, libro, etc.	General
Fecha de extracción	-	09/2019	General
Motor de búsqueda	-	IEEE Xplore	General
Técnica de representación	Técnica empleada para la representación del objeto/forma en el computador	Multi-view Images (Imágenes de vista múltiple)	P2
Enfoque del reconocimiento	Propósito del desarrollo del modelo de reconocimiento de superficies en 3D	Reconocimiento y clasificación de múltiples objetos dentro de una propia escena/imagen panorámica.	P1

		Reconstrucción de la forma de un único objeto a partir de un conjunto de imágenes.	
Problemática abordada	Problemática a partir del cual parte el trabajo de investigación.	Análisis de la viabilidad de la clasificación de múltiples objetos en una misma imagen.	P1, P2
Métrica de evaluación	La métrica empleada para evaluar el modelo o modelos desarrollado/s	Score de precisión	P1, P2

2.1.6 Datos Analizados

Los datos extraídos que se analizarán a profundidad en este caso para poder resolver las preguntas de investigación planteadas son la técnica de representación de los objetos 3D en el computador y el enfoque de reconocimiento de las superficies.

2.2 Resultados

2.2.1 Resultados de Búsqueda

Luego de aplicar los criterios de inclusión y exclusión nos quedamos con los siguientes estudios primarios obtenidos de los distintos motores de búsqueda.

Tabla 4. Lista de estudios primarios utilizados

ID	Título	Autor	Año de Publicación	Motor de Búsqueda	Tipo de Bibliografía
R01	Category-Specific Object Reconstruction from a Single Image	A. Kar, S. Tulsiani, J. Carreira, & J. Malik	2015	IEEE Xplore	CVPR

R02	Volumetric and Multi-View CNNs for Object Classification on 3D Data	Qi, C. R., Su, H., Niessner, M., Dai, A., Yan, M., & Guibas, L. J.	2016	IEEE Xplore	CVPR
R03	Object Detection in 3D Scenes Using CNNs in Multi-view Images	Qi C. R.	2016	IEEE Xplore	Paper
R04	ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes	Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T.	2017	IEEE Xplore	CVPR
R05	Learning Category-Specific Mesh Reconstruction from Image Collections	Kanazawa, A., Tulsiani, S., Efros, A. A., & Malik, J.	2018	ACM	ECCV
R06	Multi-view Convolutional Neural Networks for 3D Shape Recognition	Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E.	2015	ACM	ICCV
R07	3D ShapeNets: A Deep Representation for Volumetric Shapes	Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, & J. Xiao	2015	ACM	CVPR
R08	VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection	Y. Zhou, & O. Tuzel.	2018	MIT Libraries	CVPR
R09	Variational Autoencoders for Deforming 3D Mesh Models	Q. Tan, L. Gao, Y. Lai, & S. Xia	2018	MIT Libraries	CVPR

R10	3D-PRNN: Generating Shape Primitives with Recurrent Neural Networks	Z. Chuhang, E. Yumer, J. Yang, D. Ceylan & D. Hoiem	2017	IEEE Xplore	Paper
R11	Photometric Mesh Optimization for Video-Aligned 3D Object Reconstruction	Lin, C.-H., Wang, O., Russell, B. C., Shechtman, E., Kim, V. G., Fisher, M., & Lucey, S	2019	Springer	CVPR

2.3 Discusión

En esta sección se discutirán los resultados de la revisión sistemática enfocado a responder las preguntas de investigación planteadas en las secciones previas. Para ello, cada pregunta de investigación representará una sub-sección bajo el cual se mencionarán y discutirán los trabajos de investigación relacionados.

2.3.1 ¿Qué enfoques se buscan con la generación de superficies u objetos reales a partir de imágenes?

En el estado del arte se han identificado múltiples técnicas de generación de superficies a partir de imágenes. Entre los enfoques hallados encontramos: la estimación de la posición 3D de un objeto a partir de una única imagen, la clasificación de un objeto, la detección de múltiples objetos en una escena, la segmentación de una escena por regiones y finalmente la reconstrucción y síntesis geométrica 3D a partir de una colección de imágenes.

2.3.1.1 Estimación de posición 3D

Category-Specific Object Reconstruction from a Single Image (2014)

El ser humano es capaz de reconstruir un objeto en su mente a partir de una sola imagen. Esto lo podemos lograr gracias a los años de experiencia que tenemos viendo otros objetos de la misma clase. Con clases nos referimos a tipos de objetos (ej.: carros, aviones, motocicletas, televisores, etc.). Pueden existir muchas variaciones dentro de cada clase, lo que llamamos subclases o subcategorías.

Para lograr una reconstrucción digital se necesita construir un modelo que reciba como entrada pixeles y genere una superficie 3D de la clase y subcategoría correspondiente (A. Kar et al. 2015). Para ello se modeló un proceso donde a partir de un conjunto de imágenes pertenecientes a una misma subcategoría, se estima el punto de vista por cada imagen. El

framework de NRSfM o Non-rigid Shape from Model propuesto por Bregler C. et al. se emplea para las estimaciones de los puntos de vistas esparcidos alrededor del objeto (C. Bregler, A. Hertzmann, & H. Biermann, 2000). El siguiente paso combina las siluetas detectadas en cada imagen con los puntos de vista calculados para generar distintas formas medias en 3D (A. Kar et al. 2015). Estas formas son deformables y por ende capturan las variaciones que se presentan con las diferentes imágenes. Estas variaciones con respecto a la forma aprendida media del objeto se denominan variaciones intraclase (variación de superficie entre objetos de una misma sublcase).

En el trabajo de tesis presente encontramos que dentro de la categoría: “piezas arqueológicas” y subcategoría: “huacos” los objetos presentan múltiples diferencias, lo que complica el entrenamiento del modelo.

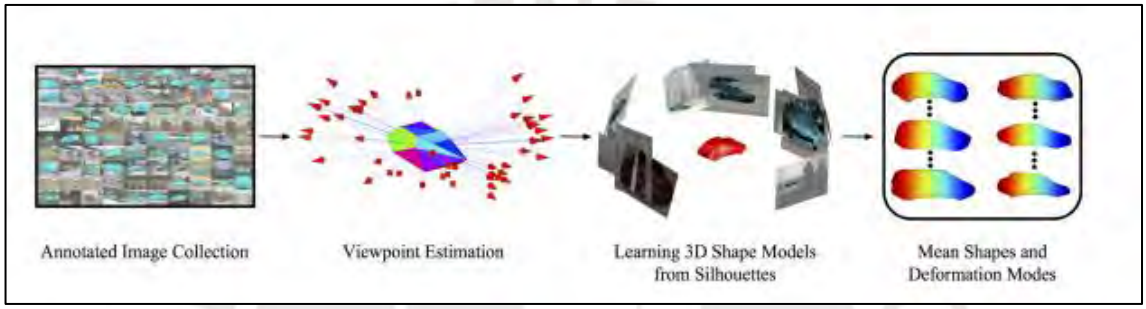

Figura 1. Flujo del proceso para la reconstrucción de la forma media de un objeto a partir de una colección de imágenes (A. Kar et al., 2015).

2.3.1.2 Clasificación de un único objeto

Volumetric and Multi-View CNNs for Object Classification on 3D Data (2016)

Las redes neuronales convolucionales (CNNs) son utilizados en trabajos de investigación para resolver problemas de clasificación de imágenes. Debido a que la captura de información 3D de objetos es desarrollada constantemente, las CNN’s juegan un papel importante en la clasificación de dichos objetos (Qi et al., 2016). Existen dos tipos de CNN’s principales en este campo: volumetric CNN’s y multi-view CNN’s, las cuales son utilizadas cuando las representaciones de los objetos en el computador son volumétricas o multi-vistas respectivamente. Se puede afirmar que ambas herramientas son el soporte del entendimiento del espacio para el proceso de clasificación (Qi et al., 2016).

Las CNN’s son principalmente ventajosas en “aprender” características sobre objetos en imágenes o vídeos RGB. Esto permite la identificación de las características principales. El uso de esta herramienta en el campo 3D empieza a partir de un conjunto de datos (representaciones en 3D de objetos) donde se detectan las características principales de los grupos de objetos, a esto se le llama fase de entrenamiento (Qi et al., 2016). Generalmente

se utiliza data de diferentes bases de datos para entrenar el modelo y datos del mundo real para las pruebas. Luego, con el modelo entrenado se puede procesar data de casos reales a través de las diferentes capas (convolutional layer, pooling layer y fully connected layer o FC layer) para la clasificación.

El trabajo presentado por Qi Charles, Su Hao et al. en 2016 pretende explorar las características de las CNN's volumétricas y multi-vistas. Ellos buscan analizar y mejorar el uso de dichas herramientas.

La representación multi-vista constituye en una forma 3D que se renderiza en múltiples imágenes mediante una o varias cámaras. Estas vistas del objeto deben captar el 100% de la superficie de ser posible (Qi et al., 2016). Luego, los atributos principales de las imágenes se extraen por cada vista generada con ayuda de la CNN. Estos atributos son procesados entre las diferentes vistas en la capa de pooling para finalmente pasar a la capa FC (fully connected layer). Por otro lado, las CNN's volumétricas codifican la representación de la forma en 3D en un tensor de valores binarios o reales (Qi et al., 2016).

Figura 2. Resultados de los experimentos realizados a las CNN's para medir su desempeño bajo las mismas condiciones (Qi et al., 2016).

Las estadísticas iniciales entre ambas redes nos indican que un CNN volumétrico es 7.3% peor que un CNN multi-vistas. Luego de una serie de experimentos sobre el desempeño de ambas redes, desde un punto de vista arquitectónico y de resoluciones 3D, los autores proponen nuevas arquitecturas que superan a las anteriores bajo las mismas condiciones y tamaño de input.

2.3.1.3 Detección de múltiples objetos

Object Detection in 3D Scenes Using CNNs in Multi-view Images (2016)

En el trabajo propuesto por Qi C. R. se busca detectar múltiples objetos en una escena reconstruida en 3D. Para lograr ello se combinaron 2 componentes básicos que ya existían en trabajos de investigación previos: la detección de cuerpos en imágenes 2D y el cálculo de profundidades en imágenes. Para lograr lo propuesto se realizó un proceso (pipeline) que recibe un video RGB y genera un mapa de calor. Este mapa de calor se divide en varios rangos espaciales pequeños donde cada uno muestra la probabilidad de que cierto objeto

se encuentre en dicho espacio. De esta forma, se logra representar la detección de múltiples objetos en una escena completa (Qi C. R. 2016).

La escena, en este caso, debe estar representada por una serie de imágenes de la misma. Primero se clasifican dichas imágenes obteniendo como resultado un cuadrilátero sobre la imagen que encierra el objeto clasificado. Esto se logra haciendo uso de las CNN's entrenados previamente con los objetos a clasificar. Además, el algoritmo también indica el score de precisión para usarlo posteriormente en el cálculo de las probabilidades.

Una vez se cuenta con todos los cuadros clasificados se recuperan las posiciones de las cámaras (donde fueron tomadas las imágenes) y las profundidades de los objetos en las imágenes (Qi C.R. 2016). Esto se logra haciendo uso de propuestas en trabajos de investigación previos.

Al contar con los cuadriláteros y las profundidades se procede a proyectar las imágenes 2D a una cuadrícula de voxeles, donde cada profundidad de la imagen corresponde únicamente a un voxel (volumetric pixel / pixel volumétrico).

Finalmente, se realizan los cálculos estadísticos: los pixeles al interior de un cuadrilátero son los píxeles pertenecientes a un objeto (teóricamente), por lo tanto, el voxel correspondiente es asignado con una probabilidad (basado en el score) que mide la presencia del objeto (Qi C. R. 2016). En este trabajo se prescindió de la colisión entre cuerpos.

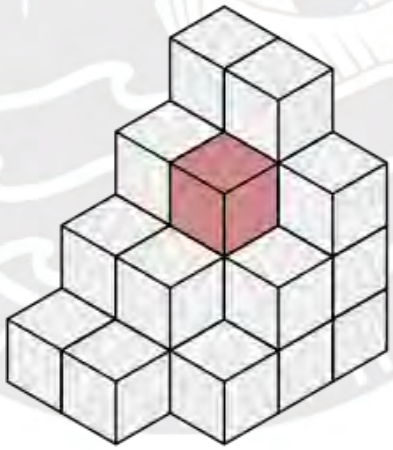

Figura 3. Conjunto de vóxeles agrupados (M. W. Toews. para Wikipedia).

2.3.1.4 Segmentación de una escena por regiones (ScanNet)

ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes (2017)

En el campo de segmentación de una escena, Dai A. et al. propusieron el uso de un dataset llamado ScanNet, que contiene cerca de 2.5 millones de vistas de 1513 escenas en videos RGB-D. Como mencionan los autores, existen muy pocos datasets que puedan ser usados en este campo para estimar posiciones 3D de cámaras, reconstrucción de superficies y

segmentaciones semánticas. ScanNet permitió (en el trabajo de investigación mencionado) la construcción de un sistema capaz de segmentar escenas con sus respectivas etiquetas (Dai et al., 2017).

En este trabajo de investigación se propone el flujo de un sistema dedicado a usuarios con nulo conocimiento sobre el tema. La utilidad principal del sistema es que el usuario pueda grabar una escena (indoor scene) y reciba como respuesta una construcción 3D de la escena grabada, pero con el valor agregado de la segmentación (refiriéndose a la clasificación de regiones pertenecientes a objetos y sus respectivas etiquetas). Por ejemplo, si en una escena de un cuarto encontramos una cama, una mesa y una silla, el output de dicho sistema debería ser la reconstrucción 3D del cuarto y con regiones indicando los objetos que pertenecen a dichas regiones, así como una región que pertenezca al espacio desocupado.

El modelo CNN que se emplea es entrenado con los 2.5 millones de imágenes pertenecientes a ScanNet y formas parciales de ShapeNet. Primero se voxeliza la escena y por cada voxel se guarda información sobre la clase de objeto al que pertenece (incluyendo espacio vacío). Luego se divide el espacio en volúmenes de 31 x 31 x 62 voxeles y son alineados con el plano del piso. Para ello, se debe verificar que la muestra entregada (escena) posee el 2% de su espacio lleno, de lo contrario es descartada.

Luego se emplea el CNN entrenado con una posibilidad de clasificar entre 20 objetos más el espacio vacío. Para la clasificación se analiza una columna de voxeles y se compara con sus columnas más cercanas. El análisis previo a la clasificación es netamente geométrico (Dai et al., 2017). El resultado final es una clasificación de todas las superficies visibles de una escena en 3D bajo 20 posibles etiquetas de objetos, una herramienta que sin duda ayuda al entendimiento digital del espacio y puede ser usado para otras investigaciones.

2.3.1.5 Reconstrucción y síntesis geométrica en 3D

Learning Category-Specific Mesh Reconstruction from Image Collections (2018)

La reconstrucción geométrica en 3D se lleva a cabo mediante 3 tipos de métodos que se basan en diferentes conceptos. Estos métodos son:

- Métodos basados en modelos paramétricos cambiantes (Parametric morphable Model-based)
- Métodos basados en el aprendizaje de plantillas (Part-based Template learning)
- Métodos de aprendizaje profundo (Deep learning methods)

En este trabajo se expone nuevamente la reconstrucción 3D de un objeto, pero además se considera el agregado de texturas. El modelo puede funcionar con múltiples formas si se cuenta con la data suficiente para el entrenamiento. La idea que proponen los autores es

que a través de un modelo entrenado (empleando deep learning) sobre una colección de imágenes (pertenecientes a un objeto que se desea reproducir) se parametriza una forma media aprendida. Dicha forma es representada por una malla deformable en 3D. Cuando se experimenta con una nueva imagen/instancia no perteneciente al conjunto de entrenamiento la forma 3D se genera a partir de la forma media aprendida y las deformaciones predichas por la instancia actual (Kanazawa, Tulsiani, Efros, & Malik, 2018).

Dos problemas principales que presenta este enfoque son:

- Por cada objeto que se desea reproducir debe aprenderse una vasta colección de imágenes, por la tanto, si no se cuenta con la data será imposible realizar la reconstrucción. Las imágenes deben tener información adjunta sobre la expansión de puntos claves (keypoints) y las máscaras de segmentación, esto los hace más difíciles de obtener.

- Existe una malla deformable 3D por objeto aprendido. Una instancia muy distinta al resto generará predicciones con mayor cantidad de deformaciones, lo que implica mayor error cuadrático medio en la forma 3D generada.

Como fue mencionado, los autores también proponen la recuperación de las texturas y ligarlos a la forma reconstruida. La siguiente imagen muestra la perspectiva general del modelo construido, el cual se divide en 3 módulos que calculan la posición de la cámara relativa al objeto, la deformación con respecto a la malla media aprendida y las texturas.

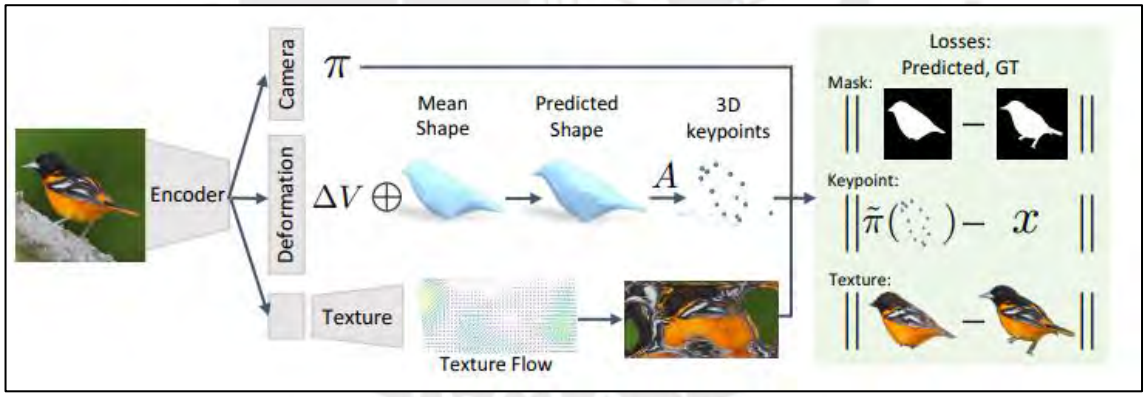

Figura 4. Representación de los módulos y flujo (Kanazawa et al., 2018).

2.3.2 ¿Cuáles son las técnicas que se utilizan para representar un objeto 3D en el computador?

Entre los artículos revisados encontramos las siguientes técnicas de representación: imágenes de vista múltiple, representación volumétrica, nube de puntos y red de polígonos.

2.3.2.1 Imágenes de vista múltiple

Multi-view Convolutional Neural Networks for 3D Shape Recognition (2015)

En este trabajo se propone el reconocimiento de una superficie a partir de 12 imágenes que capten las vistas 360° alrededor del objeto. Las imágenes se procesan mediante una CNN para extraer las principales características del objeto. Se simplifican las vistas mediante la operación de “pooling” para que sean comparadas entre ellas y procesadas mediante un segundo CNN. El output final del segundo CNN será la predicción de la clasificación del objeto (Su, Maji, Kalogerakis, & Learned-Miller, 2015).

Este trabajo en particular emplea la representación de un objeto en 3D mediante el uso de múltiples vistas en diferentes imágenes. Estas imágenes se compactan en un único contenedor llamado descriptor de forma (shape descriptor). De esta forma, la información sobre las múltiples vistas de un mismo objeto permiten su clasificación.

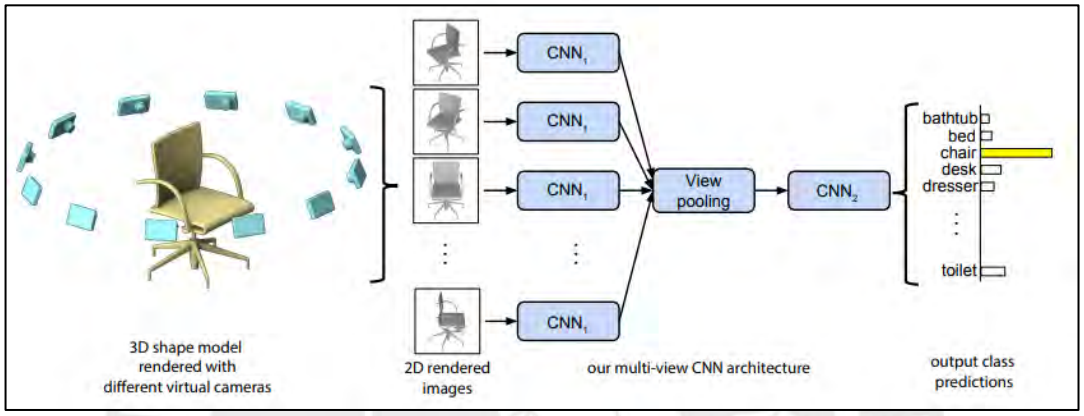

Figura 5. Pipeline del reconocimiento de una silla mediante el uso de CNN’s y la representación multi-vistas (Su et al., 2015).

2.3.2.2 Volumétrico

3D ShapeNets: A Deep Representation for Volumetric Shapes (2015)

En este caso se aprovechan los mapas de profundidad (depth maps) generados por sensores de profundidad 2.5D para la representación de objetos del mundo real en el computador. Para ello, se construye la representación como una distribución de un conjunto de variables binarias. Lo que se conoce como representación volumétrica. Estas variables se almacenan en una cuadrícula de voxeles como se muestra en el 3° paso de la imagen inferior y almacenan la probabilidad de que cierto objeto se encuentre en dicho punto en el espacio.

Los autores introducen 3D-ShapeNets, un modelo CNN que aprende la distribución del espacio que ocupa la forma y también soporta una sugerencia para el autocompletado de mapas de profundidad incompletos (Zhirong Wu et al., 2015).

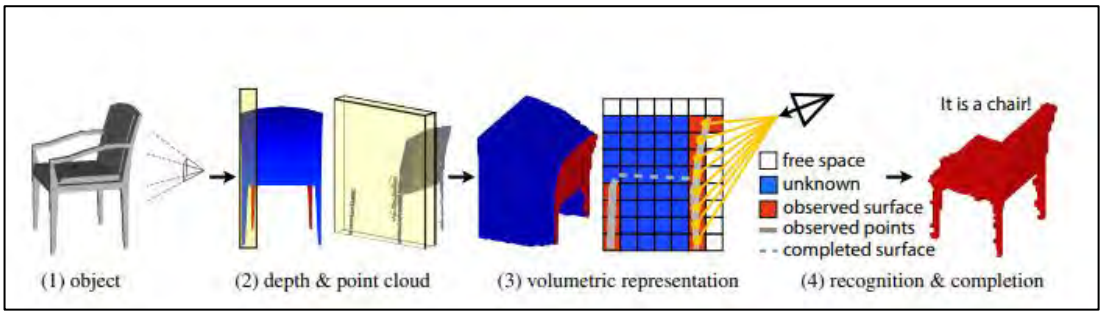

Figura 6. Flujo de un objeto en el mundo real hasta la obtención de su representación volumétrica (Zhirong Wu et al., 2015).

Lo resaltante de este trabajo es la forma en cómo se representan las formas 3D digitalmente. Como podemos apreciar en la figura, se consigue una cuadrícula consistente de voxeles con las probabilidades asociadas para regiones visibles de la superficie, así como las regiones desconocidas.

2.3.2.3 Nube de puntos

VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection (2017)

En este proyecto se propone el modelo VoxelNet, una red de detección 3D basado en un modelo de deep learning capaz de predecir/clasificar objetos, recuperar las características principales y unirlas con la generación de un recuadro alrededor de la imagen en la predicción. Además, se demuestra que VoxelNet tiene un mejor desempeño que algunos sistemas LiDAR propuestos anteriormente (Y. Zhou & O. Tuzel, 2018).

La representación 3D se da mediante una nube de puntos. Por cada conjunto de puntos perteneciente al rango de un voxel se detecta una característica principal (conocido como feature). De esta forma la nube de puntos se convierte en una representación volumétrica descriptiva (Y. Zhou & O. Tuzel, 2018).

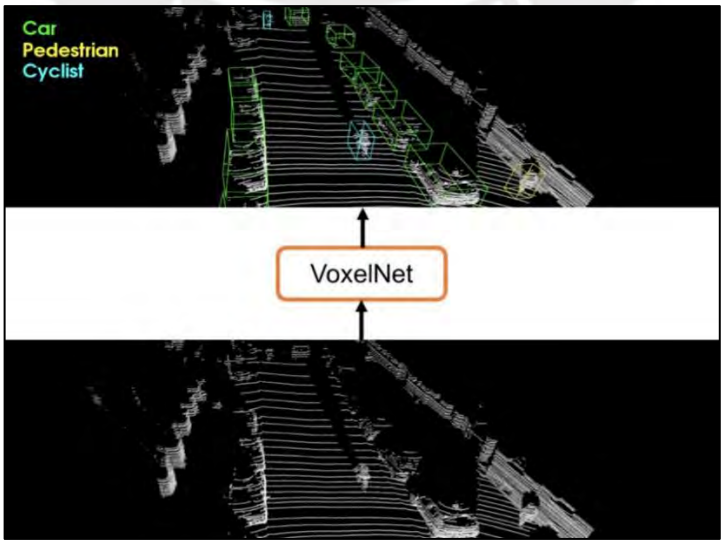

Figura 7. Input y Output de VoxelNet. Nube de puntos sin procesar (Y. Zhou & O. Tuzel, 2018).

Recuperar la superficie a partir de una nube de puntos requiere de aprendizaje de máquina debido a que debemos asociar dichos puntos a una superficie o forma. Los puntos o conjuntos de coordenadas X-Y-Z en realidad carecen de sentido alguno si no se procesan o explotan (Y. Zhou & O. Tuzel, 2018). Sin embargo, visualizar una nube de puntos podría darnos una idea rudimentaria de la densidad del objeto y su complejidad de reconstrucción. La captura de los puntos también se realiza mediante un proceso de escaneo sobre el objeto físico (Y. Zhou & O. Tuzel, 2018).

2.3.2.4 Red poligonal

Variational Autoencoders for Deforming 3D Mesh Models (2018 CVPR)

En este trabajo los autores resaltan la importancia de una representación de modelos 3D como mallas deformables. Este tipo de representación permite la modificación del objeto (en forma) luego de ser reconstruido. Esta variabilidad apoya algunas aplicaciones en medicina, así como animaciones y entretenimiento (Q. Tan, L. Gao, Y. Lai, & S. Xia, 2018).

Una malla como representación 3D se compone de polígonos adheridos que imitan la superficie del objeto. La ventaja es que al tratarse de polígonos todas las regiones de la superficie representada son convexas, por lo tanto, no existen regiones no visibles desde las múltiples vistas.

En el trabajo de investigación se propone un framework que permite el modelamiento de probabilidades de la presencia de cuerpos en un ambiente 3D representado por mallas deformables. Los autores lo denominaron modelo de malla VAE (Variational AutoEncoder) (Q. Tan et al., 2018). Se trata de una red fácil de entrenar ya que requiere de mínima data como input de formas deformables. Sin embargo, la limitación es que solo se pueden procesar mallas homogéneas.

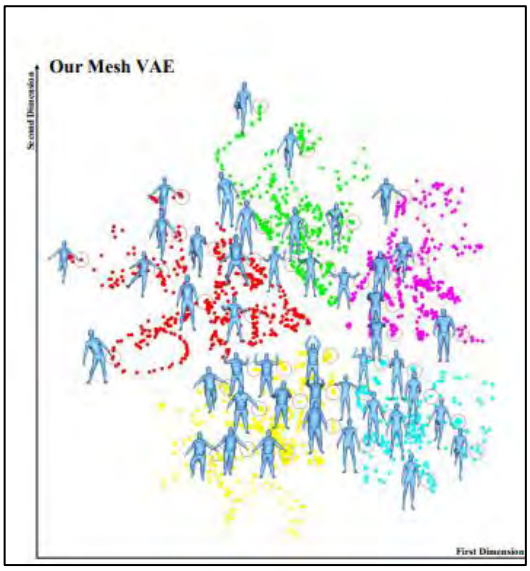

Figura 8. Dyna dataset, representaciones poligonales. Cada color representa una pose humana diferente (Q. Tan et al., 2018).

2.4 Revisión de Tesis

Para complementar las preguntas de revisión también se llevó a cabo una búsqueda de trabajos de investigación del repositorio de Tesis PUCP. Se emplearon las siguientes cadenas de búsqueda: “reconstrucción 3D”, “gráficos por computadora” y “visión computacional”. Además, se limitaron los resultados únicamente a la facultad de Ciencias e Ingeniería y de especialidad Ingeniería Informática. De los 9 resultados obtenidos solo se encontró un trabajo de investigación relevante para la revisión literaria actual que permite complementar las preguntas de investigación.

Toribio (2018) propone el desarrollo de un software que permita completar la geometría de piezas arqueológicas parcialmente dañadas mediante la predicción de simetrías en estructuras complejas. En la tesis denominada “Desarrollo de herramienta de visualización para la reparación de piezas arqueológicas basado en su simetría” se emplea la representación de las piezas como mallas de polígonos. Esto facilita la visualización y detección de simetrías en la herramienta propuesta. Esta forma de representar la pieza arqueológica además le permite modificar los polígonos que definen la malla de manera que sea posible la reparación de la superficie incompleta basándose en las simetrías encontradas.

Finalmente se puede resumir que el enfoque del autor es ayudar a completar las reconstrucciones de los objetos 3D integrando el método de aproximación de detección de simetrías en Mallas 3D (Sipiran, Gregor & Schreck, 2014) con una interfaz gráfica que le permita al usuario interactuar con el método brindando posibles ejes o planos de simetría que sirven como input para llevar a cabo las reconstrucciones de las piezas.

3 Conclusiones

En conclusión, existen varias formas de representar objetos 3D en el computador. Cada representación implica un procesamiento diferente y dependen fuertemente de las herramientas empleadas durante el proceso de escaneo 3D de los objetos. Además, encontramos que algunos tipos de representaciones como las mallas de polígonos y nube de puntos pueden generarse a partir de redes neuronales entrenadas en una extensa colección de imágenes sobre los propios objetos. Estos conjuntos de datos iniciales y su pre-procesamiento son cruciales para la obtención de resultados. Debido a que existe una dependencia importante con los datos de entrenamiento para la obtención de resultados, los autores buscan integrar la mayor cantidad de datos posibles de distintas fuentes o 'datasets' con el fin de mejorar la calidad de sus reconstrucciones.

También existen proyectos que emplean técnicas de reconstrucción con múltiples enfoques o propósitos distintos como la estimación de la posición 3D de un objeto a partir de una única imagen, la clasificación de uno o varios objetos en una escena, la segmentación de una escena por regiones o la reconstrucción y síntesis geométrica 3D partiendo de una colección de imágenes. Sin embargo, son pocos trabajos los que se dedican específicamente al escaneo de superficies altamente variables debido a la dificultad que implica el reconocimiento de dichas superficies.

En el estado del arte explorado encontramos métodos de reconstrucción 3D pasivos que permiten capturar las superficies de los objetos sin interactuar físicamente con los mismos. Esto se logra mediante herramientas como cámaras fotográficas que generan una colección extensa de imágenes las cuales son procesadas por una red neuronal entrenada. Sin duda se deben explorar a mayor detalle este tipo de tecnologías pasivas de escaneo 3D con el fin de mejorar el reconocimiento de superficies altamente variables.

4 Referencias

A. Kar, S. Tulsiani, J. Carreira, & J. Malik. (2015). Category-specific object reconstruction from a single image. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1966-1974. https://doi.org/10.1109/CVPR.2015.7298807

A. Kar, S. Tulsiani, J. Carreira, & J. Malik. (2015). Category-specific object reconstruction from a single image. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1966-1974. https://doi.org/10.1109/CVPR.2015.7298807

ACMDL (2020) Association for Computing Machinery Digital Library. Consulta: Setiembre 2019 a Diciembre 2020. https://dl.acm.org/

Blanz, V., & Vetter, T. (1999). A Morphable Model for the Synthesis of 3D Faces. Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, 187–194. https://doi.org/10.1145/311535.311556

C. Bregler, A. Hertzmann, & H. Biermann. (2000). Recovering non-rigid 3D shape from image streams. Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662), 2, 690-696 vol.2. https://doi.org/10.1109/CVPR.2000.854941

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nie\s sner, M. (2017). ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. Proc. Computer Vision and Pattern Recognition (CVPR), IEEE.

IEEE (2020) Institute of Electrical and Electronics Engineers Xplore Digital Library. Consulta: Setiembre 2019 a Diciembre 2020. https://ieeexplore.ieee.org/Xplore/home.jsp

Kanazawa, A., Tulsiani, S., Efros, A. A., & Malik, J. (2018). Learning Category-Specific Mesh Reconstruction from Image Collections. ECCV.

Lin, C.-H., Wang, O., Russell, B. C., Shechtman, E., Kim, V. G., Fisher, M., & Lucey, S. (s. f.). Photometric Mesh Optimization for Video-Aligned 3D Object Reconstruction. 10.

MIT Libraries (2020) Massachusetts Institute of Technology Digital Library. Consulta: Setiembre 2019 a Diciembre 2020. https://libraries.mit.edu/

Pears, N. & Yonghuai (2012) 3D Imaging, Analysis and Applications. York: Springer

Qi, C. R., Su, H., Niessner, M., Dai, A., Yan, M., & Guibas, L. J. (2016). Volumetric and Multi-View CNNs for Object Classification on 3D Data. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Qi C. R. (2016). Object Detection in 3D Scenes Using CNNs in Multi-view Images. Stanford University.

Q. Tan, L. Gao, Y. Lai, & S. Xia. (2018). Variational Autoencoders for Deforming 3D Mesh Models. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5841-5850. https://doi.org/10.1109/CVPR.2018.00612

Sipiran, I., Gregor, R., & Schreck, T. (Octubre de 2014). Approximate Symmetry Detection in Partial 3D Meshes. doi:10.1111/cgf.12481

Springer (2020) Springer Science+Business Media. International Publisher Science and Technology. Consulta: Setiembre 2019 a Diciembre 2020. https://www.springer.com/la

Sreenivasa Kumar Mada, Melvyn L. Smith, Lyndon N. Smith, & Prema Sagar Midha. (2003, marzo 19). Overview of passive and active vision techniques for hand-held 3D data acquisition. 4877. Recuperado de https://doi.org/10.1117/12.463773

Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. G. (2015). Multi-view convolutional neural networks for 3d shape recognition. Proc. ICCV.

Toribio, G. (Diciembre de 2018). Desarrollo de Herramienta de Visualización para la Reparación de Piezas Arqueológicas basado en su Simetría. http://hdl.handle.net/20.500.12404/13894

Y. Zhou, & O. Tuzel. (2018). VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4490-4499. https://doi.org/10.1109/CVPR.2018.00472

Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, & J. Xiao. (2015). 3D ShapeNets: A deep representation for volumetric shapes. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1912-1920. https://doi.org/10.1109/CVPR.2015.7298801