Universidad Autónoma del Estado de México

Facultad de Ingeniería

Propuesta de una taxonomía para la reconstrucción urbana de interiores y exteriores

T E S I S

QUE PARA OBTENER EL GRADO DE:

Maestro en Ciencias de la Ingeniería

PRESENTA:

Rafael Mercado Herrera

DIRIGIDA POR:

Dra. Vianney Muñoz Jiménez
Dr. Marco Antonio Ramos Corchado
Dr. José Antonio Hernández Servín

Toluca, México, Septiembre 2018

If one hopes to achieve full understanding of a system...then one must be prepared to contemplate different levels of description that are linked, at least in principle, into a cohesive whole.

David Courtnay Marr

Resumen

La reconstrucción tridimensional es el proceso de generar la forma y representación de objetos de manera tridimensional, es un campo importante para diversas disciplinas como son el Diseño Asistido por Computadora, Gráficos de Computadora, Realidad Virtual, Ingeniería Civil, etc. A pesar del progreso en el área de reconstrucción, aún existen múltiples problemas por resolver que requieren cooperación multidisciplinaria; esta abundancia de problemas a tratar, apoyada por el progreso tecnológico en los equipos de captura y procesamiento, ha causado un progreso acelerado en el área.

Hablando de la reconstrucción urbana, tópico central de este documento, en años recientes, se han propuesto diferentes taxonomías que han contribuido a mejorar la comprensión del problema y facilitar la búsqueda de los acercamientos actuales propuestos para solventarlo. Sin embargo, es precisamente por el progreso acelerado en el área, y a través de diversas disciplinas, que es necesario visualizar el estado del arte de manera pragmática, como se propone en esta investigación por medio de una taxonomía.

Se presenta una propuesta de taxonomía para la reconstrucción urbana que toma en cuenta características que comienzan a demostrar su utilidad: el conocimiento de los elementos dinámicos de la escena (tales como personas caminando por la escena) y el tipo de interacción con el usuario durante la reconstrucción (como puede ser un funcionamiento en línea con retroalimentación durante la reconstrucción).

Se presenta la propuesta de un *pipeline* de reconstrucción tridimensional para lidiar con un tema poco retomado, la reconstrucción de un modelo tridimensional del exterior e interior de una estructura a partir de imágenes digitales tomadas de la estructura en cuestión.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Rafael Mercado Herrera. Toluca, México, 2018

Reconocimientos

Esta maestría fue posible gracias al financiamiento ofrecido por el Consejo Nacional de Ciencia y Tecnología por medio de una beca proporcionada al expediente número 784441/447539.

Igualmente, se realizó una estancia investigación en la Unidad Guadalajara del Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, lo que potenció los alcances de esta investigación.

Tabla de contenido

Índice de figuras

Pág.

Índice de tablas

Pág.

Capítulo 1

Introducción

En este capítulo se presenta el problema observado, los objetivos de la investigación realizada sobre el mismo y las contribuciones resutlantes.

1.1. Presentación

Diversas tareas de las personas como el entrenamiento e incluso la preservación de patrimonios culturales, se han visto mejoradas por el empleo de entornos virtuales. La generación de estos entornos se denomina reconstrucción tridimensional ó 3D [1].

La reconstrucción tridimensional usualmente se realiza de manera manual [2], pero también abre la posibilidad de aplicar técnicas que permitan trasladar escenas del mundo real a modelos tridimensionales empleables en estos entornos virtuales.

Estas técnicas, hasta cierto punto automatizables en dispositivos, logran reducir la carga de trabajo y las barreras de conocimientos y habilidades que pueden encontrarse en un usuario, lo que democratiza el uso de estos conocimientos y aumenta la efectividad de los expertos en el tema de generación de entornos virtuales, incrementando sus opciones [3].

La reconstrucción tridimensional actual se enfrenta al problema de encontrar una técnica adecuada para cada situación, con sus objetivos y limitaciones propios [1, 3];

investigadores de diversas disciplinas, como la graficación por computadora y la ingeniería civil, realizan propuestas para responder a situaciones específicas dentro del área de reconstrucción, y mantenerse al tanto de estas propuestas se vuelve difícil sin una guía que permita agrupar trabajos para su comparación o uso complementario.

En esta investigación se presenta una taxonomía para facilitar la búsqueda de aportes multidisciplinarios al área de reconstrucción tridimensional urbana; la taxonomía toma en cuenta las siguientes características:

- La forma en que se lidia con los **objetos dinámicos** de la escena (i.e. objetos que se mueven o cambian entre observaciones de la escena).
- El **funcionamiento del sistema** (i.e. funcionamiento *online* sobre un flujo de datos u *offline* sobre un grupo de datos).
- El **tipo de reconstrucción** (i.e. guiado únicamente por los fundamentos de la fotogrametría, apoyado por presuposiciones arquitectónicas o modelado como la solución conjunta del problema de reconstrucción y de segmentación semántica).

Se realiza la propuesta de un *pipeline* de reconstrucción enfocado a resolver un problema poco retomado en el estado del arte: generar un modelo tridimensional del exterior e interior de una estructura a partir de imágenes digitales 2D tomadas de la estructura en cuestión.

1.2. Objetivo

El objetivo general de esta investigación es proponer una taxonomía que unifique conocimientos de las diferentes disciplinas tales como: fotogrametría y sensado remoto, visión por computadoras, ingeniería civil, graficación, entre otras, para la reconstrucción de entornos virtuales y que permita una comunicación interdisciplinaria.

Los objetivos específicos requeridos para alcanzar el objetivo general de esta investigación se presentan a continuación:

- Realizar una revisión extensa del estado del arte del área de la reconstrucción tridimensional urbana.

- Identificar elementos comparables entre las propuestas revisadas en cuanto a sus objetivos, métodos y aplicaciones.
- Generar una taxonomía con base en los elementos identificados.
- Clasificar las propuestas revisadas de acuerdo a la taxonomía propuesta.
- Presentar una propuesta de reconstrucción usando la taxonomía como guía.

1.3. Meta de ingeniería

Una propuesta de taxonomía que clasifique los acercamientos de reconstrucción tridimensional urbana de acuerdo a los objetivos y procesamientos de los mismos; la taxonomía debe estar basada en el estado del arte interdisciplinario del área de investigación.

1.4. Motivación

La reconstrucción tridimensional urbana tiene impacto en una gran variedad de disciplinas (i.e. ingeniería civil, visión por computadora, robótica), al mismo tiempo avanza gracias a aportes de investigadores de las mismas [4], lo que trae a la luz la necesidad de facilitar la comunicación interdisciplinaria para exponenciar el impacto de los aportes realizados [3].

Un ejemplo de la utilidad de la comunicación interdisciplinaria se encuentra en la reconstrucción de modelos de edificios. Los acercamientos actuales a la reconstrucción de modelos a partir de entornos del mundo real se enfocan por separado en exteriores o interiores de edificaciones, sin realizar una reconstrucción completa del entorno, lo que impide una navegación libre dentro [5] y fuera del mismo [6].

Se percibe que dichos acercamientos proveen suficiente información en conjunto con las imágenes de entrada para permitir la alineación de las reconstrucciones tal que se puedan integrar en un modelo completo; esta investigación está motivada por la posibilidad de cubrir esta brecha, tanto en la comunicación interdisciplinaria, como en el problema más específico aquí descrito.

1.5. Planteamiento del problema

Los avances tecnológicos han permitido la proliferación del empleo de entornos virtuales como apoyo en la realización de tareas de las personas; un ejemplo de estas tareas es el entrenamiento médico, en donde se les permite a los nuevos doctores familiarizarse con herramientas y procedimientos en entornos controlados para evitar errores que pongan en peligro el bienestar de un paciente [7].

De la misma manera, en el ámbito militar, se realizan evaluaciones por medio de un simulador en realidad virtual como complemento para el aprendizaje de las reglas de maniobra, que son aprendidas predominantemente con material no interactivo [8].

Estas aplicaciones requieren de entornos virtuales en los cuales puedan interactuar los usuarios, los cuales pueden ser generados manualmente [2] o semi-automáticamente [9]. En la Figura 1.1 se muestran ejemplos de estas últimas dos aplicaciones.

(a) Reconstrucción de un pozo romano en [10].

(b) Aplicación HouseCraft [11] para recorridos virtuales.

Figura 1.1: Ejemplos de aplicaciones de realidad virtual.

Usualmente la generación de entornos virtuales se realiza usando herramientas de diseño asistido por computadora [3]; una alternativa o apoyo a este acercamiento es la generación automática de entornos virtuales a partir de entornos del mundo real.

El uso de escáneres láser permite recolectar los datos necesarios del ambiente a reconstruir, sin embargo, éstos resultan costosos; la ventaja que ofrecen estos equipos es la alta calidad de los ambientes virtuales reconstruidos. Por otra parte, se encuentra el uso de fotografías para la reconstrucción, obteniendo un modelo de calidad comparable al obtenido con el uso de láser de acuerdo a un estudio reciente en [12]

Aunque los acercamientos actuales a la reconstrucción de modelos a partir de en-

tornos del mundo real, que ya tienen implementaciones prácticas [13, 14], se enfocan en exteriores o interiores de edificaciones, sin realizar una reconstrucción completa del entorno, con lo que se termina generando una cáscara del exterior del edificio o el modelo de los cuartos del interior del edificio sin el contexto de su exterior.

Cada uno de estos acercamientos tiene su origen en diversas disciplinas como son sensado remoto, graficación por computadora, inteligencia artificial, tratamiento de imágenes, entre otras; retoman la reconstrucción de ambientes virtuales para obtener sus fines particulares y consecuentemente aportan al área de reconstrucción propuestas congruentes con sus fortalezas particulares.

La diversidad de los aportes al área de reconstrucción urbana permiten una evolución constante y multidisciplinaria que requiere de comunicación y conciencia de las otras disciplinas, algo que en ocasiones puede tomar años en ocurrir. Un claro ejemplo del valor de la comunicación interdisciplinaria se observa con el trabajo seminal de Triggs et al. [15]; presentaron de manera exhaustiva un método de fotogrametría a la comunidad de visión de computadoras y en estos días es el método predominante para la reconstrucción [16].

Del mismo modo, las propuestas presentadas por investigadores se ven poco socorridas a causa de la multidisciplinariedad de la tarea que dificulta la transmisión de conocimientos y su fusión para resolver problemas de mayor complejidad [3, 4], con lo que se vuelve aparente la necesidad de establecer un canal de comunicación multidisciplinaria que favorezca una actualización conjunta del conocimiento.

1.6. Contribuciones

Se presenta una taxonomía actualizada para la reconstrucción tridimensional junto con una revisión de trabajos recientes clasificados de acuerdo a la misma. Usando los elementos reconocidos durante la creación de la taxonomía, se propone un *pipeline* de reconstrucción para la reconstrucción tridimensional enfocado en la reconstrucción de interiores y exteriores de entornos urbanos. Durante esta investigación se realizó la producción de dos artículos, uno publicado en un congreso y otro enviado a una revista especializada.

1.7. Estructura de la tesis

Este trabajo está dividido en cinco capítulos. En el segundo capítulo se encuentra el marco teórico que pone en contexto la investigación y presenta el estado del arte actual en la rama de la reconstrucción tridimensional.

En el tercer capítulo se presenta la taxonomía propuesta, detallando cada uno de sus elementos, y el *pipeline* propuesto, igualmente detallando las herramientas y procedimientos que envuelve.

Distintas observaciones hechas empleando la taxonomía se presentan en el cuarto capítulo, permitiendo extraer conclusiones del estado actual del área de investigación. Se muestran también resultados y situaciones interesantes encontrados en el *pipeline*.

Finalmente, en el último capítulo se presentan las conclusiones y propuesta de trabajos futuros.

Capítulo 2

Marco teórico

En este capítulo se abordan los conceptos y conocimientos generales empleados a lo largo de la tesis y se discutirán los avances en el estado del arte actual.

2.1. Antecedentes

Este tema de investigación abarca diversos temas, desde el concepto de realidad virtual hasta los conocimientos necesarios para reconstruir un modelo tridimensional a partir de un conjunto de imágenes; en esta sección se presentan los conocimientos básicos usados en esta investigación.

2.1.1. Realidad virtual

Para definir a la realidad virtual se puede partir inicialmente de los conceptos de "realidad" (algo que existe independientemente de las ideas que le conciernen) y de "virtual" (ser en esencia o efecto, pero no en hecho) [17], con lo que el término "realidad virtual" sugiere una realidad que no existe de manera física. La realidad virtual se centra en usar computadoras para crear imágenes de escenas 3D con las que se pueda navegar e interactuar [18], y en donde usualmente se busca una inmersión.

La inmersión se define como la ilusión de estar sumergido en un entorno virtual; esta inmersión se realiza por medio de un aislamiento sensorial en el entorno virtual a través de una configuración de hardware y software [19]. Esta privación de los sentidos puede causar cansancio o incomodidad, por lo que algunas aplicaciones pueden hacer un mejor uso de las ventajas de la realidad virtual si utilizan técnicas no inmersivas [18].

Un punto importante a tener en cuenta es la capacidad del usuario de navegar dentro del entorno virtual; al tratarse de una interacción abstraída del mundo real, la navegación puede tomar varias formas, por lo que Bowman et al. [20] propusieron una taxonomía para clasificar las distintas técnicas de control del movimiento, y que se muestra en la Figura 2.1. Una técnica de control se forma al tomar un método de cada una de las tres ramas de esta taxonomía y combinarlos.

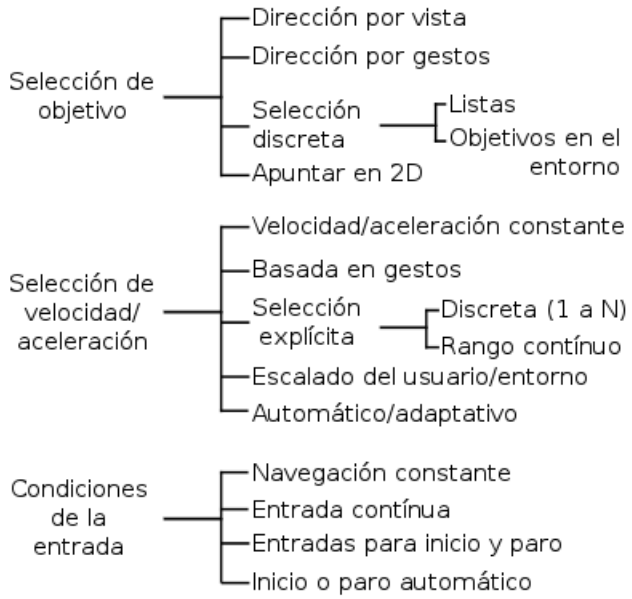


Figura 2.1: Taxonomía de técnicas del control de movimiento de Bowman et al. [20].

Un entorno virtual es un espacio imaginario manifestado en un medio, y que puede ser transmitido de tal forma que pueda ser compartido con otros. Al visualizar este entorno a través de un sistema que presenta los objetos y simulaciones del entorno de una manera interactiva, se está experimentando por medio de realidad virtual [17]. Existen diversas técnicas para crear entornos virtuales a partir de entornos del mundo real, y las fundamentadas en fotogrametría son las más prolíferas actualmente [3].

2.1.2. Fotogrametría y geometría proyectiva

La fotogrametría es la ciencia de obtener información confiable, medible e interpretable sobre las propiedades de superficies y objetos sin tener contacto con ellos [21]. Implica que una imagen está relacionada con el entorno que retrata, y las mediciones hechas en la imagen se pueden traducir a mediciones hechas en el entorno, por lo que se puede proceder a la generación de un modelo tridimensional del entorno retratado, que consiste en la representación matemática de una superficie u objeto tridimensional, este proceso se denomina reconstrucción. Hartley y Zisserman detallan esta relación en su libro *Multiple view geometry in computer vision* [16]; en este escrito se exponen los conocimientos generales necesarios para entender el contexto de la fotogrametría.

La creación de una imagen es el paso de un mundo tridimensional a una imagen bidimensional, es un proceso de proyección en el que se pierde una dimensión. Este proceso usualmente se modela como una proyección central, donde un rayo parte de un punto en el espacio tridimensional y se traslada a través de un punto fijo en el espacio denominado centro de proyección o de cámara para luego intersectar un plano específico en el espacio escogido como el plano de la imagen; se muestra gráficamente en la Figura 2.2.

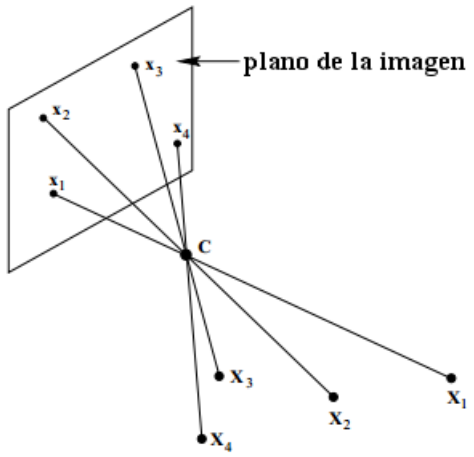


Figura 2.2: Visualización del proceso de proyección de 3 dimensiones a 2 dimensiones [16].

El modelo de proyección central más común es el de cámara pinhole [22] que se muestra en la Figura 2.3. Bajo este modelo un punto en el espacio $\mathbf{X} = (X, Y, Z)^T$ es

mapeado al plano de la imagen a través de la línea que une al punto **X** con el centro de la cámara. Por triángulos similares se obtienen las coordenadas en el plano de la imagen de la proyección de **X**:

$$(X, Y, Z)^T \mapsto (f\frac{X}{Z}, f\frac{Y}{Z}) \tag{2.1}$$

Donde f es la distancia o longitud focal de la cámara. Este modelo es expandible con el uso de coordenadas homogéneas:

$$\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} fX \\ fY \\ Z \end{pmatrix} = \begin{bmatrix} f & & & 0 \\ & f & & 0 \\ & & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{2.2}$$

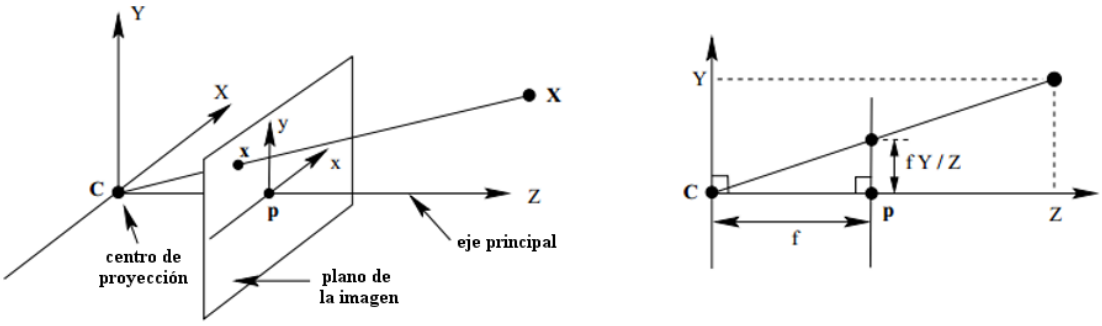


Figura 2.3: Modelo simple de la cámara pinhole [16].

Con esta representación se puede escribir de manera compacta la relación entre posiciones en el entorno y posiciones en la imagen de la siguiente manera:

$$x = P\mathbf{X} \tag{2.3}$$

Con P llamada la matriz de proyección de la cámara. Esta representación permite añadir consideraciones de las características de la cámara modelada (parámetros intrínsecos) y de su relación con el entorno (parámetros extrínsecos), que se muestran en la Figura 2.4.

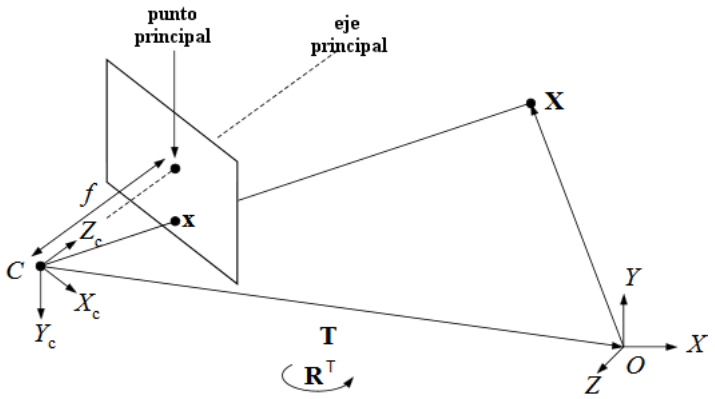


Figura 2.4: Modelo general de la cámara pinhole [22].

Las consideraciones agregadas se dividen de la siguiente manera:

$$P \propto KR\left[I\middle|-\tilde{C}\right] \tag{2.4}$$

Donde K es una matriz de 3×3 con los parámetros intrínsecos de la cámara y es llamada la matriz de calibración de la cámara. R es una matriz de 3×3 con la rotación de la cámara en las coordenadas del entorno y $\tilde{C}$ representa las coordenadas del centro de la cámara en las coordenadas del entorno. Los parámetros intrínsecos de la cámara son [22]:

- **Desplazamiento del punto principal**. Es el desplazamiento que se debe hacer dentro del plano de la cámara para llegar al punto principal como se observa en la Figura 2.5.
- **Factores de escala horizontal y vertical diferentes**. Los sensores CCD, comunes en las cámaras actuales; pueden tener una escala horizontal distinta a la vertical por diversas razones que varían desde fallas en el sensor o en el post-procesamiento de la imagen, hasta distorsiones causadas por el lente de la cámara [23]. Esta posibilidad implica que se debe tomar en cuenta cada escala al realizar mediciones en píxeles.
- **Sesgo**. El sesgo ocurre cuando los ejes x y y no son perpendiculares en el sensor CCD, o se está tomando la fotografía de una fotografía [16].

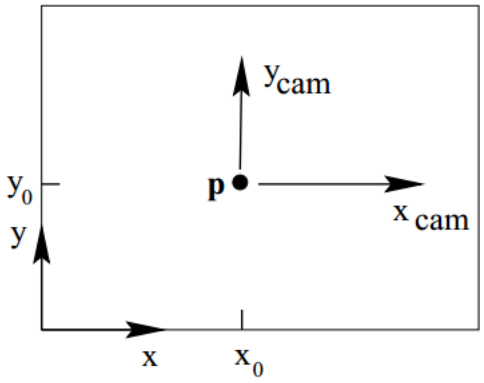


Figura 2.5: Desplazamiento del punto principal para pasar de coordenadas del plano a coordenadas de la cámara [16].

Con estas consideraciones se tiene la matriz K definida de la siguiente manera:

$$K = \begin{bmatrix} \alpha_x & s & x_0 \\ & \alpha_y & y_0 \\ & & 1 \end{bmatrix} \tag{2.5}$$

Donde $\alpha_x = fm_x$ y $\alpha_y = fm_y$ representan la distancia focal de la cámara en términos de las dimensiones de píxeles para modelar los factores de escala, $x_0 = m_x p_x$ y $y_0 = m_y p_y$ posicionan al punto principal, y s determina el sesgo en la imagen.

2.1.3. Geometría epipolar

Revisando la Figura 2.3 se nota una ambigüedad en la proyección, que todos los puntos a lo largo del rayo proyectado del punto 3D al centro de la cámara terminan proyectando en el mismo punto en el plano de la imagen. Una manera intuitiva de reducir esta ambigüedad es utilizando más cámaras durante la reconstrucción.

Al tratar con dos o más cámaras observando una escena, se requiere el conocimiento de la relación que existe entre estas cámaras, que se identifica como la transformación (i.e. traslación y rotación) entre una cámara y otra.

Esta relación se modela por medio de la geometría epipolar, el modelo se visualiza en

Figura 2.6: Visualización de la geometría epipolar, se genera un plano entre los centros de cámara y el punto de referencia.

la Figura 2.6. Los centros de cámara son unidos con una línea (denominada *baseline*) y las intersecciones de esta línea con los planos de imagen se identifican como los epípolos, todos los rayos proyectados de puntos 3D al centro de cámara de una imagen se proyectan en la otra imagen como líneas epipolares que tienen al epípolo como punto en común. En la Figura 2.7 se muestra un ejemplo de estas líneas epipolares, nótese que los epípolos se encuentran más allá de las imágenes.

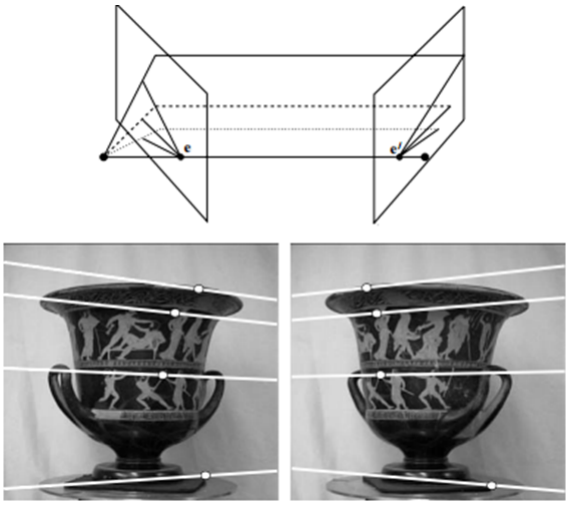


Figura 2.7: Ejemplo de líneas epipolares, las líneas visualizadas son las proyecciones de los rayos entre el punto 3D y el centro de la otra cámara [16].

Nuevamente usando álgebra lineal se modela la relación entre los puntos 2D correspondientes entre cada imagen a través de una transformación que es llamada *matriz*

Esencial (E) de acuerdo a [22]:

$$xEx' = 0 \tag{2.6}$$

Donde E implica la rotación y traslación de puntos de una imagen hacia sus puntos correspondientes en la otra imagen:

$$E \propto [T]_x R \tag{2.7}$$

Aquí R es la matriz de rotación y $[T]_x$ es la matriz de producto cruzado de la traslación:

$$[T]_x = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \tag{2.8}$$

La matriz Esencial solamente toma en cuenta el marco de referencia de cada cámara para las transformaciones, falta tomar en cuenta los parámetros intrínsecos de las cámaras (i.e. la matriz K). Ésto se hace con la siguiente ecuación:

$$u^T F u' = 0 \tag{2.9}$$

Donde u y u' son posiciones en el plano de la imagen como x, pero ahora medidas en píxeles. Finalmente, la matriz que realiza esta transformación de puntos de una imagen a puntos de la otra imagen se denomina *matriz Fundamental* y se relaciona a la matriz Esencial de la siguiente manera.

$$F \propto K^{-1T} E K^{-1} \tag{2.10}$$

La ecuación 2.9 y la proporcionalidad 2.10 son importantes porque ya se encuentran en medidas de píxeles, lo que permite el cálculo de los valores de la matriz Fundamental usando puntos que sabemos son correspondientes entre las imágenes, como se observan

los puntos señalados en la Figura 2.7, y luego usar las matrices F y K para calcular E y obtener las transformaciones necesarias para calcular P, y así tener tanto el modelo de geometría epipolar entre dos cámaras como el modelo de proyección de cada cámara para realizar la reconstrucción. En la Figura 2.8 se visualiza el comportamiento de cada una de estas tres matrices. La matriz Proyectiva mapea de puntos 3D a puntos 2D en la imagen, mientas que las otras dos mapean entre puntos 2D de dos imágenes con la Fundamental usando píxeles como unidad de medida.

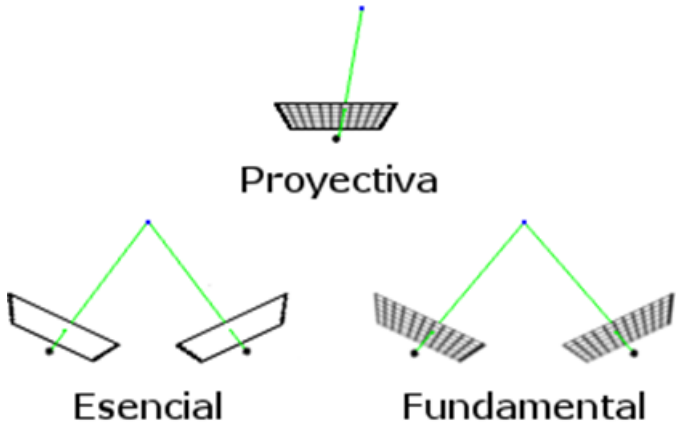


Figura 2.8: Visualización del comportamiento de las matrices expuestas.

2.1.4. Reconstrucción por fotogrametría

Contando con estos modelos se puede proceder a la reconstrucción tridimensional de un entorno a partir de fotografías. Para la reconstrucción generalmente se toman en cuenta dos problemas integrales [1]:

- Conocer la configuración de las cámaras que tomaron las fotografías en el espacio tridimensional.
- Inferir la geometría del objeto que están visualizando las cámaras tal que dicha geometría concuerde con lo que se visualiza en las fotografías y por ende con el objeto visualizado.

Se puede añadir un tercer problema con fines prácticos que es el de convertir la geometría inferida en una estructura de datos diferente para su uso; un ejemplo de ello, interesante para los fines de este trabajo, es su transformación en una malla de polígonos de la superficie reconstruida para su visualización e inclusión en un entorno virtual; a este problema se le denomina reconstrucción de superficie y se trata en la Sección 2.1.4.3.

Este método es el más común en el área de reconstrucción tridimensional. Un primer punto de vista de estos acercamientos se encuentra en las herramientas de reconstrucción. Como nota inicial cabe destacar que este tipo de reconstrucción generalmente sigue un flujo de trabajo, expuesto en [1], como el siguiente:

1. **Structure from Motion (SfM):** Obtener las posiciones 3D de las imágenes y puntos de la estructura visualizada.
2. **Multi-View Stereo (MVS):** Generar una geometría de la escena congruente con lo que visualizan las imágenes.
3. **Reconstrucción de Superficies:** Convertir la geometría en una malla utilizable para otros fines como la visualización o la interacción. Este paso es parte de MVS, pero en ocasiones es omitido de acuerdo a los requerimientos del problema.

En las siguientes subsecciones se detallarán elementos de estos métodos, donde las Figuras 2.12, 2.14 y 2.17 fueron generadas utilizando herramientas de reconstrucción, expuestas en el texto, sobre imágenes obtenidas de un video de la Pirámide de la Luna de Teotihuacán.

2.1.4.1. Configuración de las cámaras

Teniendo las cámaras modeladas matemáticamente como se presentó en la Sección 2.1.2, se procede a realizar el proceso denominado *Structure from Motion* (SfM), que es un acercamiento a la automatización de la reconstrucción. SfM es el proceso de estimar la localización de puntos 3D a partir de múltiples imágenes dado un conjunto disperso de correspondencias entre puntos característicos reconocibles entre distintas imágenes [24].

Usualmente se usa *Scale-Invariant Feature Transform* (SIFT) propuesta por Lowe [25] por la robustez que ofrece al momento de encontrar correspondencias que pueden ser asignables a distintas vistas (imágenes), en la Figura 2.9 se observan los puntos característicos encontrados por el algoritmo. Su funcionamiento se conforma de las siguientes fases:

1. **Detección de puntos extremos en el espacio de la escala** en todas las escalas y todas las ubicaciones de la imagen.
2. **Localización de puntos característicos** a partir de un modelo ajustado a cada punto extremo para determinar su ubicación y escala, seleccionando los puntos de acuerdo a su estabilidad.
3. **Asignación de orientación** u orientaciones de acuerdo a las direcciones de los gradientes locales de cada punto característico.
4. **Cálculo de descriptores de puntos característicos** a partir de los gradientes medidos en el vecindario de cada punto característico a la escala seleccionada.

Figura 2.9: Puntos característicos encontrados por SIFT, los puntos encontrados son invariantes a la escala de la imagen [25].

Un punto a tomar en cuenta en este proceso es que SfM otorga un estimado de la configuración de las cámaras y las posiciones de los puntos 3D, denotado por inconsistencias entre la reproyección de los puntos 2D y los puntos 3D que representan; estas inconsistencias, llamadas errores de reproyección, se presentan de la manera mostrada por la Figura 2.10 y representan la distancia entre el punto 2D característico (u_i en la Figura 2.10) y el punto 2D reproyectado desde la cámara de acuerdo a la estructura

estimada ($\hat{u}_i$ en la Figura 2.10). Para resolver este problema se emplea el algoritmo *Bundle Adjustment* (BA) que se emplea para la estimación conjunta de parámetros.

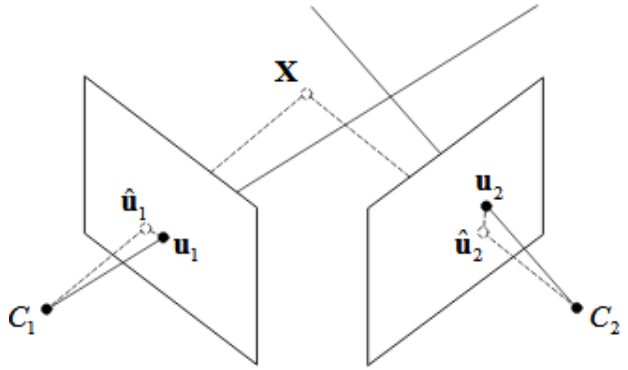


Figura 2.10: Visualización del error de reproyección, la distancia entre la medición u_i y la predicción $\hat{u}_i$ [22].

BA modela el problema como conjuntos de rayos de luz que se salen de cada característica de la estructura y convergen en cada cámara, los cuáles son ajustados conjuntamente con respecto a las características de la estructura (posición de las características en el espacio) y las configuraciones de las cámaras; con cada ajuste se actualizan tanto la estructura como las posiciones de las cámaras [15] de acuerdo al error 2.11.

$$\Delta x_{ip}(C_c, P_i, X_p) \equiv \underline{x}_{ip} - x(C_c, P_i, X_p) \tag{2.11}$$

Donde $X_p, parap = 1...n$ representa los n puntos 3D identificados y proyectados a las m imágenes por $P_i, i = 0...m$ y con otros posibles parámetros de calibración $C_c, c = 1...k$ tal que se tiene un modelo predictivo $x_{ip} = x(C_c, P_i, X_p)$; contando con mediciones inciertas $\underline{x}_{ip}$ de un subconjunto de los puntos característicos en las imágenes se obtiene la medición de error que se busca minimizar para refinar tanto la estructura como las configuraciones de las cámaras.

El algoritmo SfM sigue el funcionamiento ilustrado por Schönberger et al. en [26] se puede observar en la Figura 2.11. Inicialmente se buscan puntos característicos en las imágenes, que son emparejadas de acuerdo a la proximidad de los descriptores de estos puntos. Las parejas formadas, usualmente modeladas como un grafo de la escena, son introducidas, incrementalmente, en el modelo reconstruido por medio de la

triangulación de puntos, filtrado de outliers y refinado de la reconstrucción usando BA.

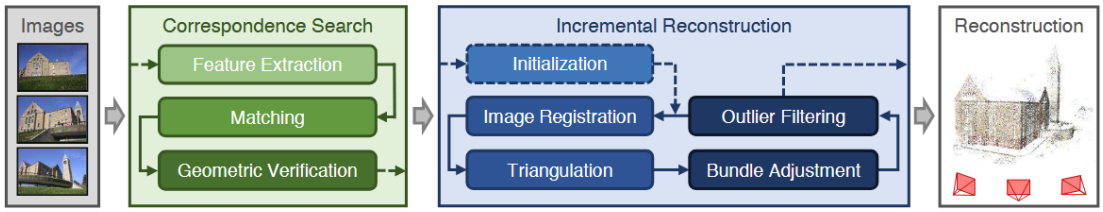


Figura 2.11: Funcionamiento usual del algoritmo Structure from Motion [26].

SfM entrega como resultado las posiciones de las cámaras en el espacio tridimensional del entorno reconstruido y el modelo geométrico del entorno en la forma de una nube de puntos, un conjunto de puntos orientados que no presenta ninguna cara o superficie. Este modelo puede cambiar en topología y densidad conforme va evolucionando durante el proceso [24].

En la Figura 2.12 se muestra el resultado del algoritmo SfM propuesto por Wu en 2013 [27], ofrecido en la aplicación VisualSfM [28], aplicado a un conjunto de fotografías de la pirámide de la Luna en Teotihuacán. La nube de puntos representa la estructura visualizada, mientras que los triángulos que se observan en la parte superior son los frustums (i.e. campos de visión) de las cámaras posicionados de acuerdo al resultado de SfM.

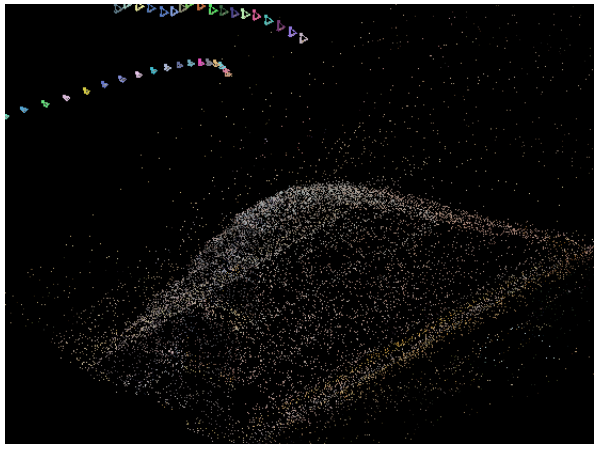


Figura 2.12: Nube de puntos resultante de aplicar la variante del algoritmo SfM propuesta por Wu [27, 28].

Cabe destacar que la matriz Fundamental (F), que transforma puntos de una imagen a otra, puede ser estimada usando 7 correspondencias [16]; las correspondencias encontradas por métodos como SIFT cuentan con grados variados de precisión, además de incluir correspondencias erróneas causadas por el cambio de punto de vista de la escena [25]. Estas características causan la necesidad de algoritmos eficientes de estimación de F. La estimación automática de la matriz usualmente ser realiza por medio de RANSAC (*RAndom SAmple Consensus*) [16]. RANSAC es un paradigma usado para ajustar un modelo a datos experimentales cuando se espera un porcentaje significativo de datos erróneos [29]. La estimación de la matriz Fundamental de acuerdo a RANSAC se realiza de la siguiente manera

1. De las correspondencias, seleccionar un conjunto (correspondencias putativas) basado en la proximidad y similaridad de las intensidades de sus vecindarios.
2. Repetir con N muestras
   a) Seleccionar aleatoriamente una muestra de 7 correspondencias y calcular F (de acuerdo al algoritmo del cálculo se pueden obtener una o tres posibles matrices).
   b) Calcular la distancia d de cada correspondencia putativa.
   c) Calcular el número de correspondencias consistentes con F dado $d < t$ píxeles.
   d) Si hay tres posibles F, seleccionar la que tenga mayor número de correspondencias consistentes.
3. Escoger F con el mayor número de correspondencias consistentes; si hay empate, elgir la de menor desviación estándar.

Para calcular la matriz fundamental a partir de 7 puntos se parte de la ecuación $x_i'^T F x_i = 0$, que se convierte en un conjunto de ecuaciones de la forma $Af = 0$. La solución de las ecuaciones $Af = 0$ tiene la forma $\alpha F_1 + (1-\alpha) F_2$ donde α es una variable escalar y las matrices F_1 y F_2 se obtienen de los generadores f_1 y f_2 del espacio nulo a la derecha de A. Añadiendo la restricción $detF = 0$ se obtiene $det(\alpha F_1 + (1-\alpha) F_2) = 0$ que es un polinomio cúbico de α, por lo que se tienen tres soluciones reales o una real y dos complejas que son descartadas.

2.1.4.2. Geometría del objeto

La nube de puntos proporcionada por SfM puede ser suficiente para algunas aplicaciones, pero si se desea reconstruir con mayor fidelidad el objeto o entorno, entonces se requieren de otros métodos. Uno de estos métodos es *Multi-View Stereo* (MVS), que es un grupo de técnicas que usan la correspondencia entre más de dos imágenes como pista para extraer la geometría del objeto o entorno que se busca reconstruir [1].

MVS emplea distintas técnicas dependiendo de los objetivos de la reconstrucción; la Figura 2.13, tomada del tutorial de Furukawa y Hernández [1] muestra un diagrama de flujo con la aplicación de distintos conjuntos de técnicas para obtener un modelo tridimensional de un objeto. El diagrama de flujo muestra las posibles etapas y algoritmos utilizados durante el proceso MVS.

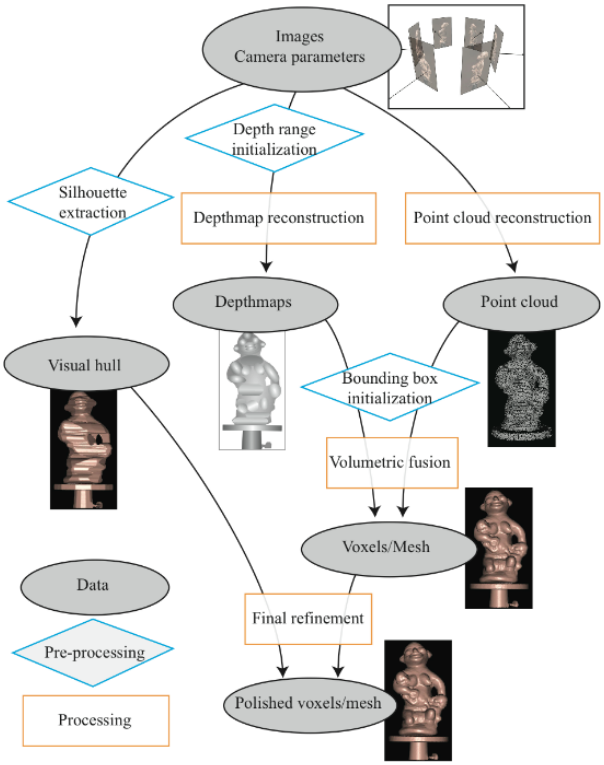


Figura 2.13: Diagrama de flujo de las diversas técnicas que pueden ser aplicadas al emplear MVS [1].

Es importante tomar en cuenta los datos de entrada, las imágenes y las configuraciones de las cámaras; por lo general es necesario un proceso inicial de SfM para conseguir las configuraciones de las cámaras. De la misma manera, cabe destacar que el diagrama de flujo no implica que todos los algoritmos sean obligatorios en cada reconstrucción por MVS; dependiendo de los objetivos que se tengan de la reconstrucción, puede ser necesario sólo un subconjunto de los algoritmos. Junto con el diagrama, Furukawa y Hernández ofrecen ejemplos de aplicaciones y sus necesidades: el renderizado desde un punto de vista libre puede funcionar con nubes de puntos y, en cambio, la representación de escenas se ve ayudada por una malla poligonal de alta calidad.

En la Figura 2.14, se presenta el resultado de aplicar el acercamiento a MVS, llamado PMVS, propuesto por Furukawa et al. en 2010, implementado por ellos y modificado por Pierre Moulon para funcionar en VisualSfM, el cual se basa en la consistencia entre regiones de las imágenes para poblar con un conjunto denso de puntos el área correspondiente y obtener la nube de puntos [30].

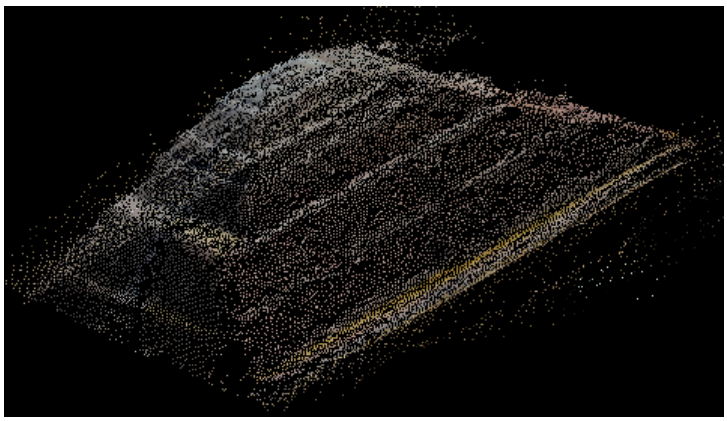


Figura 2.14: Resultado de aplicar MVS [27, 30] sobre el resultado visualizado en la Figura 2.12. Se observa una reconstrucción densa.

2.1.4.3. Reconstrucción de superficie

Al contar con una nube de puntos densa, resultado de MVS, para facilitar la visualización de esta reconstrucción se puede generar la superficie dictada a partir de la nube de puntos, como proponen Kazhdan et al. [31], que llaman reconstrucción Poisson de superficies. El algoritmo es ilustrado en la Figura 2.15.

Parten del modelado de una función indicadora que se aplica sobre el espacio tridimensional y tiene valor 1 en el volumen interno de la superficie y 0 en su exterior. De esta manera, el gradiente de la función indicadora tiene valor distinto de 0 sólo cerca de la superficie, donde su valor es el de la normal interna de la superficie. La nube de puntos se puede considerar, entonces, como un conjunto de muestras de ese gradiente.

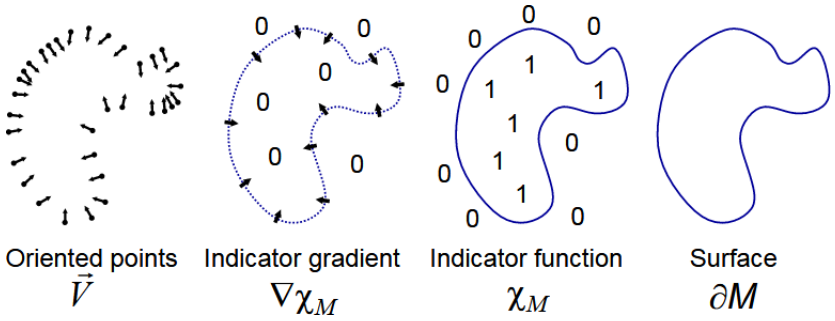


Figura 2.15: Ilustración del algoritmo de reconstrucción de superficies de Poisson [31].

Para conseguir la función indicadora χ a partir de la nube de puntos, se busca la mejor aproximación del gradiente de χ al campo vectorial $\vec{V}$ muestreado por la nube de puntos.

$$min_{\chi}\|\nabla\chi - \vec{V}\| \tag{2.12}$$

Aplicando el operador de divergencia, el problema se convierte en calcular la función escalar χ cuyo Laplaciano (divergencia del gradiente, $\nabla \cdot \nabla$) es igual a la divergencia del campo vectorial.

$$\Delta\chi \equiv \nabla \cdot \nabla\chi = \nabla \cdot \vec{V} \tag{2.13}$$

Para finalizar la reconstrucción se usa el proceso llamado "Cubos Marchantes", se divide el volumen en cubos de los que se obtienen triángulos que serán parte de la superficie; los triángulos adquieren distintas formas dependiendo de qué vértices del cubo se encuentran dentro o fuera de la superficie de acuerdo a la función indicadora, en la Figura 2.16 se observan los triángulos generados, los vértices que se encuentran dentro de la superficie se encuentran marcados con un punto.

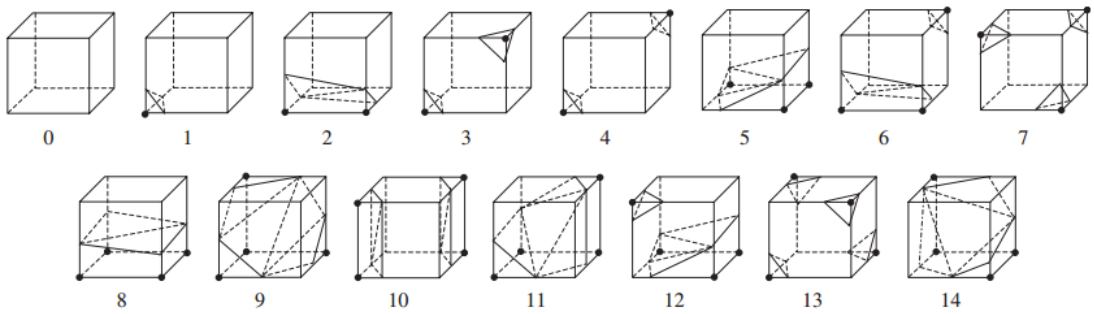


Figura 2.16: Configuraciones básicas de los cubos y sus triángulos generados para Cubos Marchantes [32].

En la Figura 2.17 se muestra una reconstrucción por este método. Se tomó como entrada la nube de puntos de la Figura 2.14 y se realizó un pre-proceso manual de filtrado de los puntos innecesarios en la reconstrucción para el buen funcionamiento del algoritmo, se usó la implementación del algoritmo disponible en la aplicación MeshLab de Cignoni et al. [33].

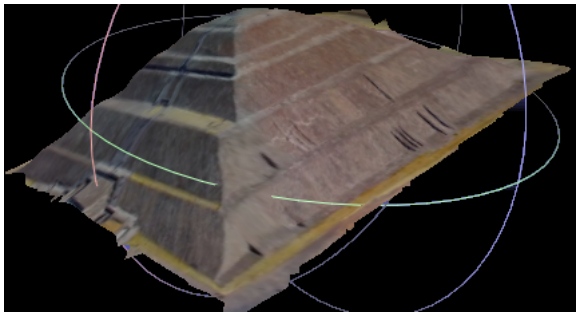


Figura 2.17: Reconstrucción de la superficie a partir de la nube de puntos de la Figura 2.14 usando la reconstrucción de superficies Poisson [31, 33].

2.2. Estado del arte

En los años recientes se ha tenido un avance considerable en la tarea de reconstrucción tridimensional de un entorno urbano. En esta sección se presenta una revisión general

de los enfoques actuales para resolver este problema haciendo énfasis en la información de entrada que reciben y el tipo de resultado que generan. Además se hace referencia a las revisiones del estado del arte realizadas con anterioridad, a fin de tener una visión global de los diferentes enfoques que se tienen hoy en día para resolver el problema de reconstrucción tridimensional.

Varios laboratorios, entre ellos el grupo de Visión por computadora y Geometría de ETH Zürich [34] y el área de Gráficos y Visión en el Departamento de Ciencia de la Computación e Ingeniería de la Universidad de Washington en San Luis [35], abordan el problema de la reconstrucción tridimensional, buscando mejorar el rendimiento en cuanto a la complejidad de los algoritmos [27], así como la precisión geométrica de la reconstrucción [36].

2.2.1. Revisión de trabajos previos

La reconstrucción urbana ha tenido un continuo avance y ha sido tratada por numerosos investigadores de diferentes áreas. Una característica que los diferencia son los distintos datos de entrada. En este trabajo, de acuerdo al tipo de dato de entrada, se clasifican de la siguiente manera:

- Conjunto de imágenes (multi-view).
- Una imagen (single-view).
- Imágenes panorámicas.
- Imágenes RGB-D.
- Plano arquitectónico.
- Información inercial.

Esta diversidad en los datos de entrada obedece a las necesidades y restricciones del problema que cada propuesta busca tratar, además de la manera más conveniente que los investigadores consideran de abordar el problema. Así mismo, las soluciones y productos finales igualmente varían de acuerdo al acercamiento y pueden ser clasificadas en:

- Navegación.
- Modelado tridimensional.
- Modelado con segmentación semántica.
- Representación de espacio libre.

Usando estas clasificaciones se presentan los artículos en el estado del arte organizados tal que sus fuentes y fines sean similares.

2.2.1.1. Multi-view

Es el método más común en el área de reconstrucción tridimensional. Un primer punto de vista de estos acercamientos se encuentra en las herramientas de reconstrucción. Como nota inicial cabe destacar que este tipo de reconstrucción generalmente sigue un flujo de trabajo, expuesto en [1] y expuesto en la Sección 2.1.4. Varios investigadores ofrecen herramientas para la reconstrucción genérica con el fin de acelerar el trabajo de otros.

Wu presentó una herramienta de reconstrucción a partir de fotografías en 2013 [27] que implementa el método de Bundle Adjustment presentado en [28] para SfM; además está diseñada para usar una implementación de MVS desarrollada por [30], tal que su resultado final es una nube de puntos densa como reconstrucción.

Fuhrmann et al. en 2014 [37] ofrecen a la comunidad un software de reconstrucción geométrica multi-view. Recibe fotografías de una escena como entrada y realiza la reconstrucción MVS de acuerdo a la metodología de reconstrucción multi-view de Goesele et al. [38], donde genera mapas de profundidad de cada una de ellas con respecto a sus vecinos, para luego fusionarlos en una nube de puntos de toda la escena. Realizan la reconstrucción de superficie de acuerdo al algoritmo *Floating Scale Surface Reconstruction* (FSSR) [39], un acercamiento virtualmente libre de parámetros y capaz de lidiar con muestras a distintas escalas en un mismo modelo. Este flujo de trabajo se muestra en la Figura 2.18.

Schönberger et al. en 2016 [26, 40] presentan y ofrecen un sistema multi-view stereo para modelado denso y robusto a partir de colecciones no estructuradas de fotografías,

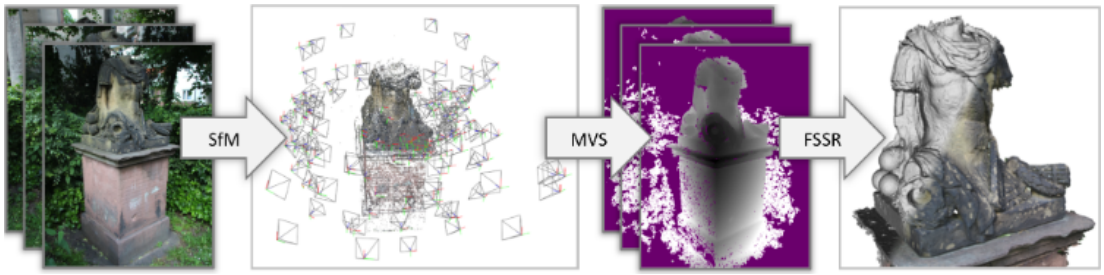


Figura 2.18: Flujo de trabajo del software Multi-View Environment ofrecido por Fuhrmann et al. [37].

con estimación conjunta de información de profundidad y normales, selección de vistas a nivel de pixel, y un término de consistencia geométrica para el refinado.

Estas herramientas se pueden comparar de acuerdo a los algoritmos que implementan y sus resultados. Existen benchmarks como el ofrecido por Seitz et al. [36] para objetos, y el ofrecido por Strecha et al. [41] para escenas en exteriores; estos benchmarks permiten una comparación cuantitativa de los resultados.

Por otro lado, utilizar presuposiciones mejora el rendimiento en situaciones específicas. Furukawa et al. en 2009 [42] definen Manhattan-world stereo, donde todas las superficies están alineadas con las tres direcciones $\vec{i}, \vec{j}, \vec{k}$ definidas por Coughlan et al. [43]. Agregar estas presunciones mejora el rendimiento de los acercamientos de reconstrucción en entornos urbanos, gracias a la proliferación de elementos planares en las construcciones humanas.

Holzmann et al. [44] en 2017 proponen un método de reconstrucción tridimensional híbrida; combinan la reconstrucción volumétrica y plane fitting, son capaces de reducir el ruido y compactar al modelo resultante. Inicialmente, ellos detectan planos usando RANSAC y los usan para identificar puntos en el modelo; tras proyectar los puntos 3D a sus planos respectivos, los tetraedros son creados usando triangulación Delaunay etiquetados usando corte de grafos con una minimización de energía que toma en cuenta la visibilidad, adherencia a las presuposiciones de mundo Manhattan y el nivel de detalle deseado. Habiendo probado su acercamiento en su pipeline SfM, demostraron que es más flexible que métodos anteriores y con rendimiento comparable.

Yang et al. [45] en 2016 proponen un acercamiento divide-y-vencerás y fusión a la reconstrucción de entornos a gran escala; dividen la escena en múltiples subsecciones

superpuestas, asegurándose de que las superposiciones contienen lo que ellos llaman "imágenes ancla" pertenecientes a ambas subsecciones. Ellos proponen un esquema RANSAC para coser subsecciones adyacentes, y otro esquema RANSAC para mejorar el cierre de bucle de las subsecciones en el modelo global. Sus experimentos demuestran una mejora en la eficiencia de tiempo sobre los acercamientos no-distributivos.

Locher et al. [46] en 2018 presentaron un pipeline para SfM progresivo, diferenciado de los pipelines incrementales y globales tradicionales. El pipeline evita mínimos locales y la dependencia en el orden de las imágenes de entrada, ambos problemas de los pipelines incrementales; el pipeline también puede trabajar con un flujo de imágenes de entrada y entregar resultados intermedios, imposible para los pipelines globales. Su acercamiento fue diseñado para escenarios multi-usuario con múltiples flujos de entrada y sin restricción al orden de las imágenes.

Alcantarilla et al. en 2013 [47] proponen un método de reconstrucción densa a gran escala de entornos. Partiendo de cámaras calibradas en estéreo y cuyo movimiento es conocido, generan mapas de disparidad, los que fusionan en un modelo global usando asociaciones basadas en consistencia geométrica y fotométrica. Finalmente realizan un filtrado para lidiar con los requerimientos de almacenamiento de modelos a gran escala. Su acercamiento descarta automáticamente obstáculos en movimiento en la escena, lo que brinda robustez en reconstrucciones estáticas.

Shan et al. [48] en 2014 emplean contornos ocluyentes para identificar espacio libre en una escena y usan esta información como una nueva restricción para la reconstrucción de superficies, mejorando la calidad de las superficies y los contornos reconstruidos. Su acercamiento hace uso de una nube de puntos semi-densa proporcionada por PMVS [30], la cual usan para interpolar mapas de profundidad y luego generar un volumen de espacio libre, cortando puntos 3D espurios en el proceso. Ellos extienden el algoritmo Screened Poisson Surface Reconstruction [49] con una restricción del espacio libre para la reconstrucción.

Fouhey et al. en 2014 [50] exponen un método que explota la presencia de personas en las imágenes. Infieren pistas de la escena (i.e. espacio para sentarse, o espacio para caminar) a partir de las acciones de las personas. Con las pistas mapeadas en la escena, extraen restricciones geométricas y funcionales. La Figura 2.19 muestra los pasos de su flujo de trabajo; comienzan detectando a las personas y sus acciones en la escena, estiman la funcionalidad de secciones de la escena y generan hipótesis de la geometría de la escena, finalmente, combinan las hipótesis y las acciones detectadas para seleccionar la hipótesis e inferir el espacio ocupado.

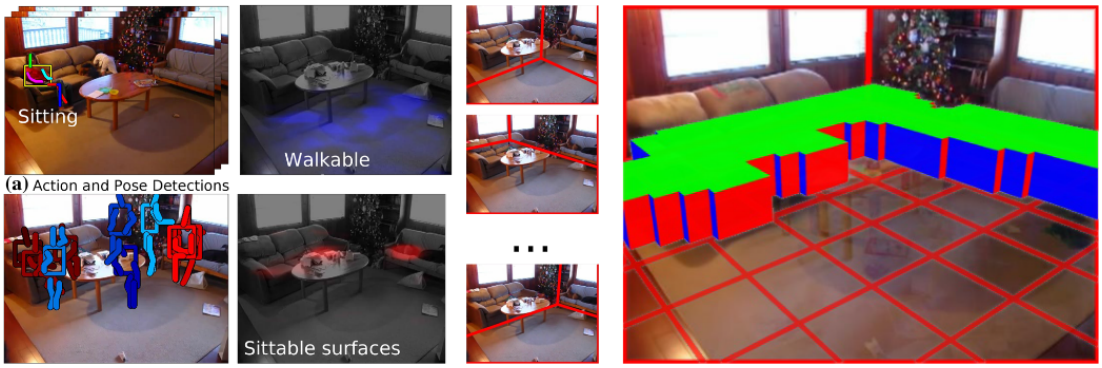


Figura 2.19: Flujo de trabajo de la propuesta de Fouhey et al. [50].

Dong et al. en 2015 [5] construyen un sistema de apoyo a la navegación en interiores a partir de datos recabados del público. Toman como entrada fotografías 2D de las que generan una malla de navegación de los modelos 3D generados. Igualmente que Martin-Brualla et al. [51], integran el flujo recopilado de las personas; pero aquí para apoyar en la generación de indicaciones de navegación.

Lafarge y Mallet [52] en 2012 introducen un método para reconstruir una escena urbana segmentada semánticamente, diferenciando entre piso, edificio, vegetación y obstrucciones. Su acercamiento toma una nube de puntos como entrada, obtenida por láser o por MVS, y calcula las características (i.e. no-planaridad, elevación, dispersión y agrupamiento) para ayudarse durante la clasificación. Se extraen primitivas geométricas (i.e. cilindros, esferas, planos) de los puntos de "edificio;" finalmente, a través de un proceso híbrido de pareo de modelos y generación de mallas, se recupera un modelo 3D de la escena. Demuestran una alta generalización y una reconstrucción compacta de los edificios, además de una descripción semántica de la escena.

Lafarge et al. [53] en 2013 proponen un método híbrido para la reconstrucción 3D, ellos combinan primitivas geométricas con parches de mallas 3D para obtener una escena compacta y preservar detalles; las primitivas cubren estructuras regulares mientras que las mallas cubren las irregularidades en el modelo. Su hibridación se divide en dos etapas: la primera segmenta una malla inicial del modelo completo de acuerdo a su curvatura local; durante la segunda etapa toman muestras, en la forma de primitivas y parches, de la malla segmentada usando un algoritmo basado en Jump Diffusion.

Un factor que raramente se toma en cuenta es el efecto del tiempo en una escena. El mundo en que vivimos es dinámico y se ve modificado con el paso del tiempo, ya

sea de manera natural o con intervención humana. Schindler et al. entre 2007 y 2010 [54, 55] proponen técnicas para detectar cambios en el tiempo y ordenar temporalmente conjuntos de imágenes hasta obtener un modelado 4D (3D cambiando en el tiempo) de una ciudad [56].

De una manera específica Matzen y Snavely en 2014 [57] proponen un método para reconocer el momento en que puntos 3D individuales existieron en una escena urbana reconstruida. Usan la afinidad espacial y temporal de estos puntos para segmentar objetos espacio-temporales, en su caso anuncios y señalamientos.

Martin-Brualla et al. en 2015 [58] generan un time-lapse tridimensional de una escena. Obtienen un conjunto de fotografías tomadas de internet, las filtran y tratan de acuerdo a su trabajo en [59]. El time-lapse generado es de toda la escena, sin embargo, cambios demasiado grandes en la escena causan fallas en el proceso.

Radenovic et al. en 2016 [60] presentaron un acercamiento que toma en cuenta los cambios de iluminación presentados en una escena entre el día y la noche de tal manera que existe una correspondencia entre ellos. Obtienen dos modelos de la escena que luego fusionan o editan con las características del otro.

Similarmente, incrementar los factores a tomar en cuenta puede mejorar la calidad de la solución obtenida. Un ejemplo, recientemente tratado, es la conjunción de la segmentación semántica con la reconstrucción tridimensional, donde las soluciones parciales de un problema sirven de información adicional al otro problema y viceversa.

Duan y Lafarge [61] proponen en 2016 un pipeline para la reconstrucción de modelos de ciudades a gran escala, compactos y segmentados semánticamente a partir de un par de imágenes satelitales. Su pipeline consiste en dividir ambas imágenes en polígonos convexos, determinar su clase semántica y elevación, y mezclar ambas particiones en una representación planimétrica de la elevación para generar un modelo tridimensional. Su modelo resultante es segmentado en tres clases: tierra, techo y fachadas; el nivel de detalle de su reconstrucción es bajo y los techos más complicados (i.e. domos) son únicamente aproximados.

Sengupta et al. presentan en 2013 [62] un algoritmo que genera una reconstrucción tridimensional con un etiquetado semántico asociado. Toman imágenes en estéreo obtenidas de una cámara montada en un vehículo, con las que generan mapas de profundidad que fusionan a un modelo tridimensional. Las imágenes son etiquetadas, y estas etiquetas se fusionan para anotar el modelo tridimensional. Su implementación se

puede considerar online; el flujo de datos es continuo y el modelo se va construyendo con éste, pero su velocidad de actualización impide la interacción.

Son Häne et al. en 2013 [63] quienes reconocen que la segmentación y la reconstrucción densa 3D se aportan mutuamente información valiosa. Proponen un sistema para resolver de manera conjunta las dos tareas. Para 2017, Häne et al. [64], proponen un marco matemático para formular y resolver de manera conjunta un problema de segmentación y reconstrucción. La clase semántica de una geometría da pistas para su probable superficie; mientras que la superficie provee información de la probabilidad de su clase semántica. Con su formulación logran un modelo cuya segmentación es consistente con las imágenes usadas para la segmentación.

Kundu et al. en 2014 [65] realizan la segmentación y reconstrucción 3D conjuntas a partir de video. Usan los indicadores específicos de semántica donde las restricciones multi-view son débiles y viceversa. Realizan sus pruebas en secuencias de video monocular a gran escala moviéndose hacia adelante.

Vineet et al. en 2015 [66] construyen sobre las técnicas de reconstrucción y fusión semántica a gran escala; logran obtener un rendimiento tal que dicen haber obtenido el primer sistema que puede realizar la reconstrucción semántica, densa y a gran escala de una escena en exteriores. Además de ésto, presentan un acercamiento de fusión semántica que les permite manejar objetos dinámicos de manera más efectiva. El sistema parte de imágenes estéreo para generar los mapas de profundidad, estimar la pose de la cámara, y generar el modelo tridimensional. La segmentación semántica se realiza directamente en tres dimensiones, lo que reduce el costo computacional e impone consistencia temporal. En la Figura 2.20 se muestran sus resultados, donde logran reconstruir la escena y segmentarla semánticamente, distinguiendo entre caminos, pavimento, edificios, etc.

Savinov et al. en 2016 [67] proponen un acercamiento a la reconstrucción semántica 3D usando potenciales sobre rayos de visión, combinados con penalización de área de la superficie continua. Reformulan el problema de reconstrucción semántica tal que garantiza convergencia y un manejo exacto de la visibilidad; éstos siendo problemas identificados en la reconstrucción semántica hasta esos momentos.

Cherabier et al. en 2016 [68] dividen la escena en bloques donde generalmente sólo un subconjunto de etiquetas se encuentra activo; esta táctica reduce el costo en memoria del procesamiento. El conjunto de etiquetas activas se va actualizando durante el proceso iterativo de optimización.

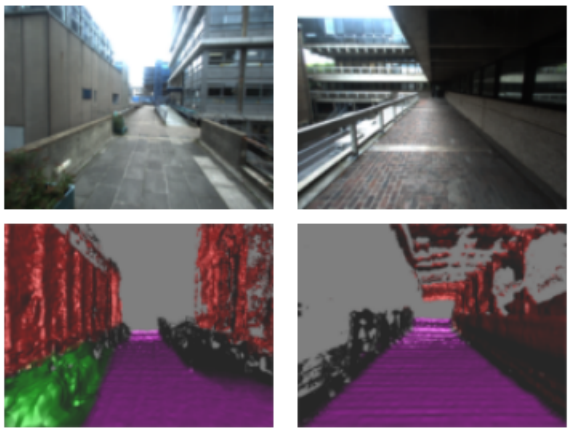


Figura 2.20: Escena reconstruida y segmentada por Vineet et al. [66].

Bláha et al. en 2016 [69] plantean la reconstrucción semántica 3D formulada con resolución múltiple adaptativa para escenas de gran escala. Su planteamiento genera un esquema jerárquico que refina la reconstrucción sólo en regiones que probablemente contengan una superficie.

Bláha et al. en 2017 [70] declaran presentar el primer método de refinamiento de superficies informado por semántica. Maximiza foto-consistencia para preservar detalles finos; explota información semántica para maximizar la consistencia de etiquetas en las imágenes; restringe la reconstrucción con supuestos previos dictados por la etiqueta local.

Un enfoque multi-view, con entradas monoculares o estereoscópicas, permite que múltiples tareas se ejecuten conjuntamente (i.e. segmentaciones semántica, temporal y de iluminación). Fusionar el trabajo de estas tareas y sus restricciones abre la puerta a diferentes acercamientos capaces de generar reconstrucciones más robustas.

2.2.1.2. Single-view

Emplear una imagen, aunque usualmente no permite generar un modelo completo de la escena, sintetiza el problema de reconstrucción a uno de comprensión de la escena, donde se reconoce el espacio libre y los objetos que contiene.

Ladický et al. [71] en 2014, tras haber identificado la codependencia entre la apariencia visual de las clases semánticas y su profundidad geométrica, proponen un clasificador de profundidad semántico; el clasificador es robusto a tendencias en los datos de entrada, relaciona los píxeles a una profundidad canónica y reduce el impacto de la perspectiva en la geometría, a costa de un mayor procesamiento y debilidad a imágenes de baja resolución o clases semánticas de varianza alta.

Hedau et al. en 2009 [72] recuperan el diseño espacial de escenas en interiores a partir de una imagen. Su propuesta presenta robustez contra oclusiones; modelan el espacio global de la habitación con una "caja" 3D paramétrica, localizando objetos de manera iterativa y ajustando la caja. La Figura 2.21 muestra un ejemplo de esta caja, se observa el diseño de caja y las etiquetas de superficies donde localizan objetos visibles de la escena.

Figura 2.21: Caja 3D paramétrica propuesta por Hedau et al. [37].

Para 2012, ellos mismos [73] logran recuperar el espacio libre de una escena en interiores a partir de una imagen. Entre sus presuposiciones se encuentra la estructura geométrica casi de caja de los muebles y las restricciones proveídas por la escena.

Gupta et al. en 2010 [74] modelan la interacción 3D entre espacio libre y objetos. Representan paramétricamente los objetos en 3D, y los usan para incorporar restricciones volumétricas a la escena.

La popularidad actual de las redes neuronales las ha traído al área de la reconstrucción como una nueva forma de inferir la profundidad en una vista y estimar el layout de la escena visualizada.

Eigen et al. [75] en 2014 consiguen generar un mapa de profundidad a partir de una

imagen usando dos redes neuronales; la primera red se enfoca en predecir profundidades toscas de la escena de manera global y, después, la segunda red refina las profundidades a través de múltiples convoluciones tomando en cuenta objetos en la escena y los bordes de las paredes. Ellos también proponen una métrica de error invariante a la escala que les ayuda a eliminar la ambigüedad inherente de las profundidades predichas.

En 2015 Eigen y Fergus [76] logran extraer más información de la escena de tres formas: profundidad, normales de la superficie y etiquetas semánticas. Su arquitectura es una red neuronal profunda multi-escala de tres escalas: la primera escala predice características para la imagen entera; la segunda escala hace predicciones a media resolución, entregando un número de canales dependiente de la tarea a realizar; la tercera escala trae las predicciones a una resolución mayor para obtener salidas detalladas que son también coherentes en el espacio. Aplican la misma arquitectura para las tres tareas cambiando la función de pérdida usada y la estructura de la salida.

Lee et al. [77] en 2017 desarrollaron una arquitectura de red, de inicio a fin, que estima el layout de una habitación. Ven a la tarea como una tarea de localización de puntos clave, lo que evade ambigüedades en el etiquetado de paredes encontrado usualmente en los clasificadores a pixel. Ellos configuran una red neuronal convolucional para trabajar en conjunto con la regresión de los puntos y la clasificación del tipo de layout. La arquitectura tiene un mejor rendimiento que propuestas previas en cuanto a los errores en píxeles y puntos clave además de la velocidad de procesamiento.

Liu et al. [78] construyeron en 2018 una red neuronal profunda para reconstruir un mapa de profundidad a partir de una imagen RGB; su mapa de profundidad es planar a partes y ellos llaman a su arquitectura la primera con dicho objetivo. Ellos subdividen el problema en tres subproblemas (i.e. predecir los parámetros de los planos, generar un mapa de profundidad estándar para estructuras no planares y generar máscaras de segmentación para cada plano predicho y para el mapa de profundidad no planar). Sus predicciones permiten hasta K planos (10 para sus experimentos), controlando el número de planos al poner en 0 sus máscaras de segmentación.

Las reconstrucciones single-view son útiles para el manejo de escenas urbanas, ya sea de interiores o exteriores, gracias a conocimiento a priori como las restricciones Manhattan. Los elementos identificados en una escena, en conjunto con este conocimiento, ayudan a establecer restricciones volumétricas para la reconstrucción de la escena.

2.2.1.3. Imágenes panorámicas

El empleo de imágenes panorámicas se presta a la reconstrucción de entornos en interiores; se tiene información del cuarto entero sin la necesidad de alinear varias imágenes, esta tarea es complicada por la falta de texturas en las paredes y la gran cantidad de objetos ocluyentes.

Cabral y Furukawa en 2014 [79] crean un sistema para reconstruir planos arquitectónicos compactos a partir de imágenes panorámicas de interiores. Proyectan a un plano los puntos 3D y las cámaras, y dividiendo el plano en celdas obtienen: la evidencia de espacio libre de acuerdo a la visibilidad de los puntos 3D; la evidencia de pared de acuerdo a la cantidad de puntos 3D en la celda. Generan un plano como el camino más corto entre celdas que contenga el espacio libre; en la Figura 2.22, se presenta el modelo resultante de este acercamiento.

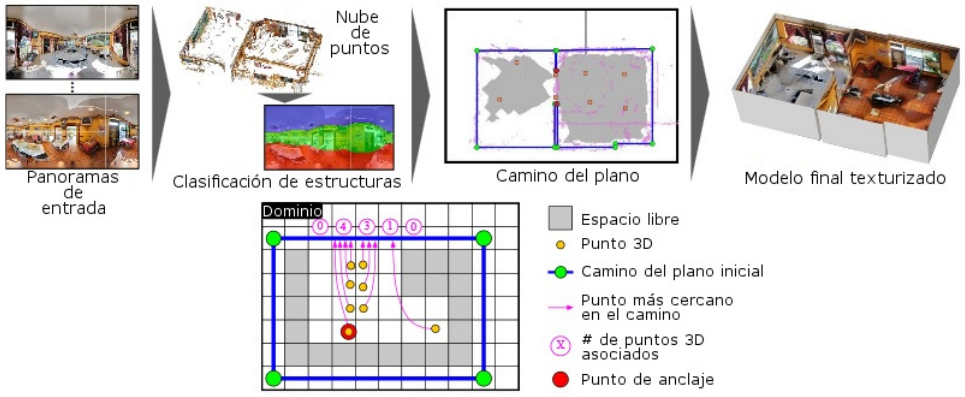
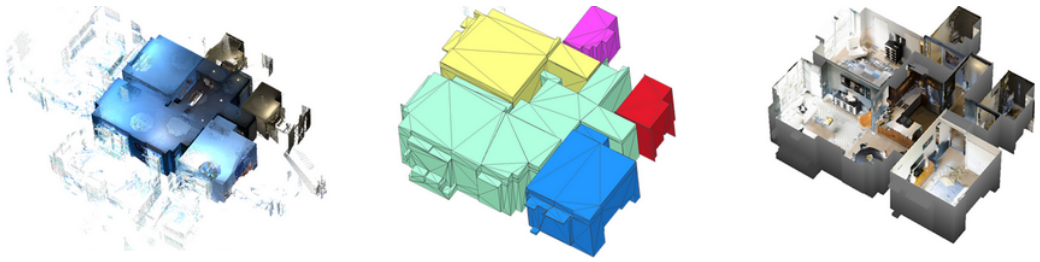


Figura 2.22: Visualización del algoritmo del acercamiento de Cabral et al. [79].

Ikehata et al. en 2015 [80] reconstruyen una escena como un modelo, estructurado como grafo de acuerdo a una gramática que ellos mismos proponen, a partir de imágenes panorámicas RGB-D. Toman la propuesta de Cabral y Furukawa [79] como guía y obtienen las evidencias de espacio libre. La escena es representada como un grafo, donde los nodos corresponden a elementos estructurales como cuartos, paredes y objetos. En la Figura 2.23 se muestran sus modelos finales.

Yang y Zhang en 2016 [81] proponen un método para recuperar la forma de una habitación 3D de un panorama de vista completa. Infieren la forma 3D a partir de facetas de superpíxeles parcialmente orientados y segmentos de línea. El peso de su



(a) Nube de puntos de entrada. **(b)** Mallado resultante. **(c)** Ejemplo de renderizado.

Figura 2.23: Modelo de interiores obtenido por Ikehata et al. [80].

propuesta radica en la generación de un grafo de restricción; usa los superpíxeles y las líneas como vértices, y su relación geométrica como ejes.

Pintore et al. en 2016 [82] capturan, reconstruyen y exploran estructuras de interiores con múltiples cuartos, comenzando a partir de imágenes panorámicas. Se enfocan en mapear la estructura para navegación, dejando de lado los detalles 3D.

Ikehata et al. en 2016 [83] proponen la reconstrucción de un cuarto a partir de un conjunto de imágenes tomadas para formar una imagen panorámica. Su enfoque permite al usuario tomar las fotografías sin la necesidad de realizar un barrido donde la cámara sólo rote sin trasladarse. Utilizan relaciones de coplanaridad entre líneas de la escena para apoyar al proceso de SfM.

Zou et al. [84] proponen en 2018 un algoritmo para la reconstrucción del layout de una escena en interiores a partir de una imagen panorama. Su algoritmo opera en tres pasos: alinear la imagen al piso, identificar las esquinas y los límites de las paredes en mapas, y realizar una regresión del layout 3D obtenido de los mismos mapas. Ellos prueban que su sistema se compara favorablemente al estado del arte tanto en velocidad como en precisión. La característica que exaltan es la capacidad de su algoritmo de modelar habitaciones no cuboides, con lo que relajan un limitante común para la reconstrucción de escenas en interiores.

Cada uno de los acercamientos revisados lidia con un problema en particular; el primordial siendo determinar el diseño de la escena, identificando el espacio libre para generar reconstrucciones compactas.

2.2.1.4. Imágenes RGB-D

Las imágenes RGB-D ofrecen información extra tridimensional reconocida como la profundidad, o distancia de cada pixel a las cámaras. Identificando la posición de las imágenes en la escena, es posible realizar la reconstrucción con la fusión de los valores de profundidad de cada imagen.

Silberman et al. [85] en 2012 presentan un clasificador de píxeles sobre una imagen RGB-D capaz de identificar los elementos estructurales de una habitación (i.e. piso, estructuras permanentes, mobiliario y accesorios). Ellos usan esta clasificación para mejorar la segmentación y la identificación de la estructura de soporte de un objeto. Tras alinear los puntos 3D de entrada a las direcciones Manhattan usando líneas rectas de la imagen, generan planos potenciales usando RANSAC y determinan correspondencias pixel-plano a través de corte de grafos con expansión alfa.

Nießner et al. presentan en 2013 [86] un sistema en línea para la reconstrucción a pequeña y gran escala; la reconstrucción resuelve el problema básico de obtener mapas de profundidad, de sensores RGB-D, y fusionarlos incrementalmente en un modelo 3D. Su contribución se enfoca en proponer el uso de una estructura de datos eficiente tanto en velocidad como en memoria, que llaman voxel hashing.

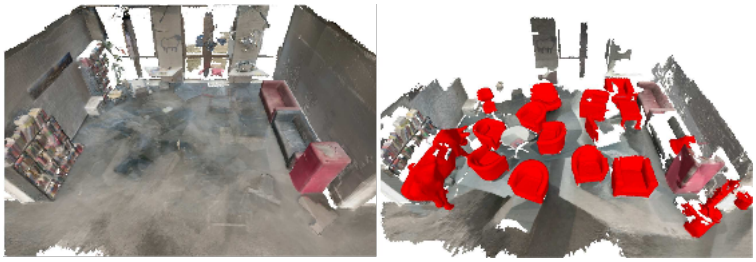
Su aportación mejora el uso de memoria y la complejidad del procesamiento, con lo que se liberan recursos para otros procesamientos, como demuestran Cavallari y Di Stefano [87]. Presentan una fusión del etiquetado semántico de las imágenes con el modelo tridimensional reconstruido; el voxel hashing libera suficientes recursos y tiempo para que el proceso extra de etiquetado semántico de las imágenes, que es un proceso costoso, se pueda llevar a cabo y sea fusionado con el modelo reconstruido, manteniendo un funcionamiento online.

Choi et al. en 2015 [88] realizan la reconstrucción de interiores a partir de video RGB-D. Registran geométricamente fragmentos de la escena con una optimización global. Su método de optimización deshabilita los alineamientos erróneos incluso cuando superan numéricamente a los correctos.

Fehr et al. [89] en 2017 presentan un algoritmo planteado como una extensión de una Truncated Signed Distance Function (TSDF). Dividen al entorno tridimensional en voxels que contienen el valor TSDF, que indica la distancia del voxel a una superficie de la escena. Su algoritmo permite un refinado constante del mapa estático, así como la reconstrucción de objetos dinámicos 3D. El mapa estático se elabora con base en la

primera observación del entorno; las observaciones subsecuentes se usan para refinar este mapa y encontrar los objetos dinámicos, que son identificados como clústers y etiquetados como dinámicos. Finalmente, va agregando los objetos dinámicos a una base de datos, y mejora sus modelos con cada instancia encontrada de los mismos. El modelo final que obtienen es la reconstrucción de la escena estática y los objetos dinámicos descubiertos y se muestra en la Figura 2.24.

(a) Escena reconstruida.



(b) Parte estática. **(c)** Elementos dinámicos.

Figura 2.24: Reconocimiento de elementos dinámicos en una escena hecho por Fehr et al. [89].

Whelan et al. [90] en 2016 presentan un sistema para la reconstrucción tridimensional usando visual Simultaneous Localization And Mapping con funcionalidad online. Ellos logran convertir una entrada RGB-D en un mapa consistente de entornos a escala de una habitación, y detectar fuentes de luz discretas. Su acercamiento usa la arquitectura típica para visual SLAM denso, alternando entre mapeo y localización; mantienen un área activa del modelo que representa donde esperan estar trabajando, y en cada toma tratan de registrar una porción del modelo activo al modelo inactivo para generar un bucle.

La propuesta de Dai et al. en 2017 [91] es un sistema de reconstrucción 3D en línea con una estimación de pose optimizada local y globalmente sobre un video RGB-

D. Ofrece un rendimiento mayor al estado del arte de reconstrucciones online y una calidad a la par con métodos offline. Un punto importante a tomar en cuenta es que sólo necesita de la información RGB-D; ésto a cambio de un mayor costo en el procesamiento, lo que requiere de un equipo de cómputo apropiado para el procesamiento.

Liu et al. [92] proponen en 2018 un framework para la reconstrucción de planos arquitectónicos llamado FloorNet. El framework trabaja con un video RGB-D con posiciones conocidas de las cámaras y utiliza tres redes neuronales profundas, ya existentes en el estado del arte, en una arquitectura híbrida de tres ramas intercomunicadas. Logran convertir el video en la representación de un plano arquitectónico que habían propuesto anteriormente en [93]. Una rama encuentra características en la nube de puntos, la segunda encuentra características en la imagen; ambas ramas entregan estas características a la tercera rama, la cual produce la geometría y características semánticas del plano arquitectónico.

Este tipo de entrada es efectivo para tareas de reconstrucción gracias a la información de profundidad que ofrecen, su utilidad es evidente en reconstrucciones de interiores. Se puede seleccionar este tipo de entrada por múltiples razones, desde la optimización de recursos hasta el refinamiento de una reconstrucción.

2.2.1.5. Plano arquitectónico

Al reconstruir escenas hechas por el hombre, generalmente se cuenta con un plano arquitectónico o mapa de la escena. Esta información permite obtener un modelo fiel de la escena, o una comprensión de la misma y los elementos que contiene, conforme se requiera.

Martin-Brualla et al. en 2014 [51] analizan mapas anotados de un sitio, en conjunto con fotografías de internet, para reconstruir interiores de sitios turísticos. Utilizan posición, orientación e indicaciones de forma extraídos del mapa. Optimizan un objetivo global para recuperar el layout de las piezas reconstruidas. Analizan el flujo recopilado de las personas en el sitio, y recuperan la orientación de las geometrías 3D reconstruidas.

Liu et al. en 2015 [94] generan el recorrido virtual 3D de un departamento a partir de un conjunto de imágenes monoculares de diferentes cuartos y un plano arquitectónico. Utilizan el plano arquitectónico para restringir el diseño de cada cuarto, tal que se simplifica la alineación de las imágenes dentro del modelo, como se muestra en la Figura

2.25, también se observan las imágenes usadas para texturizar las paredes reconstruidas.

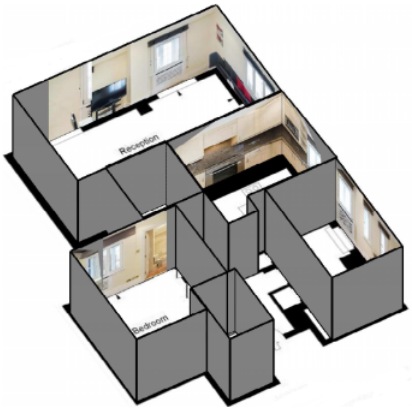


Figura 2.25: Reconstrucción realizada por Liu et al. [94].

Chu et al. en 2016 [11] usan anuncios publicitarios y Google Street View para generar un modelo 3D del exterior del edificio texturizado. Parametrizan el interior del edificio de acuerdo a su plano arquitectónico, que también obtienen del anuncio publicitario; en la Figura 1.1b se muestra un resultado de su acercamiento.

Wang et al. en 2015 [95] explotan información a priori geográfica para la comprensión de escenas en exteriores usando OpenStreetMaps. Razonan de manera conjunta la detección de objetos 3D, estimación de posiciones, segmentación semántica y reconstrucción de profundidad en una imagen.

Liu et al. [93] generaron en 2017, una representación vectorial de una imagen rasterizada de un plano. Ellos usan una red neuronal convolucional para extraer información geométrica y semántica de la imagen, la información se extrae en la forma de uniones de paredes y etiquetas semánticas por pixel; usando Integer Programming, integran esta información de bajo nivel en primitivas, que usan para generar el plano vectorizado en un post-proceso donde los errores de alineamiento son corregidos y las habitaciones son creadas de acuerdo a la etiqueta asignada.

Wijmans y Furukawa [96] exploraron en 2017 el alineamiento de imágenes panorámicas RGB-D en un plano arquitectónico 2D; modelan el problema como un Markov Random Field donde minimizan tres potenciales: consistencia escaneo-plano, aplicada a un escaneo con una posición específica en el plano y medida usando pistas semánticas y geométricas; consistencia escaneo-escaneo, donde se miden las consistencias geométrica

y fotométrica entre todos los escaneos dada su posición; cobertura del plano, que mide el número de píxeles cubiertos por todos los escaneos en su configuración actual.

Estos acercamientos se enfocan en fusionar información para obtener una reconstrucción fiel de la escena. Una característica importante a tomar en cuenta es que algunos de estos acercamientos logran una reconstrucción completa de una escena con interior y exterior.

2.2.1.6. Información inercial

El uso de información de movimiento generalmente implica el procesamiento en dispositivos móviles. Cada observación que se realiza con un sensor, visual o inercial, tiene su propia frecuencia de muestreo, por lo que es necesario realizar una fusión de sensores para obtener las imágenes posicionadas en tres dimensiones.

Tanskanen et al. en 2013 [97] presentan un pipeline completo de reconstrucción 3D para dispositivos móviles monoculares; ofrecen retroalimentación para el usuario en tiempo real, con un retardo de unos segundos. Ocupa la cámara y los sensores inerciales del teléfono (acelerómetro, giroscopio).

Tomando ese pipeline como guía, Kolev et al. en 2014 [98] exponen un esquema para la integración de mediciones de profundidad basadas en estéreo. A cada profundidad se le asigna un peso con base en confianza de acuerdo a la orientación geométrica local, configuración de la cámara y evidencia fotométrica. Integran esta propuesta al pipeline de Tanskanen et al. [97] mejorando su precisión a cambio de una carga de un segundo extra en la reconstrucción.

Dryanovski et al. exploran en 2017 [99], con un trabajo preliminar presentado en 2015 [100], el problema de la reconstrucción a gran escala en tiempo real en dispositivos móviles. Sus mayores desafíos siendo el ruido y la baja frecuencia de los datos de profundidad, además de la limitante de recursos computacionales. Usando únicamente el CPU del dispositivo logran generar una malla del modelo tridimensional; lo logran dividiendo la escena en volúmenes que discriminan de acuerdo a la visiblidad actual, para así poder enfocar los recursos de procesamiento y la memoria. En la Figura 2.26 se muestra una de sus reconstrucciones de una escena en exteriores con la trayectoria del dispositivo; también se muestra la trayectoria recuperada del dispositivo al momento de la captura.

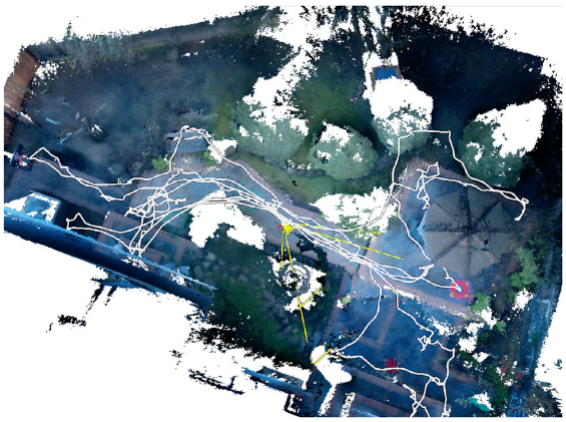


Figura 2.26: Reconstrucción obtenida por Dryanovski et al. [99].

Por otro lado, Schöps et al. entre 2015 y 2017 [101, 102] se enfocan en la reconstrucción de escenas con un dispositivo móvil con sensores inerciales, usando la versión preliminar el trabajo de Dryanovski et al. [100]; logran una reconstrucción en tiempo real. En contraste con la reconstrucción de objetos pequeños; en escenarios grandes las mediciones de espacio libre son menos efectivas para suprimir outliers.

Usando únicamente información inercial, Yan et al. [6] logran proponer la doble integración de los datos de una unidad de medición inercial en un celular, estimando la trayectoria de una persona. Para lograrlo hacen uso de un número de suposiciones; la suposición principal es la suposición de que el sujeto está caminando en una superficie plana; también dictan cuatro posibles configuraciones del celular (i.e. en la pierna, en una bolsa, en la mano, pegado al cuerpo). Usando estas suposiciones, proponen regresiones específicas a cada opción, y logran producir trayectorias similares a las de su referencia obtenida en un teléfono Google Project Tango.

Este conjunto de entradas por lo general ayuda a obtener un funcionamiento online en la reconstrucción. La fusión de información, no necesariamente sincronizada, de la escena, en conjunto con las restricciones de procesamiento inherentes a dispositivos móviles, vuelven a la eficiencia de los algoritmos una característica fundamental para este tipo de reconstrucción.

2.2.1.7. Trabajos complementarios

Los avances logrados hasta el momento comienzan a ser incorporados en herramientas para un público tanto general como especializado (algunos ejemplos de estas herramientas son ReCap [14], Pix4D [13], VisualSFM [27, 28]), pero hasta ahora estos acercamientos se enfocan principalmente en la reconstrucción, ya sea de un objeto o escena (en exteriores o interiores) y no en la generación de la escena completa con interior y exterior, ésto por la dificultad de encontrar correspondencias entre imágenes del exterior y del interior. A continuación, se presentan trabajos que proponen y emplean diversas técnicas para subsanar este problema en distintas situaciones.

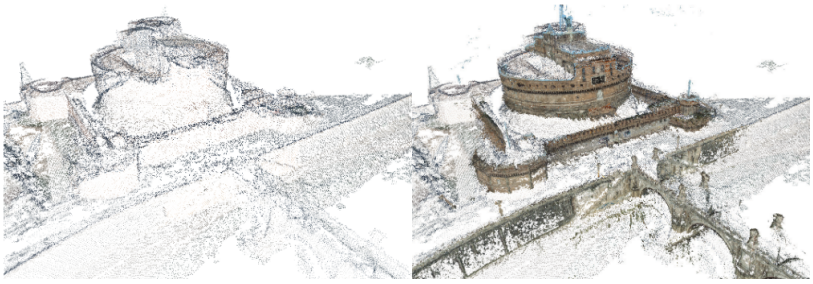
Strecha et al. publicaron en 2014 un reporte técnico [103] donde lograron la integración de la reconstrucción aérea y de interiores del castillo Chillon en Suiza combinando tecnología de posicionamiento global y taquimetría con imágenes tomadas con cámaras comerciales, en la Figura 2.27 se muestra una sección vertical de su reconstrucción. Cabe mencionar que el reporte técnico fue realizado a nombre de la empresa Pix4D [13], de la que Strecha es fundador, y se presentó como un *tour de force* para demostrar la flexibilidad y capacidades de la fotogrametría.

Figura 2.27: Sección vertical con interior y exterior de la reconstrucción realizada por Strecha et al. [103].

En 2014, [104] Xiao et al. realizaron la reconstrucción de museos a partir de una nube de puntos obtenida de escáneres láser, la convierten en primitivas bidimensionales (líneas y rectángulos) que se apilan para formar primitivas tridimensionales (cuboides) que delimitan el volumen del entorno; el perímetro de esa estructura se convierte en

un modelo de paredes que carga por medio de un plugin de Google Earth para su visualización en el mapa.

Shan et al. [105] abordaron en 2014 el problema de alineación de modelos reconstruidos con puntos de referencia diferentes, que son uno terrestre y uno aéreo. Su acercamiento permite alinear un modelo denso terrestre de un punto prominente con un modelo disperso de todo el entorno, siendo el alineamiento afectado en menor medida que acercamientos anteriores gracias al empleo de transformaciones dependientes del punto de vista. La Figura 2.28 muestra un ejemplo del alineamiento de las nubes de puntos.



(a) Nube obtenida de manera aérea.

(b) Nube obtenida de manera terrestre ya alineada.

Figura 2.28: Alineación de dos nubes de puntos de [105].

Es hasta recientemente que se ha empezado a abordar la generación de un modelo completo por medio de la alineación de los modelos del exterior y el interior. Por una parte, se encuentra el acercamiento de Koch et al. [106] que usa la reconstrucción por líneas en 3D de Hofer y Maurer [107] para convertir los modelos tridimensionales de SfM de nubes de puntos a conjuntos de líneas; a partir de las líneas genera hipótesis de los planos que contienen la fachada interior o exterior en los modelos y con estas hipótesis realiza el pareo entre las hipótesis del interior y del exterior para encontrar las posibles posiciones en donde se alinearían los modelos.

Cohen et al., en 2016 [108] utiliza semántica para realizar el alineado de ventanas, usando un clasificador por-pixel y estima las posiciones 3D de las ventanas a partir de los puntos SfM. Con las ventanas identificadas en 3D se registran modelos disjuntos con base en las correspondencias entre ventanas, generando un modelo por cada correspondencia uno-a-uno entre las ventanas y explorando todas las posibles combinaciones calificándolas de acuerdo al número de ventanas que no tienen correspondencia en el

modelo; en caso de ambigüedad defiere al usuario para seleccionar la transformación adecuada entre las mejores obtenidas por el método.

Esta revisión trae a la luz una clara necesidad de organizar los trabajos existentes para facilitar las comparaciones y colaboraciones entre investigadores; a continuación, se presentan las clasificaciones actuales ofrecidas para este fin.

2.2.2. Clasificaciones previas

En el estado del arte se encuentran revisiones que ofrecen clasificaciones de los tipos de reconstrucción para organizar los acercamientos a la reconstrucción tridimensional. Este tipo de trabajos es claramente una necesidad debido a la evidente interdisciplinariedad que se ha detectado en el área, y que corre el riesgo de que los avances en una disciplina no trasciendan a otras de manera oportuna.

En 2013 Musialski et al. [3] presentaron una clasificación de la reconstrucción urbana a partir de imágenes y datos de LiDAR (Light Detection And Ranging). Su clasificación se enfoca en los resultados esperados de los métodos, desde las reconstrucciones más generales fundamentadas en fotogrametría y nubes de puntos y llegando a reconstrucciones masivas de cuadras y ciudades; su clasificación a grosso modo es la siguiente:

- *Nubes de puntos y cámaras*: Sistemas estéreo basados en imágenes; se encuentra en un estado maduro, como se ve en [37], y usualmente se usa como un paso inicial para otros acercamientos.
- *Edificios y semántica*: Uso de modelos parametrizados para aprovechar características encontradas en las construcciones humanas, un ejemplo siendo la tendencia a la planaridad [42].
- *Fachadas*: Extracción y representación de fachadas [109], una parte importante del modelado de áreas urbanas.
- *Cuadras y ciudades*: Reconstrucción a gran escala a partir de múltiples tipos de datos de entrada, como son imágenes aéreas oblicuas y nadir (i.e. en un ángulo recto con la tierra) [69].

Esta clasificación hace un gran trabajo mapeando el amplio espectro de la reconstrucción urbana; cumple una función didáctica explicando conceptos importantes del

área de una manera constructiva incremental. Su artículo es muy recomendable para nuevos investigadores, sin embargo, el área es actualmente muy dinámica con avances innovadores y nuevos acercamientos. Estos cambios invitan a realizar una actualización o bien proponer una nueva taxonomía que los tome en cuenta.

Un problema importante que no se aborda es la reconstrucción de interiores; Chen et al. estudiaron este tema en 2015 [110]. En su estudio exponen los datasets de reconstrucciones de interiores introducidos en el estado del arte de los acercamientos que utilizan imágenes RGB-D. Tomando en cuenta este tipo de entrada, proponen su clasificación en técnicas de modelado geométrico y modelado semántico, descritas brevemente a continuación:

- *Modelado geométrico*: Digitalizar la figura de objetos 3D es un problema subdividido en la fase de registro, donde se agregan las imágenes RGB-D a un mismo marco de referencia y la fase de fusión, donde se unen las imágenes para obtener un modelo 3D.
- *Modelado semántico*: Uso de conocimiento a priori de la escena (i.e. planaridad y simetría) para inferir información faltante y etiquetar elementos de la escena para aplicaciones de nivel más alto. Subdividen este tipo de modelado en dos acercamientos: los basados en primitivas donde se construyen los elementos de la escena agregando primitivas (i.e. esferas y cilindros), y los basados en modelos donde se tienen modelos 3D de objetos , como lámparas y sillas, que son transformados para acoplarse a la escena.

La peculiaridad que salta a la vista al revisar el tema de reconstrucción de interiores es que la gran mayoría de las propuestas existentes usan imágenes RGB-D o imágenes panorámicas de un cuarto para modelarlo [79, 81, 91].

Por su parte, Berger et al. en 2016 [111] se enfocan en estudiar la reconstrucción de superficies a partir de nubes de puntos, dejando la obtención de dicha nube de puntos fuera de su enfoque. Recopilan las imperfecciones presentes en las nubes de puntos, causadas por ruido o falta de información; estas imperfecciones ocurren a causa de las tecnologías de recopilación usadas, que usualmente dictan también el tipo de superficie que se busca adquirir. Tomando en cuenta estas características de las nubes de puntos, se ocupan presuposiciones (i.e. suavidad de la superficie o visibilidad) para obtener una salida de manera robusta a alguna imperfección. Aunado a estas características presentan diversos criterios de evaluación de las superficies reconstruidas, los cuales son descritos a continuación:

- *Precisión geométrica*: Comparar directamente la geometría reconstruida con mediciones de la superficie reconstruida (*ground truth*).
- *Precisión topológica*: Evaluar la recuperación de información de alto nivel de la estructura como es el genus o la topología de su esqueleto; esta evaluación generalmente es cualitativa.
- *Recuperación de la estructura*: Obtener entidades geométricas, relaciones (i.e. paralelismo) y regularidades.
- *Reproducibilidad*: Que, con las mismas entradas, el algoritmo presente la misma salida. Con el aumento de complejidad de los algoritmos aumenta la importancia de su reproducibilidad para tener mayor confianza en la estabilidad de los mismos.

Observando a la reconstrucción de edificios desde el ámbito de la ingeniería de edificios, se encuentra el estudio de Gimenez et al. [112]. Ellos estudian los acercamientos a la generación de un modelo tridimensional a partir de datos del sitio o la documentación del edificio; describen la usabilidad y limitaciones de cada dato de entrada de la siguiente manera:

- *Fotografía aérea*: Permite un procesamiento rápido usando poca información a costa de un nivel bajo de detalle.
- *Fotografía terrestre*: Permite procesar fachadas para obtener reconstrucciones de mayor detalle usando procesos más complejos como SfM; el modelo resultante seguirá estando falto de interiores.
- *Escaneo láser 3D*: El acercamiento más confiable para recuperar la geometría de un edificio, a un costo monetario mayor y con significativamente más ruido dependiendo del diseño del edificio. Este acercamiento puede, al contrario de los anteriores, capturar información de interiores.
- *Aplicaciones móviles*: El opuesto al escaneo láser, a un precio bajo se obtiene información aproximada del edificio, en la forma de planos arquitectónicos, que debe considerarse en conjunto con otros acercamientos para brindar utilidad a usuarios profesionales.
- *Bosquejos de arquitectura del edificio*: Los bosquejos usualmente contienen un bajo nivel de detalle dado que se usan sólo para ilustrar conceptos de diseño. Usualmente se procesan como imágenes o a través de un software CAD durante su creación.

- *Planos 2D escaneados*: Se procesan como imágenes y usualmente requieren un pre-procesamiento extenso para eliminar la información innecesaria y extraer información estructurada. Al igual que con los bosquejos, errores en el dibujo pueden causar inconsistencias en la reconstrucción.
- *Planos CAD*: Estos documentos ya se encuentran estructurados y compuestos de primitivas, por lo que su conversión a un modelo tridimensional es casi trivial. Su desventaja es que los edificios viejos nunca existieron como planos CAD, y de hecho un objetivo común es lograr la conversión de imágenes de bocetos o planos a este formato [93].

Se enfocan en la tarea de reconstrucción a partir de dibujos 3D de edificios con el propósito de incentivar la investigación encaminada a generar modelos de información de edificios (BIMs) de edificios ya existentes para facilitar las tareas de renovación; identifican también la naturaleza fragmentada y específica a la aplicación de los acercamientos actuales.

Como una demostración del valor de esta área de investigación en la ciencia aplicada, uno puede observar el trabajo de Volk et al. [4] en 2014; ellos presentan un survey que identifica de manera exhaustiva los problemas encontrados al generar BIMs para edificios existentes; examinan journals de diversas disciplinas como son la ingeniería civil, el sensado remoto, la visión por computadora y la administración de desperdicio. Ellos identifican problemas en la funcionalidad, interoperabilidad, y los aspectos técnicos y organizacionales. Con este estudio se visualiza un ecosistema donde la reconstrucción de edificios sólo es una parte y aún así puede ser usada de manera extensiva.

Beneš et al. [113] publicaron en 2017 un estudio del realismo en el modelado procedural. Aunque su acercamiento se enfocó en la generación procedural de edificios, sus hallazgos pueden ser aplicados en la reconstrucción de edificios cuando menos de manera limitada, lo que nos provee de revelaciones sobre las posibles necesidades asociadas con la calidad de las reconstrucciones. El estudio se enfoca en el rol de los detalles, tanto finos como burdos, en la percepción de realismo, para identificar sus factores contribuyentes. Ellos identificaron los factores que influyen en el realismo tras entrevistar a los participantes de su estudio; a continuación, se muestran los factores aplicables a la reconstrucción encontrados en su lista:

- *Detalles pequeños*: Las imperfecciones, como son las fisuras en las paredes, se ven como indicadores de realismo.

- *Textura*: La precisión, resolución y localización de texturas son consideradas factores importantes para determinar si una imagen es de un modelo generado o de un edificio real.
- *Ventanas*: Los autores identificaron que los participantes usualmente confundían las texturas en las ventanas con reflexiones. Por su parte, los participantes insistieron en la importancia de la información presentada por las ventanas: los interiores, la posición de las cortinas y la consistencia de las reflexiones.
- *Luz, color y sombras*: Un color uniforme y sombras nítidas brindan poco realismo a la apariencia de los edificios, al punto de que los participantes pensaron que edificios de este tipo no podrían existir en el mundo real.

En una línea similar, Lafarge [114] presentó en 2015 un artículo que lista la naturaleza del estado del arte de forma general, describiendo retos actuales, tendencias y posibles direcciones futuras de la investigación. Respecto a los retos, identifica como las tareas más importantes la robustez a la adquisición, la calidad de los modelos y la automatización completa. Sobre las tendencias, cada una con diferentes niveles de importancia y desarrollo en las disciplinas, lista la reconstrucción y generación, la adquisición terrestre y aérea, imágenes y láser, geometría y semántica, modelos de forma libre y estructurados, y estrategias locales y globales. Las direcciones nuevas de la investigación que visualiza son la exploración en el espacio de la escala, la funcionalidad, los datos de la comunidad y los entornos urbanos dinámicos.

Se observa que estas revisiones son complementarias, con enfoques específicos que ayudan a comprender distintas partes del problema de reconstrucción 3D. Pero aún no toman en cuenta los acercamientos recientes, ya sea por su fecha de publicación o porque no es parte de su enfoque. Aunque estos acercamientos no cuentan con gran popularidad actualmente, los trabajos que los retoman han demostrado su aplicabilidad y utilidad, por lo que es necesario comenzar a prestarles atención, como lo hemos hecho en esta revisión, y lo haremos en la taxonomía presentada a continuación.

Capítulo 3

Contribuciones

En este capítulo se presenta una nueva taxonomía para la reconstrucción urbana con un énfasis en características, anteriormente obviadas [3], que ahora comienzan a ser abordadas gracias a los avances en el área de reconstrucción. Además, se presenta un pipeline de reconstrucción para un tema poco retomado que es la reconstrucción de un edificio con fachada e interiores.

3.1. Taxonomía propuesta

La taxonomía propuesta sigue en espíritu a la taxonomía de Musialski et al. [3], aunque ve como punto principal de la clasificación el proceso que se lleva a cabo durante la reconstrucción, desde la obtención de los datos hasta la manera en que son usados y su dinamismo temporal.

La taxonomía se muestra en la Figura 3.1. La figura es simétrica, en el nivel más bajo se observa el tipo de reconstrucción obtenido (Genéricas, Edificios y Semántica), mientras que en los niveles superiores se observan las características de los datos de entrada y su uso. Una explicación de nuestra propuesta es la siguiente:

- **Dinámicas** implica que las reconstrucciones de alguna manera son sensibles a cambios ocurridos por variaciones en el tiempo. Martin-Brualla et al. [58] gene-

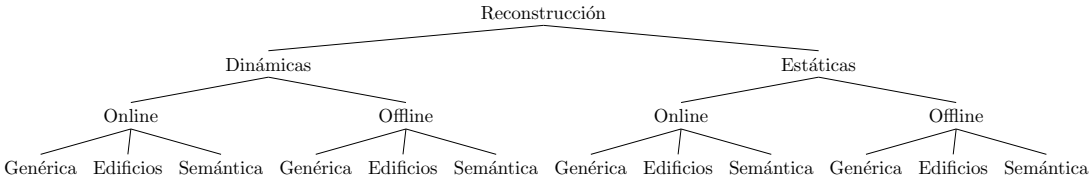


Figura 3.1: Taxonomía propuesta, lidia de manera jerárquica con el tipo de entrada, su manejo y los métodos de reconstrucción usados.

ran time-lapse que ilustran esta característica de dinamismo. Ellos recopilaron imágenes tomadas durante diferentes etapas de la construcción de la torre Goldman Sachs; mediante el tratamiento de estas imágenes, ellos logran obtener un modelo tridimensional que evoluciona en el tiempo de la construcción de la torre.

- **Estáticas** presupone que toda la información a usar es de un momento particular en el tiempo, por lo que se busca ignorar los objetos dinámicos.
- **Online** implica que los datos obtenidos se procesan en tiempo real; permite dar retroalimentación al usuario para obtener un modelo completo. Un ejemplo de esta utilidad la muestran Schertler et al. [115] para la reconstrucción con escáneres láser. En su implementación se presenta el modelo reconstruido en tiempo real; donde se señalan los huecos en el modelo para que el usuario escanee esa parte.
- **Offline** implica que los datos de entrada ya han sido obtenidos de antemano y sólo se tienen que procesar.
- **Genéricas** se refiere a los acercamientos genéricos de MVS (multi-view stereo) que últimamente los investigadores ofrecen como herramientas para acelerar la investigación de otros; toman como enfoque a la sección de nubes de puntos y cámaras de [3].
- **Edificios** engloba los acercamientos de reconstrucción de edificios y semántica, y de fachadas de [3].
- **Semántica** se refiere al proceso en el cual tanto el proceso de reconstrucción tradicional como el de segmentación semántica de la escena contribuyen en el proceso de reconstrucción 3D. Este enfoque puede ser atribuido a Häne y Pollefeys en el 2013 [63, 116].

El uso de esta taxonomía comienza con la primera disyunción donde se decide si se requiere conocimiento dinámico de la escena o si ignorarlo es benéfico. En la segunda

disyunción se escoge la forma en que se alimentará al sistema, si será online u offline; un punto importante para la interacción con el sistema y posiblemente la forma de recopilación de datos. Finalmente se escoge el tipo de procesamiento que se hará a esa entrada; entre un modelado genérico, un modelado restringido por presuposiciones, o un modelado conjuntado con segmentación semántica. A continuación se describe cada nivel de la taxonomía a mayor detalle.

3.1.1. Reconstrucción estática/dinámica

La reconstrucción estática fue una presuposición de los acercamientos iniciales, e incluso actualmente se considera una robustez que la reconstrucción pueda ignorar hasta cierto punto elementos dinámicos de la escena (i.e. personas) [30].

Un ejemplo de la utilidad de discriminar entre los elementos estáticos y dinámicos de la escena, en este caso en la robótica, es el algoritmo de reconstrucción 3D presentado por Fehr et al. [89], donde consiguen la discriminación de la escena estática, y un mejor modelado de los objetos dinámicos tras múltiples observaciones.

Se observa en los acercamientos dinámicos encontrados, que cada propuesta toma en cuenta alguna característica de la escena: cambios en los edificios a lo largo del tiempo; cambios en la iluminación de la escena; cambios paulatinos en una escena (i.e. vegetación). Este nivel de la taxonomía implica una decisión del nivel de interés en los cambios en la escena, si eliminarlos o reconocerlos.

3.1.2. Reconstrucción online/offline

La distinción entre online y offline ocurre sobre la manera en que se alimentan los datos al proceso para la reconstrucción. Una reconstrucción online recibe un flujo de datos que utiliza para una reconstrucción incremental, permitiendo retroalimentación al usuario para mejorar el modelo [115], y que es de vital importancia para tareas en tiempo real como la navegación de vehículos [66] y robots [89]. Una reconstrucción offline utiliza los datos para generar el modelo y se presta al uso de datos recopilados de internet. Un punto importante a tomar en cuenta es que ambos tipos de reconstrucción tienen sus aplicaciones ventajosas, como se observa en las reconstrucciones offline a escala de edificio [38], ciudad [117, 118] o incluso mundial [119] utilizando fotografías recopiladas

de internet. Ocupar imágenes recopiladas evita la necesidad de preparar la escena o el equipo de captura, ésto a cambio de un trabajo extra en la recopilación y discriminación de las imágenes útiles.

Varios de los acercamientos encontrados que recaen en esta clase ocupan tanto información visual como inercial de un dispositivo, generalmente un celular. Contar con menor poder de procesamiento implica que la búsqueda de algoritmos más eficientes se está volviendo un tema de interés; ésto aunado a las nuevas posibilidades abiertas por los dispositivos móviles, además de la entrada al mercado de dispositivos de captura RGB-D costeable para el público en general.

3.1.3. Reconstrucción genérica

Aunque la reconstrucción genérica ya es un tema maduro con avances considerables al punto de que se emplea de manera comercial con productos como ReCap™[14], Agisoft PhotoScan [120] o Pix4D [13], existen incontables problemas abiertos (i.e. la iluminación en exteriores [60, 121]) que permiten una gran variedad de propuestas y herramientas.

Por otro lado, se siguen haciendo propuestas para mejorar aspectos de la reconstrucción genérica.

Romanoni et al. en 2016 [122] proponen incrementalmente inicializar una superficie para la reconstrucción 3D automática a partir de imágenes. Inicializan automáticamente una malla 3D lo más cercana a la solución final a partir de los puntos 3D obtenidos por SfM. Usan lo que llaman *mesh sweeping*, donde generan imágenes de la malla de acuerdo a las cámaras y calculan la consistencia entre píxeles; con los píxeles de mayor consistencia generan nuevos puntos 3D para reinicializar la malla. Experimentalmente demuestran una mejora en la precisión respecto al popular acercamiento de Furukawa y Ponce [30].

Schneider et al. en 2016 [123] presentan el upsampling de datos dispersos de profundidad a partir de imágenes de alta resolución. Usan indicadores de intensidad, indicadores de bordes y etiquetado semántico de la escena para guiarse.

La reconstrucción genérica vista como un flujo de trabajo completo se ve madura para aplicaciones de ingeniería. No obstante su madurez, se observan márgenes de mejora. Uno de los aspectos que actualmente ha cobrado popularidad es el de la reconstrucción

online, donde el rendimiento se vuelve crucial para dispositivos móviles.

3.1.4. Reconstrucción de edificios

Esta clase incluye la reconstrucción de escenas urbanas tanto de interiores como de exteriores con base en presuposiciones de las características de las construcciones. En esta revisión se dio mayor enfoque a la reconstrucción de interiores puesto que Musialski et al. [3] realizaron una revisión de los acercamientos para exteriores de edificios.

La reconstrucción de la escena en interiores se ve dificultada por las características propias de la escena. Existen problemas como entornos desordenados, o la falta de textura en las paredes; estos problemas reducen considerablemente el rendimiento de los acercamientos usuales. Para contrarrestar estos problemas, diversos aportes hacen uso de distintos datos, modelos o presuposiciones.

Cabe mencionar que existen acercamientos que trabajan sobre las reconstrucciones o las ocupan en su proceso; un caso es el de Hermans et al. en 2014 [124]. Discuten la segmentación semántica de interiores. Utilizando imágenes RGB-D, generan una nube de puntos 3D, y realizan la segmentación semántica de las imágenes. Posteriormente transfieren esta segmentación a la nube de puntos, tal que terminan con una nube de puntos segmentada semánticamente para apoyo a la comprensión de escenas tridimensionales.

Resumiendo, si se sabe qué es lo que se va a reconstruir, se pueden generar presuposiciones que faciliten y mejoren la calidad de la reconstrucción, como se muestra en los artículos revisados. Usando presuposiciones para la reconstrucción de edificios, se obtiene una reconstrucción robusta a falta de información del entorno; ésto gracias a la regularidad existente en las construcciones humanas, tanto la prevalencia de planos en las mismas, como la simetría con que cuentan.

3.1.5. Reconstrucción semántica

La reconstrucción puede requerir de más información dependiendo del contexto en que se use. En casos como el de los vehículos autónomos y la navegación de robots autónomos, es necesario reconocer tanto la posición actual del robot en la escena,

como tener una comprensión de los elementos que la conforman, como pueden ser otros vehículos, obstáculos o personas. La manera en que se obtiene esta comprensión es con una segmentación semántica que se realiza en el modelo tridimensional. La segmentación semántica consiste en agregar una etiqueta semántica a todos los puntos 3D de una reconstrucción [124].

Recientemente se propuso la reconstrucción semántica para resolver los problemas de segmentación y reconstrucción de manera conjunta tal que las fortalezas de un acercamiento apoyen a las debilidades del otro y viceversa, obteniendo una reconstrucción fiel del entorno con un etiquetado semántico completo y consistente con la escena visualizada.

Se observa que este nuevo acercamiento abre la puerta a nuevas propuestas con un punto de vista diferente el cual es colaborativo, pero es importante no menospreciar los acercamientos anteriores, ya que éstos siguen siendo competitivos como demuestran Vineet et al. [66]. El empleo de un planteamiento conjunto de los problemas de reconstrucción y segmentación finalmente permite una formalización [64] de la relación que se vislumbraba de manera empírica.

3.2. Pipeline propuesto

El pipeline propuesto para la investigación toma como base el flujo de trabajo típico presentado por Furukawa et al. en [1]. Se propone combinar los acercamientos existentes de reconstrucción para producir un modelo que combine el interior y el exterior de un edificio. Se trabaja por separado la reconstrucción de cada modelo y luego éstos son alineados y combinados para generar un sólo modelo. A continuación se describen los pasos en cada etapa. En la Figura 3.2 se muestra el diagrama con el flujo propuesto para la reconstrucción.

3.2.1. Obtener fotografías

La obtención de fotografías del entorno a reconstruir es una etapa crucial, ya que los algoritmos que obtienen la estructura 3D a partir de ellas pueden generar estructuras erróneas por falta de características distinguibles en el entorno (un caso usual en el

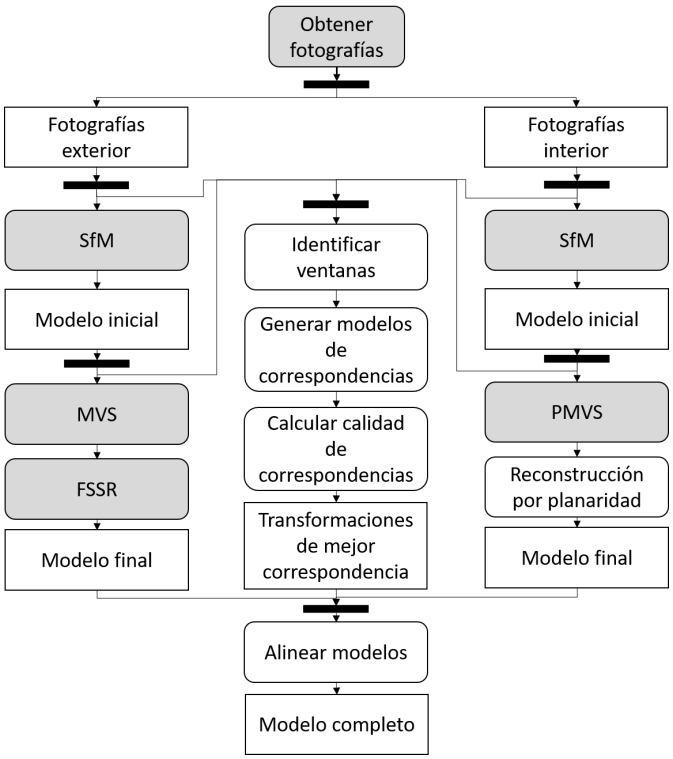


Figura 3.2: Diagrama del flujo propuesto para la reconstrucción. Aparecen sombreados los algoritmos ya existentes como herramientas ofrecidas por otros investigadores.

interior de edificios).

Este problema generalmente es resuelto añadiendo marcas al entorno antes de comenzar a tomar fotografías. Estas marcas deben ser reconocidas por el algoritmo, por lo que se pueden usar diseños como los de Schweiger et al. [125]; por otro lado, si ya se han tomado las fotografías del entorno, se pueden seleccionar puntos de enlace entre ellas para usar como marcas virtuales, como es el caso de la aplicación Pix4D.

3.2.2. Herramientas de reconstrucción

Gracias a aportes de varios investigadores y ofertas comerciales, actualmente se encuentran diversas herramientas de reconstrucción con diferentes características, a continua-

ción se describen algunas de ellas.

VisualSfM

Una herramienta de reconstrucción ofrecida por Changchang Wu en 2013 [27], disponible en la página http://ccwu.me/vsfm/, desarrollada para sistemas operativos Windows. Cuenta con algoritmos implementados para la reconstrucción por fotogrametría, desde la búsqueda de correspondencias entre imágenes hasta la generación del modelo tridimensional por SfM, con la opción de usar algoritmos que trabajan en CPU ó GPU.

La herramienta se encuentra adaptada para hacer uso del acercamiento CMVS/PMVS de Furukawa et al. [30, 126], igualmente con funcionamiento en CPU ó GPU, por medio de los archivos binarios para Windows desarrollados por Pierre Moulon y disponibles en el repositorio https://github.com/pmoulon/CMVS-PMVS.

La implementación de SfM que utiliza [27] tiene un funcionamiento casi lineal en cuanto al número de imágenes que utiliza a cambio de una mayor sensibilidad al par inicial de imágenes seleccionado para la reconstrucción. La herramienta mantiene todos los modelos parciales obtenidos en un proyecto; se obtienen modelos parciales de un conjunto de imágenes cuando no existen suficientes correspondencias entre las imágenes para calcular las matrices fundamentales entre ellas, cuando eso ocurre se genera un nuevo modelo parcial con el otro conjunto de imágenes; el proyecto puede ser incrementado con nuevas imágenes que salven las brechas entre modelos parciales para que éstos se logren fusionar en un sólo modelo.

Multi-View Environment (MVE)

Una suite de algoritmos de reconstrucción ofrecida por Fuhrmann et al. en 2014 [37], disponible en el repositorio https://github.com/simonfuhrmann/mve como código y en la página https://www.gcc.tu-darmstadt.de/home/proj/mve/ como archivos binarios para Windows. Cuenta con algoritmos para la búsqueda de correspondencias entre imágenes, SfM, MVS y reconstrucción de superficies.

A diferencia de VisualSfM, la implementación de SfM en MVE sólo mantiene un

modelo, descartando las imágenes que no pudieron ser agregadas al mismo. Su implementación de MVS [38] genera como datos intermedios los mapas de profundidad de cada una de las imágenes, que pueden ser visualizados dentro de la aplicación. Su implementación para la reconstrucción de superficies, *Floating Scale Surface Reconstruction* (FSSR) [39], cuenta con pocos parámetros y las superficies que genera no son necesariamente *watertight*, es decir que la superficie no busca cubrir un volumen cerrado; este tipo de superficie tiene menos artefactos de reconstrucción, ésto al no intentar cubrir el volumen, pero deja huecos en la superficie.

Pix4D

Un producto comercial para reconstrucción tridimensional y sensado remoto por medio de fotogrametría [13] disponible a través de suscripción que ofrece múltiples aplicaciones y un entorno completo desde la captura por medio de drones y cámaras; además de procesamiento local y en la nube con múltiples aplicaciones de visualización, reconstrucción y clasificación.

Un punto útil de esta herramienta durante la reconstrucción es la posibilidad de seleccionar puntos de enlace (i.e. correspondencias obtenidas manualmente) entre las imágenes, una característica que subsana la falta de correspondencias que genera múltiples modelos parciales en VisualSfM ó un modelo incompleto en MVE.

MeshLab

Una herramienta de código abierto para el procesamiento de datos tridimensionales [33] disponible en `http://www.meshlab.net/` como código fuente o instaladores para Windows, MacOS y Linux. La herramienta cuenta con implementaciones de diversos algoritmos disponibles en el estado del arte, entre ellos se encuentran los algoritmos de reconstrucción de superficies Poisson y Screened Poisson de Kazhdan et al. [31, 49] que reconstruyen una superficie *watertight*.

3.2.3. Reconstrucción del modelo interior

Se emplea la aplicación VisualSFM ofrecida por Wu [27, 28] para obtener la estructura tridimensional y las posiciones en el espacio de las cámaras por medio de SfM; el modelo de la estructura es alimentado a la implementación PMVS de Furukawa y Ponce [30] para obtener una nube de puntos densa del interior.

A la nube de puntos y las imágenes se les aplica una implementación basada en la propuesta de Cabral y Furukawa[79], con la que se genera el plano arquitectónico de la nube de puntos obtenida, y con éste crear el modelo tridimensional del entorno.

Inicialmente se rota el marco de referencia para que los ejes XY se encuentren alineados con las direcciones Manhattan del modelo. Los puntos son filtrados de acuerdo al valor de la vertical de sus normales que se modela de la siguiente manera:

$$-1 \leq n_z \leq 1 \rightarrow -90^\circ \leq \theta_z \leq 90^\circ \tag{3.1}$$

$$f(\theta_z) = \begin{cases} 1, & \text{si } |\theta_z| < 72^\circ \\ 0, & \text{de otro modo} \end{cases}$$

Se agregó este margen de error por el ruido que existe durante la reconstrucción, en el análisis de resultados se mostrará cómo este valor permite una segmentación entre el piso/techo y las paredes. Para los fines de esta etapa sólo se toman en cuenta los puntos de las paredes, donde $f(\theta_z) = 1$. Los puntos son proyectados hacia el plano XY y el plano se divide en celdas de las que se calcula la evidencia de pared y de espacio libre

La evidencia de pared de una celda se usa para determinar si la celda contiene una pared y se calcula como el número de puntos que han sido proyectado hacia ella.

La evidencia de espacio libre de una celda se usa para determinar si la celda debe o no ser contenida por el plano arquitectónico, se calcula proyectando los centros de cámara de las imágenes usadas al mismo plano XY y lanzando rayos de estos centros a todos los puntos que son visibles en la imagen; la evidencia es el total de rayos que pasan por la celda.

El plano dividido en celdas se considera como un grafo y se resuelve el camino más corto que rodee a la estructura sin pasar sobre celdas con una evidencia de pared alta. Este camino se convierte en la localización de las paredes del modelo del interior, las paredes se generan extendiendo estas celdas del plano hacia la tercera dimensión para tener el modelo final del interior.

3.2.4. Reconstrucción del modelo exterior

Se emplea VisualSFM para obtener la estructura y las posiciones de las cámaras. La salida de VisualSFM se exporta a la aplicación MVE ofrecida por Fuhrmann et al. [37] donde se aplica MVS basado en la propuesta de Goesele et al. [38] de generar un mapa de profundidad por cada fotografía. Los mapas de profundidad son triangulados y coloreados de acuerdo a su imagen de entrada para obtener una nube de puntos que es tomada como un conjunto de muestras por el algoritmo FSSR de Fuhrmann et al.[39] para construir una función implícita que representa a la superficie reconstruida, y que es convertida en una malla por el método Marching Cubes modificado de Kazhdan et al. [127].

3.2.5. Alineación de los modelos

Aplicando la propuesta de Cohen et al. [108] se toman como entrada los modelos separados del exterior y el interior obtenidos por SfM, y son alineados a través de información semántica. Se detectan ventanas para generar correspondencias entre ellos, y las correspondencias son usadas para calcular su alineación. Teniendo esta alineación, se sustituyen por los resultados finales de las etapas anteriores para obtener el modelo completo.

El primer paso es la detección de ventanas en las imágenes para su posterior proyección al modelo tridimensional. La detección se realiza por medio de un clasificador a nivel de pixel con el método de aprendizaje supervisado *Associative hierarchical random fields* de Ladický et al. [128], con el que se obtienen las imágenes segmentadas.

Se enlistan todos los puntos 3D del modelo inicial que se encuentran proyectados en el área definida por las ventanas; consecuentemente se calcula el plano que mejor se acople a estos puntos, y las ventanas se proyectan a este plano. En este punto se

cuenta con las ventanas 3D de una imagen, se realiza el mismo proceso para las demás imágenes.

Contando con las ventanas 3D de todas las imágenes, éstas se agrupan en clústers de ventanas que traslapan y se encuentran en el mismo plano. En cada clúster se calcula una caja contenedora B que encierre todas las ventanas 3D del clúster.

En B se calcula el consenso de las ventanas como la suma de calificaciones del clasificador, por un lado la calificación de ventana, y por el otro la de pared. Se genera la ventana de consenso V_c como el rectángulo dentro de B que maximiza la suma de calificaciones de ventana menos las calificaciones de pared que contiene.

Tras este proceso se cuenta con los conjuntos de ventanas consenso de interior V_{ci} y del exterior V_{ce}; se realiza una búsqueda exhaustiva en las posibles correspondencias entre ventanas del exterior y del interior, donde una correspondencia implica una transformación de similaridad T de un modelo al marco de referencia del otro modelo.

Las transformaciones son calificadas de acuerdo al número de ventanas con correspondencia y a la cantidad de espacio libre que es violada a causa de una intersección de los modelos.

El radio de intersección entre modelos γ se mide como la proporción de puntos 3D de un modelo que se encuentran dentro del otro modelo.

Una mejor calificación implica mayor correspondencia y menor intersección entre los modelos.

Capítulo 4

Evaluación de las propuestas

En este capítulo se detallan y discuten los resultados obtenidos. Se presenta la taxonomía aplicada a la revisión del estado del arte realizada, clasificando los trabajos para identificar oportunidades de investigación y elementos poco retomados del problema. Además, se presentan los resultados de aplicar las herramientas de reconstrucción al pipeline propuesto.

4.1. Observaciones de la taxonomía

Existe una gran diversidad de datos de entrada para la reconstrucción tridimensional; aunque el enfoque de este estudio son las imágenes, vale la pena revisar qué datos utilizan los investigadores y qué tipo de reconstrucción realizan de acuerdo a la taxonomía para poder identificar patrones.

4.1.1. Clasificación de acuerdo a datos de entrada

En la Tabla 4.1 se presentan las agrupaciones de datos de entrada de acuerdo a cada una de las clases identificadas en la taxonomía.

Tabla 4.1: Número de artículos que usan cada tipo de datos de acuerdo a la taxonomía.

	Dinámica						Estática					
	Online			Offline			Online			Offline		
	G	B	S	G	B	S	G	B	S	G	B	S
Multi-view	0	0	1	3	4	1	0	0	0	24	4	7
Single-view	0	0	0	0	0	0	2	1	0	2	4	0
Panorama	0	0	0	0	0	0	0	0	0	0	4	0
RGB-D	0	0	0	1	0	0	4	0	0	0	5	0
Plano arquitectónico	0	0	0	0	0	0	0	0	0	0	5	1
Información inercial	0	0	0	0	0	0	7	0	0	0	0	0

G Reconstrucción genérica.
B Reconstrucción de edificios usando conocimiento a priori.
S Reconstrucción semántica.

La Tabla 4.1 claramente muestra lo poco que han sido retomados los acercamientos de reconstrucción **Dinámica Online**. Ésto puede ser atribuido a la falta de poder de procesamiento o información inadecuada para la tarea, casos recientemente subsanados gracias a la creciente disponibilidad de imágenes en comunidades de internet, el mayor poder de procesamiento por avances tecnológicos y la democratización de dispositivos de captura como los teléfonos celulares y sensores, han permitido que esta categoría se mueva a la vanguardia de la investigación.

La reconstrucción dinámica, de acuerdo a los artículos encontrados en esta revisión, es realizada ya sea con celulares [97, 98, 102], o con sensores montados en un dispositivo móvil [91]. Igualmente, la reconstrucción dinámica, con aportes encontrados desde 2007 [54], se ve potenciada por la gran cantidad de imágenes con etiquetas de tiempo disponibles en internet [58, 59].

4.1.2. Clasificación de acuerdo a datos de salida

La aplicabilidad de la reconstrucción tridimensional es tan variada como las disciplinas que la soportan; en la Tabla 4.2 se presentan ejemplos de los resultados obtenidos de los trabajos clasificados.

Tabla 4.2: Número de artículos con cada salida esperada de acuerdo a la taxonomía.

	Dinámica						Estática					
	Online			Offline			Online			Offline		
	G	B	S	G	B	S	G	B	S	G	B	S
Navegación	0	0	0	0	0	0	1	0	0	0	3	0
Modelo 3D	0	0	0	0	3	0	10	1	0	22	10	0
Modelo 3D con segmentación semántica	0	0	1	4	1	1	2	0	0	4	4	8
Representación de espacio libre	0	0	0	0	0	0	0	0	0	0	5	0

G Reconstrucción genérica.
B Reconstrucción de edificios usando conocimiento a priori.
S Reconstrucción semántica.

Los acercamientos a la reconstrucción de modelos segmentados semánticamente cuentan en estos momentos con una gran popularidad, aunada a la búsqueda de automatizar la navegación de vehículos autónomos y robots. Principalmente se buscan procesos capaces de un funcionamiento online.

La navegación y la representación de espacio libre contribuyen a la comprensión de la escena, su tratamiento puede tener la misma utilidad o verse auxiliado de la reconstrucción con segmentación semántica.

La mayoría de los acercamientos de reconstrucción con segmentación semántica revisados en este trabajo se enfocan en exteriores, comprensible por la llegada de vehículos autónomos, pero también existen trabajos enfocados en interiores, que permiten el uso para navegación y aplicaciones de realidad aumentada o virtual.

4.1.3. Clasificación de acuerdo a funcionalidad esperada

En la Tabla 4.3 se muestran los artículos revisados, identificando si funcionan en tiempo real y las características de la escena que manejan: escenas a gran escala, dinámicas, de interiores y de exteriores.

Tabla 4.3: Clasificación de los artículos revisados de acuerdo a su dinamismo, capacidad y uso esperado (escenas de interiores o exteriores).

Clases		**Entradas**		**Salidas**	
Entrada	**Ent**	Multi-View	**M**	Navegación	**N**
Salida	**Sal**	Single-view	**S**	Modelo 3D	**M**
Reconstrucciones a gran escala	**GE**	Panorama	**P**	Modelo 3D con segmentación semántica	**S**
Reconstruye interiores	**Int**	RGB-D	**R**		
Reconstruye exteriores	**Ext**	Un plano	**F**	Representación de espacio libre	**F**
		Información de movimiento	**I**		

Artículo	**Ent**	**Sal**	**GE**	**Int**	**Ext**
Dinámica-Online-Semántica					
Vineet et al. [66]	M	S	●	○	●
Dinámica-Offline-Genérica					
Fehr et al. [89]	R	M	○	●	○
Martin-Brualla et al. [58]	M	M	○	○	●
Martin-Brualla et al. [59]	M	M	○	○	●
Radenovic et al. [60]	M	M	●	○	●
Dinámica-Offline-Edificios					
Schindler et al. [54]	M	M	○	○	●
Schindler et al. [55]	M	M	●	○	●
Schindler et al. [56]	M	M	●	○	●
Matzen et al. [57]	M	M	○	○	●
Dinámica-Offline-Semántica					
Sengupta et al. [62]	M	S	●	○	●
Estática-Online-Genérica					
Nießner et al. [86]	M	M	●	●	●
Tanskanen et al. [97]	I	M	○	●	●
Kolev et al. [98]	I	M	●	●	●
Klingensmith et al. [100]	I	M	●	●	●

Artículo	Ent	Sal	GE	Int	Ext
Cavallari et al. [87]	R	S	●	●	○
Schöps et al. [101]	I	M	●	○	●
Schöps et al. [102]	I	M	●	○	●
Dryanovski et al. [99]	I	M	●	●	●
Eigen et al. [75]	S	M	○	●	●
Eigen y Fergus [76]	S	S	○	●	○
Whelan et al. [90]	R	M	●	●	○
Yan et al. [6]	I	N	●	●	●
Dai et al. [91]	R	M	●	●	○
Estática-Online-Edificios					
Liu et al. [78]	S	M	○	●	○
Estática-Offline-Genérica					
RECAP [14]	M	M	●	○	●
Pix4D [13]	M	M	●	○	●
PhotoScan [120]	M	M	●	○	●
Goesele et al. [38]	M	M	○	○	●
Frahm et al. [117]	M	M	●	○	●
Agarwal et al. [118]	M	M	●	○	●
Heinly et al. [119]	M	M	●	○	●
Wu [27]	M	M	●	○	●
Wu et al. [28]	M	M	●	○	●
Furukawa et al. [30]	M	M	○	○	●
Fuhrmann et al. [39]	M	M	●	○	●
Fuhrmann et al. [37]	M	M	●	○	●
Schönberger et al. [26]	M	M	●	○	●
Schönberger et al. [40]	M	M	●	○	●
Alcantarilla et al. [47]	M	M	●	○	●
Lafarge y Mallet [52]	M	S	●	○	●
Lafarge et al. [53]	M	M	●	○	●

Artículo	Ent	Sal	GE	Int	Ext
Shan et al. [48]	M	M	○	○	●
Duan y Lafarge [61]	M	S	●	○	●
Yang et al. [45]	M	M	○	●	○
Holzmann et al. [44]	M	M	○	○	●
Locher et al. [46]	M	M	●	○	●
Ladický et al. [71]	S	S	○	●	●
Liu et al. [78]	S	S	○	●	○
Romanoni et al. [122]	M	M	○	○	●
Estática-Offline-Edificios					
Hermans et al. [124]	R	S	○	●	○
Ikehata et al. [83]	M	M	○	●	○
Choi et al. [88]	R	M	●	●	○
Cabral et al. [79]	P	M	○	●	○
Ikehata et al. [80]	P	M	●	●	○
Yang y Zhang [81]	P	M	○	●	○
Martin-Brualla et al. [51]	F	M	●	●	●
Chu et al. [11]	F	M	○	●	●
Liu et al. [94]	F	N	○	●	○
Pintore et al. [82]	M	N	●	●	○
Dong et al. [5]	I	N	●	●	○
Furukawa et al. [42]	M	M	○	●	●
Fouhey et al. [50]	M	F	○	●	○
Zou et al. [84]	P	M	○	●	○
Hedau et al. [72]	S	F	○	●	○
Hedau et al. [73]	S	F	○	●	○
Liu et al. [93]	F	S	●	●	○
Wijmans y Furukawa [96]	F	M	●	●	○
Silberman et al. [85]	S	S	○	●	○
Liu et al. [92]	R	S	●	●	○

Artículo	Ent	Sal	GE	Int	Ext
Gupta et al. [74]	S	F	○	●	○
Estática-Offline-Semántica					
Häne et al. [63]	M	S	●	○	●
Kundu et al. [65]	M	S	○	○	●
Savinov et al. [67]	M	S	○	○	●
Cherabier et al. [68]	M	S	●	○	●
Bláha et al. [69]	M	S	●	○	●
Häne et al. [64]	M	S	●	○	●
Wang et al. [95]	F	S	●	○	●

Clases		**Entradas**		**Salidas**	
Entrada	**Ent**	Multi-View	**M**	Navegación	**N**
Salida	**Sal**	Single-view	**S**	Modelo 3D	**M**
Reconstrucciones a gran escala	**GE**	Panorama	**P**	Modelo 3D con segmentación semántica	**S**
Reconstruye interiores	**Int**	RGB-D	**R**		
Reconstruye exteriores	**Ext**	Un plano	**F**	Representación de espacio libre	**F**
		Información de movimiento	**I**		

4.2. Detalles de pipeline propuesto

Durante la aplicación del pipeline propuesto se encontraron los siguientes detalles.

4.2.1. Imágenes usadas

Se toma el dataset ofrecido por Koch et al. [129] para mostrar la aplicación de los algoritmos, este dataset cuenta con imágenes de una misma escena en distintas modalidades, como son fotografías tomadas con drones y con cámaras de mano. En la Figura 4.1 se muestran algunas de las imágenes que contiene.

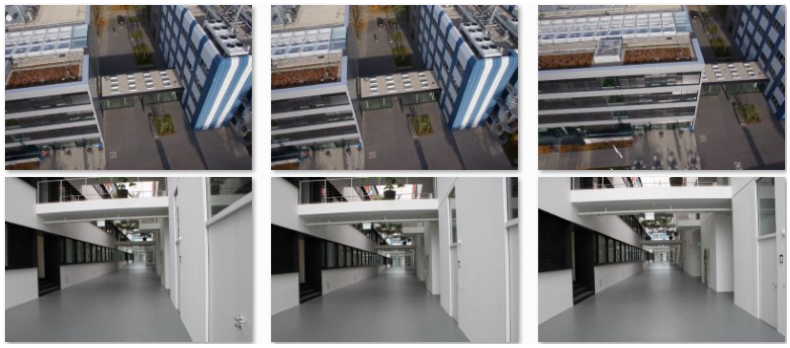


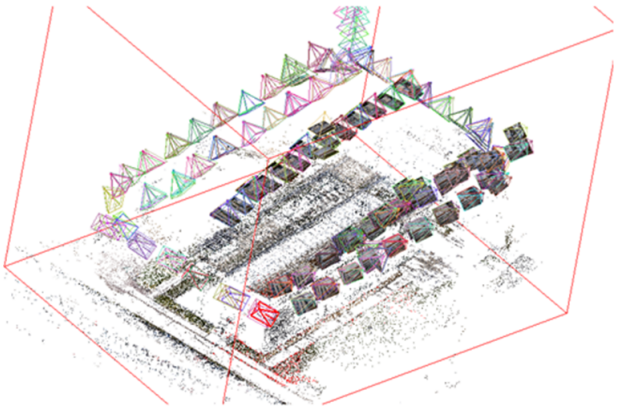
Figura 4.1: Tres de las imágenes tomadas del exterior y del interior del edificio ofrecidas por el dataset TUM-DLR de Koch et al. [129].

Para interiores se toman la imágenes del pasillo interior del mismo dataset, imágenes de una sala del Edificio G de la Facultad de Ingeniería de la UAEMex y un subconjunto de imágenes para una habitación del dataset de interiores de museos presentado en [130].

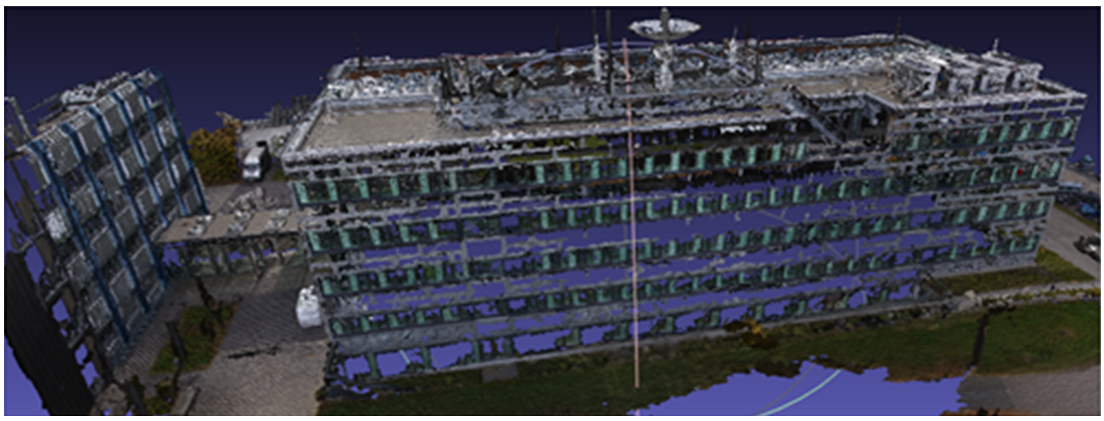
4.2.2. Reconstrucción de exterior

Del dataset de Koch et al. [129] se usan las fotografías aéreas del exterior del edificio observado, y las fotografías de un pasillo del interior del mismo edificio. En la Figura 4.2 se muestra el resultado de aplicar SfM, MVS y FSSR, conforme al pipeline propuesto,

sobre las imágenes del exterior del edificio del dataset descrito anteriormente. En el resultado intermedio de aplicar SfM, se observan las posiciones de las cámaras y una nube de puntos dispersa de la estructura visualizada en las imágenes. Subsecuentemente, la malla tridimensional es obtenida al aplicar los algoritmos MVS y FSSR.



(a) Resultado SfM



(b) Resultado MVS y FSSR

Figura 4.2: Resultado de aplicar los algoritmos SfM, MVS y FSSR seleccionados para el pipeline.

Esta reconstrucción fue posible por la calidad de la información del dataset de Koch et al., que ofrece una vista aérea de la escena, lo que permite a los algoritmos distinguir elementos de la escena y no perderse en elementos con una gran frecuencia de repetición como son las ventanas.

La obtención de fotografías del entorno a reconstruir es una etapa crucial, ya que los algoritmos que obtienen la estructura 3D a partir de ellas pueden generar estructuras erróneas por falta de características distinguibles en el entorno (un caso usual en el interior de edificios).

Este problema generalmente es resuelto añadiendo marcas al entorno antes de comenzar a tomar fotografías. Estas marcas deben ser reconocidas por el algoritmo, por lo que se pueden usar diseños como los de Schweiger et al. [125]; por otro lado, si ya se han tomado las fotografías del entorno, se pueden seleccionar puntos de enlace entre ellas para usar como marcas virtuales, como es el caso de la aplicación Pix4D.

Además de la falta de características distinguibles se debe tomar en cuenta la naturaleza de los algoritmos, SfM busca recuperar la profundidad implícita encontrada en las imágenes para volver de proyecciones 2D a estructuras 3D, por lo que se tiene un espacio de búsqueda que involucra tanto las posiciones de las imágenes en el espacio tridimensional y la profundidad de los puntos que se observan en ellas.

El problema de definición de la profundidad de un punto se ve altamente influenciado por la distancia entre las cámaras que se usan para definirlo, esta distancia es llamada *baseline* y en la Figura 4.3 se observa el espacio de búsqueda de profundidades que puede tomar un pixel en cada imagen; el área marcada incluye todos los puntos 3D que proyectan al pixel observado en las imágenes. Una baseline pequeña implica un mayor error de profundidad, mientras que una baseline grande dificulta la búsqueda de características comunes, ésto porque mientras mayor sea la baseline será más difícil en la práctica poder agrupar las imágenes debido a oclusiones causadas por otros objetos en la escena.

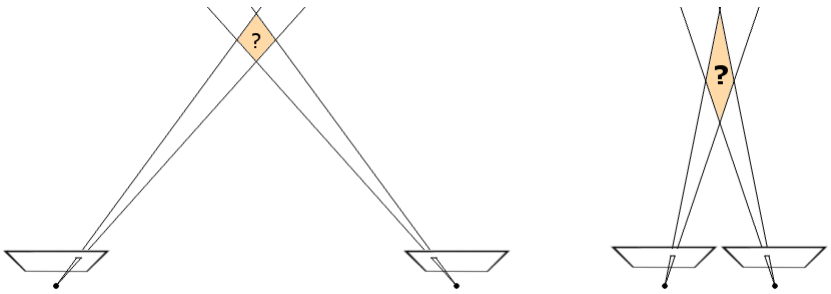


Figura 4.3: Diferentes espacios de búsqueda de acuerdo a la distancia entre las cámaras o baseline.

Este error conlleva a una mala reconstrucción, usando el dataset *yard* que presenta [131] se obtiene la reconstrucción de la Figura 4.4; las imágenes fueron tomadas tal que sólo se realizó una rotación al momento de tomar una nueva fotografía, pero en la reconstrucción se observa que se calculan erróneamente traslaciones entre las imágenes durante SfM (los puntos puntos de color en la imagen superior representan las posiciones de las cámaras).

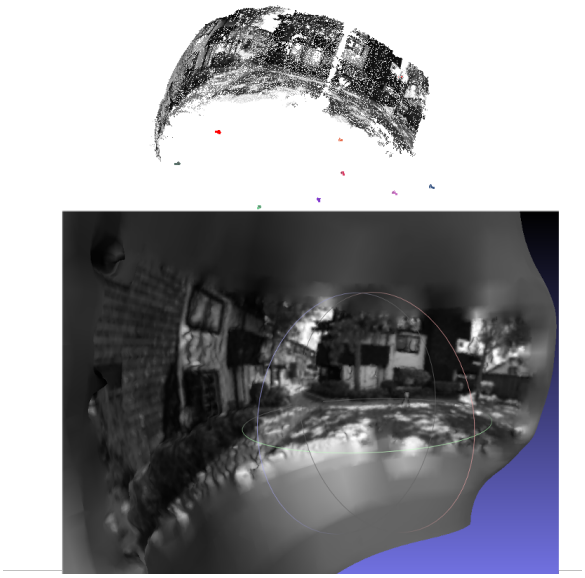


Figura 4.4: Reconstrucción de un conjunto de imágenes que se diferencian sólo con rotación.

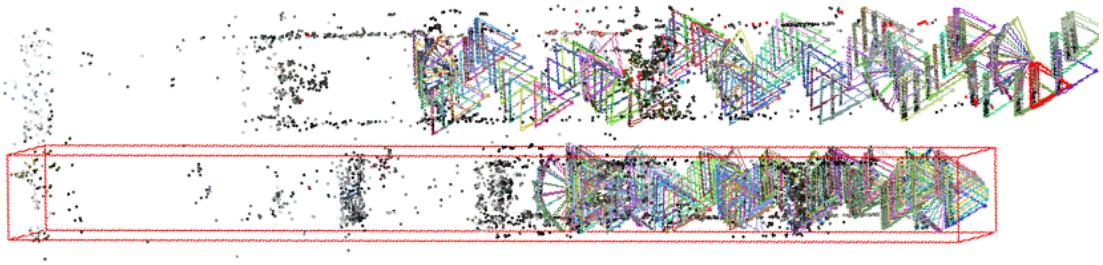
4.2.3. Reconstrucción de interior

Durante la reconstrucción de interiores se presentan diversos problemas causados por la repetibilidad de elementos de la escena, la falta de características distinguibles (necesarias para el buen funcionamiento de los algoritmos genéricos) y la presencia de obstáculos en la escena

En la Figura 4.5 se puede visualizar el problema de falta de información para recuperar la configuración de las cámaras, se tomaron las imágenes del interior ofrecidas en [129], estas imágenes son de un pasillo tomadas en ambos sentidos y en dos puntos

un conjunto de imágenes tomadas en los 360° sin desplazarse. Las imágenes del pasillo no cuentan con suficientes puntos de interés en común, lo que causa que el algoritmo implementado en VisualSfM las separe en dos modelos distintos.

(a) Posiciones esperadas de las cámaras [129].



(b) Falla por falta de información para alinear cámaras.

(c) Vista de ambos sentidos del pasillo [129].

Figura 4.5: Ejemplo de una falla en la reconstrucción por SfM de un pasillo.

El problema que se observa es que estas imágenes no son suficientes para que el proceso de SfM pueda conectarlas, lo que causa que sólo se use un subconjunto de ellas (i.e. una dirección del pasillo) para crear el modelo. La aplicación VisualSfM es capaz de mantener modelos separados cuando es incapaz de formar un único modelo, es por esta razón que se ven dos pasillos reconstruidos, cada uno con las imágenes del pasillo

en una u otra dirección.

En la Figura 4.6 se observa tanto la presencia de obstáculos en la escena como la falta de características distinguibles (i.e. textura) en las paredes, lo que causa que se pierda la información de las dimensiones del cuarto y sólo se cuente con puntos en los elementos de la habitación.

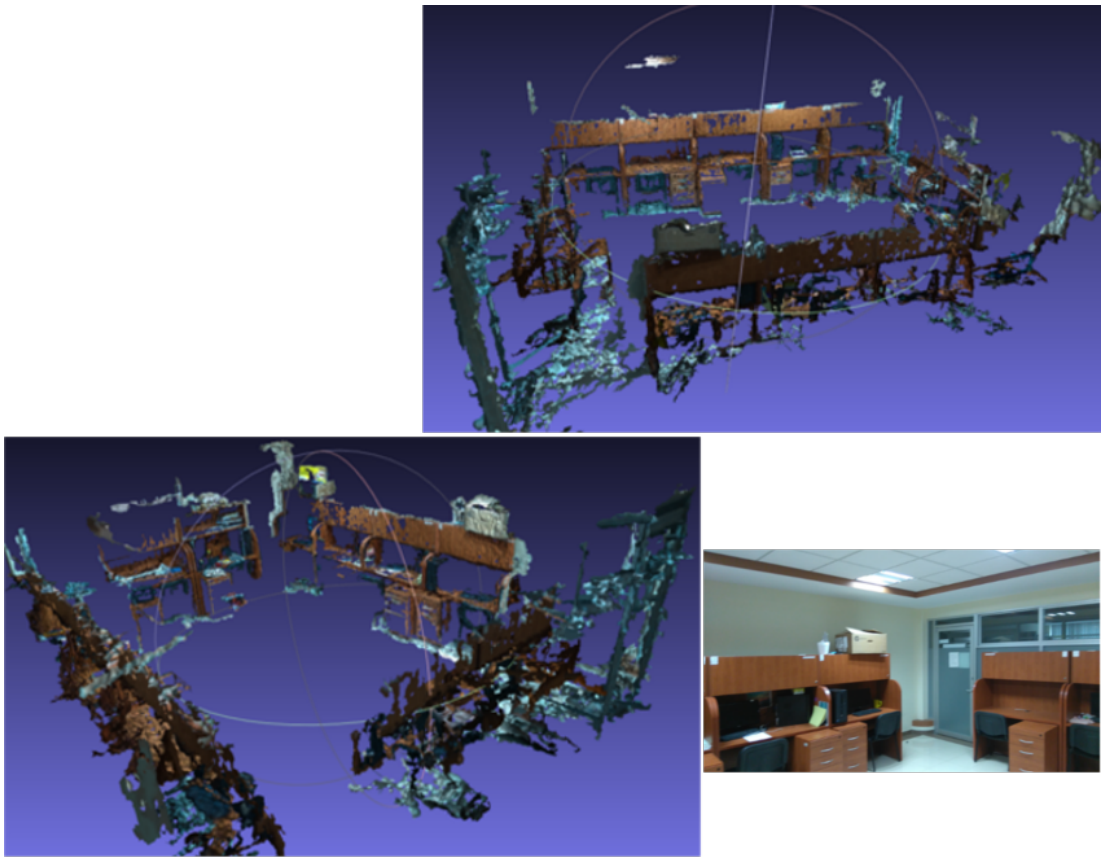


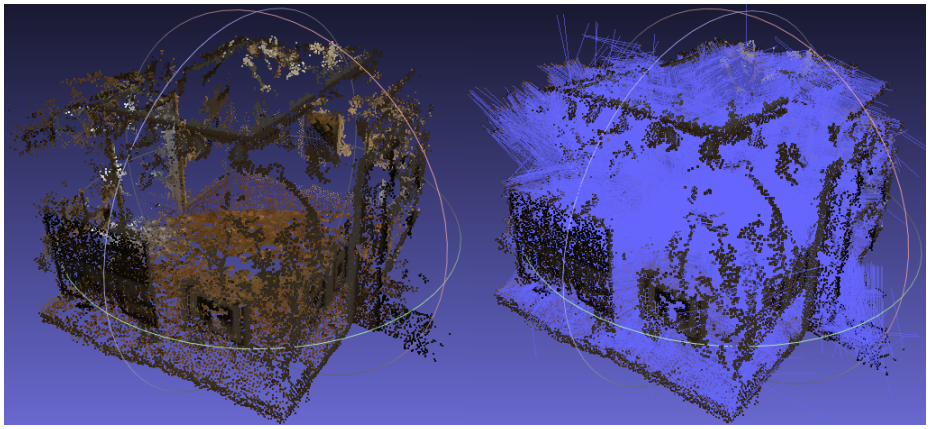
Figura 4.6: Los algoritmos SfM y MVS funcionan con los obstáculos en la escena (i.e. los escritorios), pero fallan al manejar las paredes por falta de textura.

Se observa que la falta de textura evitó que se reconstruyeran el piso, las paredes y el techo, lo que implica que el acercamiento tomado debe ayudarse de otro proceso, ya sea durante la captura de las imágenes con proyección de luz estructurada, o durante la reconstrucción usando alguno de los acercamientos de comprensión de escenas [73] para identificar la posición de las paredes y piso, de otro modo se tiene como limitación

que el pipeline únicamente funciona en escenas ricas en textura, lo que contradice la utilidad propuesta del mismo.

A continuación se presenta un dataset más amigable con los métodos de reconstrucción por fotogrametría. Tomando el conjunto de imágenes del dataset [130] se obtiene la reconstrucción mostrada en la Figura 4.7, donde se muestra la nube de puntos obtenida de los procesos SfM [27] y PMVS [30] donde cada punto es considerado como un pequeño plano o "parche" y la normal indica la orientación del plano.

(a) Ejemplo del dataset

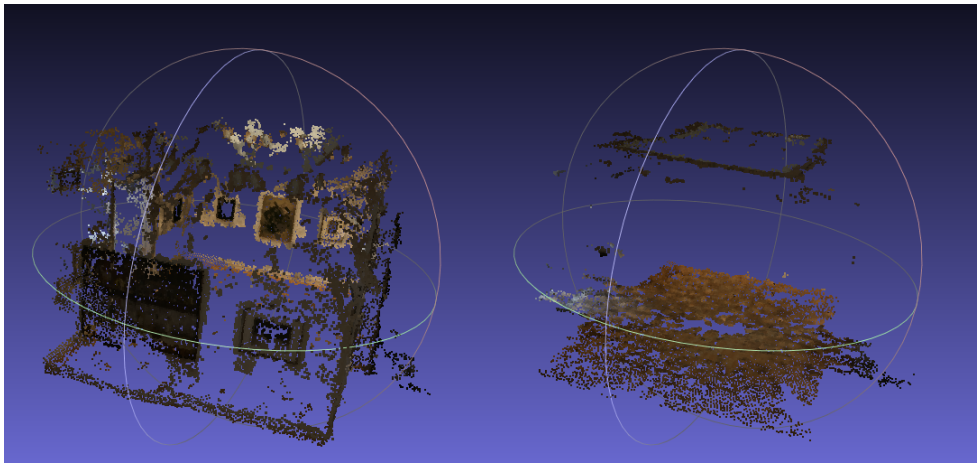


(b) Nube de puntos **(c)** Normales de los puntos

Figura 4.7: Nube de puntos de un cuarto.

Se observa que los cuadros colgados en las paredes proveen la textura necesaria para tener una idea de las dimensiones del cuarto, además, la textura del piso permite un buen funcionamiento de los algoritmos. PMVS entrega además las normales de cada punto, indicando su orientación. Usando las normales se puede distinguir entre paredes, pisos y techos, en la Figura 4.8 se muestra la segmentación usando las condiciones expuestas en la sección anterior. En los puntos identificados como pared se observa que

se recuperan los cuadros, los segmentos de pared cercanos a cada cuadro, los zoclos y el marco del pasillo de entrada. Por otra parte, los puntos identificados como piso o techo corresponden al piso del cuarto y la estructura de soporte de las luces en el techo.



(a) Paredes **(b)** Pisos y techos

Figura 4.8: Separando los puntos de acuerdo al valor de sus normales para distinguir entre paredes y techos o pisos.

Tras aplicar esta segmentación resulta más aparente la presencia de ruido en la nube de puntos (i.e. puntos que parecen estar flotando en el espacio donde no debería existir nada), ésto se atribuye a una iluminación no uniforme de la escena, lo que causa que los algoritmos imaginen geometrías donde sólo hay sombras, por lo que puede ser necesario un proceso intermedio de filtrado de estas geometrías espurias, un proceso que se realizó manualmente durante esta investigación.

Capítulo 5

Conclusiones

En este capítulo se discuten los resultados obtenidos de esta investigación, se presentan las conclusiones alcanzadas y se identifican posibles rutas futuras para la continuación del tratado de este tema.

Actualmente existe un claro interés en la reconstrucción 3D urbana, ésto es atribuible a su aplicabilidad en la solución de problemas de naturaleza variada, por ejemplo, mantener un registro de la información de edificios para la administración de su ciclo de vida, desde su conservación hasta su deconstrucción, la creación de ciudades inteligentes y la producción en el entretenimiento como son escenas de películas y juegos serios.

Aún falta consenso sobre la forma de manejar los datos de entrada, tanto el tipo como la cantidad, a causa de la naturaleza específica de cada problema; obtener una reconstrucción 3D aceptable es enteramente dependiente del resultado requerido, los recursos disponibles y las características del entorno, por lo que es muy probable que las técnicas de reconstrucción sigan siendo tan variadas como los problemas que buscan resolver, y esto no debe ser visto como una debilidad de los acercamientos, sino como una característica inherente del área de investigación.

El dinamismo en esta área de investigación es evidente. Es necesario mantener un estado del arte actualizado, tomando en cuenta los aportes de las diversas disciplinas, para incrementar el impacto de los esfuerzos realizados en la reconstrucción 3D urbana. Esta investigación presentó una revisión detallada del estado del arte y, a partir de la

evidencia de la misma, se propuso una taxonomía de la reconstrucción 3D urbana.

Hablando de las técnicas de reconstrucción por sí mismas, en los últimos años las redes neuronales se han comenzado a posicionar como herramientas capaces de subsanar las carencias de los algoritmos actuales (logrando identificar características de escenas en interiores [92]), y en algunos casos mostrarse como substitutos aceptables (llegando a la generación single-view de mapas de profundidad [78] y modelos tridimensionales [84]). Aunque estas técnicas de reconstrucción son relativamente jóvenes, habiendo sido propuestas en los últimos años [78, 84, 92] han demostrado resultados impresionantes ya que logran identificar patrones donde anteriormente era necesario modelar restricciones; la conjunción de estas técnicas con una creciente comprensión del tema por parte de investigadores traerá en los años venideros grandes avances en el área, tanto en la investigación como en los productos disponibles al público en general.

La taxonomía propuesta busca presentar las características que actualmente se esperan de las técnicas de reconstrucción 3D urbana para facilitar la comunicación interdisciplinaria y permitir la resolución de problemas más complejos. Uno de los objetivos de esta investigación es ayudar a investigadores a identificar acercamientos similares/-complementarios a los que proponen para mejorar sus soluciones propuestas.

Al revisar los acercamientos que gozan de popularidad actualmente, se pueden descubrir nuevas oportunidades de investigación, ya sea identificando la posible utilidad de nuevas tecnologías o la viabilidad de nuevos puntos de vista. Clasificando a los artículos de acuerdo a la taxonomía se identifican ciertos aspectos que, hasta recientemente, han sido poco retomados.

La revisión del estado del arte identifica un interés por la reconstrucción dinámica, a causa de la combinación de un creciente poder de cómputo, mayor variedad de sensores, algoritmos eficientes [66, 86, 91] y escalables [117, 118, 119]. El problema sigue siendo abierto y aún requiere de un gran esfuerzo en la investigación.

El reconocimiento del dinamismo en una escena tiene aplicaciones demostradas tanto en la identificación de objetos dinámicos [66, 89] como en la historia de una escena [56, 57]; aun dada esta utilidad, existen situaciones donde ignorar dicho dinamismo es preferido [30, 47].

La utilidad de la reconstrucción genérica para los exteriores de edificios se vio ejemplificada durante la aplicación del pipeline propuesto, pero también se vieron sus faltas durante la reconstrucción de un modelo de interiores. A lo largo de la investigación

se identificó un grave problema en la forma de persianas que cubren gran parte de la pared donde se encuentra una ventana.

5.1. Trabajos futuros

Durante la reconstrucción de interiores se identificaron serias faltas en el acercamiento multi-view genérico; el acercamiento original de Cabral y Furukawa [79] lo compensaba con una clasificación estructural de techo, pared y piso de la imagen panorámica. De manera similar se pueden usar acercamientos de comprensión de escena como el de Hedau et al. [73] para identificar la "caja contenedora" del cuarto y reconstruir la escena de acuerdo a ella.

Es necesario también reformular el problema de alineación de los modelos de interior y exterior tal que se tome en cuenta la existencia de obstáculos de las ventanas como son las persianas. Un posible acercamiento es una búsqueda conjunta de las ventanas, identificando desde el exterior los tipos/formas de las ventanas y usar esta información como restricciones para el modelo del interior, definiendo ventanas candidatas en las paredes donde hay persianas cerradas.

Los avances actuales en el uso de redes neuronales indican que es posible que la taxonomía actual deba modificarse para clasificarlos de manera independiente; hasta ahora se observan como propuestas que usan los modelos y conocimientos tradicionales, pero la evolución y cruza de estos acercamientos los puede volver cada vez más independientes de los acercamientos tradiciones. Un ejemplo de esta independencia se encuentra en Liu et al. [92], quienes combinaron únicamente distintas redes neuronales en una arquitectura que genera un plano arquitectónico a partir de imágenes. En el futuro, ya con una mejor comprensión del rumbo que están tomando los acercamientos de redes neuronales, la taxonomía deberá ser actualizada para clasificarlos de manera adecuada.

5.2. Publicaciones

En este apéndice se anexan las publicaciones realizadas y enviadas durante el transcurso de la investigación. Un artículo fue publicado en el *International Multidisciplinary Congress from the XX Anniversary of our University: Centro Universitario UAEM Valle de México 2016*:

> Gómez-Martínez, D. G., Mercado-Herrera, R., Muñoz-Jiménez, V., & Ramos-Corchado, M. A. (2016). Generación de lenguaje corporal en agentes virtuales usados en realidad virtual inmersiva. In Desarrollo Multidisciplinario en Investigación y Docencia del Centro Universitario UAEM Valle de México Capítulo 1 Investigación y Desarrollo Multidisciplinario en Ingeniería. (pp. 21-26). Universidad Autónoma del Estado de México Centro Universitario UAEM Valle de México.

Se envió también un artículo a la revista *Computer Graphics Forum* de la *European Association for Computer Graphics*, el cual se encuentra en proceso de revisión al momento de la impresión de este documento:

> Rafael Mercado, Vianney Muñoz-Jiménez, Félix Ramos, and Marco Ramos. 2018. Urban 3D reconstruction: A survey and an updated taxonomy.

Referencias

[1] Y. Furukawa and C. Hernández, "Multi-View Stereo: A Tutorial," *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 1-2, pp. 1–148, 2015.

[2] X. Yin, P. Wonka, and A. Razdan, "Generating 3d building models from architectural drawings: A survey," *IEEE computer graphics and applications*, vol. 29, no. 1, 2009.

[3] P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. Gool, and W. Purgathofer, "A Survey of Urban Reconstruction," *Comput. Graph. Forum*, vol. 32, pp. 146–177, Sept. 2013.

[4] R. Volk, J. Stengel, and F. Schultmann, "Building Information Modeling (BIM) for existing buildings—Literature review and future needs," *Automation in construction*, vol. 38, pp. 109–127, 2014.

[5] J. Dong, Y. Xiao, M. Noreikis, Z. Ou, and A. Ylä-Jääski, "iMoon: Using Smartphones for Image-based Indoor Navigation," in *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, SenSys '15, (New York, NY, USA), pp. 85–97, ACM, 2015.

[6] H. Yan, Q. Shan, and Y. Furukawa, "RIDI: Robust IMU Double Integration," *arXiv preprint arXiv:1712.09004*, 2017.

[7] R. S. Haluck, R. W. Webster, A. J. Snyder, M. G. Melkonian, B. J. Mohler, M. L. Dise, and A. Lefever, "A virtual reality surgical trainer for navigation in laparoscopic surgery," *Studies in health technology and informatics*, pp. 171–176, 2001.

[8] N. Cooke and R. Stone, "RORSIM: a warship collision avoidance 3D simulation designed to complement existing Junior Warfare Officer training," *Virtual Reality*, vol. 17, no. 3, pp. 169–179, 2013.

[9] C. A. Vanegas, D. G. Aliaga, P. Wonka, P. Müller, P. Waddell, and B. Watson, "Modelling the appearance and behaviour of urban spaces," in *Computer Graphics Forum*, vol. 29, pp. 25–42, Wiley Online Library, 2010.

[10] J. De Reu, G. Plets, G. Verhoeven, P. De Smedt, M. Bats, B. Cherretté, W. De Maeyer, J. Deconynck, D. Herremans, P. Laloo, *et al.*, "Towards a three-dimensional cost-effective registration of the archaeological heritage," *Journal of Archaeological Science*, vol. 40, no. 2, pp. 1108–1121, 2013.

[11] H. Chu, S. Wang, R. Urtasun, and S. Fidler, *HouseCraft: Building Houses from Rental Ads and Street Views*, pp. 500–516. Cham: Springer International Publishing, 2016.

[12] A. Koutsoudis, B. Vidmar, G. Ioannakis, F. Arnaoutoglou, G. Pavlidis, and C. Chamzas, "Multi-image 3D reconstruction data evaluation," *Journal of Cultural Heritage*, vol. 15, no. 1, pp. 73–79, 2014.

[13] Pix4D, "Pix4d - drone mapping software." https://pix4d.com/. Accessed: 2017-10-24.

[14] Autodesk®, "Recap — reality capture and 3d scanning software." http://www.autodesk.com/products/recap/overview. Accessed: 2017-10-24.

[15] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment–a modern synthesis," in *International workshop on vision algorithms*, pp. 298–372, Springer, 1999.

[16] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2 ed., 2003.

[17] W. R. Sherman and A. B. Craig, *Understanding Virtual Reality: Interface, Application, and Design*. The Morgan Kaufmann Series in Computer Graphics, Morgan Kaufmann, 1st ed., 2002.

[18] J. Vince, *Introduction to virtual reality*. Springer-Verlag London Limited, 2004.

[19] M. Heim, *The Metaphysics of Virtual Reality*. Oxford University Press, 1993.

[20] D. A. Bowman, D. Koller, and L. F. Hodges, "Travel in immersive virtual environments: An evaluation of viewpoint motion control techniques," in *Virtual Reality Annual International Symposium, 1997., IEEE 1997*, pp. 45–52, IEEE, 1997.

[21] T. Schenk, "Introduction to Photogrammetry." GS400.02 Department of Civil and Environmental Engineering and Geodetic Science The Ohio State University, 2005.

[22] D. P. Robertson and R. Cipolla, *Practical Image Processing and Computer Vision*, ch. Structure from Motion, p. 49. John Wiley, Hoboken, NJ, USA, 2009.

[23] K. Simek, "Sightations a computer vision blog," Aug 2013.

[24] R. Szeliski, *Computer Vision: Algorithms and Applications*. Texts in Computer Science, Springer, 2013.

[25] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[26] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] C. Wu, "Towards linear-time incremental structure from motion," in *2013 International Conference on 3D Vision-3DV 2013*, pp. 127–134, IEEE, 2013.

[28] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3057–3064, IEEE, 2011.

[29] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[30] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[31] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, vol. 7, 2006.

[32] T. S. Newman and H. Yi, "A survey of the marching cubes algorithm," *Computers & Graphics*, vol. 30, no. 5, pp. 854–879, 2006.

[33] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia, "Meshlab: an open-source mesh processing tool.," in *Eurographics Italian Chapter Conference*, vol. 2008, pp. 129–136, 2008.

[34] ETHZ, "Computer vision and geometry group at ethz." http://cvg.ethz.ch/index.php.

[35] WUSTL, "Graphics and vision - computer science and engineering,." https://cse.wustl.edu/research/areas/Pages/Graphics-Vision.aspx.

[36] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 519–528, June 2006.

[37] S. Fuhrmann, F. Langguth, and M. Goesele, "MVE-A Multi-View Reconstruction Environment," in *GCH*, pp. 11–18, 2014.

[38] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," in *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, IEEE, 2007.

[39] S. Fuhrmann and M. Goesele, "Floating scale surface reconstruction," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 46, 2014.

[40] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, *Pixelwise View Selection for Unstructured Multi-View Stereo*, pp. 501–518. Cham: Springer International Publishing, 2016.

[41] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[42] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Manhattan-world stereo," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1422–1429, IEEE, 2009.

[43] J. M. Coughlan and A. L. Yuille, "Manhattan world: Compass direction from a single image by bayesian inference," in *Proceedings of the Seventh International Conference on Computer Vision, IEEE*, vol. 2, pp. 941–947, IEEE, 1999.

[44] T. Holzmann, M. Oswald, M. Pollefeys, F. Fraundorfer, and H. Bischof, "Plane-based Surface Regularization for Urban 3D Reconstruction," in *28th British Machine Vision Conference*, 2017.

[45] Y. Yang, M.-C. Chang, L. Wen, P. Tu, H. Qi, and S. Lyu, "Efficient large-scale photometric reconstruction using Divide-Recon-Fuse 3D Structure from Motion," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 180–186, IEEE, 2016.

[46] A. Locher, M. Havlena, and L. V. Gool, "Progressive Structure from Motion," *CoRR*, vol. abs/1803.07349, 2018.

[47] P. F. Alcantarilla, C. Beall, and F. Dellaert, "Large-Scale Dense 3D Reconstruction from Stereo Imagery," in *5th Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV13)*, 2013.

[48] Q. Shan, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, "Occluding contours for multi-view stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4002–4009, 2014.

[49] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 3, p. 29, 2013.

[50] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic, "People Watching: Human Actions as a Cue for Single View Geometry," *International Journal of Computer Vision*, vol. 110, pp. 259–274, Dec 2014.

[51] R. Martin-Brualla, Y. He, B. C. Russell, and S. M. Seitz, *The 3D Jigsaw Puzzle: Mapping Large Indoor Spaces*, pp. 1–16. Cham: Springer International Publishing, 2014.

[52] F. Lafarge and C. Mallet, "Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation," *International journal of computer vision*, vol. 99, no. 1, pp. 69–85, 2012.

[53] F. Lafarge, R. Keriven, M. Brédif, and H.-H. Vu, "A hybrid multiview stereo algorithm for modeling urban scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 5–17, 2013.

[54] G. Schindler, F. Dellaert, and S. B. Kang, "Inferring temporal order of images from 3D structure," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, IEEE, 2007.

[55] G. Schindler and F. Dellaert, "Probabilistic temporal inference on reconstructed 3D scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1410–1417, June 2010.

[56] G. Schindler and F. Dellaert, "4D Cities: Analyzing, Visualizing, and Interacting with Historical Urban Photo Collections," *JOURNAL OF MULTIMEDIA*, vol. 7, no. 2, p. 125, 2012.

[57] K. Matzen and N. Snavely, *Scene Chronology*, pp. 615–630. Cham: Springer International Publishing, 2014.

[58] R. Martin-Brualla, D. Gallup, and S. M. Seitz, "3d time-lapse reconstruction from internet photos," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1332–1340, 2015.

[59] R. Martin-Brualla, D. Gallup, and S. M. Seitz, "Time-lapse Mining from Internet Photos," *ACM Trans. Graph.*, vol. 34, pp. 62:1–62:8, July 2015.

[60] F. Radenovic, J. L. Schonberger, D. Ji, J.-M. Frahm, O. Chum, and J. Matas, "From dusk till dawn: Modeling in the dark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5488–5496, 2016.

[61] L. Duan and F. Lafarge, "Towards large-scale city reconstruction from satellites," in *European Conference on Computer Vision*, pp. 89–104, Springer, 2016.

[62] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. S. Torr, "Urban 3D semantic modelling using stereo vision," in *2013 IEEE International Conference on Robotics and Automation*, pp. 580–585, May 2013.

[63] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys, "Joint 3D Scene Reconstruction and Class Segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[64] C. Häne, C. Zach, A. Cohen, and M. Pollefeys, "Dense Semantic 3D Reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1730–1743, Sept 2017.

[65] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg, *Joint Semantic Segmentation and 3D Reconstruction from Monocular Video*, pp. 703–718. Cham: Springer International Publishing, 2014.

[66] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, and P. H. S. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 75–82, May 2015.

[67] N. Savinov, C. Hane, L. Ladicky, and M. Pollefeys, "Semantic 3D Reconstruction With Continuous Regularization and Ray Potentials Using a Visibility Consistency Constraint," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[68] I. Cherabier, C. Häne, M. R. Oswald, and M. Pollefeys, "Multi-Label Semantic 3D Reconstruction Using Voxel Blocks," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 601–610, Oct 2016.

[69] M. Blaha, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler, "Large-Scale Semantic 3D Reconstruction: An Adaptive Multi-Resolution Model for Multi-Class Volumetric Labeling," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[70] M. Blaha, M. Rothermel, M. R. Oswald, T. Sattler, A. Richard, J. D. Wegner, M. Pollefeys, and K. Schindler, "Semantically Informed Multiview Surface Refinement," *arXiv preprint arXiv:1706.08336*, 2017.

[71] L. Ladický, J. Shi, and M. Pollefeys, "Pulling Things out of Perspective," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, (Washington, DC, USA), pp. 89–96, IEEE Computer Society, 2014.

[72] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering the spatial layout of cluttered rooms," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 1849–1856, Sept 2009.

[73] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering free space of indoor scenes from a single image," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2807–2814, June 2012.

[74] A. Gupta, M. Hebert, T. Kanade, and D. M. Blei, "Estimating Spatial Layout of Rooms using Volumetric Reasoning about Objects and Surfaces," in *Advances in Neural Information Processing Systems 23* (J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, eds.), pp. 1288–1296, Curran Associates, Inc., 2010.

[75] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, pp. 2366–2374, 2014.

[76] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2650–2658, 2015.

[77] C.-Y. Lee, V. Badrinarayanan, T. Malisiewicz, and A. Rabinovich, "RoomNet: End-To-End Room Layout Estimation," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[78] C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa, "PlaneNet: Piecewise Planar Reconstruction from a Single RGB Image," *arXiv preprint arXiv:1804.06278*, 2018.

[79] R. Cabral and Y. Furukawa, "Piecewise planar and compact floorplan reconstruction from images," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 628–635, IEEE, 2014.

[80] S. Ikehata, H. Yang, and Y. Furukawa, "Structured Indoor Modeling," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1323–1331, 2015.

[81] H. Yang and H. Zhang, "Efficient 3D Room Shape Recovery From a Single Panorama," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[82] G. Pintore, F. Ganovelli, E. Gobbetti, and R. Scopigno, *Mobile Mapping and Visualization of Indoor Structures to Simplify Scene Understanding and Location Awareness*, pp. 130–145. Cham: Springer International Publishing, 2016.

[83] S. Ikehata, I. Boyadzhiev, Q. Shan, and Y. Furukawa, "Panoramic Structure from Motion via Geometric Relationship Detection," *arXiv preprint arXiv:1612.01256*, 2016.

[84] C. Zou, A. Colburn, Q. Shan, and D. Hoiem, "LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image," *arXiv preprint arXiv:1803.08999*, 2018.

[85] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*, pp. 746–760, Springer, 2012.

[86] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3D reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 169, 2013.

[87] T. Cavallari and L. Di Stefano, *On-Line Large Scale Semantic Fusion*, pp. 83–99. Cham: Springer International Publishing, 2016.

[88] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust Reconstruction of Indoor Scenes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[89] M. Fehr, F. Furrer, I. Dryanovski, J. Sturm, I. Gilitschenski, R. Siegwart, and C. Cadena, "TSDF-based change detection for consistent long-term dense reconstruction and dynamic object discovery," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5237–5244, IEEE, 2017.

[90] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "ElasticFusion: Real-time dense SLAM and light source estimation," *The International Journal of Robotics Research*, vol. 35, no. 14, pp. 1697–1716, 2016.

[91] A. Dai, M. Nießner, M. Zollöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Reintegration," *ACM Transactions on Graphics 2017 (TOG)*, 2017.

[92] C. Liu, J. Wu, and Y. Furukawa, "FloorNet: A Unified Framework for Floorplan Reconstruction from 3D Scans," *arXiv preprint arXiv:1804.00090*, 2018.

[93] C. Liu, J. Wu, P. Kohli, and Y. Furukawa, "Raster-to-Vector: Revisiting Floorplan Transformation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2195–2203, 2017.

[94] C. Liu, A. G. Schwing, K. Kundu, R. Urtasun, and S. Fidler, "Rent3D: Floor-Plan Priors for Monocular Layout Estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[95] S. Wang, S. Fidler, and R. Urtasun, "Holistic 3D Scene Understanding From a Single Geo-Tagged Image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[96] E. Wijmans and Y. Furukawa, "Exploiting 2D Floorplan for Building-scale Panorama RGBD Alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 308–316, 2017.

[97] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys, "Live Metric 3D Reconstruction on Mobile Phones," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[98] K. Kolev, P. Tanskanen, P. Speciale, and M. Pollefeys, "Turning Mobile Phones into 3D Scanners," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[99] I. Dryanovski, M. Klingensmith, S. S. Srinivasa, and J. Xiao, "Large-scale, real-time 3D scene reconstruction on a mobile device," *Autonomous Robots*, vol. 41, pp. 1423–1445, Aug 2017.

[100] M. Klingensmith, I. Dryanovski, S. Srinivasa, and J. Xiao, "Chisel: Real Time Large Scale 3D Reconstruction Onboard a Mobile Device using Spatially Hashed Signed Distance Fields.," in *Robotics: Science and Systems*, vol. 4, 2015.

[101] T. Schöps, T. Sattler, C. Häne, and M. Pollefeys, "3D Modeling on the Go: Interactive 3D Reconstruction of Large-Scale Scenes on Mobile Devices," in *2015 International Conference on 3D Vision*, pp. 291–299, Oct 2015.

[102] T. Schöps, T. Sattler, C. Häne, and M. Pollefeys, "Large-scale outdoor 3D reconstruction on a mobile device," *Computer Vision and Image Understanding*, vol. 157, pp. 151 – 166, 2017. Large-Scale 3D Modeling of Urban Indoor or Outdoor Scenes from Images and Range Scans.

[103] C. Strecha, M. Krull, and S. Betschart, "The Chillon Project: Aerial / Terrestrial and Indoor Integration," tech. rep., Pix4D, June 2014.

[104] J. Xiao and Y. Furukawa, "Reconstructing the world's museums," *International journal of computer vision*, vol. 110, no. 3, pp. 243–258, 2014.

[105] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernández, and S. M. Seitz, "Accurate geo-registration by ground-to-aerial image matching," in *2014 2nd International Conference on 3D Vision*, vol. 1, pp. 525–532, IEEE, 2014.

[106] T. Koch, M. Korner, and F. Fraundorfer, "Automatic Alignment of Indoor and Outdoor Building Models Using 3D Line Segments," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 10–18, June 2016.

[107] M. Hofer, M. Maurer, and H. Bischof, "Line3d: Efficient 3d scene abstraction for the built environment," in *German Conference on Pattern Recognition*, pp. 237–248, Springer International Publishing, October 2015.

[108] A. Cohen, J. L. Schönberger, P. Speciale, T. Sattler, J.-M. Frahm, and M. Pollefeys, *Indoor-Outdoor 3D Reconstruction Alignment*, pp. 285–300. Cham: Springer International Publishing, 2016.

[109] A. Cohen, A. G. Schwing, and M. Pollefeys, "Efficient Structured Parsing of Facades Using Dynamic Programming," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[110] K. Chen, Y.-K. Lai, and S.-M. Hu, "3D indoor scene modeling from RGB-D data: a survey," *Computational Visual Media*, vol. 1, pp. 267–278, Dec 2015.

[111] M. Berger, A. Tagliasacchi, L. M. Seversky, P. Alliez, G. Guennebaud, J. A. Levine, A. Sharf, and C. T. Silva, "A Survey of Surface Reconstruction from Point Clouds," *Computer Graphics Forum*, vol. 36, no. 1, pp. 301–329, 2017.

[112] L. Gimenez, J.-L. Hippolyte, S. Robert, F. Suard, and K. Zreik, "Reconstruction of 3D building information models from 2D scanned plans," *Journal of Building Engineering*, vol. 2, pp. 24–35, 2015.

[113] J. Beneš, T. Kelly, F. Děchtěrenko, J. Křivánek, and P. Müller, "On realism of architectural procedural models," in *Computer Graphics Forum*, vol. 36, pp. 225–234, Wiley Online Library, 2017.

[114] F. Lafarge, "Some new research directions to explore in urban reconstruction," in *Urban Remote Sensing Event (JURSE), 2015 Joint*, pp. 1–4, IEEE, 2015.

[115] N. Schertler, M. Tarini, W. Jakob, M. Kazhdan, S. Gumhold, and D. Panozzo, "Field-aligned Online Surface Reconstruction," *ACM Trans. Graph.*, vol. 36, pp. 77:1–77:13, July 2017.

[116] C. Häne and M. Pollefeys, "An overview of recent progress in volumetric semantic 3D reconstruction," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 3294–3307, Dec 2016.

[117] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, *et al.*, "Building rome on a cloudless day," in *European Conference on Computer Vision*, pp. 368–381, Springer, 2010.

[118] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[119] J. Heinly, J. L. Schonberger, E. Dunn, and J.-M. Frahm, "Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset)," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3287–3295, 2015.

[120] "Agisfot photoscan." http://www.agisoft.com/. Accessed: 2017-11-15.

[121] D. Hauagge, S. Wehrwein, P. Upchurch, K. Bala, and N. Snavely, "Reasoning about Photo Collections using Models of Outdoor Illumination," in *Proceedings of BMVC*, 2014.

[122] A. Romanoni, A. Delaunoy, M. Pollefeys, and M. Matteucci, "Automatic 3D reconstruction of manifold meshes via delaunay triangulation and mesh sweeping," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–8, March 2016.

[123] N. Schneider, L. Schneider, P. Pinggera, U. Franke, M. Pollefeys, and C. Stiller, *Semantically Guided Depth Upsampling*, pp. 37–48. Cham: Springer International Publishing, 2016.

[124] A. Hermans, G. Floros, and B. Leibe, "Dense 3D semantic mapping of indoor scenes from RGB-D images," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2631–2638, May 2014.

[125] F. Schweiger, B. Zeisl, P. F. Georgel, G. Schroth, E. G. Steinbach, and N. Navab, "Maximum Detector Response Markers for SIFT and SURF," in *International Workshop on Vision, Modeling and Visualization (VMV)*, 2009.

[126] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards internet-scale multi-view stereo," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1434–1441, IEEE, 2010.

[127] M. Kazhdan, A. Klein, K. Dalal, and H. Hoppe, "Unconstrained isosurface extraction on arbitrary octrees," in *Symposium on Geometry Processing*, vol. 7, pp. 256–263, 2007.

[128] L. Ladickỳ, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical random fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1056–1077, 2014.

[129] T. Koch, P. d'Angelo, F. Kurz, F. Fraundorfer, P. Reinartz, and M. Korner, "The TUM-DLR Multimodal Earth Observation Evaluation Benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 19–26, 2016.

[130] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Reconstructing building interiors from images," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 80–87, Sept 2009.

[131] J. Brandt, "Transform coding for fast approximate nearest neighbor search in high dimensions," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1815–1822, June 2010.