Universidad
Carlos III de Madrid

TESIS DOCTORAL

DESARROLLO Y EVALUACIÓN DE ESTRATEGIAS PARA APLICACIONES DE PERCEPCIÓN 3D USANDO CÁMARAS DE RANGO

Autor:
Silvia Rodríguez Jiménez
Director:
Mohamed Abderrahim

DEPARTAMENTO DE INGENIERÍA DE
SISTEMAS Y AUTOMÁTICA

Leganés, Septiembre 2015

TESIS DOCTORAL (THESIS)

DESARROLLO Y EVALUACIÓN DE ESTRATEGIAS PARA APLICACIONES DE PERCEPCIÓN 3D USANDO CÁMARAS DE RANGO

Autor (Candidate): Silvia Rodríguez Jiménez

Director (Adviser): Mohamed Abderrahim

Tribunal (Review Committee)

Presidente (Chair): _____

Vocal (Member): _____

Secretario (Secretary): _____

Suplente (Substitute): _____

Título (Grade): Doctorado en Ingeniería Eléctrica, Electrónica y Automática

Calificación: _____

Leganés, 29 Septiembre de 2015

Esta tesis ha sido parcialmente financiada por el proyecto europeo HANDLE, dentro del Séptimo Programa Marco de la Comunidad Europea (FP7/2007-2013) en virtud del acuerdo de subvención ICT 231640. También ha sido parcialmente financiada por una beca de Formación de Personal Investigador de la Universidad Carlos III de Madrid (PIF-UC3M ref. 03-1213).

Abstract

The evolution of low cost depth sensors in recent years has allowed a more widespread use of range cameras. They provide depth information in real time so their application has been extended to various areas of object perception where three-dimensional (3D) knowledge of the environment is mandatory to interact with it. For this reason, there is a growing demand for visual perception systems to exploit the information from range devices. However, most existing approaches are limited in unstructured environments, which are common in the real world.

The objective of this thesis is to contribute to the progress of 3D object perception in these environments, addressing open challenges like lack of prior information or picture noise. A methodology to develop strategies for complex environment has been designed to solve the challenges, exploiting the characteristics of the scene. The strategy is validated following a methodology based on scenarios, so three environments have been selected with different complexity. These are sufficiently significant and complementary. Each context is part of a group of perception unstructured scenarios, classified by environmental conditions and the degree of information known a priori. For each scenario, a strategy is proposed, which is developed and evaluated thoroughly in an experimental context. This has also taken into account that hardware and software should be appropriated to the requirements and environmental needs. The approach described in this thesis decreases uncertainty by focusing on a particular context. Furthermore the state space is reduced without affecting the task to be performed.

Resumen

La evolución de sensores de profundidad de bajo coste en los últimos años ha permitido un uso más generalizado de las cámaras de rango. Su aplicación, gracias a que proporcionan información de profundidad en tiempo real, se ha ampliado a diversos ámbitos de percepción de objetos donde es necesario el conocimiento del entorno en tres dimensiones (3D) para poder interaccionar con él. Por este motivo, existe una demanda creciente de sistemas de percepción visual capaces de explotar la información procedente de los dispositivos de rango. Sin embargo, la mayoría de los enfoques existentes presentan limitaciones en entornos poco estructurados, comunes en el mundo real.

El objetivo de esta tesis es contribuir al progreso de sistemas eficaces de percepción 3D de objetos en este tipo de entornos, abordando diversos retos aún sin resolver como la falta de información *a priori* o ruido en las imágenes. Para ello se ha diseñado una metodología que permite desarrollar estrategias para entornos complejos con el objetivo de dar solución a los desafíos de percepción planteados, explotando al máximo las características de la escena. Para validar la estrategia planteada se ha seguido una metodología guiada por escenarios, por lo que para su evaluación se han seleccionado tres entornos con diferente complejidad, y lo suficientemente significativos a la vez que complementarios entre sí. Cada contexto se enmarca dentro de un grupo de escenarios semiestructurados de percepción visual, clasificados según condiciones ambientales así como grado de información conocida *a priori*. Para cada escenario se propone una estrategia, la cual se desarrolla y se evalúa exhaustivamente en un contexto experimental, seleccionando el hardware así como el software a implementar más adecuado a los requisitos y necesidades del entorno. El enfoque descrito en esta tesis permite reducir la incertidumbre centrándose en un determinado contexto para plantear la estrategia a seguir, reduciendo el espacio de estados sin afectar a la tarea a desempeñar.

A mis abuelos.

El mérito no es ser un luchador incansable, sino estar cansado y seguir luchando.
Mafalda.

Agradecimientos

El desarrollo de esta tesis ha sido una carrera de fondo, con diferentes etapas en las que sin el apoyo de las personas que han estado a mi alrededor no hubiese sido posible llegar a la meta. En primer lugar, me gustaría agradecer a Mohamed su apoyo y guía durante todo el desarrollo de esta tesis así como en los proyectos en los que hemos participado.

Aunque la tesis se ha desarrollado en la UC3M, tengo que remontarme a mis años de investigación en UNIOVI ya que, como en toda carrera, siempre hay una fase de calentamiento. Nacho, siempre te agradeceré la oportunidad que me diste de trabajar en tu grupo. Mi motivación por la visión por computador empezó durante mi proyecto fin de carrera contigo. Gracias a ti, y a compañeros como Chema, Ricardo, y especialmente Jorge, asenté una importante base de conocimientos de visión y óptica. Durante esa fase tuve la ocasión de hacer una estancia en TU Delft bajo la supervisión del Prof. Albert J.P. Theuwissen. Albert, Gayathri, Bernhard and Ning Xie, thanks for those marvellous months!

Tras esa etapa, empieza la maratón en la UC3M. Muchísimas gracias a mis compañeros del Departamento. Lo mejor de esa fase no sólo es lo mucho que he aprendido, sino también, las grandes personas que he conocido y cuya amistad espero que perdure durante años. Nico, te mereces una mención especial, eres un crack! Desde el primer día en que nos conocimos siempre has estado apoyándome y enseñándome, incluso a miles de kilómetros de distancia... gracias por todo!

Además, en este tiempo he tenido la gran oportunidad de trabajar en diversos proyectos, por lo que me gustaría aprovechar estas líneas para agradecer a todos los miembros del consorcio del proyecto HANDLE, PROSAVE2 y SARBOT su gran colaboración. Gracias especialmente los profesores Luis M. Bergasa y Manuel Ocaña, de la UAH, por acogernos durante un mes para la competición robótica. Y por supuesto, a mis compañeros del equipo por convertir aquellos días en memorables.

En el último período, trabajando en una empresa, nunca he caminado sola. Gracias a Adán, Ravín y Octavio por el apoyo recibido, fundamentalmente para dar el último empujón. A los integrantes de la Cátedra y del B105 Lab por todos los ánimos. Alvaro, tu figura ha sido clave...muchísimas gracias por tu ayuda!

Por último, aunque no menos importante, me gustaría agradecer a mis amigos y especialmente a mi familia su apoyo desde el kilómetro cero, sin descanso. Gracias a mis padres, a mi hermano y a Marcos por estar a mi lado en todo momento, sin vosotros no hubiese sido posible. Gracias de todo corazón!

Abreviaturas

2D	**2 D**imensiones
3D	**3 D**imensiones
AR	**A**erial **R**efueling – Repostado en Vuelo
AUC	**A**rea **U**nder the **C**urve – Área Bajo la Curva
CDF	**C**umulative **D**istribution **F**unction – Función de Distribución Acumulativa
DoF	**D**egrees **o**f **F**reedom – Grados de Libertad
DRC	**D**arpa **R**obotics **C**hallenge
EtherCAT	**E**thernet for **C**ontrol **A**utomation **T**echnology
FPR	**F**alse **P**ositive **R**ate – Tasa de Falsos Positivos
FPS	**F**rames **P**or **S**egundo
GPS	**G**lobal **P**ositioning **S**ystem – Sistema de Posicionamiento Global
KCL	
NAR	**N**on-**A**mbiguity **R**ange – Rango sin Ambigüedad
NFA	**N**úmero de **F**alsas **A**larmas
NTP	**N**etwork **T**ime **P**rotocol
OCU	**O**perator **C**ontrol **U**nit – Unidad de Control del Operador
PCL	**P**oint **C**loud **L**ibrary
PDF	**P**robability **D**ensity **F**unction – Función de Densidad de Probabilidad
PFA	**P**robabilidad de **F**alsas **A**larmas
PROSAVE2	**PRO**yecto de investigación en **S**istemas **A**vanzados para a**V**iones más **E**co-Eficientes
RANSAC	**RAN**dom **S**ampling **C**onsensus
RGB	**R**ed **G**reen **B**lue – Color (Rojo Verde Azul)
RGB-D	**R**ed **G**reen **B**lue - **D**epth – Color - Profundidad

ROC	**R**eceiver **O**perating **C**haracteristic – Característica Operativa del Receptor
ROS	**R**obot **O**perating **S**ystem
SARBOT	**S**earch **A**nd **R**escue ro**BOT**
ToF	**T**ime-**o**f-**F**light – Tiempo de Vuelo
TPR	**T**rue **P**ositive **R**ate – Tasa de Verdaderos Positivos
UC3M	**U**niversidad **C**arlos **III** de **M**adrid
UPMC	**U**niversidad **P**ierre y **M**arie **C**urie de París
VRC	**V**irtual **R**obotics **C**hallenge
W	**W**att – Vatios

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

1.1. Contexto

La creciente demanda de contenido en tres dimensiones (3D) de la industria del entretenimiento y la mejora en la tecnología están potenciando el crecimiento del mercado de las cámaras de rango, como se desprende del artículo sobre las tendencias globales de estos dispositivos 3D [1]. En 2020 se espera que el mercado global de estas cámaras alcance 7.661,8 millones de dólares (Fig. 1.1). De hecho, numerosas empresas han identificado las imágenes 3D como una forma de innovación en su oferta de productos en el mercado de cámaras digitales (Nikon, Go Pro, Sony Corp., Canon, Panasonic Corp., LG Electronics Inc., Samsung Electronics Corp., Fujifilm Corp., Kodak y Faro Technologies).

Las cámaras de rango han sido utilizadas para diversas aplicaciones de visión por computador durante años, pero el alto precio y la baja calidad en algunos casos de estos dispositivos han limitado su aplicabilidad. La incursión y evolución de sensores de profundidad de bajo coste en los últimos cinco años (Fig. 1.2), como la Kinect o cámaras de Tiempo de Vuelo (*Time of Flight*, ToF), han permitido un uso más generalizado de la tecnología [2]. Debido a que este tipo de dispositivos genera imágenes de rango que proporcionan información de profundidad en tiempo real, se ha potenciado su aplicación a diversos ámbitos donde es necesario el conocimiento del entorno en 3D para poder interaccionar con él: navegación autónoma, interacción hombre-máquina, robótica médica e industrial, manipulación y agarre robótico, entre otros.

Actualmente hay un importante auge para capturar el mundo en 3D con el objetivo de almacenar, intercambiar y analizar información visual 3D. Los fabricantes de cámaras se centran fundamentalmente en las tecnologías de ToF, la visión

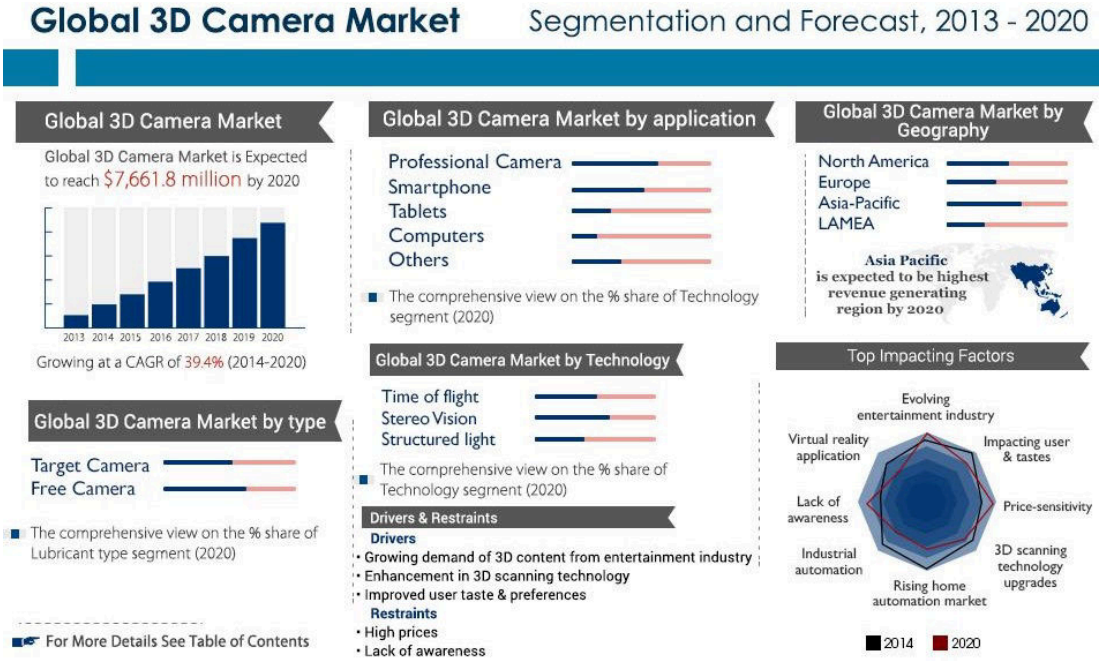


Figura 1.1: *Infografía del reporte sobre las tendencias globales de las cámaras 3D [1].*

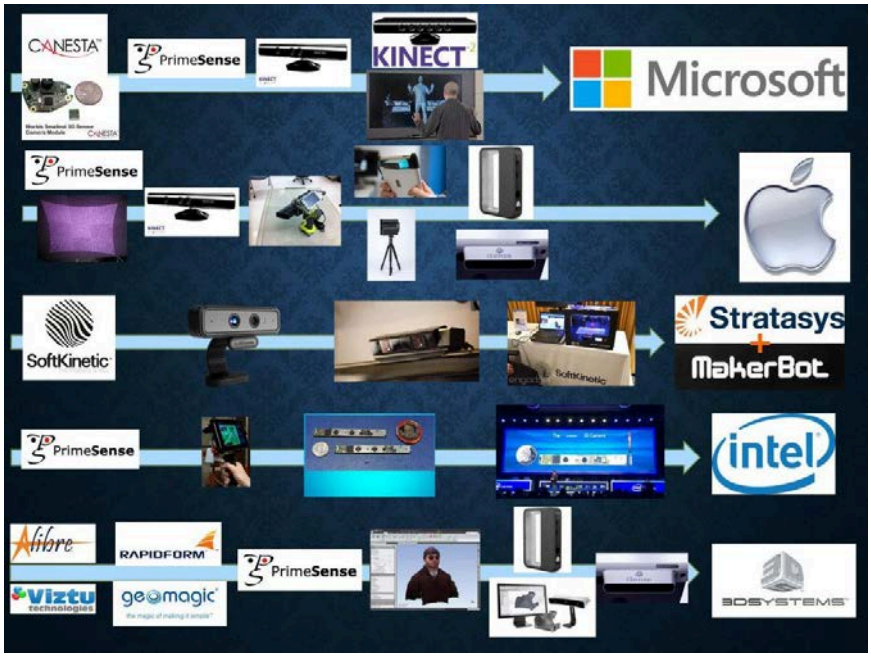


Figura 1.2: *Evolución de diversos dispositivos de rango de bajo coste [3].*

estéreo y luz estructurada (o codificada). Además, se espera una creciente adopción en smartphones y tablets (Fig. 1.1), con una estimación de implantación de la tecnología al 80 % de smartphones en 2018 [1]. Hay varios ejemplos representativos de esta tendencia. Occipital recaudó en 2013 un millón de dólares en su campaña de Kickstarter para desarrollar su sensor Structure para tablets [4]. Google está inmersa en el Proyecto Tango, con el que pretende incorporar sensores de profundidad en los móviles [5] y Apple Inc. adquirió el año pasado la compañía PrimeSense, que desarrolló el sensor de la Kinect [6].

Por tanto, es evidente el crecimiento de esta tecnología y esta tendencia en la evolución de los dispositivos de rango es muy prometedora para la percepción 3D de objetos. Sin embargo, para entender el mundo que rodea al sistema, no basta sólo con las imágenes de profundidad sino que una vez adquiridos los datos es necesario segmentar la escena para detectar, reconocer y estimar la posición del objeto con el que interaccionar. En este sentido, existe una demanda creciente de sistemas inteligentes capaces de estructurar y explotar esa información procedente de los dispositivos de rango. Esto es debido a que, a pesar de los avances significativos en sistemas de percepción visual 3D, la mayoría de los enfoques existentes presentan limitaciones en entornos poco estructurados, comunes en el mundo real.

1.2. Motivación y Alcance

La percepción visual de objetos ha progresado enormemente desde los primeros trabajos realizados en la década de los 60 [7]. Desde entonces se ha trabajado enormemente con diversos enfoques para mejorar la robustez de la solución [8, 9, 10, 11]. Estas técnicas suelen basarse en métodos de extracción de características a partir de dos dimensiones (2D), 3D y la combinación de ambos. Para 2D caben destacar los algoritmos de características fotométricas *Scale-Invariant Feature Transform* (SIFT) [12], *Speeded-Up Robust Features* (SURF) [13] y *Histograms of Oriented Gradients* (HOG) [14]. Debido a que se basan en propiedades de apariencia del objeto (color, textura, intensidad, esquinas) no funcionan correctamente en determinadas situaciones, como por ejemplo objetos poco texturizados. En este caso, los métodos de extracción de características a partir de modelos 3D solventan estas dificultades mediante descriptores geométricos como formas, líneas, superficies normales, curvaturas y planos [15]. También existen descriptores de objetos globales, como *Viewpoint Feature Histogram* (VFH) [16] que es una extensión de *Fast Point Feature Histograms* (FPFH) [17], e integra la información del punto de vista en las características geométricas 3D. Sin embargo, le falta la componente de color por lo que puede seguir existiendo ambigüedad. Para lograr mayor robustez y eficacia se desarrollaron descriptores que combinan la información geométrica y de

apariencia como el *Viewpoint Oriented Color-Shape Histogram* (VCSH) [15].

Teniendo en consideración estos enfoques, la robustez actual de las cámaras de profundidad, así como la sincronización, en muchas de ellas, de la información de profundidad junto con la de color, abre nuevas oportunidades para resolver los problemas fundamentales de la visión por computador, incluyendo reconocimiento, detección, seguimiento de objetos, localización, etc. Sin embargo, cuando se quiere trabajar en el mundo real, la mayoría de los desarrollos disponibles presentan dificultades en entornos poco estructurados [15, 18] donde por ejemplo, apenas hay información conocida *a priori*. Dado el estado actual de la técnica, los siguientes desafíos puedan identificarse como abiertos (aunque no son los únicos):

- Eficacia frente a información parcial de la escena.
- Obtención del modelo de objetos desconocidos.
- Percepción de objetos independientemente del grado de textura.
- Robustez ante escenas con condiciones dinámicas.
- Tiempo de respuesta acorde al ámbito de ejecución.
- Detección eficaz en imágenes con ruido.

Por tanto, en estos ambientes poco estructurados, donde no se tiene pleno conocimiento sobre el entorno, la percepción en sí misma se convierte en uno de los principales retos debido a la incertidumbre inherente.

Motivada por estos desafíos, esta tesis tiene como objetivo contribuir al progreso de sistemas eficaces de percepción 3D de objetos con especial énfasis en entornos poco estructurados. El alcance de esta tesis se enmarca en abordar los retos comentados aún sin resolver, mediante la implementación de estrategias para aplicaciones de percepción bajo un modelo global de sistema visual con cámaras de rango. Esta investigación busca proporcionar la forma de superar las limitaciones de estos ambientes tan complejos dotando de la solución hardware y software que mejor se adapte a los diferentes niveles de estructuración, de forma eficaz y robusta.

1.3. Estructura

Para el desarrollo de esta tesis se sigue una metodología de investigación que engloba primeramente un marco teórico. Éste establece la base para validar las estrategias desarrolladas mediante diversas etapas de propuesta, implementación y evaluación ante los retos de percepción. Siguiendo esta metodología, esta tesis

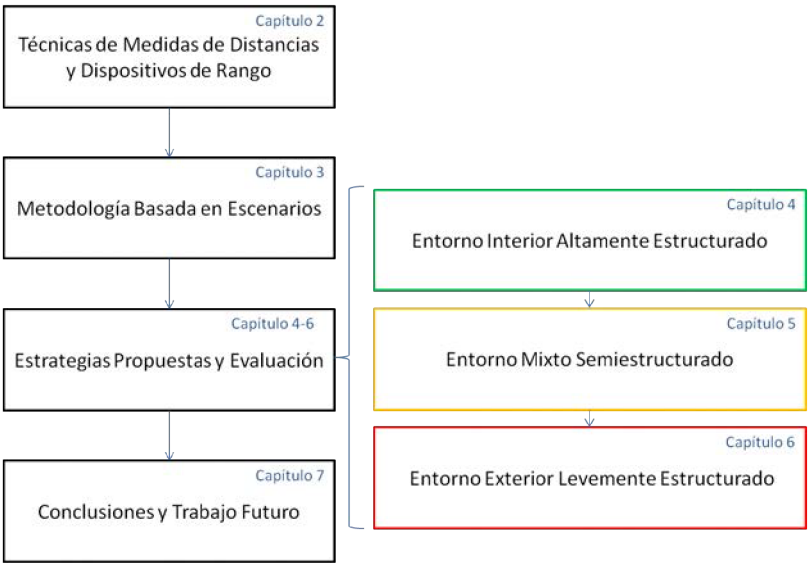


Figura 1.3: *Diagrama de esta tesis.*

está organizada en siete capítulos incluyendo al actual. La Fig. 1.3 muestra los seis capítulos principales cuya distribución valida las fases de base teórica así como propuesta, desarrollo y evaluación de las estrategias.

Para poder interaccionar con el entorno es necesario primeramente adquirir los datos que permitan conocer el escenario. El Capítulo 2 resume las diferentes técnicas de medida de distancia centrándose en aquellos dispositivos ópticos de profundidad cuyo principio de funcionamiento, rango y precisión son más acordes para tareas de interacción con objetos. Estas cámaras de rango podrán emplearse en multitud de contextos pero no son válidas para todas las aplicaciones. Por tanto, es necesario conocer los requisitos y necesidades del entorno para la elección del dispositivo hardware más adecuado así como del software a implementar.

El Capítulo 3 analiza estos entornos para diseñar la metodología que permita desarrollar estrategias para entornos complejos con el objetivo de dar solución a los desafíos de percepción aún abiertos. Por ello, se sigue una estrategia que permita explotar al máximo las características de los escenarios poco estructurados. En los siguientes tres capítulos (4, 5 y 6) se evalúan las estrategias propuestas para cada entorno bajo estudio, incrementándose el nivel de complejidad según las condiciones ambientales cambiantes así como la estructurada conocida *a priori*. Por último, en el Capítulo 7 se recopilan las conclusiones de las estrategias desarrolladas para superar este gran desafío en la visión por computador para aplicaciones de percepción 3D. Además, se resumen una lista de posibles modos de mejorar las estrategias propuestas como futuras líneas de investigación.

Capítulo 2

Técnicas de Medidas de Distancias y Dispositivos de Rango

En esta tesis se aborda el desarrollo de estrategias de visión 3D para diversos ámbitos de aplicación. Las características y requisitos del entorno influyen tanto en la elección del hardware como en los algoritmos software a implementar.

La estimación de profundidad en escenas dinámicas es un desafío para la visión por computador. Para poder interaccionar con el entorno es necesario disponer de la información relevante de los objetos de la escena. Para ello, el primer paso a realizar es la adquisición de los datos que permitan conocer el escenario. La información de profundidad se puede obtener mediante técnicas con o sin contacto [19]. Los métodos basados en contacto habitualmente tienen una punta montada sobre un brazo robótico o directamente en un sistema de posicionamiento tridimensional muy preciso. Las coordenadas 3D se determinan tocando la superficie mediante la punta. Debido a que su localización espacial está determinada con mucha precisión, estos métodos son muy precisos pero, sin embargo, son lentos, suelen tener un alto coste y no son válidos para objetos frágiles. En cambio, en los últimos años los métodos sin contacto han mejorado notablemente en precisión, como se explicará en la siguiente sección.

2.1. Técnicas de Medida Sin Contacto

Los métodos sin contacto se dividen en técnicas basadas en trasmisión (recogen el tipo de radiación trasmitida por el objeto) y en reflexión (recogen el tipo de radiación que emitieron), comúnmente más usadas estas últimas. Como se puede ver en la taxonomía de la Fig. 2.1 [20], las técnicas basadas en reflexión (*reflective*) a

Figura 2.1: *Clasificación de las técnicas de medidas de distancia [20].*

su vez se pueden clasificar en ópticas y no ópticas, atendiendo a la longitud de onda empleada. Debido a las limitaciones de difracción de las primeras, las técnicas de microondas y sonar tienen propiedades de resolución angular bastante limitadas, por lo que habitualmente se usan para aplicaciones de Global Positioning System (GPS) o Synthetic Aperture Radar (SAR) [21]. En comparación con estas técnicas, los métodos ópticos son más fiables para medir distancias en términos de alta precisión y resolución [22].

Las técnicas ópticas pueden obtener la información emitiendo y recibiendo algún tipo de radiación (activas) o detectando luz ambiente visible, sin ningún tipo de emisión (pasivas). Los métodos pasivos determinan distancias con algoritmos basados en propiedades de los objetos en la imagen, como distancias entre puntos de un objeto, profundidad de enfoque/desenfoque, formas a partir de sombras o siluetas, etc.. Aunque la aplicación de los métodos activos está restringida a escenarios donde puede haber iluminación especial, esto simplifica algunos pasos en el proceso de captura 3D. En la Fig. 2.2 se muestra la diferencia entre técnicas de triangulación activa y pasiva.

Como se ha visto, diferentes tipos de técnicas ópticas para obtener medidas de distancias están disponibles en la literatura y su clasificación no es única [23]. Teniendo en cuenta su principio de funcionamiento, rango y precisión pueden clasificarse en las siguientes tres categorías [24] (Tabla 2.1):

- **Triangulación**: detección de profundidad por medio de la medición del ángulo entre dispositivos por geometría.

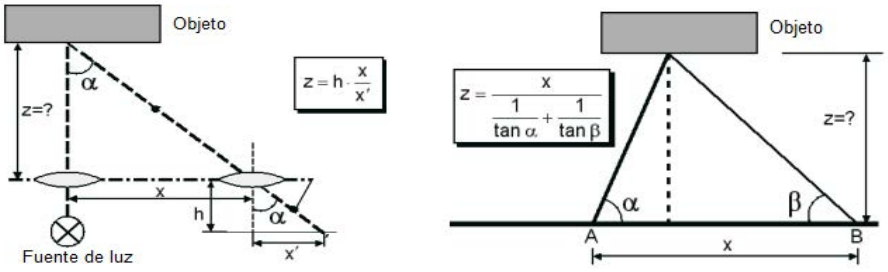


Figura 2.2: *Triangulación: (a) activa; (b) pasiva.*

Técnica	Triangulación	Retraso de tiempo	Pasivo	Activo
Triagulación láser	X			X
Luz estructurada	X			X
Visión estéreo	X		X	
Interferometría		X		X
Tiempo de Vuelo		X		X

Tabla 2.1: *Clasificación de técnicas ópticas de rango.*

- **Interferometría**: detección de profundidad por medio de la medición óptica coherente.
- **Tiempo de Vuelo** (*Time of Flight* (ToF)): detección de profundidad por medio de la modulación de medición óptica de tiempo de vuelo.

En relación a esta clasificación se han propuesto muchas soluciones para obtener la estimación de profundidad total de una escena. La interferometría permite una precisión muy alta en la estimación de la distancia que está básicamente relacionada con la longitud de onda de la fuente de luz coherente que sirve para iluminar la escena activamente. Sin embargo, los sistemas interferométricos requieren una configuración óptica voluminosa y precisa, y su rango de profundidad es limitado centrándose en objetos de menos 1 cm. Las técnicas de triangulación (por ejemplo, la visión estéreo, triangulación láser, y luz estructurada) ofrecen un excelente rendimiento en términos de precisión, pero el rango de profundidad máximo está limitado ya que está determinada por la longitud de línea de base de triangulación. Las técnicas ToF se basan en la medición (directa o indirecta) del tiempo

Figura 2.3: *Comparativa de técnicas de visión estéreo, luz codificada y ToF.*

necesario para una señal óptica para viajar desde un origen a un destino y volver a un sensor.

Debido a que las técnicas interferométricas se centran más en aplicaciones de inspección superficial o calidad [25], con el objetivo de abordar técnicas más comunes para interaccionar con objetos se van a explicar en detalle las técnicas de visión estéreo, luz codificada (basado en luz estructurada) y cámaras de ToF (Fig. 2.3), que se explicarán en detalle en las próximas subsecciones.

2.1.1. Visión Estéreo

Las técnicas de visión estéreo se basan en el análisis de dos o más imágenes de la misma escena desde diferentes perspectivas. Debido a que se conocen las posiciones de las cámaras entre sí, se pueden establecer relaciones geométricas entre los puntos 3D y sus proyecciones sobre las imágenes en 2D. Estas relaciones se basan en la restricción geométrica epipolar y en la suposición de que las cámaras pueden ser aproximadas por el modelo *pinhole*, que relaciona los puntos 3D con una imagen 2D a través de parámetros extrínsecos e intrínsecos.

La gran ventaja de esta técnica es el bajo coste de hardware sin embargo suele presentar problemas en zonas con poca textura. Para solventar los problemas ante objetos poco texturizados han aparecido las cámaras estéreo con patrón proyectado. Estos dispositivos integran dos sensores CMOS o CCD y un proyector que genera un patrón de puntos aleatorio con una textura estática y muy contrastada. Por tanto, facilita la detección de estructuras no visibles al disponer de varias perspectivas y permite realizar capturas de superficies sin textura [26].

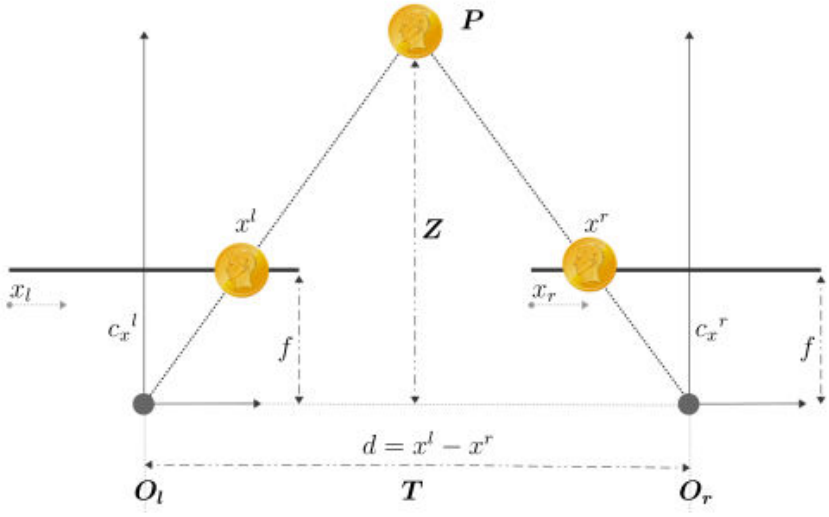


Figura 2.4: *Principio de visión estéreo [27].*

2.1.1.1. Principio de Funcionamiento

Para determinar la restricción geométrica y obtener la información en 3D es necesario que los parámetros de la cámara se obtengan mediante calibración. Si los puntos coincidentes entre varias imágenes son conocidos, el punto en 3D puede ser reconstruido a partir de dos o más imágenes en 2D mediante triangulación. Como se aprecia en la Fig. 2.4, si los puntos de las cámaras izquierda y derecha se conocen, x^l y x^r respectivamente, sus líneas de proyección son también conocidas. Si esos dos puntos corresponden al mismo punto 3D, P, entonces las líneas de proyección se cruzaran precisamente en P. La distancia Z vendrá determinada por $Z = \frac{fT}{d}$ siendo d la disparidad, f la distancia focal y T la distancia entre los centros de las cámaras.

2.1.1.2. Tipos de Imágenes Adquiridas

Las imágenes que proporciona una cámara estéreo generalmente son (Fig. 2.17):

- **Profundidad**: teniendo en cuenta el principio de funcionamiento explicado en la subsección anterior, cada píxel proporciona una distancia de la cámara al objeto observado.
- **Color**: cada píxel representa el color correspondiente en la escena. Se suele dar la opción de mostrar la imagen a color o en escala de grises. Debido a que se tienen dos cámaras a color, se podría utilizar la imagen que el usuario desee.

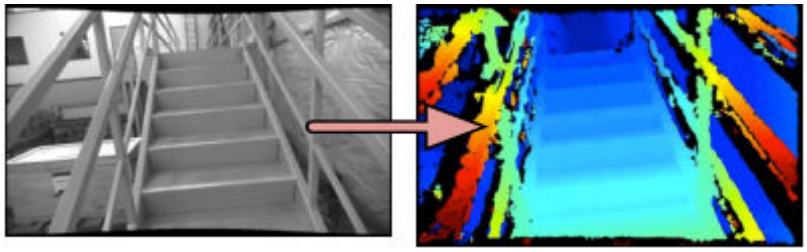


Figura 2.5: *Imágenes de la cámara estéreo de la cabeza sensora MultiSense-SL [28]: imagen a color en escala de grises y la imagen de profundidad a su derecha.*

El procesamiento de los datos mediante algoritmos estéreos transforma la imagen de las cámaras izquierda y derecha en mapas de profundidad. Muchas de las cámaras comerciales ya tienen este procesamiento integrado para proporcionar la nube de puntos a color RGB-D en tiempo real [28].

2.1.2. Tiempo de Vuelo

Las cámaras ToF son dispositivos compactos que proporcionan mapas de profundidad densos en tiempo real, calculando la distancia hacía un objeto por medio de evaluación directa o indirecta del tiempo, desde que la luz es emitida por el sistema hasta que se recibe [29].

Los dispositivos de ToF emiten luz infrarroja (IR) y usan un sensor de luz (CCD o CMOS) para medir el tiempo que la luz tarda en ser reflejada por un objeto y volver de vuelta a la cámara [30]. Dado que la velocidad de la luz es conocida y constante, la distancia de los objetos a la cámara se puede estimar directamente. Para medir este tiempo de vuelo existen diferentes tecnologías [31]:

- **Indirecta** (*phase-shift*): emiten una onda IR modulada y miden el cambio de fase de la señal para cada píxel al reflejarse en los objetos de la escena. La profundidad del píxel se puede deducir directamente a partir de este desfase. La mayor parte de las cámaras ToF corresponden a este tipo.
- **Directa**: emiten un pulso IR y utilizan obturadores ópticos rápidos para medir la cantidad de luz recibida durante un período de tiempo corto. En este caso, la cantidad de luz es directamente proporcional a la distancia del objeto. Esta tecnología, usando *optical shutter*, era utilizada por las cámaras comercializadas anteriormente por 3DV Inc. [32] y Canesta 3D [33]. Ambas compañías fueron adquiridas por Microsoft con el fin de desarrollar el dispositivo Kinect que se explicará en detalle en la próxima subsección.

La mayoría de las cámaras también proporcionan simultáneamente imágenes de intensidad de niveles de gris e incluso imágenes en color para algunos de ellos.

Las imágenes de intensidad pueden ser directamente calculadas a partir del módulo de señal modulada en métodos basados en fase. En el caso de los métodos de obturador óptico necesitan un sensor adicional, generalmente una cámara de color clásica. En este caso, técnicas de registro específicas son necesarias para calcular las correspondencias entre los píxeles de color y profundidad.

Los dispositivos de Tiempo-de-Vuelo comparten las siguientes propiedades:

- **Sensibilidad a variaciones de luz**: como las cámaras ToF emiten su propia luz IR, no requieren ninguna iluminación externa. Sin embargo, si los objetos son demasiado distantes u oscuros, la luz emitida puede ser demasiado baja para asegurar una buena relación señal a ruido (SNR). Para tales aplicaciones, algunas cámaras integran módulos de iluminación adicionales. Dado que la estimación de profundidad sólo se basa en la luz IR, todos los dispositivos son bastante insensibles a las variaciones de iluminación de interior. Para aplicaciones al aire libre, la luz infrarroja emitida por el sol es un grave problema, por lo que se requieren algoritmos específicos de supresión de la luz de fondo.

- **Desenfoque de movimiento**: este problema es más específico en los métodos de fase debido a que se requieren varias mediciones sucesivas para estimar la fase de la señal reflejada. Si el objeto se mueve demasiado rápido, pueden aparecer inconsistencias entre las muestras.

- **Interferencias**: si se utilizan varias cámaras al mismo tiempo, las señales IR emitidas puede interferir. Los recientes sistemas basados en fase pueden soportar múltiples cámaras mediante el uso de diferentes frecuencias de modulación.

A continuación se explica en mayor detalle el principio de funcionamiento de cámaras con tecnología directa e indirecta, comentando las características técnicas de alguna de las más representativas.

2.1.2.1. Cámara ToF Directa: 3D Flash LIDAR

Las cámaras *Direct ToF*, también conocidas como *Light Detection and Ranging* (LIDAR), miden con un sensor infrarrojo (IR) cuánto tiempo tarda en volver el pulso de luz que se refleja en un objeto [30] (Fig. 2.6).

La distancia, d, objeto$-$cámara se puede estimar directamente del tiempo entre emisión y detección del pulso, t, puesto que la velocidad de la luz, c, es conocida y constante:

$$d = \frac{ct}{2} \tag{2.1}$$

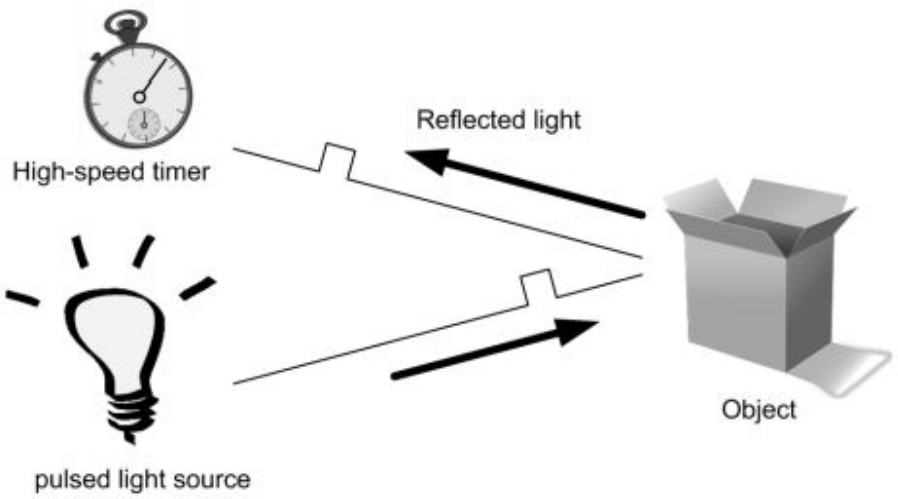


Figura 2.6: *Método de medida de distancia para cámaras ToF Directa.*

Figura 2.7: *Modelos de 3D Flash LIDAR de la empresa ASC 3D: (a) Dragoneye; (b) TigerEye; (c) Portable; (d) Peregrine; (e) Tigercub; (f) Goldeneye.*

El número de cámaras comerciales disponibles es reducido debido a que son de reciente aparición. A finales del 2011 la empresa Advanced Scientific Concepts Inc. (ASC 3D) [34] anunció tres modelos 3D Flash LIDAR y actualmente tiene otros 3 disponibles (Fig. 2.7).

Como características generales caben destacar que cada píxel actúa independientemente y proporciona un valor de profundidad e intensidad, empleando un láser de 1570 nm. Además, tiene una configuración denominada SULAR, que permite obtener datos aunque haya niebla o humo. Esto es debido a que se reduce drásticamente el ruido al eliminar el trigger inicial y producirse el muestreo del pulso simultáneamente en todos los píxeles a unos incrementos específicos. La gran ventaja de esta tecnología es que funciona en exteriores hasta 1500 m, con una tasa de hasta 30 fps, siendo el campo de visión de 45º para distancias inferiores de 60 m. La precisión es de 5 cm a una distancia de 1m, y de en torno a 60cm a 4km, siendo el sensor de 128 x 128 píxeles.

Se han empleado de forma exitosas en diversos entornos [34], tanto aéreos

Figura 2.8: *Modelos de cámara ToF indirectas: (a) D-IMager de Panasonic, (b) FOTONIC-C70 by Fotonic, (c) DepthSense 311 de Optrima-SoftKinetic, (d) PMD[vision] CamCube 3.0 de PMDTechnologies, (e) SwissRanger 4000 by MESA Imaging, (f) 3D MLI Sensor by IEE S.A., (g) TOFCam Stanley P-301DM, (h) TriDiCam Application kit.*

como terrestres, aunque aún en aplicaciones muy específicas debido a sus características.

2.1.2.2. Cámara ToF Indirecta: PMD CamCube

Estas cámaras de Tiempo de Vuelo Indirecto (*Indirect ToF*) experimentaron un uso extendido a partir de 2010 cuando numerosas compañías empezaron a desarrollarlas gracias a los avances tecnológicos. En esa fecha existían numerosos dispositivos basados en fase de las empresas PMDtec [35], Mesa [36], IEE [37], Optrima [38] y Canesta [39] (Fig. 2.8). Estos dispositivos se basan en el cálculo de la distancia a la que se encuentra el objeto mediante la evaluación indirecta del tiempo que transcurre desde que se envía la luz hasta que se recibe procedente de la escena.

La influencia del sol sobre la cámara PMD CamCube es inferior a la que se tienen en otros dispositivos ToF debido a que su tecnología patentada de supresión de fondo de iluminación [40] le aporta mayor robustez. Por otra parte, tiene una gran flexibilidad en cuanto a la potencia de iluminación ya que permite la adición de módulos. Estas particularidades han permitido que la cámara siga en el mercado, evolucionando en los últimos 5 años. Por este motivo, se detallan diversos aspectos sobre el principio de funcionamiento de este tipo de cámaras utilizando la PMD Camcube como modelo representativo. Se explicarán además características técnicas claves del dispositivo.

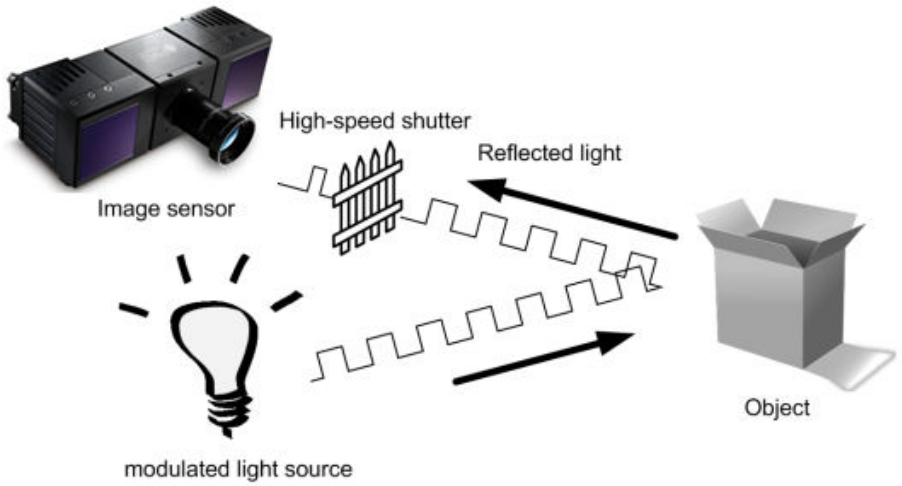


Figura 2.9: *Método de medida de distancia para cámaras AMCW.*

2.1.2.2.1 Principio de Funcionamiento

En las técnicas ToF Indirectas, existen diferentes formas para calcular la distancia una vez que la luz emitida es reflejada por la escena. Los métodos *Shuttered Light Pulse* y *Range Gating Segmentation* miden la amplitud del pulso recibido. El método denominado Onda Continua de Amplitud Modulada, *Amplitude-Modulated Continuous-Wave* (AMCW), al que corresponde la cámara PMD CamCube, mide el cambio de fase de la señal para cada píxel del receptor tras emitir una onda continua de amplitud modulada (Fig. 2.9).

Este desfase, φ, está relacionado con la distancia, d, por medio de la frecuencia de modulación, f_{mod}, y la velocidad de la luz, c, que son constantes [41]:

$$d = \frac{c\varphi}{4\pi f_{mod}} \tag{2.2}$$

En concreto, la PMD CamCube obtiene los valores de profundidad de la escena, iluminándola mediante una luz modulada por una señal cuadrada de amplitud y frecuencia constante proporcionada por módulos de 42 diodos de luz, *Light-Emitting Diodes* (LEDs), de infrarrojo cercano, *Near-Infrared Light* (IR), de 870 nm de longitud de onda. La señal emitida es reflejada por los objetos, y llega hasta el sensor receptor con la misma frecuencia. Sin embargo, tiene una amplitud transformada en función de la reflectividad del objeto y de la luz ambiente así como una fase modificada en función de la distancia entre el objeto y el emisor. Este desfase de la señal reflejada se estima a partir del desfase de activación de los dos acumuladores de luz que hay en cada píxel inteligente en función de la diferencia de la cantidad de luz acumulada en cada uno (Fig. 2.10). Éstos se activan con la

Figura 2.10: *Principio de funcionamiento de la PMD CamCube.*

frecuencia de modulación, durante un tiempo llamado tiempo de integración. El rango soportado por la cámara es de entre 12 y 50000 μs. Si es demasiado bajo, la luz del emisor no tendrá tiempo para rellenar los acumuladores y si es demasiado alto, se podrían saturar.

2.1.2.2.2 Tipos de Imágenes Adquiridas

La cámara PMD CamCube proporciona imágenes de profundidad, amplitud e intensidad:

- **Profundidad**: cada píxel proporciona una distancia de la cámara al objeto observado con una precisión de centímetros, teniendo en cuenta el principio de funcionamiento explicado en la subsección anterior.
- **Amplitud**: esta imagen proporciona una información muy útil ya que valores altos de amplitud corresponden a valores más fiables de distancia. Por tanto, si la cámara observa una escena con buena reflectividad en espectro infrarrojo, entonces los valores de amplitud son altos. Por otro lado, las amplitudes serán cercanas a cero cuando el objeto a evaluar tiene poca reflectividad ya que produce valores de distancia muy ruidosos. Por tanto, para asegurar que los valores de distancia son fiables, se pueden ignorar valores por debajo de un determinado valor de amplitud.

- **Intensidad**: es similar a una simple imagen en escala de grises de una cámara tradicional 2D, donde se muestra la textura y el brillo de la escena. La implicación es que cuanta más luz llegue a un píxel, mayor valor de intensidad tiene. Sin embargo, al disponer de la tecnología de supresión de luz de fondo, que se explica en la subsección 2.1.2.2.4, el valor de intensidad de cada píxel disminuye a pesar de que aumente la cantidad de luz que llega al sensor. Debido a ello, si se quiere mostrar la imagen en escala de grises es más recomendable usar la imagen de amplitud.

2.1.2.2.3 Distancia: Píxeles Inteligentes

Una vez emitida una luz modulada, el array de sensores debe ser capaz de medir con precisión el desfase de la señal que llega a cada píxel. En la cámara PMD CamCube, cada píxel tiene su propio hardware para medir el desfase, y por eso les denominan *smart pixels*.

Un *smart pixel* tiene dos acumuladores de luz, controlados por una puerta (*gate*). La puerta activa un receptor durante la primera mitad del período de modulación, y el otro durante la segunda mitad. De esta forma, los dos acumuladores tienen un desfase de 180°, y midiendo la diferencia entre los dos, se puede estimar el desfase de la señal reflejada. Este dispositivo mide el cambio de fase utilizando cuatro medidas de la onda que regresa, separadas un cuarto de longitud de onda, para eliminar la luz adicional del fondo. Cada medida se hace de dos en dos (primero a 0º y 180º, posteriormente a 90º y 270º), con los dos acumuladores que existen por píxel. Teniendo en cuenta que A_1, A_2, A_3 y A_4 representan la cantidad de luz tomada a intervalos de 0º, 90º, 180º y 270º, respectivamente, el desfase φ entre la señal emitida y la reflejada se puede expresar como la siguiente función de autocorrelación (ACF) [41]:

$$\varphi = \arctan \frac{A_1 - A_3}{A_2 - A_4} \tag{2.3}$$

Este desfase, φ, está relacionado con la distancia, d, por medio de la frecuencia de modulación, f_{mod}, y la velocidad de la luz *299.792.458 m/s*, c, que son constantes, como se indico en la Ecuación 2.2.

A partir de las medidas A_i, también se pueden sacar dos cantidades numéricas [42]

- La amplitud a de la señal, que aporta información sobre la calidad del desfase medido:

$$a = \frac{1}{2}\sqrt{(A_1 - A_3)^2(A_2 - A_4)^2} \tag{2.4}$$

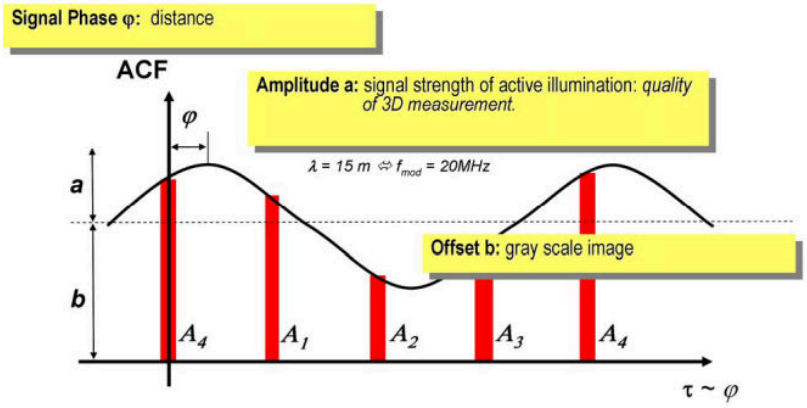


Figura 2.11: *Función de autocorrelación (ACF), desfase de la señal, amplitud and offset. Imagen obtenida del artículo de T. Ringbeck [42].*

- El *offset* b, relacionado directamente con la reflectividad del objeto y representa el valor en escala de grises de cada píxel:

$$b = \frac{\sum_1 + \sum_2 + \sum_3 + \sum_4}{4} \tag{2.5}$$

Si la amplitud es alta, significa que una gran parte de la señal reflejada corresponde a la señal modulada emitida.

2.1.2.2.4 Tecnología de Supresión de la Iluminación de Fondo

Las cámaras ToF habitualmente usan luz infrarroja (IR) para iluminar la escena, por lo que la luz del sol influye en las medidas de distancia dadas por el sistema debido a que tiene más potencia que la iluminación activa. Esto tiene dos impactos negativos [42]:

- Ruido en las medidas de distancia se incrementa.
- El sensor puede saturarse debido a que una gran cantidad de su dinámica es ocupada por la luz del sol.

Por esta razón, este tipo de dispositivos requieren algoritmos de supresión de luz ambiental infrarroja para poder funcionar en exteriores. Trabajos previos [43] demostraron que la cámara PMD CamCube es más robusta frente al sol que otros dispositivos ToF basados en fase. Esto es gracias a su sensor PMD 41k-S2 [44] con tecnología de supresión de la iluminación de fondo llamada SBI (*Supression of Background Illumination*).

La energía infrarroja de la luz del sol es muy superior a lo que se puede emitir con LED. Para reducirlo, se puede añadir un filtro que elimine las radiaciones IR que no corresponden a la longitud de onda de la señal IR emitida, pero aun así, el ratio de potencia es del orden de 400 usando 100 LEDs como emisores (83 W/m^2 frente a 0,2 W/m^2). Esta influencia disminuye la precisión de las medidas, e incluso puede saturar los acumuladores e impedir la estimación de desfase. La tecnología SBI de la cámara PMD CamCube permite reducir varios órdenes de magnitud la influencia de la luz ambiental, haciendo posible obtener medidas 3D en condiciones de luz de sol de hasta 150klux [41]. El SBI añade unidades hardware específicas que acumulan la luz constante recibida y eliminan de los acumuladores todas las señales no correladas instantáneamente durante el proceso de integración. Este proceso evita que se saturen, aumentando el contraste de la señal modulada recibida.

2.1.2.2.5 Rango Máximo Sin Ambigüedad

Al tratarse la fuente de luz de una señal periódica, existe ambigüedad en dicha señal a partir de una cierta distancia en función de la frecuencia utilizada. De esta forma, objetos lejanos aparecen erróneamente a una distancia más cercana de la que están realmente. Por tanto, un cálculo más preciso de la distancia debe tener en cuenta el rango máximo sin ambigüedad, d_u, siendo k un entero:

$$d = d_u\left(\frac{\varphi}{2\pi} + k\right) siendo\, d_u = \frac{c}{2f_{mod}} \tag{2.6}$$

Teniendo en cuenta que el desfase está comprendido entre 0º y 360º, la fase se determina de forma única sólo hasta la mitad de la longitud de onda ya que la iluminación tiene que recorrer el camino dos veces (ida y vuelta). Por tanto, se puede expresar el alcance máximo del sistema en función de la frecuencia de modulación:

$$d_{max} = \frac{c}{2f_{mod}} = \frac{\lambda_{mod}}{2} \tag{2.7}$$

El rango de frecuencias de modulación estándar es de 18 a 21 MHz, siendo la de 20 MHz la que se suele utilizar por defecto. Para una f_{mod} de 20 MHz, por ejemplo, la longitud de onda λ_{mod} es 15 metros por lo que el rango de trabajo típico sin ambigüedad es hasta 7,5 metros. Este dispositivo, frente a otras cámaras ToF presenta la gran ventaja de soportar un amplio margen de frecuencias de modulación. En la Tabla 2.2 se muestra el alcance máximo teórico para cada frecuencia.

La precisión de la estimación de distancia de este dispositivo depende de la frecuencia de modulación, la relación señal a ruido y el contraste de modulación.

f_{mod} (MHz)	40	35	30	25	21	20	19	18	17	16
d_{max} (m)	3,75	4,28	5	6	7,14	7,5	7,89	8,33	8,82	9,37
f_{mod} (MHz)	15	14	13	12	11,5	11	10,5	10	9,5	9
d_{max} (m)	10	10,71	11,54	12,5	13,04	13,63	14,28	15	15,79	16,66
f_{mod} (MHz)	8	7	6	5	4	3	2	1	0,5	0,1
d_{max} (m)	18,75	21,43	25	30	37,5	50	75	150	300	1500

Tabla 2.2: *Tabla con los parámetros de frecuencia de modulación y alcance máximo soportados por la cámara.*

Por tanto, hay que tener en cuenta que:

- A mayor f_{mod}, más difícil es obtener un contraste de luz alto entre los cambios de amplitud.
- A menor f_{mod}, mayor distancia pero se necesita más potencia de iluminación y se tiene peor resolución del alcance, como se verá en la próxima sección.

Además, hay que considerar que la precisión de la medida depende de la frecuencia utilizada, pero de manera inversa a como lo hace el alcance, por lo cual estos conceptos son complementarios, y se debe encontrar un punto de compromiso entre ambos, o bien utilizar medidas de la misma escena con frecuencias de modulación distintas (*multi-frequency-ranging*). Con este modo multi frecuencia, el rango sin ambigüedad efectivo se incrementa el mínimo común múltiplo del máximo rango sin ambigüedad de las frecuencias de modulación usadas, tomando una medida a cada frecuencia [45]. Este método es ilustrado en la Fig. 2.12.

2.1.2.2.6 Precisión del Alcance

La fiabilidad de la medida está directamente influenciada por la cantidad de luz activa que llega al sensor. La precisión de las medidas de rango, $d\,R$, se obtiene [42]:

$$d\,R = \frac{1}{\sqrt{N_{phase}}} \frac{1}{k_{tot}\frac{S}{N}} \frac{\lambda_{mod}}{\sqrt{8}\pi} \tag{2.8}$$

Donde k_{tot} es el contraste de modulación, S el número de electrones de señal, N es el número equivalente de electrones debido al ruido (incluyendo todas las fuentes: luz ambiente, emitida, de fondo, ruido térmico y corriente de oscuridad, etc.), N_{phase} es el número de medidas y λ_{mod} es la longitud de onda de modulación de la señal en metros.

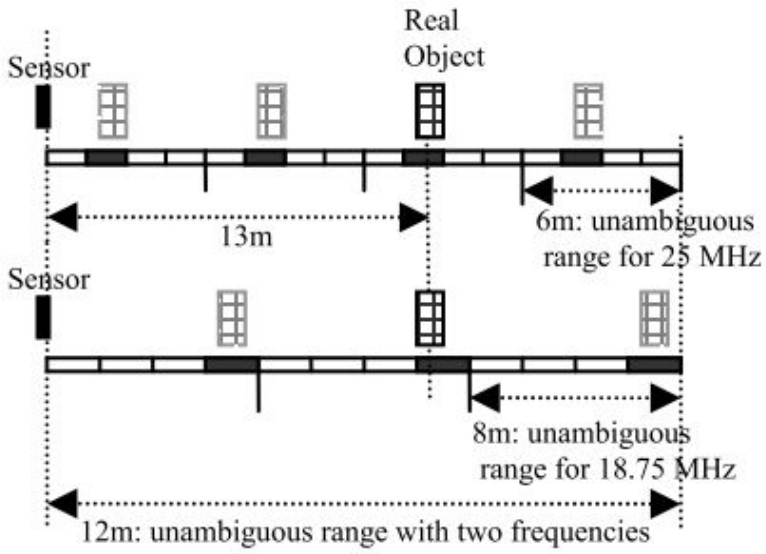


Figura 2.12: *Reducción de ambigüedad usando dos frecuencias de modulación diferentes.*

La fuente principal de ruido procede de las corrientes de oscuridad, que se debe a la generación aleatoria de electrones y huecos que son arrastrados por el campo eléctrico. El ruido debido a la luz ambiente se puede reducir con filtros de banda, solo dejando pasar las ondas de infrarrojo cercano.

El contraste de modulación corresponde al alcanzable entre los dos acumuladores, multiplicado por el contraste de la señal emitida por la de los LEDs. Un alto contraste y modulación mejora la precisión. Sin embargo, aunque la optimización del contraste no afecta a la relación señal-ruido, la precisión disminuye por un factor $\frac{1}{k_{tot}}$. Por tanto, a la hora de modificar el contraste hay que tener en cuenta que éste depende de varios elementos:

- Frecuencia de modulación: cuanto más alta es la f_{mod}, más difícil es obtener un buen contraste.
- El tamaño de los píxeles: cuanto más grande es el píxel, mejor es el poder de separación entre los acumuladores. La arquitectura de PMD permite alargar el tamaño de los píxeles en función de la aplicación.
- La luz ambiente: el contraste está definido como la resta entre los dos acumuladores dividida entre la suma de ambos. Este último valor depende fuertemente de la cantidad de luz ambiente.

Los valores típicos de la precisión $d\,R$ son de 1 cm para f_{mod} de 20 MHz en entornos interiores, y de 10 cm para f_{mod} de 7,5 MHz en entornos exteriores.

2.1.2.2.7 Resolución del Alcance y Factores Limitantes

Es habitual que en las medidas existan algunos errores sistemáticos como la iluminación no homogénea de la escena. En este caso, en el centro se suele concentrar más potencia de iluminación disminuyendo a medida que se aleja del centro de la escena observada. De esta forma, la resolución de la distancia, R_d, viene determinada por el número de divisiones en las que se puede discretizar el alcance sin ambigüedad. Según Gokturk et al. [45] puede ser descrito por la siguiente ecuación:

$$R_d = \frac{C}{2f_{mod}}\sqrt{\frac{A}{P_{laser}k_{opt}q_e\rho T}} \tag{2.9}$$

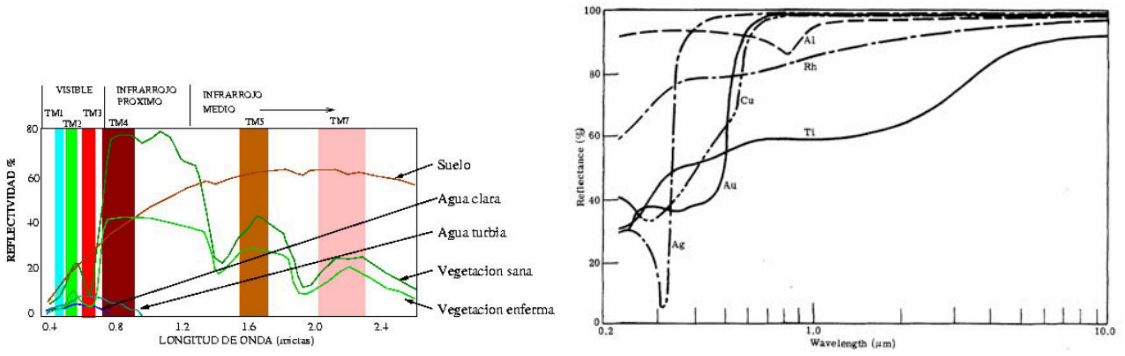
donde f_{mod} es la frecuencia de modulación, q_e la eficiencia cuántica, T el tiempo de integración, P_{laser} la potencia de iluminación, ρ reflectividad del objeto, A el área iluminada, k_{opt} es una constante determinada por las propiedades ópticas del sistema (lentes, difusor, tamaño del píxel) y C una constante. Esta ecuación revela que la resolución del alcance se ve influenciada por la cantidad de luz incidente en el sensor, dependiente de la relación señal-ruido así como de la reflectividad de los objetos a la longitud de onda de trabajo en la dirección del sensor. Por tanto, también influye el ángulo de inclinación de la luz sobre los objetos. Teniendo estos aspectos en cuenta, la resolución puede ser mejorada:

- Incrementando el tiempo de integración, considerando el valor máximo de saturación del sensor.
- Incrementando la potencia de iluminación.
- Minimizando el área iluminada.
- Incrementando la frecuencia de modulación, aunque esto reduce el rango sin ambigüedad, dando lugar a un alto grado de *aliasing*.

2.1.2.2.8 Desviación Estándar del Alcance

Para cuantificar la habilidad que tiene el sensor de la cámara para diferenciar dos objetos a distinta distancia se utiliza el término varianza que es inversamente proporcional a la irradiancia de la luz recibida [46]. En general en todos los dispositivos AMCW se expresa como:

$$\theta_r = \frac{\lambda d^2}{\rho \cos\alpha} \tag{2.10}$$



(a) **(b)**

Figura 2.13: *Curvas espectrales para diversos elementos. (a) Superficie terrestre. (b) Plata (Ag), oro (Au), cobre (Cu), aluminio (Al), rodio (Rh) y titanio (Ti). Imagen tomada de [47]).*

donde θ_r es la desviación estándar del alcance d, λ la longitud de onda (en este caso, 870 nm), ρ la reflectividad del objeto, y α el ángulo de incidencia.

Cada material tiene una curva de reflectancia espectral, donde se expresa su reflectividad en función de la longitud de onda. En la Fig. 2.13 se muestran diversas gráficas de elementos que se pueden encontrar en las imágenes de contextos aéreos, siendo de interés el valor de reflectancia para 870 nm.

Por tanto, la reflectividad de los objetos a la longitud de onda de trabajo en la dirección del sensor influye en la cantidad de luz recibida, y por tanto, en el alcance máximo al que pueden ser vistos por una cámara ToF. Este hecho ha sido observado y evaluado en otras cámaras ToF de tecnología AMCW [48], comprobando que ciertos materiales provocan problemas de saturación o de luminosidad en las cámaras ToF [49], así como exceso de ruido, siendo la información de medida más débil.

2.1.2.2.9 Comparativa de Modelos

La compañía PMDTechnologies GmbH ha desarrollado dos versiones de la cámara PMD CamCube: 2.0 y 3.0. Aunque ambas comparten la forma de obtener los valores de profundidad de la escena, el modelo 3.0 posee un nuevo sensor que permite una alta precisión incluso con tiempos de integración bajos. De esta forma, se consiguen valores adecuados a mayores distancias, y a una mayor tasa de adquisición, por lo que es más adecuado el nuevo modelo para el escenario bajo estudio y en general, para aplicaciones en exterior.

Además, en la versión 3.0 tanto el firmware como el software están actualizados por lo que se pueden conseguir en un solo disparo hasta 4 frames con diferentes

tiempos de integración y frecuencias, obteniendo así una gran variedad de datos. La lente de la versión 3.0 tiene una relación focal (F) menor que el de la 2.0, lo que significa que su apertura es mayor, y por tanto, permite que llegue mayor cantidad de luz al sensor. En la Tabla 2.3 se muestran las características generales de ambas versiones.

Parámetros	**CamCube 2.0**	**CamCube 3.0**
Sensor PhotonICs (SBI)	PMD 41k-S (45 x 45 μm/píxel)	PMD 41k-S2 (45 x 45 μm/píxel)
Tamaño de imagen (píxeles)	204 x 204	200 x 200
Repetibilidad (θ)	<3mm@d=2m, 90 % reflexión	<3mm@d=4m, 75 % reflexión
Frames por segundo (3D)	25 fps	40fps@200x200 píxeles;
	-	60 fps@176 x 144 píxeles;
	-	80 fps@160 x 120 píxeles
Lente montura CS	f = 12,8 mm; F = 1,4	f = 12,8 mm; F=1,1
Temperatura soportada (ºC)	0º-50º	0º-50º
Campo de Visión	40º x 40º	40º x 40º

Tabla 2.3: *Tabla comparativa de las dos versiones de la PMD CamCube.*

2.1.3. Luz Codificada

Este tipo de dispositivos de luz codificada o estructurada fueron inicialmente desarrollados en el 2002 [50] y también se les conoce como luz texturizada. En estos sistemas un transmisor emite un patrón preestablecido, proyectándolo como una textura sobre la escena real. El receptor registra las deformaciones del patrón sobre los objetos de la escena, que proporcionan la composición espacial de la escena así como las distancias a las que se encuentran los objetos.

Para evitar que los usuarios vean el patrón proyectado y reducir las interferencias producidas por fuentes de luz externas en interiores se ilumina la escena con un patrón infrarrojo (IR) especial, y una cámara IR calcula la posición de cada elemento del patrón. Esta tecnología es muy prometedora para aplicaciones en interiores pero de momento no puede funcionar en entorno exterior debido a que la componente infrarroja de la luz del sol interfiere. Existen diferentes cámaras como Kinect [51] de Microsoft y Wavi Xtion [52] de Asus (sin cámara color), basadas en la tecnología de PrimeSense [53].

Debido al bajo coste de la Kinect y a la disponibilidad de un kit de desarrollo software, se popularizó rápidamente su uso en campos de investigación de visión por computador y concretamente en aplicaciones robóticas [54]. Por tanto, ya que la Kinect fue la cámara de este tipo de tecnología que revolucionó el mercado, se van a explicar sus características a modo representativo para las demás.

(a) **(b)**

Figura 2.14: *Cámara Kinect. (a) Elementos fundamentales. (b) Detalle de todos los elementos.*

2.1.3.1. Cámara Comercial: Kinect

La cámara Microsoft Kinect se engloba en el grupo de dispositivos de rango de patrón proyectado o luz estructurada. Esta cámara salió al mercado como periférico de la consola Xbox360 de Microsoft en noviembre de 2010 como resultado de las investigaciones realizadas durante el proyecto Natal [55].

Además de una matriz de múltiples micrófonos, los elementos fundamentales de la Kinect son (Fig. 2.14):

- **Soporte móvil**: se trata de un pie que soporta la cámara y cuya inclinación es graduable con el fin de mejorar y facilitar el enfoque de la escena.
- **Sensor de profundidad 3D**: permite obtener información 3D de la escena mediante la combinación de un proyector de luz infrarroja (IR) y una cámara IR cuyo sensor es CMOS MT9M001 de Aptina Imaging. Los dos componentes del sensor de profundidad se encuentran alineados con ejes ópticos paralelos a lo largo del eje horizontal del dispositivo, a una distancia de 75 mm.
- **Cámara RGB**: cámara digital estándar de color, con sensor CMOS MT9M112 de Aptina Imaging, que proporciona información de color a los datos obtenidos con la cámara de profundidad.

La combinación de la información procesada procedente de la cámara de color y profundidad proporcionará una nube de puntos a color RGB-D, como se explicará en la próxima subsección.

2.1.3.1.1 Principio de Funcionamiento

El cálculo de los datos de profundidad se basa en un principio similar al de triangulación activa entre emisor y cámara. Sin embargo, en este caso se basa en luz

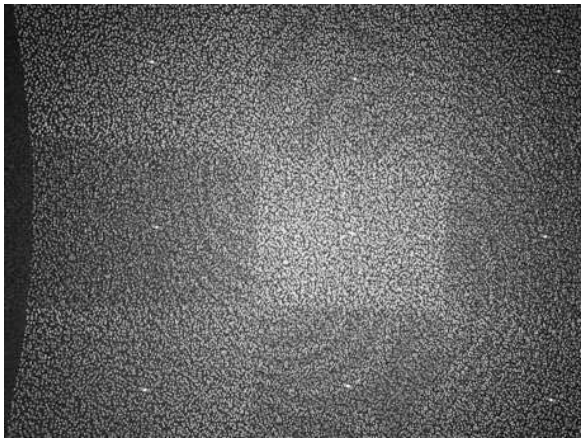


Figura 2.15: *Patrón proyectado de la Kinect.*

estructurada ya que utiliza un patrón de luz infrarroja conocido para proyectarlo sobre la escena. Por tanto, el proceso combina la luz estructurada y el principio estéreo (*depth from stereo*) como se explica a continuación [56]:

- **Calibración**: el proyector de luz infrarroja proyecta un patrón de puntos específico (Fig. 2.15) sobre un plano a una distancia conocida del sensor. Mediante la cámara de IR, este patrón de referencia se almacena en la memoria del dispositivo. Este paso se realiza en fábrica por lo que la Kinect ya viene calibrada.

- **Funcionamiento**: para adquirir la imagen de profundidad se compara el patrón proyectado sobre la escena con el de referencia. De esta forma el cálculo de profundidad se simplifica a un problema de visión estéreo con configuración ideal (misma cámara infrarroja, ejes alineados y separación conocida), estimando la profundidad por un procedimiento de correlación simple de imágenes a partir de la disparidad entre los puntos IR proyectados con los de referencia.

  La cámara puede adquirir información desde 0,3 hasta 5 metros pero el rango de funcionamiento recomendado está entre 1 y 3 metros aproximadamente. Como se aprecia en la Fig. 2.16, el error se incrementa cuadráticamente desde unos pocos milímetros a 0,5 m hasta 4 cm a la máxima distancia alcanzable por el sensor [56].

  Las cámaras de color e infrarrojo están calibradas, siendo sus posiciones conocidas en el dispositivo, por lo que se asigna el color correspondiente a cada dato de profundidad. De esta forma, se obtiene una información 3D a color de la escena.

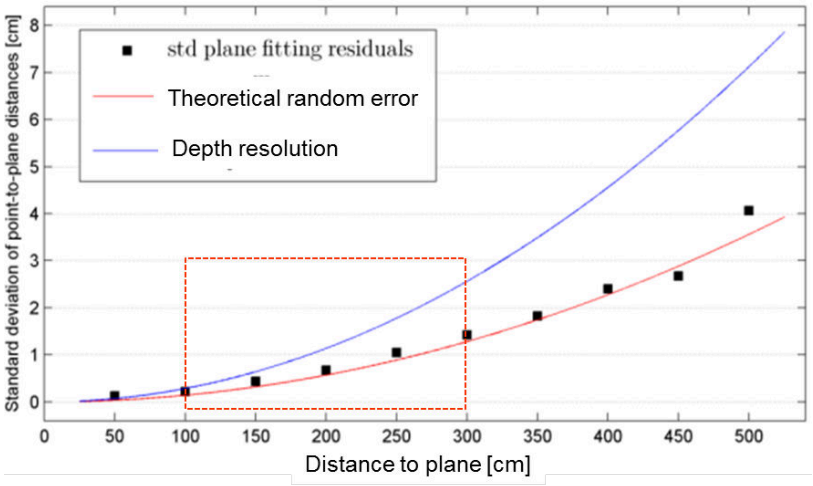


Figura 2.16: *Gráfica de relación entre la distancia de funcionamiento y la desviación.*

2.1.3.1.2 Tipos de Imágenes Adquiridas

La cámara Kinect proporciona dos tipos de imágenes (Fig. 2.17):

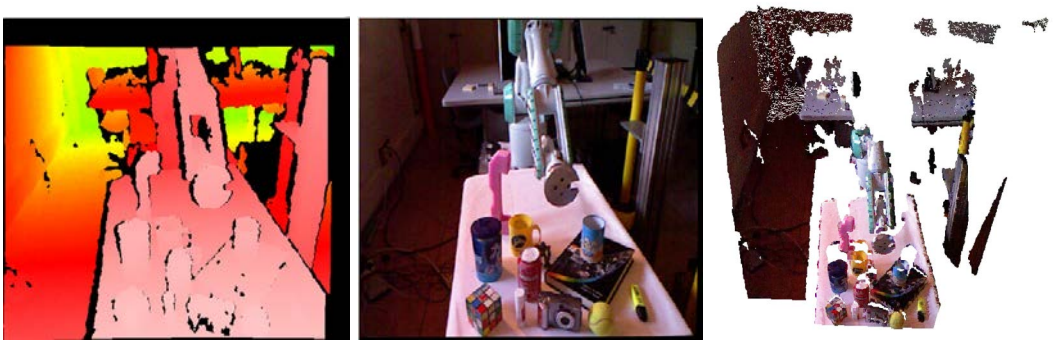
- **Profundidad**: teniendo en cuenta el principio de funcionamiento explicado en la subsección anterior, cada píxel proporciona una distancia de la cámara al objeto observado con una precisión de centímetros.
- **Color**: cada píxel representa el color correspondiente en la escena.

El procesamiento de los datos de ambas cámaras proporciona la nube de puntos a color RGB-D que se muestra en la Fig. 2.17c. Como se puede apreciar, sólo se combinan los datos de las zonas de la imagen de color y profundidad que se solapan. Por otra parte, los píxeles negros en la imagen de profundidad corresponden a aquellas zonas donde no ha sido posible obtener datos de profundidad. Un claro ejemplo es en los bordes de los objetos ya que en ellos es muy complicado que se proyecte un punto entero, por lo que el sistema no es capaz de identificar ese punto proyectado en las imágenes de referencia.

2.1.3.1.3 Características Técnicas

Las propiedades técnicas más relevantes de la cámara Kinect son:

- Flujo de datos:
  - Cámara de profundidad: 320 x 240 @30 fps (16 bits); 640 x 480 @30 fps.



(a) (b) (c)

Figura 2.17: *Cámara Kinect. (a) Elementos fundamentales. (b) Detalle de todos los elementos.*

  - Cámara de color: 640 x 480 píxeles @30 fps (32-bit de color). ´
  - Micrófonos: audio a 16 KHz.
- Campo de visión (Horizontal x Vertical x Diagonal): 57º x 43º x 70º (HxVxD)
- Resolución espacial (X x Y x Z): 3 x 3 x 10 mm.
- Rango de detección y funcionamiento: 1,2m−3,5m (Xbox 360); 0,4m−3,5m (kit de desarrollo SDK).
- Rango de inclinación física: +/−27º.
- Detección de uniones para el seguimiento de personas: hasta 20 uniones.

A continuación se compararán estas características con el siguiente modelo de reciente aparición.

2.1.3.1.4 Comparativa de Modelos

La cámara Microsoft Kinect explicada en esta subsección fue lanzada en Noviembre de 2010 y sucesivas versiones de desarrollo *Kinect for Windows SDK* fueron publicadas posteriormente. A mediados de 2014 se lanzó la Kinect v2.0 para Xbox One, cuyo principio de funcionamiento se basa en la tecnología ToF explicada en la subsección anterior. En la Tabla 2.4 se muestran las características generales de ambas versiones.

A pesar de que la Kinect v2.0 tiene grandes mejoras, debido a la relativa reciente aparición en el mercado, la mayor parte de las investigaciones relacionadas con la Kinect, como esta tesis, siguen enmarcándose en la primera versión.

Parámetros	Kinect	Kinect v2.0
Tecnología	Patrón proyectado	ToF
Tamaño de imagen color (píxeles)	640 x 480@30fps	1920 x 1080@30fps
Tamaño de imagen profundidad (píxeles)	320 x 240	512 x 424
Máxima distancia de profundidad (m)	En torno a 4,5	En torno a 4,5
Mínima distancia de profundidad (cm)	40	50
Campo de visión (H x V)	57º x 43º	70º x 60º
Motor de inclinación	Sí	No

Tabla 2.4: *Tabla comparativa de las versiones de la Kinect.*

2.2. Discusión

Para poder interaccionar con el entorno es necesario conocer la escena. Para ello, en este capítulo se han visto numerosas técnicas sin contacto. En cuanto a los sistemas de visión estéreo, son relativamente baratos y ampliamente usados gracias a los avances relacionados con la mejora en la calidad de la estimación de la geometría. Sin embargo, los resultados no son aún completamente satisfactorios cuando la información de textura en la escena es limitada. Además, la precisión de la profundidad alcanzable cae rápidamente con la distancia al objeto.

La introducción de las cámaras de Tiempo de Vuelo y de las cámaras de luz proyectada es más reciente. En los últimos años están recibiendo un gran interés debido a que superan algunas limitaciones de técnicas clásicas, como por ejemplo, funcionar correctamente con objetos poco texturizados. Los sistemas ToF son capaces de estimar con suficiente precisión en tiempo real la geometría 3D de la escena, pero también tienen algunas limitaciones como la baja resolución espacial, la incapacidad para hacer frente a las superficies poco reflectantes, y el alto nivel de ruido en sus mediciones. Por otra parte, las cámaras ToF pueden hacer frente a iluminación variable, tanto de día como de noche. Por tanto, estos dispositivos son adecuados para una amplia gama de aplicaciones, desde interfaces hombre−máquina hasta detección de obstáculos para vehículos inteligentes.

La introducción de cámaras de luz proyectada a la industria de los videojuegos, como la Kinect, ha permitido una distribución a gran escala y una disminución de precios significativa en un corto período de tiempo. De todas formas, este tipo de cámaras no valen para todo tipo de aplicaciones debido a la incertidumbre en ciertas zonas de los objetos y su funcionamiento sólo válido en interiores.

Los puntos débiles y fuertes de cada técnica comentados en los párrafos anteriores se resumen en la Tabla 2.5 [23, 57].

Por tanto, para proponer una estrategia de visión 3D en diversos ámbitos de

Técnica	Puntos fuertes	Puntos débiles	Objetos sin textura	Interior	Exterior
Visión estéreo	Simple. Bajo coste. Alta precisión en objetivos definidos. Alta resolución de imagen.	Computación exigente. Limitado a escenas bien definidas. Tasa de datos según cámara.		X	X
Luz codificada	Velocidad alta de adquisición de datos. Volumen de medición intermedio. Bajo coste.	Rendimiento depende generalmente de la luz ambiente. Faltan datos con oclusiones y sombras.	X	X	
Tiempo de Vuelo	Rango de medición mayor. Buena velocidad de adquisición de datos. Rendimiento generalmente independiente de luz ambiental en interior.	Menor resolución. Problemas con objetos poco reflectantes. Ruido.	X	X	Ver cámara

Tabla 2.5: *Comparativa de las técnicas de visión estéreo, luz codificada y ToF.*

aplicación es necesario conocer los requisitos del escenario para seleccionar la cámara de rango más acorde a las características del entorno. Las cámaras detalladas en este capítulo podrán emplearse como dispositivo hardware en multitud de contextos y su elección para la configuración del sistema dependerá de los requisitos y necesidades del entorno bajo evaluación. Teniendo en cuenta este punto, en el próximo capítulo se aborda la metodología basada en escenarios desarrollada en esta tesis.

Capítulo 3

Metodología Basada en Escenarios

Para el diseño de los sistemas de percepción visual es necesario tener en cuenta la diferenciación de ambientes a los que se incorporan. El análisis de estos entornos determinará el diseño de la metodología a aplicar, teniendo en cuenta que las estrategias para entornos complejos suponen un gran desafío en el campo de visión por computador y más concretamente en el contexto robótico. Por ello, esta tesis abordará en las próximas secciones el análisis de diversos ambientes, proponiendo por último qué estrategia seguir para explotar al máximo las características de los escenarios seleccionados.

3.1. Clasificación General de Entornos

Los sistemas de visión por computador se utilizan en diversos ámbitos de aplicación para incorporar información visual a vehículos no tripulados, manipulación robótica y agarre, navegación autónoma, interacción hombre-máquina, entre otros. Las cámaras de profundidad están siendo ampliamente usadas en estas aplicaciones, gracias fundamentalmente a la nueva generación de sensores de profundidad de bajo coste. Sin embargo, como se explicó en el anterior capítulo de esta tesis, debido a las características de cada cámara no todas son idóneas para los mismos escenarios, por lo que es necesario identificar el entorno de trabajo para elegir la más adecuada.

Los sistemas de visión para percepción, principalmente en el campo de robótica, se suelen diferenciar en dos ámbitos según el ambiente: interior y exterior. Cada uno de estos entornos presenta unos requisitos y necesidades diferentes, que repercute tanto en software como en la elección del hardware. Sin embargo, en muchos casos esta distinción no es suficiente por lo que se propone tener en cuenta además cómo de complejo es el entorno de funcionamiento [58]:

- **Entorno estructurado**: pertenecen a esta categoría aquellos escenarios controlados. En general, hacen referencia a la mayoría de ambientes interiores como líneas de producción en fábricas, o ambientes exteriores diseñados específicamente. En este tipo de situaciones no existen ambigüedades y las características del ambiente están perfectamente identificadas.

- **Entorno semiestructurado**: en este caso sólo se pueden realizar ciertas hipótesis sobre el ambiente, ya que existen incertidumbres a las que el sistema visual debe ser capaz de hacer frente. Un claro ejemplo corresponde a entornos urbanos debido a que las carreteras y objetos de circulación son conocidos pero existen objetos dinámicos. En el contexto robótico, la incorporación de robots para tareas domésticas o colaborativas con un humano implica que el robot debe poder actuar de forma lo más autónoma posible sin tener un completo conocimiento del entorno que le rodea. Sin embargo, se pueden hacer ciertas suposiciones para explotar el conocimiento de una determinada tarea o la escena para reducir la complejidad.

- **Entorno no estructurado**: no se puede hacer ningún tipo de suposición por lo que son escenarios muy complejos. Un claro ejemplo es cuando un robot tiene que adquirir la información necesaria del entorno para realizar un conjunto de tareas de forma autónoma, sin ninguna hipótesis ni conocimiento previo de la escena, pudiendo ser incluso dinámica.

De un entorno a otro existen muchos ámbitos intermedios, esto es debido a que incluso entornos no estructurados contienen una cantidad significativa de "estructura" que puede ser explotada.

3.2. Diseño de Metodología

En escenarios poco estructurados no se tiene pleno conocimiento sobre el entorno. Por tanto, la percepción del ambiente se convierte en uno de los principales retos a consecuencia de la alta dimensionalidad del espacio de estados, así como de la incertidumbre inherente.

Una tendencia es realizar sistemas de visión específicos para una tarea haciendo complicado reutilizar ideas en diferentes disciplinas como explican Kragic et al.[59]. Sin embargo, como se comenta en ese mismo artículo, muchos de los sistemas visuales tienen que hacer frente a problemas comunes de segmentación de escena, detección y reconocimiento de objetos, así como su estimación de posición, por lo que es natural contemplar la posibilidad de definir un modelo integrado de sistema.

Por otra parte, a pesar de la complejidad aparente de los entornos no estructurados, es posible seleccionar determinadas características e identificar estructuras relevantes para reducir el espacio de estados sin afectar a la tarea a desempeñar como muestran Katz et al. [60]. Por ejemplo, en ese mismo artículo se plantean dos directrices para el uso de robots en estos escenarios de forma robusta y competente:

- Habilidades centradas en una tarea específica, considerándose todas las áreas técnicas relevantes relacionadas.
- El desarrollo/diseño/aprendizaje de las tareas elementales primeramente para luego desarrollar las complejas.

Por tanto, combinando las ideas propuestas por Kragic et al. [59] y Katz et al. [60] es posible definir los bloques que apoyen un diseño unificado e integrado de sistema visual, adaptando cada bloque a los problemas concretos de cada aplicación. Estos puntos motivan el planteamiento desarrollado en esta tesis de proponer un conjunto de estrategias visuales para cada escenario según la tarea o la información del contexto. Todo ello bajo un mismo modelo global de sistema de percepción para poder ser usado en escenarios con tareas similares.

Por otra parte, Katz et al. [60] confirmaron en un contexto robótico que para tener éxito en entornos no estructurados se tiene que seleccionar cuidadosamente las características específicas de la tarea e identificar la estructura relevante del mundo real.

Teniendo en cuenta estas premisas, se concluye que para esta tesis la metodología guiada por escenarios es adecuada debido a que la elección de entornos suficientemente significativos permitirá validar la estrategia propuesta para el entorno bajo estudio. Este análisis de escenarios permite identificar estrategias que podrán estandarizarse sirviendo estos escenarios como referencia [61].

Partiendo del planteamiento de modelos para el diseño de sistemas tecnológicos de ingeniería [62], se propone el proceso mostrado en la Fig. 3.1 para la definición de los escenarios a evaluar. En esta tesis se aborda el proceso de generación del sistema y su evaluación, dejando las fases de mejora y revisión como trabajo futuro. Por tanto, como se muestra en la Fig. 3.1, la metodología a seguir en esta tesis es:

- Análisis de las condiciones del entorno: se detallará cada condición en la Subsección 3.2.1.
- Selección del escenario según las condiciones del entorno: se explicará cómo se elige cada escenario y sus características generales en la Subsección 3.2.2.

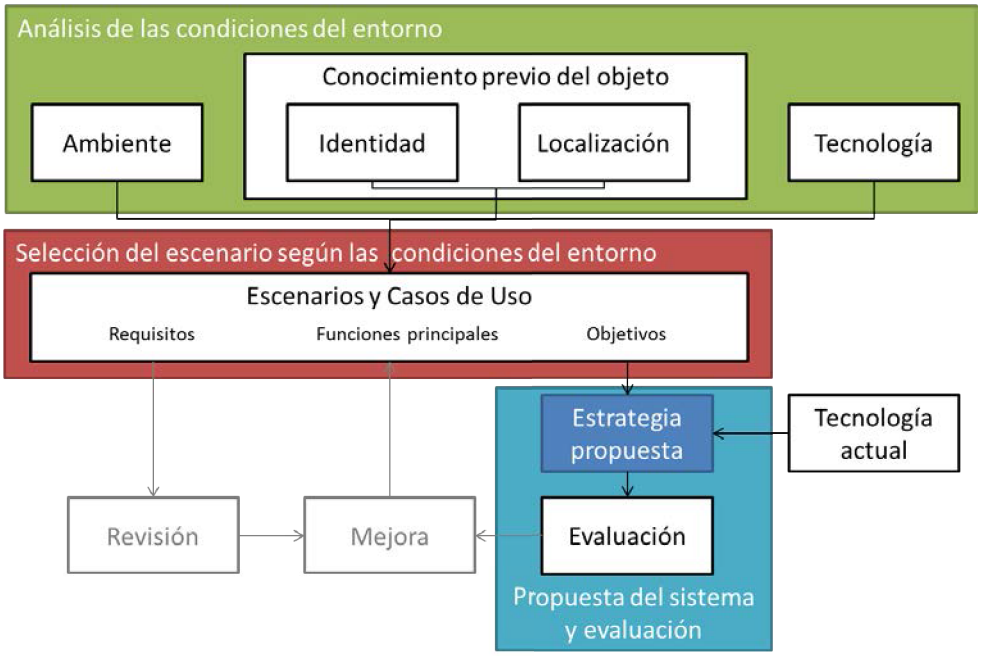


Figura 3.1: *Metodología y proceso de análisis de requisitos del sistema propuesto para la evaluación basada en escenarios. Planteamiento explicado en la Sección 3.2 de forma general, profundizando en cada bloque en las siguientes secciones.*

- Propuesta de la estrategia y evaluación: una vez seleccionados los escenarios, en los próximos capítulos se verá en detalle cada estrategia propuesta y su evaluación experimental.

Como en el caso de Kragic et al. [59], el sistema de visión propuesto para todos los escenarios se basará en uno común que contiene varios bloques. Este planteamiento se ha realizado para que, aunque el sistema se valide para un contexto dentro de un tipo de escenarios, sea lo más general posible para ser usado en otras aplicaciones de percepción visual. Por tanto, se plantean los bloques mostrados en la Fig. 3.2, que incluyen aspectos típicos de un sistema de visión:

- **Calibración**: si no se conoce la posición de la cámara respecto a la plataforma en la que se encuentra, es necesario estimar su posición.
- **Segmentación**: separación de los datos provenientes de la cámara en grupos de puntos u objetos.
- **Detección**: detección de objetos, identificando y extrayendo los píxeles o puntos 3D correspondientes.

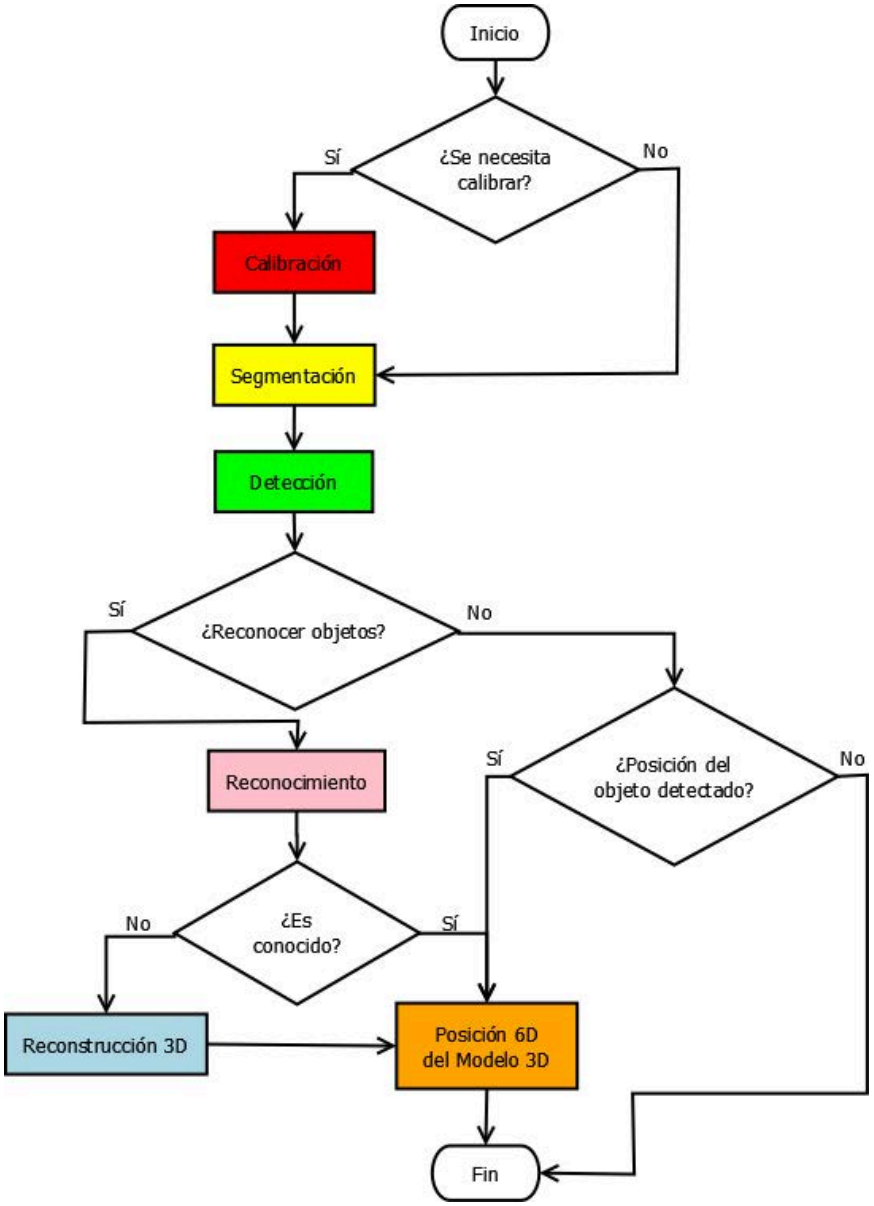


Figura 3.2: *Diagrama general de bloques integrados en un diseño unificado de sistema visual común para los escenarios.*

- **Reconocimiento**: reconocimiento de los objetos según los modelos e imágenes disponibles en una base de datos.

- **Estimación de posición 6D**: estima la posición 6D (vector de traslación y matriz de rotación) del objeto para poder interaccionar con él.

- **Reconstrucción 3D** : en el caso de que no se tenga el modelo 3D del objeto, se reconstruye y se guardará en la base de datos.

Cada uno de estos bloques se detallará en la propuesta del sistema para el entorno bajo estudio, explicando qué bloques son necesarios y planteando tanto el software como el hardware más adecuado para reducir el nivel de complejidad de la escena. De esta forma se abordan los desafíos que presentan los entornos poco estructurados reduciendo la complejidad gracias al planteamiento propuesto.

3.2.1. Análisis de las Condiciones del Entorno

Debido a que en los entornos semiestructurados existen diversas incertidumbres a las que el sistema visual debe ser capaz de hacer frente, es necesario centrarse en un determinado contexto para plantear la estrategia a seguir y simplificar la complejidad. Según la metodología presentada en la sección anterior se evalúan las condiciones claves del entorno:

- **Ambiente**: las variantes ambientales, por ejemplo, la luz, son diferentes según se esté en interiores o exteriores por lo que el sistema debe diseñarse según el fin de la aplicación. La distinción entre ambientes interiores y exteriores es muy clara, por lo que los tres tipos de escenarios a evaluar serán interiores, exteriores y mixtos. Este último caso se debe a que hay aplicaciones, como las relacionadas con robots móviles, que podrían requerir tareas que engloban tanto interior como exterior de un edificio por lo que hay que contemplar cómo abordar este tipo de situaciones.

- **Conocimiento previo del objeto**: considerando como referente ambientes comunes en manipulación robótica, esta tesis se focaliza en los escenarios cuya ambigüedad viene dada por la falta de conocimiento *a priori* acerca de la localización e identificación de los objetos. La combinación de ambas incertidumbres proporciona cuatro grupos de escenarios, cuyo resumen se muestra en la Fig. 3.3 que recoge un caso concreto sobre "con qué interaccionar" y "dónde está" un objeto:

    1. *Esta caja*: la representación interna del objeto ya es conocida y además, se conoce su localización.
    2. *Este objeto*: la localización del objeto es aproximadamente conocida pero no se sabe qué es.
    3. *La caja*: el objeto a interaccionar es conocido pero se desconoce dónde se encuentra.

Percepción de Objetos		Localización: ¿Dónde?	
		Conocido	Desconocido
Identidad: ¿Qué?	Conocido	*Esta caja*	*La caja*
	Desconocido	*Este objeto*	*Algo*

Figura 3.3: *Tabla de combinación de incertidumbres sobre localización e identidad del objeto, basada en la propuesta por Kragic et al. [59]. Se destacan los tres grupos menos estructurados.*

4. *Algo*: sólo se conoce que se desea interaccionar con algo pero no se conoce qué es ni donde se encuentra.

Teniendo en cuenta estos cuatros casos, el primero de todos se refiere a entornos prácticamente estructurados por lo que se descarta su estudio debido a que ha sido ampliamente abordado a lo largo de los últimos años [58]. Por tanto, esta tesis se centra en los otros tres grupos para abordar los desafíos inherentes a la interpretación de entornos semiestructurados. Para ello es necesario plantear un conjunto diferente de estrategias visuales a seguir teniendo en cuenta el conocimiento previo disponible de la identidad del objeto y su ubicación. La superación de estos desafíos así como la toma de decisiones ha motivado la elección de un contexto experimental para validar la estrategia propuesta en cada escenario.

- **Tecnología**: teniendo en cuenta el estado actual de la técnica, esta tesis se focaliza en las cámaras de rango para los escenarios evaluados. La elección del dispositivo hardware vendrá dado por el entorno y se tendrá en cuenta a la hora de proponer el sistema.

A diferencia de la distinción entre ambientes interiores y exteriores, la clasificación en cuanto a nivel de estructuración del entorno no está estrictamente bien definida [58]. A pesar de ello, en ambos casos se han destacado tres tipos de escenarios de interés para abordar en esta tesis. Por tanto, para lograr una selección más completa de los escenarios a evaluar, en la siguiente subsección se realiza una clasificación inicial ambiental, que se combina para cada caso con diversos niveles de complejidad estructural asociados a la combinación de identidad y localización de los objetos.

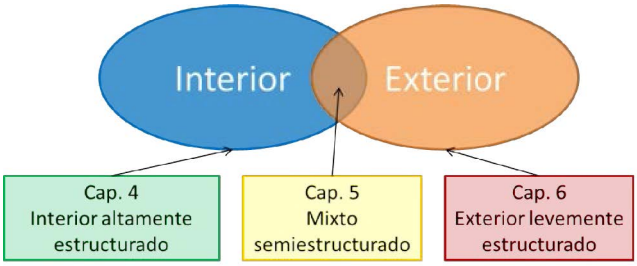


Figura 3.4: *Diagrama de relación entre los capítulos de escenarios y las características del ambiente.*

3.2.2. Selección de Escenarios Según las Condiciones del Entorno

Los siguientes tres capítulos de la tesis hacen referencia a los sistemas implementados para abordar escenarios según aspectos clave ambientales (Fig. 3.4) y el nivel de conocimiento del objeto con el que interaccionar (Fig. 3.5). Este último punto está directamente relacionado con el nivel de estructuración del entorno. Como se puede apreciar, los siguientes tres capítulos se abordan en orden de complejidad de estructuración desde interiores a exteriores. El nivel de complejidad estructural suele aumentar en exteriores debido a que es habitual un mayor número de variables dinámicas, incluyendo condiciones ambientales cambiantes.

Cada sistema que se explica a continuación se engloba dentro de los grupos de escenarios semiestructurados de percepción visual mostrados en la Fig. 3.3, destacados en la Fig. 3.6:

- **Interior altamente estructurado (Capítulo 4): Sistema de Visión para Manipulación Robótica Usando una Sola Vista RGB-D**. Este contexto engloba los escenarios en los que existen objetos tanto conocidos como desconocidos, cuya localización es aproximadamente conocida ya que se sabe que se encuentran encima de una mesa cuya posición es fija respecto al sistema. Este tipo de escenarios son bastante comunes en entornos domésticos, y es frecuente que sólo una parte de los objetos sea visible, ya sea porque la escena contiene muchos elementos o porque sólo se dispone de una cámara sin apenas movimiento.

  La mayoría de los sistemas robóticos en estos casos utilizan la visión para detectar la posición de objetos previamente conocidos, almacenados en una base de datos. Sin embargo, una extensa base de datos de modelos de objetos no siempre es accesible o práctica, por lo que esta estrategia no podrá hacer frente a objetos desconocidos cuya forma debe ser determinada y perderá robustez incluso con objetos conocidos, una vez que ya estén en la mano

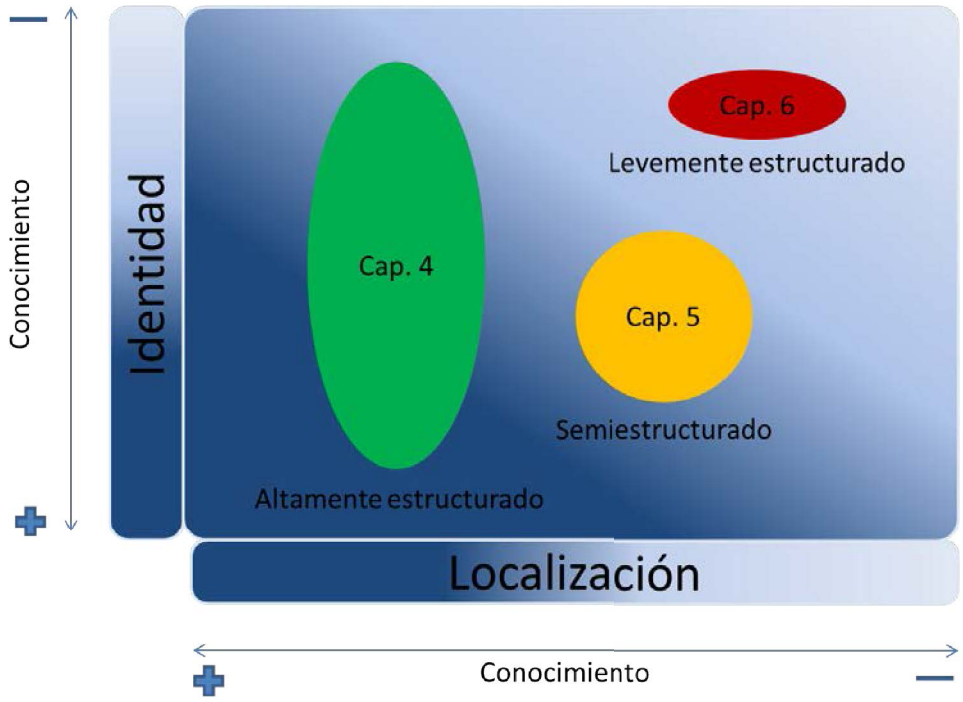


Figura 3.5: *Diagrama de relación entre los capítulos de escenarios y el nivel de conocimiento de la percepción de objetos (su identidad y localización).*

debido a que la propia mano robótica genera oclusiones. Por ese motivo se aborda el desafío de enfrentarse a objetos desconocidos.

Por tanto, el objetivo a abordar en este escenario es mejorar la autonomía del sistema proporcionando una solución a los retos de percepción de objetos con y sin textura, desconocidos y todo ello disponiendo de información parcial al tener sólo una vista de la escena.

- **Mixto semiestructurado (Capítulo 5): Sistema de Percepción 3D para Tareas de Agarre en un Sistema Robótico Móvil**. Este entorno engloba escenarios en los que se tiene que realizar una tarea tanto en interiores como en exteriores, por lo que el gran desafío es tanto el software como el hardware para su correcto funcionamiento en tiempo de ejecución. Por ese motivo es fundamental caracterizar la tarea a realizar. En este caso se engloban escenarios en los que se conoce con qué objeto interaccionar pero no se tiene perfectamente caracterizado donde se encuentra respecto al robot. Además, se añade más complejidad ya que el objeto no está completamente modelado, por lo que sólo se conocen determinadas características que permiten diferenciarlo de otros pero no es posible recurrir a una base de datos donde su modelo esté

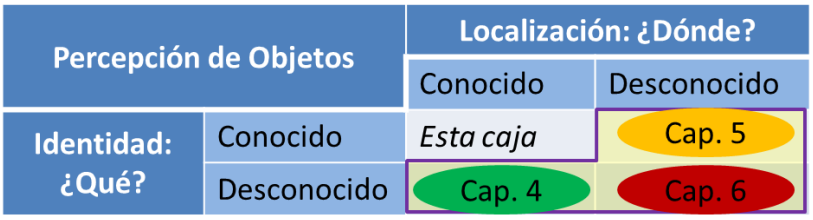


Figura 3.6: *Clasificación de cada capítulo dentro de la tabla de los grupos de escenarios combinando incertidumbres sobre localización e identidad del objeto.*

accesible.

En este caso, la estrategia debe dar solución a los desafíos inherentes a la percepción en condiciones dinámicas y objetos cuyas características no son del todo conocidas, considerando en todo momento las restricciones temporales impuestas para su ejecución.

- **Exterior levemente estructurado (Capítulo 6): Sistema de Percepción con Cámara de Tiempo de Vuelo para Detección de Objetos en Exteriores**. Este contexto hace referencia a los escenarios más complejos debido a que los ambientes exteriores son muy dinámicos por lo que añaden un gran número de variables no controlables. Por tanto, para abordar este caso se debe realizar un estudio de cómo influyen los parámetros en los datos obtenidos por la cámara. Por otra parte, este entorno engloba aquellos escenarios en los que el nivel de desconocimiento es elevado ya que apenas se tiene información del objeto con el que se quiere interaccionar.

  Estas características del escenario están estrechamente relacionadas con ambientes con ruido, condiciones dinámicas cambiantes y donde apenas se tienen datos *a priori* del objeto. Por tanto, dando solución a este escenario se abordan todos los retos abiertos mencionados en el Capítulo 1.

Los dos primeros casos son los más representativos de manipulación robótica [59] por lo que se escoge ese contexto para su validación. Por otra parte, para el caso de mayor complejidad en cuanto a estructuración se escoge un contexto aéreo debido a que es lo suficientemente representativo de exteriores y muestra un claro desafío ante el reto de detectar bajo ruido un objeto cuyas características no se pueden modelar *a priori*.

Estos escenarios se han seleccionado teniendo en cuenta las condiciones del entorno. Debido a que son lo suficientemente significativos y a la vez complementarios entre sí, en conjunto los resultados son más consistentes ante los desafíos de percepción. En los siguientes tres capítulos se van a evaluar en detalle las estrategias propuestas para cada escenario bajo estudio, incrementándose el nivel de

complejidad en cuanto a la combinación de información estructurada conocida *a priori* así como las condiciones ambientales cambiantes.

Capítulo 4

Entorno Interior Altamente Estructurado

4.1. Escenario: Sistema de Visión para Manipulación Robótica Usando una Sola Vista RGB-D

En el ámbito del uso de robots para tareas domésticas, los robots deben poder interactuar de forma autónoma y con destreza con los objetos que les rodean para poder formar parte de nuestra vida diaria. En estos casos, no todas las condiciones están controladas ya que tienen que hacer frente a nuevos objetos constantemente. Por tanto, es necesario que tengan la habilidad de actuar en entornos poco estructurados, agarrando y manipulando objetos tanto conocidos como desconocidos. Un problema habitual que se presenta en estos casos es cómo obtener el modelo 3D de objetos desconocidos sobre una mesa, por lo que el planteamiento de una estrategia para abordarlo es de gran interés para este tipo de aplicaciones. Concretamente, la reconstrucción de objetos 3D a partir de un único punto de vista es un tema de gran importancia en la visión por computador.

La estrategia de percepción visual propuesta para este escenario permite al robot hacer frente a entornos poco estructurados de agarre y manipulación, por lo que, si el objeto situado sobre una mesa no se encuentra almacenado en la base de datos, reconstruye automáticamente el objeto desconocido a partir de una sola vista de color y profundidad RGB-D (Red Green Blue - Depth). De esta forma, se obtiene un modelo 3D lo suficientemente preciso para agarrarlo y obtener tanto su posición como orientación en tiempo real.

La literatura sobre la reconstrucción de objetos utilizando múltiples puntos de vista es muy amplia, pero la obtención del modelo 3D a partir de una única vista ha

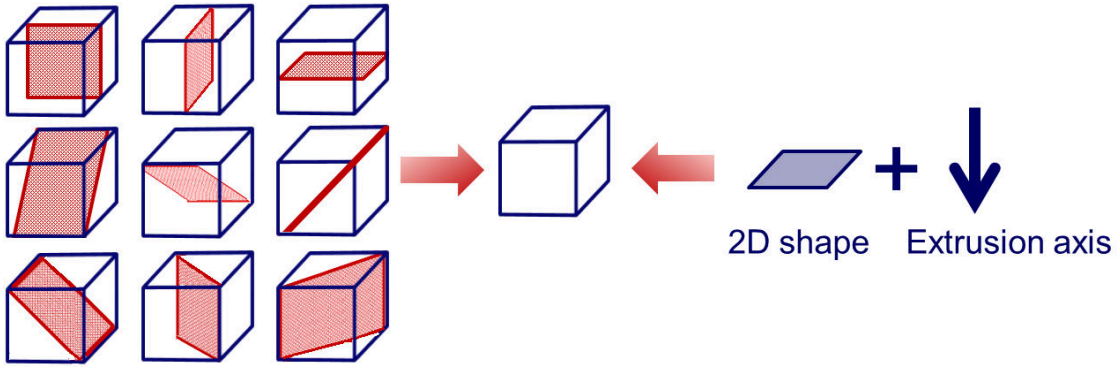


Figura 4.1: *Proceso de obtención de un cubo: (izquierda) planos de simetría; (derecha) extrusión lineal.*

recibido un interés significativo recientemente, motivado principalmente por aplicaciones de agarre mediante sistemas robóticos. Una primera categoría de métodos asume que los objetos que se van a modelar tienen una forma bastante simple, por lo que se trata de adaptar a un conjunto predefinido de formas primitivas [63] (esferas, cilindros, conos o cajas) o una combinación de ellos [64]. Este planteamiento se generalizó en otros trabajos como [65] y [66], mediante el uso de una base de datos de objetos con formas conocidas y un módulo de reconocimiento.

Cuando una extensa base de datos de modelos de objetos no está disponible o no resulta práctica, se requieren suposiciones a priori más genéricas. El supuesto más común es basarse en que la mayor parte de los objetos de la vida real son simétricos [67]. El problema entonces se convierte en encontrar la naturaleza de las simetrías en la nube de puntos parcial de la que se dispone utilizando una sola vista. Estos planos de simetría son difíciles de estimar en la práctica debido al gran espacio de búsqueda y a que los datos son limitados, lo que lleva por ejemplo, a limitar el conjunto de hipótesis a un eje del plano vertical en un área restringida [68], o centrarse en simetrías rotacionales [69].

El modelado de objetos 3D usando la suposición de simetría es un planteamiento bastante común porque muchos de los objetos son simétricos, pero también, una gran clase de objetos cotidianos se pueden generar mediante la extrusión de una forma 2D a lo largo de una determinada trayectoria, ya sea una línea (lineal) o un círculo (rotación) (Fig. 4.1).

La extrusión lineal es un proceso ampliamente utilizado por los diseñadores e ingenieros para generar modelos 3D partiendo de un boceto 2D (Fig. 4.2). Esta forma 2D corresponde a la superficie extruida y sus normales deben ser ortogonales a la dirección de extrusión [70]. Este enfoque se adapta perfectamente a la reconstrucción de objetos que descansan sobre una mesa plana, que es un escenario común en robótica, ya que la normal al plano de la mesa proporciona un eje

Figura 4.2: *Ejemplo del proceso de extrusión para generar modelos 3D de una caja de puntas de pipetas partiendo de un boceto 2D.*

de extrusión natural. De esta forma, el sistema propuesto se basa en esta propiedad para reconstruir de una forma rápida y eficaz las partes ocultas mediante una extrusión de la vista superior de los objetos.

Todas las etapas que engloban el proceso de reconstrucción 3D propuesto para un entorno interior poco estructurado se explicarán en detalle en la Sección 4.2 así como el sistema global propuesto de percepción visual usando una sola vista RGB-D. Un caso habitual de escenario semiestructurado interior de manipulación es aquel donde un robot tiene que ser capaz de realizar diversas aplicaciones domésticas, interactuando de forma autónoma con todos los objetos que se encuentren sobre una mesa. Por tanto, para validar la estrategia propuesta de percepción visual se han escogido como contexto experimental varias plataformas robóticas, cada una de ellas anexada a una mesa donde se colocan los objetos desconocidos a agarrar y manipular. Las características de este ámbito de aplicación se explican en la Sección 4.3. Posteriormente, en la Sección 4.4 se evalúa el método propuesto de reconstrucción. Además, se valida el sistema completo mediante casos prácticos de agarre y manipulación robótica. Por último, en la Sección 4.5 se resumen las principales conclusiones del sistema propuesto así como su validación en varios contextos.

4.2. Estrategia Propuesta

En una plataforma robótica real, el robot tendrá que agarrar y manipular objetos cuya forma tendrá que determinar. Con la disponibilidad de cámaras RGB-D a precios asequibles, tales como la cámara Microsoft Kinect [51], se puede adquirir en tiempo real información del color y la profundidad de la escena con una buena precisión a distancias cortas. Por lo tanto, si se orienta una de estas cámaras para observar la escena desde una zona superior, con solo una imagen RGB-D se tendría una gran cantidad de información, sin embargo una única imagen sólo proporciona la geometría de las partes visibles (Fig. 4.3). Debido a las auto-oclusiones, estas

Figura 4.3: *Ejemplo de una nube de puntos procedente de la Kinect: (izquierda) vista de las partes visibles de objetos del día a día, colocados encima de una mesa; (derecha) misma nube de puntos de la vista superior, donde los agujeros pertenecen a las partes y regiones ocluidas.*

partes ocultas crean vacíos que tienen que ser estimados utilizando un conocimiento a priori. Por tanto, la estrategia a seguir en este entorno se centra fundamentalmente en cómo adquirir modelos 3D usando una sola imagen RGB-D debido a que es el mayor reto del escenario bajo estudio.

El sistema propuesto solventa las problemáticas comentadas anteriormente estimando las zonas ocultas de los objetos desconocidos sobre una mesa plana a partir de la extrusión de la vista superior de los objetos. Este trabajo contribuye con esta nueva técnica como estrategia a este tipo de escenarios, combinando además profundidad y color para una mejor segmentación del objeto de interés.

Teniendo en cuenta que los objetos de interés siempre están colocados sobre una mesa, el sistema propuesto de percepción usando una sola vista RGB-D engloba seis partes fundamentales (Fig. 4.4):

- **Calibración** (remarcado en rojo en Fig. 4.4): se estima la posición de la cámara respecto a la base de la plataforma robótica.
- **Segmentación** (remarcado en amarillo en Fig. 4.4): segmentación de la nube de puntos para obtener de entre todos ellos el plazo horizontal dominante sobre el suelo. Dicho plano se considerará que será la mesa y, del resto de puntos de la nube, se detectarán los objetos en el siguiente paso.
- **Detección** (remarcado en verde en Fig. 4.4): detección de objetos sobre una mesa (*table-top object detector*), identificando y extrayendo grupos de puntos

3D pertenecientes a la mesa y a cada objeto. Cada conjunto será un clúster y se añadirá al mapa de colisiones para evitar que la plataforma robótica colisione con cualquiera de los objetos, ya sean conocidos o desconocidos.

- **Reconocimiento** (remarcado en rosa en Fig. 4.4): cada clúster se evalúa y se compara con los modelos guardados en una base de datos.
- **Estimación de posición 6D** (remarcado en naranja en Fig. 4.4): estima la posición 6D (vector de traslación y cuaternión de rotación) del objeto para poder agarrarlo así como para tareas de manipulación, hacer un seguimiento de la posición del objeto una vez agarrado.
- **Reconstrucción del modelo 3D para objetos desconocidos** (remarcado en azul en Fig. 4.4): una vez reconstruido, se guardará el modelo así como su posición en la base de datos.

A continuación se va a desarrollar cada una de las fases en detalle.

4.2.1. Calibración

Debido a que el fin es integrar el sistema de visión en una plataforma robótica multi-modal para poder agarrar y manipular un objeto con destreza, es necesario representar dicho objeto de interés en el sistema de coordenadas de la plataforma. Por ello, el sistema de visión incorpora una primera etapa de calibración donde se estima la posición de la cámara respecto a la base de la plataforma robótica. Para calcular su transformada se utiliza un algoritmo implementado en la librería OpenCV [71] que, usando una sola vista, detecta un tablero de ajedrez y calcula su posición 6D (rotación y traslación) respecto a la cámara. Teniendo en cuenta que la posición del patrón respecto a la base del robot se conoce previamente, la obtención de la transformada entre la cámara y el robot es directa. En este caso se han considerado conocidos los parámetros intrínsecos de la camera de profundidad y de la de color, así como sus parámetros extrínsecos (transformación 3D entre las dos cámaras). En el caso de que fuesen desconocidos, se podría usar el método de Bouguet [72], también implementado en OpenCV con un patrón similar a un tablero de ajedrez.

Para el algoritmo de detección del patrón se utiliza un tablero de 9x6 con cuadrados de 25mm. La calibración utiliza los vértices interiores del tablero por lo que el patrón tiene 10 cuadrados de ancho y 7 de alto. Como se ve en la Fig. 4.5, el sistema detecta las esquinas interiores del tablero usando la imagen de color, pero su posición se refina utilizando la imagen de profundidad Kinect. Para el refinamiento, la pose se optimiza mediante el uso de mínimos cuadrados considerando un

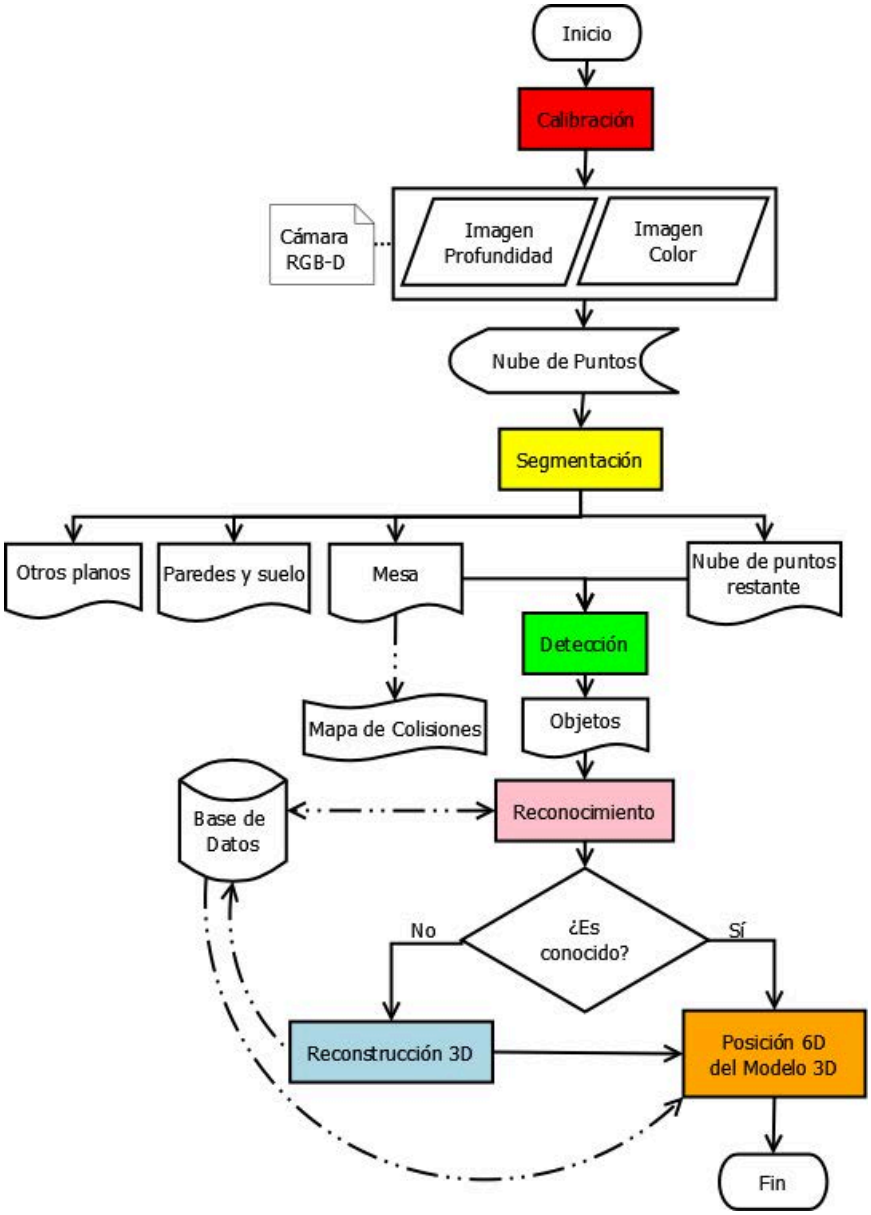


Figura 4.4: *Diagrama general del sistema de percepción propuesto.*

valor de 0,01 de importancia relativa de la imagen RGB con respecto a la imagen de profundidad.

Hay que tener en cuenta que aunque la calibración es optimizada mediante la imagen de profundidad, pueden existir errores debido a la precisión de la cámara en sí, el uso de una sola imagen o incluso posibles desviaciones en el modelo cinemático del robot. A pesar de ello, esta fase de calibración del sistema multimodal

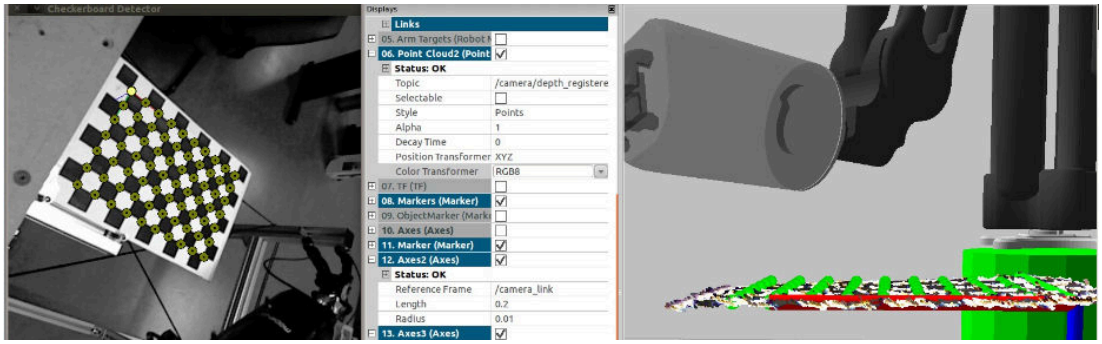


Figura 4.5: *Calibración usando un tablero de ajedrez como patrón: (izquierda) detalle de las esquinas detectadas en la imagen RGB; (derecha) en verde, se muestran en 3D las esquinas detectadas usando la información de profundidad.*

y sus diversos componentes aporta un valor fundamental al sistema de percepción visual permitiendo su adaptabilidad a diferentes plataformas.

4.2.2. Segmentación

El primer paso a completar es la identificación de la superficie de apoyo para separar el fondo de los objetos de primer plano. Las escenas en interior comprenden un gran número de planos, por ejemplo, mesas, paredes y suelo. Este tipo de planos de fondo pueden perturbar la selección de objetos por lo que el primer paso es su localización y exclusión de los puntos asociados.

En la visión por computador, el método de segmentación más ampliamente usado para localización de planos es el algoritmo *RANdom Sampling Consensus* (RANSAC) [73], que busca el consenso entre un conjunto de puntos. Este método ha sido probado con éxito para detección de planos tanto en 2D como en 3D [74]. En el caso 3D, se selecciona un subconjunto aleatorio de puntos a partir de una nube de entrada y se calcula el modelo de regresión planar que mejor encaje con el subconjunto.

Por tanto, primeramente el sistema propuesto encuentra los planos en la nube de puntos de la escena mediante RANSAC, discriminando los planos verticales (paredes de alrededor) y el plano horizontal a menor altura (perteneciente al suelo). Posteriormente, se selecciona el soporte plano horizontal 3D dominante cuya envolvente convexa se identificará como la mesa en la que se apoyan los objetos [75]. El sistema permite añadir la mesa al mapa de colisiones para el futuro cálculo de trayectorias de agarre del objeto de interés.

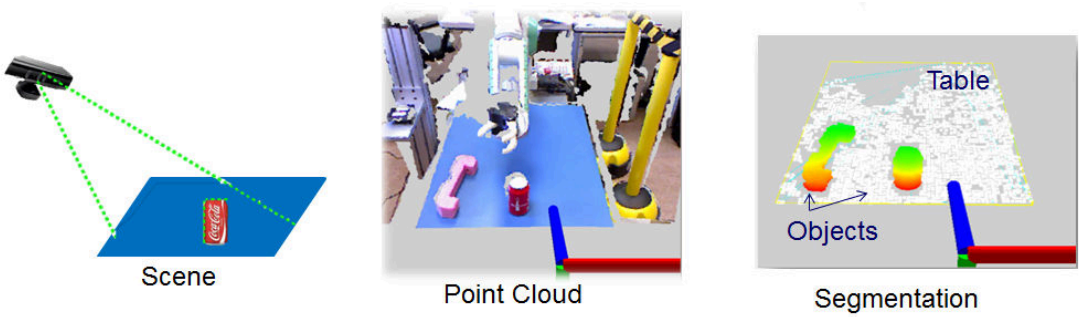


Figura 4.6: *Segmentación. De izquierda a derecha: escena, nube de puntos, segmentación de la mesa y objetos.*

4.2.3. Detección

Una vez segmentada la mesa es necesario detectar los objetos que se encuentran colocados encima. Uno de los enfoques más comunes se basa en operar directamente sobre la nube de puntos gracias a su mejor adaptación a objetos desconocidos sobre planos. En este sentido, Rusu et. al [16] propuso un algoritmo para detección de objetos sobre superficies planas, (*table-top object detector*), extrayendo agrupaciones de nubes de puntos sobre superficies planas horizontales como es el caso de una mesa. De esta forma, se considera objeto aquella agrupación de puntos contiguos que está encima del plano y cuya proyección sobre la imagen de profundidad esté dentro de los límites de la mesa. Teniendo en cuenta este enfoque, el algoritmo de detección del sistema propuesto sigue los siguientes pasos (Fig. 4.6):

- **Eliminación de información de fondo o fuera de la mesa**: los puntos que están fuera de un prisma en torno al plano de la mesa son eliminados.
- **Clusterización**: los puntos restantes cuyas proyecciones caen dentro de la delimitación del polígono 2D correspondiente a la mesa, se agrupan en clústeres utilizando distancias euclidianas con umbrales fijos. Los grupos de puntos pertenecientes a un determinado objeto se denominan clúster. Las agrupaciones de puntos que son demasiado pequeñas o no tocan la mesa se eliminan.

Esta detección es fiable para objetos que están separados al menos la mitad del radio mínimo de cada uno [16].

4.2.4. Reconocimiento

Existen una gran cantidad de algoritmos disponibles en la librería *Point Cloud Library* (PCL) [76] para abordar el reconocimiento de objetos usando una base de

datos [77]. En este Capítulo, debido a que se buscan estrategias para poder hacer frente a nuevos objetos, se han usado algoritmos ya desarrollados de reconocimiento y estimación de posición que han sido probados con éxito en escenarios similares [16, 78, 79]. Por tanto, se han incorporado al sistema de percepción para complementarlo tanto en la fase previa de agarre como durante la manipulación.

En tareas domésticas de manipulación robótica es común el cálculo de características (*features*) para el reconocimiento de objetos. Las características más comunes incluyen el uso de descriptores y puntos de interés locales basados en apariencia visual así como descriptores 3D de histogramas obtenidos de imágenes de rango [80]. En el sistema propuesto se integran ambos casos para buscar la correspondencia de cada objeto detectado con los modelos disponibles en la base de datos. Los métodos usados de extracción de características en imágenes y en modelos 3D se basan en la extracción de puntos clave (*keypoints*) para luego proceder a la construcción del descriptor que se utilizará para la búsqueda de correspondencias con el modelo de referencia. Como se verá a continuación, se usará un método u otro según las características del objeto y la situación en la que se encuentre:

- **Descriptor 2D**: cuando un objeto con textura está ocluido parcialmente, ya sea por otros objetos o por estar agarrado, las técnicas de reconocimiento más populares son aquellas basadas en características locales [81]. Esto es debido a que cuando el objeto tiene textura suficiente, estas técnicas son muy robustas ya que pueden encontrar similitudes en las zonas visibles del objeto.

  Para el sistema propuesto se ha utilizado un método derivado del *Scale-Invariant Feature Transform* (SIFT) propuesto por Lowe et at [11]. El descriptor SIFT se basa en un vector de 128 dimensiones que, cuantificando mediante histogramas, codifica los valores del gradiente de brillo en torno al punto de interés. El algoritmo SIFT se basa en computar la similitud de características entre los puntos SIFT extraídos de la imagen de color bajo análisis y los puntos SIFT de las imágenes de color de los modelos de referencia. La mejor correspondencia es aquella cuya similitud de los puntos SIFT es mayor. Por tanto, para cada asociación de puntos se dispone de una posible transformada 2D entre la vista actual y la de la base de datos por lo que se elige la transformada más idónea según un criterio estadístico.

  El método propuesto tiene como base el algoritmo SIFT y lo combina con una medida de similitud de regiones robusta [78] utilizando información 3D y el histograma de color. De esta forma, para mejorar los resultados se realiza una pre-segmentación de las regiones candidatas comparando el histograma y conociendo la profundidad de los puntos de su interior, las regiones se pueden restringir a aquellos píxeles que tienen una profundidad similar.

Por tanto, se utiliza la información de la imagen de profundidad para dar robustez al descriptor creado a partir de la imagen de color.

- **Descriptor 3D**: cuando el objeto se encuentra sobre la mesa y sin oclusiones, sus propiedades geométricas pueden ser utilizados para el reconocimiento. En este caso, el método utilizado en el sistema es el descriptor *Viewpoint Feature Histogram* (VFH) que ofrece buenos resultados para reconocimiento y estimación de posición 6D [16] donde una segmentación previa es posible, como es el escenario bajo estudio. Este descriptor, extensión del *Fast Point Feature Histogram* (FPFH) [17], codifica tanto la forma del objeto como el punto de vista desde el que se toma la nube de puntos. Para cada vista se tiene un descriptor por lo que se debe tener varias vistas almacenadas de un mismo objeto. Como se aprecia en la Fig. 4.7, el método VFH computa para cada vista:

  - El punto de vista V_p se calcula a partir de un histograma de los ángulos α que hacen las direcciones de los puntos de vista con cada normal n_i.
  - El giro relativo, los ángulos de inclinación y la oscilación se calculan midiendo la dirección del punto de vista con el punto central p_i y cada una de las normales de la superficie n_i.

  El descriptor VFH es un histograma compuesto por:

  - Cuatro distribuciones angulares diferentes de normales a una superficie: 45 subdivisiones para cada uno de los 3 valores extendidos de FPFH (α, θ, Φ) así como 128 subdivisiones para el valor angular referido a la componente del punto de vista.
  - Una distribución relativa a la geometría intrínseca: 45 subdivisiones para la distribución de la forma que corresponde con las distancias entre cada punto y su centroide.

Ambos métodos analizan el objeto comparándolo con una base de datos cuyas características se explicarán en la siguiente subsección.

4.2.4.1. Base de Datos

Para entornos domésticos de aplicaciones de agarre y manipulación robótica existen diversas bases de datos de modelos 3D y agarres [82]. De entre todas ellas, para la detección de objetos a agarrar sobre una mesa, cabe destacar la *Willow Garage Household Objects Database* [83]. Ha sido ampliamente usada en estos escenarios robóticos debido a que fue desarrollada en el framework ROS con el objetivo de

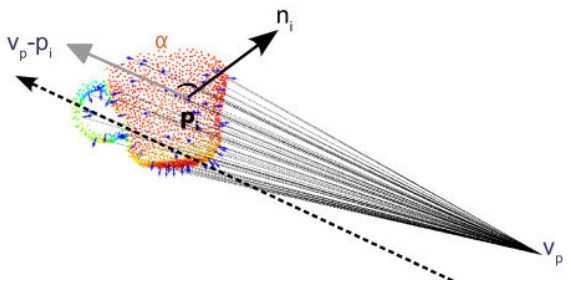


Figura 4.7: *Detalle de una nube de puntos de una taza con los parámetros relacionados con el descriptor de VFH. Imagen tomada del artículo [16].*

permitir a los investigadores replicar los experimentos realizados por *Willow Garage* y poder obtener sus propios resultados. Por tanto, la base de datos que se ha implementado se basa en dicha base de datos de objetos del hogar, partiendo de la ya disponible pero añadiendo ciertas funcionalidades, objetos y campos para adaptarla al escenario bajo estudio.

La base de datos desarrollada puede ser modificada por lo que se han implementado los algoritmos necesarios para interaccionar mediante SQL (*Structured Query Language*). De entre las funciones que se pueden realizar con la base de datos caben destacar las siguientes: insertar, actualizar, eliminar y descargar un modelo así como hacer consultas, por ejemplo, ver qué modelos corresponden a una determinada categoría.

El esquema de base de datos está organizada por tablas, referenciadas entre ellas de forma cruzada mediante una clave externa o *foreign key*. La principal diferencia entre la base de datos desarrollada y la *household objects database* radica en la definición de la clave externa. La base de datos de este Capítulo permite la actualización y el borrado en cascada, es decir, cuando se borran o actualizan las filas de la tabla maestra, las respectivas filas referenciadas de la tabla hija se borrarán o actualizarán si su clave externa coincide.

Para cada objeto, la base de datos guarda el modelo 3D de su superficie, características, puntos de agarre y enlaces a archivos externos en los diferentes campos de las principales tablas (*original model*, *scaled model*, *mesh*, *grasp*, *file path*...). Cuando un modelo se guarda, se genera automáticamente un identificador único (*original model id*) en la tabla *original model*. Teniendo en cuenta que todos los agarres en la base de datos se computan en versiones escaladas de modelos originales, cada identificador del modelo original tiene su correspondiente identificador del modelo escalado en la tabla *scaled model* como se puede apreciar en la Fig. 4.8.

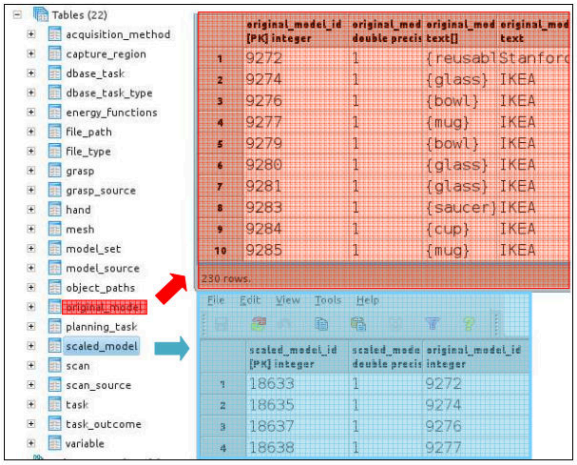


Figura 4.8: *Base de datos, resaltando las tablas original model y scaled model con sus identificadores cruzados para varios objetos.*

4.2.5. Estimación de Posición 6D

Para estimar la posición 6D (vector de traslación y matriz de rotación) en el sistema propuesto se han combinado dos técnicas de estimación de posición basados en imágenes y en modelos 3D. Se usará un método u otro según la situación en la que se encuentre:

- **Mínimos cuadrados usando características 2D junto con profundidad**: se calcula la transformada entre el objeto reconocido y el de referencia almacenado en la base de datos. Para calcular la posición, los enfoques más comunes son RANSAC y reproyección de error de características por mínimos cuadrados. RANSAC suele funcionar correctamente cuando hay datos contaminados (*outliers*), pero requiere un alto número de coincidencias de características. Debido a que sólo se tienen 2 o 3 coincidencias del algoritmo de detección, se usa un proceso iterativo de mínimos cuadrados para minimizar el error de reproyección. Debido a que este método lineal es muy sensitivo a falsas correspondencias y poco fiable ante pocos puntos, se añade la información de profundidad para calcular un error de reproyección 3D [78]. Dicho error se computa teniendo en cuenta, $H(F_i$, las coordenadas 3D de la característica i proyectada sobre la imagen actual según la posición H y C_i las coordenadas (x, y, z) del punto característico detectado en la imagen bajo análisis.

$$err = \sum_{i=0}^{n} \Delta(H(F_i) + C_i)^2 \tag{4.1}$$

Cuando la profundidad está disponible, la función de distancia Δ se computa teniendo en cuenta las coordenadas (x_1, y_1, d_1) y (x_2, y_2, d_2) de los puntos 3D p_1 y p_2 respectivamente, así como las distancias focales f_x, f_y de la cámara asumiendo el modelo *pin-hole*:

$$\Delta(p_1, p_2) = \left[\frac{(x_1 - x_2)}{f_x}\right]^2 + \left[\frac{(y_1 - y_2)}{f_y}\right]^2 + (d_1 + d_2)^2 \tag{4.2}$$

Debido a su eficacia en situaciones de oclusiones, este método se empleará tanto para obtener la posición del objeto reconocido usando el descriptor 2D anteriormente explicado como para hacer un seguimiento de la posición del objeto agarrado para tareas de manipulación.

- **Método basado en *Fast Iterative Closest Point* (ICP)**: en el caso de los objetos que han sido reconocidos mediante VFH se suele aplicar una siguiente fase de postprocesamiento para mejorar el resultado de la posición dada tras el reconocimiento. Habitualmente se aplica el algoritmo iterativo de punto más cercano, ICP [84], a la hipótesis de reconocimiento con el objetivo de refinar la posición 6D. Este método permite minimizar el error de alineamiento entre el modelo 3D de referencia y la nube de puntos detectada mediante una técnica de mínimos cuadrados que calcula las desviaciones por rotación y traslación entre ambas nubes de puntos. Este método permite minimizar la distancia entre puntos, repitiendo el proceso de forma iterativa.

4.2.6. Reconstrucción 3D Para Objetos Desconocidos

Tras haber realizado la detección de objetos sobre una mesa (*table-top object detector*), se realiza el proceso de reconstrucción 3D de un objeto desconocido que integra las siguientes dos etapas principales para luego añadirlo a la base de datos. Este nuevo método de reconstrucción se ha denominado **Xtru3D**, y está formado por dos etapas que a su vez están compuestas por varios pasos que se ilustran en la Fig. 4.9:

1. **Cálculo del volumen inicial** a partir del clúster del objeto desconocido así como de las imágenes de color y profundidad que proporciona la cámara Kinect (Sección 4.2.6.1). Esta etapa incluye dos fases fundamentales:

    - **Relleno de vóxeles por extrusión**: los puntos que corresponden a la vista superior del objeto detectado son considerados como el perfil de extrusión, y se les extruye hacia la mesa para llenar un volumen formado por voxels alrededor del *cluster* de interés.

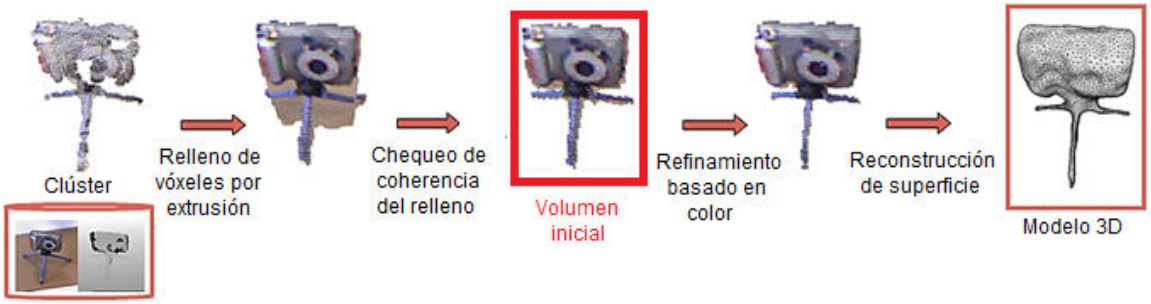


Figura 4.9: *Resumen del proceso de reconstrucción propuesto.*

- **Chequeo de la coherencia del relleno**: las concavidades de los objetos podrían rellenarse durante la etapa de extrusión, por lo que hay que compensarlos mediante la comprobación de la consistencia del voxel calculado frente a la imagen de profundidad.

2. **Refinamiento basado en color** (Sección 4.2.6.2): las imágenes de profundidad de las cámaras RGB-D de bajo coste suelen ser imprecisas en torno a los bordes del objeto, y con frecuencia tienen agujeros debido a las reflexiones u otros efectos ópticos. Dado que la imagen de color no se ve afectada por estos aspectos, en el segundo paso, se perfeccionan los bordes de los objetos utilizando segmentación basada en color, dando un conjunto refinado de voxels como entrada al algoritmo de mallado final.

3. **Reconstrucción de superficie** (Sección 4.2.6.3): para obtener un modelo 3D mallado válido para tareas de agarre y manipulación es necesario aplicar un postprocesamiento a la nube de puntos. Para ello se aplica Poisson [85] sobre la nube de puntos voxelizada para crear la superficie mallada del objeto.

4.2.6.1. Cálculo del Volumen Inicial

En la primera etapa de reconstrucción se considera como entrada del desarrollo propuesto el cluster de interés. Dicho cluster será aquel que corresponde a un objeto desconocido cuyos puntos 3D se obtuvieron al aplicar el *table-top object detector*. De esos puntos, los que corresponden a la parte superior del objeto se consideran como el perfil a extruir hacía la mesa para rellenar el volumen en torno al cluster de interés. En este paso de extrusión podrían rellenarse concavidades del objeto, por lo que se compensa este proceso chequeando la consistencia de los vóxeles con la imagen de profundidad. A continuación se explican estos pasos en más detalle.

Figura 4.10: *Resumen de la extracción del cluster.*

4.2.6.1.1 Preprocesamiento del *Cluster*

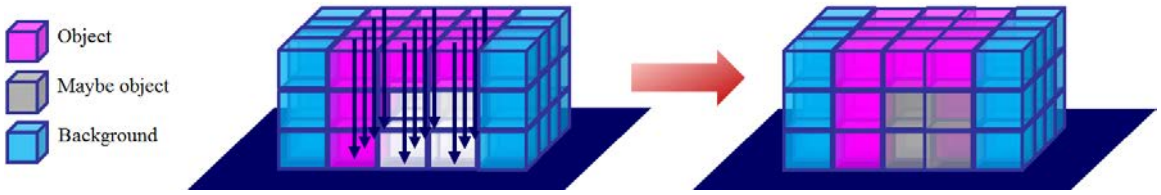
El *cluster* de interés será aquel objeto que no haya sido reconocido en las etapas previas, tras haber aplicado el *table-top object detector* (Fig. 4.10). Para hacer que el procesamiento 3D sea más rápido y que obtenga de forma natural la vecindad entre puntos 3D, se inicializa un volumen de vóxeles de tamaño fijo alrededor del *clúster*, etiquetando como ”objeto” a los voxels que corresponden a un punto del *cluster*. El tamaño de voxel es un parámetro definido por el usuario en función del equilibrio entre precisión y velocidad deseado. Los voxels de todas las reconstrucciones que se muestran en este trabajo son de 3mm.

4.2.6.1.2 Relleno de los Vóxeles por Extrusión

El objetivo de este paso es rellenar las partes ocluidas basándose en el supuesto de que el objeto puede ser aproximado por un proceso de extrusión. Teniendo en cuenta que la normal del plano de la mesa proporciona el eje de extrusión natural para la mayoría objetos, no es necesario calcular el eje de objeto para obtener la dirección de extrusión, consiguiendo acelerar el proceso. Por tanto, se considera la normal del plano como la dirección de extrusión de la cara superior del objeto, pudiendo resumir el algoritmo en los siguientes dos puntos (Fig. 4.11 (a)):

1. Para cada voxel considerado como “objeto”, se calcula el segmento de recta que va desde el voxel al plano a lo largo de la normal al plano.

2. Todos lo voxels que son intersectados por ese segmento de recta se etiquetan como “quizás objeto”.



(a)

(b)

Figura 4.11: *Relleno de voxels por extrusión. (a) Descripción general del algoritmo propuesto. (b) Izquierda: nube de puntos de una caja. Medio: malla de voxels del cluster. Derecha: malla de voxels después de la extrusión hacia el plano de la mesa. Los voxels grises corresponden a las partes que no se ven debido a las auto-oclusiones.*

El resultado de este paso es una estimación aproximada del volumen del objeto. Posteriormente, el modelo es ligeramente suavizado mediante la ejecución de un cierre morfológico para hacer frente a las incertidumbres alrededor de los bordes del objeto que se producen en la imagen de profundidad. El tamaño óptimo del elemento estructural depende del tamaño del voxel y las propiedades de los datos de profundidad. Para voxels de 3 mm y una cámara Kinect, empíricamente un cubo de 3x3x3 es un elemento estructural satisfactorio. Un ejemplo de esta salida se muestra en la Fig. 4.11 (b).

4.2.6.1.3 Chequeo de la Coherencia del Relleno

La etapa de extrusión puede llenar regiones del objeto que corresponden a agujeros o concavidades. Esto se puede corregir mediante la comprobación de la consistencia de lo voxels etiquetados como "quizás objeto" contra la imagen de profundidad. Para ello, se reproyecta cada voxel en la imagen de profundidad, y se compara la profundidad proyectada con la dada por la imagen.

Si la diferencia es mayor que un umbral δ_d, el voxel se etiqueta como "fondo". El umbral depende de la precisión del sensor de profundidad, y se ha ajustado a 3 mm en todos los experimentos mostrados. La salida de este proceso se ilustra en la Fig. 4.12.

Figura 4.12: *Chequeo de la coherencia del relleno para tallar huecos y concavidades del objeto. Izquierda: imagen de color. Medio: voxels coloreados después de la extrusión. Derecha: voxels restantes después de la comprobación de coherencia. Los agujeros y concavidades que fueron erróneamente rellenos por el algoritmo de extrusión se eliminan si son visibles.*

4.2.6.2. Refinamiento Basado en Color

Después de los pasos anteriores, el modelo 3D obtenido puede contener aún zonas sin profundidad o irregularidades debido a datos incorrectos o falta de información de profundidad en el frame RGB-D. Los píxeles incorrectos en la imagen de profundidad por lo general pertenecen a bordes y/o áreas de objetos especulares, transparentes o reflectantes. Estas particularidades no afectan habitualmente a la imagen de color, por lo que se propone mejorar la calidad del modelo 3D refinando la segmentación usando primeramente la imagen en color. Posteriormente, se rellenan los valores de profundidad perdidos o incorrectos mediante restauración de la imagen (proceso denominado *inpainting* en inglés).

4.2.6.2.1 Segmentación del Objeto Basada en Color

La segmentación de imágenes se basa en la obtención de una región de interés en un entorno. Hay muchas técnicas existentes para la segmentación basada en color, pero sigue siendo un problema sin resolver para los casos generales. Sin embargo, cuando es posible una buena inicialización, las técnicas basadas en grafos [86] han demostrado ser muy eficaces para la segmentación de la región de interés y fondo [87] (denominado *foreground* y *background* en inglés, respectivamente). En particular, la variante GrabCut [88] combina recortes de grafos con modelos mixtos de Gaussianas, pudiendo partir de una máscara proporcionada por el usuario. Por lo tanto, está particularmente adaptado al perfeccionamiento de una primera segmentación.

En los últimos años, GrabCut se ha extendido al uso de información de profundidad mediante su combinación con los canales RGB con un factor de ponderación [89]. En esta tesis, en lugar de fusionar esta información, se propone ejecutar GrabCut sólo en la imagen en color, pero inicializando la máscara a partir de la información de profundidad. Este enfoque consigue obtener lo mejor de cada técnica basándose en su complementariedad, ya que la imagen de profundidad es errónea o carece de datos cerca de las bordes de objeto; mientras que la información de color es más sensible para una segmentación inicial ante un fondo similar al objeto de interés.

De esta forma, la inicialización de la máscara se realiza partiendo del modelo inicial que se obtiene como salida del algoritmo de la Sección 5.3. Se usa sólo la información de profundidad, reproyectando cada punto 3D del modelo en la imagen de profundidad. Entonces, los píxeles son etiquetados como objeto (*foreground*), fondo (*background*) o desconocidos (si su profundidad proyectada no es consistente o carecen de dicha información). Teniendo en cuenta estas etiquetas como punto de partida, se crea la máscara de inicialización. GrabCut puede tomar cuatro valores diferentes de inicialización según píxeles pertenecen a la zona de interés, fondo, probablemente de interés o probablemente fondo. El algoritmo no cambiará de etiqueta aquellos píxeles que se han considerado de interés y fondo por lo que se garantiza una buena solidez ante errores de segmentación errores. Para controlar la incertidumbre asociada a los píxeles de bordes de la imagen de profundidad, sólo los píxeles que no están en un borde se marcarán con etiquetas definitivas. El algoritmo GrabCut es entonces aplicando una sola iteración sobre la imagen en color utilizando la máscara computada.

Debido a la precisión de la máscara inicial, GrabCut realiza bien la segmentación incluso si el objeto y el fondo tienen una distribución de color parecida o si el fondo contiene mucha información, como se muestra Fig. 4.13.

4.2.6.2.2 Relleno de Agujeros Mediante Restauración de Profundidad

La segmentación obtenida en el apartado anterior es precisa pero algunos píxeles que han sido clasificados como objeto después del refinamiento del color carecen de profundidad. La mayoría de los métodos de relleno de agujeros utilizan interpolación de imagen o técnicas de restauración (*inpainting* en inglés) para llenar los agujeros existentes mediante la información de los píxeles vecinos. Para mejorar los datos de profundidad proporcionados por la Kinect, un enfoque *cross-modal* de visión estéreo fue presentada en [90]. Sin embargo, este no beneficia a la segmentación entre fondo y el objeto de interés. Por otra parte, un método de relleno de agujeros usando *inpainting* basada en profundidad para vídeo 3D fue propuesto

Figura 4.13: *Ejemplo de segmentación refinada, incluso cuando el objeto de interés es similar al fondo. Izquierda: imagen en color, el objeto de interés es un frasco de almacenamiento encima de un póster. Medio: segmentación inicial. Los píxeles se marcan como: desconocido (negro), objeto (blanco) y fondo (gris). Derecha: segmentación final del objeto, después de GrabCut. Los píxeles se marcan como: objeto (blanco) y fondo (negro).*

en [91]. Pero existe un método ampliamente usado, más rápido y menos complejo en el caso de poder usar una máscara de inicialización de segmentación, es la técnica de Telea [92] basada en marcado rápido.

Siguiendo esta línea, en este trabajo se propone la restauración de la imagen para completar los valores de profundidad usando una máscara de segmentación para rellenar los píxeles utilizando sólo los valores de aquellos de la misma categoría que están alrededor. Por lo tanto, los agujeros de los píxeles clasificados como objeto se llenan sólo con información de profundidad procedente de otros píxeles objeto, mientras que los de fondo se rellenan con la de los píxeles circundantes clasificados como fondo. Para ello se emplea la implementación en OpenCV de Telea. Se toma como entrada la imagen de profundidad original y una máscara especificando los píxeles a rellenar. Para restaurar la imagen de profundidad, se utilizan dos máscaras dependiendo de la clase de píxeles:

1. Objeto: el área objetivo a rellenar corresponde a los píxeles sin valor de profundidad o con valores de profundidad inconsistentes de la Sección 4.2.6.2.1. Los píxeles que pertenecen al fondo son también marcados como área objetivo para evitar que influyan en la restauración.

2. Fondo: el área objetivo a rellenar corresponde a los píxeles etiquetados como fondo sin datos de profundidad. Similar al caso previo, los píxeles de la categoría objeto son marcados también como área objetivo.

Una vez que la imagen de profundidad ha sido refinada y restaurada, se ejecuta de nuevo el algoritmo de la Sección 4.2.6.1.3. En la imagen de la Figura 4.14 se

(a) (b) (c) (d) (e)

Figura 4.14: *Resultado del refinamiento basado en color, usando un libro como ejemplo ilustrativo. Las imágenes de la cámara Kinect: (a) Imagen de color y (b) la imagen de profundidad. (c) Segmentación inicial reproyectando cada punto 3D del volumen obtenido tras los pasos de la Sección 4.2.6.1.3. Los píxeles se marcan como: desconocido (negro), objeto (blanco) y fondo (gris). (d) Segmentación final del objeto después GrabCut de acuerdo a la Sección 4.2.6.2.1. Los píxeles se marcan como: objeto (blanco) y fondo (negro). (e) La imagen de profundidad tras restaurar las zonas sin datos.*

puede apreciar la mejora obtenida después de refinar la segmentación y restaurar la profundidad, rellenando los píxeles sin información y aquellos pertenecientes a bordes cuya profundidad no era correcta.

4.2.6.3. Reconstrucción de Superficie

El modelo final del objeto se obtiene utilizando reconstrucción de la superficie mediante Poisson [85] sobre la nube de puntos voxelizada para crear la superficie mallada del objeto.

Teniendo en cuenta que el procedimiento de reconstrucción de la superficie debe ser en sí mismo un método de reducción de ruido, los algoritmos basados en funciones implícitas son los más adecuados. Por este motivo se ha escogido el algoritmo de Poisson, cuya resistencia al ruido ha sido probada en otros trabajos previos [93]. Este algoritmo permite aproximar la malla de puntos con información de la normal, considerando todos los puntos al mismo tiempo, sin tener que recurrir a la compartimentación espacial heurística. Para este caso se ha utilizado el cálculo de normales e implementación de Poisson disponible en la librería PCL (versión 1.6).

4.3. Contexto Experimental

El sistema de percepción que se evalúa en este capítulo se enmarca en el proyecto integrado de gran escala titulado HANDLE, *Developmental pathway towards autonomy and dexterity in robot in-hand manipulation*, (2007-2013), financiado por la

Unión Europea dentro del Séptimo Programa Marco (FP7). La Universidad Pierre y Marie Curie de París coordina el consorcio formado por nueve socios de seis países europeos:

- Francia: Comisión de Energía Atómica (CEA) y Universidad Pierre y Marie Curie de París (UPMC).
- Reino Unido: Shadow y Universidad de King's College London (KCL).
- España: Universidad Carlos III de Madrid (UC3M).
- Portugal: Instituto Superior Técnico (IST) de la Universidad de Lisboa y el Instituto de Sistemas y Robótica de la Universidad de Coimbra (FCTUC).
- Suecia: Universidad de Örebro (ORU).
- Alemania: Universidad de Hamburgo (UHAM).

El objetivo de HANDLE es comprender cómo realizan los humanos la manipulación de objetos para reproducir los movimientos de agarre con una mano artificial antropomórfica articulada. Para lograr movimientos naturales y eficaces con autonomía se busca dotar al sistema robótico con capacidades avanzadas de percepción y control, por lo que el proyecto aborda de forma global los siguientes puntos:

- Caracterización de propiedades de los objetos.
- Aprendizaje de tareas y actividades mediante la imitación de seres humanos.
- Optimización y desarrollo de las capacidades del robot a través de la interacción.
- Manipulación diestra autónoma.
- Desarrollo de manos artificiales similares a dispositivos de fácil instalación y uso (plug-in).

El método desarrollado dotará a la plataforma robótica real de la percepción visual necesaria para agarrar y manipular objetos tanto conocidos como nuevos y cuya forma debe ser determinada. Por tanto, el contexto bajo estudio permite validar el sistema de percepción propuesto para manipulación robótica usando una sola vista.

4.3.1. Ámbito de Aplicación

El campo de la manipulación robótica se compone de numerosos sistemas robóticos integrados (móviles y no móviles) diseñados para la realización de tareas relacionadas con el agarre y manipulación. Las plataformas robóticas de manipulación comparten diversas similitudes y diferencias en el diseño de hardware, pero principalmente se distinguen por su arquitectura de software, la variedad de tareas de manipulación que pueden cumplir, y su nivel de autonomía [94]. Para recrear los movimientos de una mano humana con un sistema robótico hay que resolver varios problemas complejos desde el punto de vista mecánico (integración de todos los actuadores), sensorial e inteligencia para poder actuar de forma autónoma.

Cada una de las plataformas existentes incorporan diferentes aspectos de manipulación robótica autónoma para llevar a cabo tareas específicas: percepción, planificación y ejecución del movimiento previsto. Con el objetivo de desarrollar capacidades fiables en el área de la percepción para manipulación, este Capítulo aborda las técnicas visuales para manipulación con una sola mano de objetos que se encuentran encima de una mesa, que es un escenario bastante común en robótica [95].

Este Capítulo hace énfasis en el método de percepción visual que se ha desarrollado y se validará en diversas plataformas robóticas no móviles, integrándolo con sus sistemas de agarre y manipulación para su validación.

4.3.2. Configuración

El sistema de percepción visual propuesto para interaccionar con objetos encima de una mesa engloba el siguiente *setup*:

- Una mano artificial antropomórfica articulada.
- Un brazo robótico industrial situado al lado de una mesa.
- Una cámara RGB-D orientada para ver la parte superior de los objetos sobre una mesa.

Debido a que el sistema propuesto se ha validado en el contexto de un proyecto europeo, el sistema ha sido probado tanto en la plataforma robótica disponible en la UC3M (Fig. 4.15) como en la de UPMC (Fig. 4.16) con fines de agarre y manipulación, respectivamente. El laboratorio del coordinador, UPMC, es el demostrador final de HANDLE por lo que en su plataforma se integran los avances científicos y tecnológicos de cada institución.

Los detalles de los elementos de cada plataforma robótica no móvil se explican a continuación:

Figura 4.15: *Plataforma robótica de la UC3M: (1) La cámara Kinect;(2) Mano antropomórfica de la compañía Shadow; (3) Brazo robótico denominado PA-10. A la izquierda se muestra una infografía y a la derecha la plataforma real.*

- Una mano artificial antropomórfica articulada: la mano diestra robótica utilizada es de la compañía Shadow Robot [96] (Fig. 4.17). Dicha mano antropomórfica posee 24 grados de libertad (DoF, *Degrees of Freedom*) así como sensores de posición, fuerza y presión.

    - El modelo de mano utilizado es el C6M2 que es un sistema autónomo que integra de forma compacta todos los actuadores, sensores, sistemas de control y comunicaciones (EtherCAT, Ethernet for Control Automation Technology). Su cinemática es lo más cercano a la de una mano humana (diagrama mostrado a la derecha de la Fig. 4.17). Como se puede observar, el pulgar tiene 5 DoF y 5 uniones, mientras que el resto de dedos tiene 3 DoF y 4 uniones.
    - Además, en el caso de la plataforma de la UPMC cuenta con sensores ATI Nano17 [97] de fuerza y par de 6 ejes en cada una de las yemas de los dedos (Fig. 4.18). Estos sensores proporcionan gran información de contacto con los objetos [98] midiendo tan solo 17 mm de diámetro. Con el fin de hacer frente a los algoritmos de alto nivel, los sensores están equipados con una elipsoide adaptada a la yema de los dedos.

- Un brazo robótico situado al lado de una mesa y al que se ensambla la mano: en el caso de la UC3M, se utiliza el robot industrial Mitsubishi PA-10 (Mitsubishi Heavy Industries, Ltd.) [99] con 7 grados de libertad, que ofrece un alto nivel de destreza gracias a su redundancia (Fig. 4.19). Por otra parte, en el caso de la plataforma de UPMC se dispone del brazo 4-DoF de Shadow Robot Company, comparable a la parte superior del brazo humano, el codo y antebrazo. Se ha diseñado para funcionar como sistema de base de la mano

Figura 4.16: *Plataforma robótica de la UPMC: (1) La cámara Kinec;(2) Mano antropomórfica de la compañía Shadow; (3) Brazo robótico de Shadow; (4) Sensores ATI Nano17. A la izquierda se muestra una infografía y a la derecha la plataforma real.*

de Shadow, y contiene un distribuidor de válvulas como músculos.

- Una cámara RGB-D: en este caso una Microsoft Kinect [51] se monta en el brazo robótico y se orienta para ver la parte superior de los objetos sobre una mesa. En el caso de la UC3M la cámara se encuentra en posición horizontal a un lateral de la mesa, mientras que en el caso de la UPMC se sitúa en el lado izquierdo del brazo robótico.

Teniendo en cuenta que el sistema de visión propuesto tiene como objetivo integrarse en diversas plataformas de manipulación robótica, todos los algoritmos se han desarrollado bajo el framework de desarrollo de software para robots denominado *Robot Operating System* (ROS) [100]. ROS permite abstraerse del hardware, intercambiar mensajes entre procesos y controlar dispositivos a bajo nivel entre otras funcionalidades. En este trabajo, se han utilizado las siguientes versiones en Ubuntu 11.10: ROS Electric y la librería PCL 1.6.

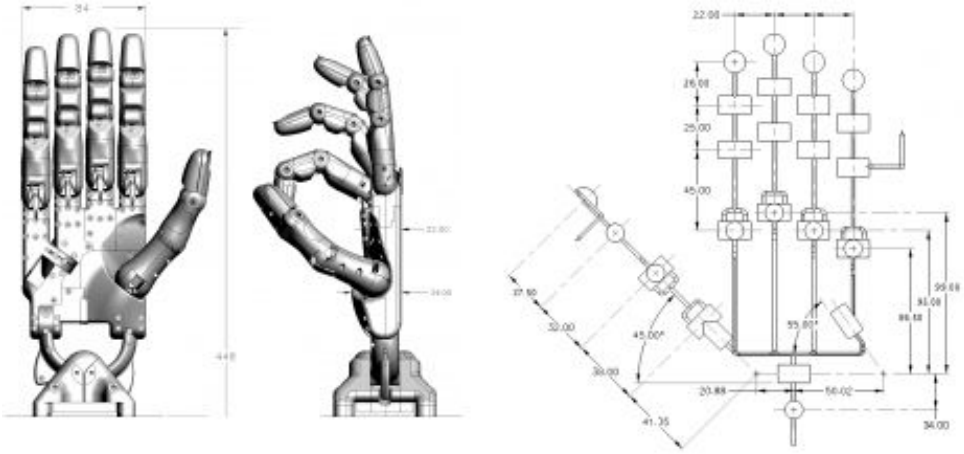


Figura 4.17: *Mano robótica de la compañía Shadow Robot [96]: (izquierda) dimensiones; (derecha) cinemática.*

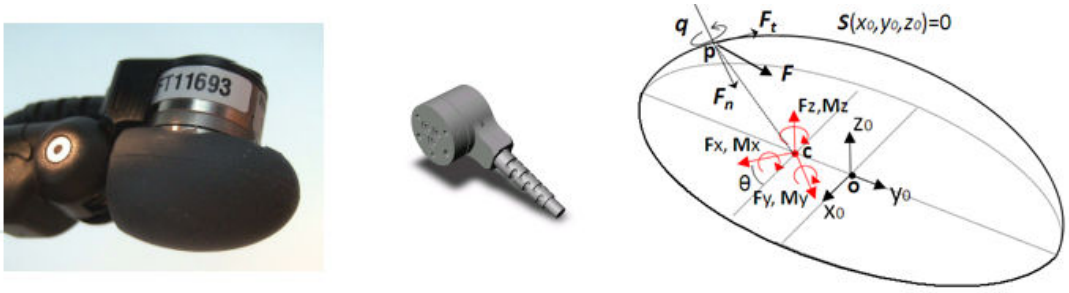


Figura 4.18: *Detalle del sensor ATI Nano17 [97]. De izquierda a derecha se muestra una imagen real del sensor sobre la mano, detalle del sensor sin elipsoide y del diagrama de fuerza/par del elipsoide.*

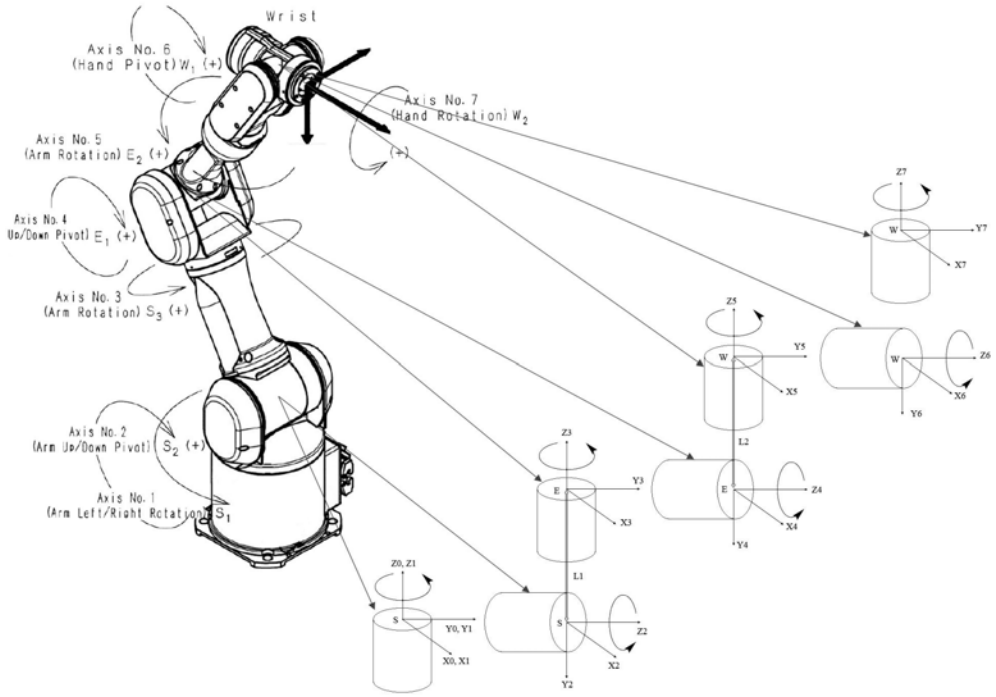


Figura 4.19: *Sistema de coordenadas del brazo robótico industrial PA-10 7 DoF [99].*

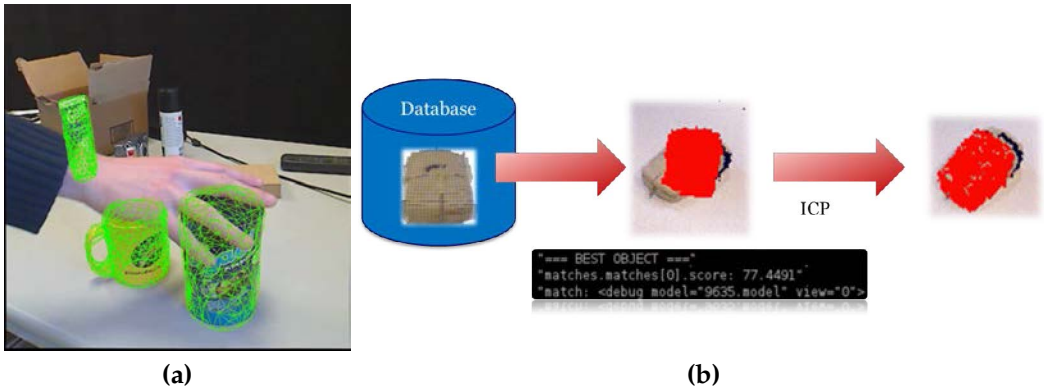


Figura 4.20: *Resultados representativos de los métodos propuestos de reconocimiento usando: (a) descriptor 2D; (b) descriptor 3D.*

4.4. Resultados Experimentales

Debido a que el sistema de percepción visual propuesto se enfoca a dotar a una plataforma robótica de autonomía y destreza suficiente para tratar con objetos desconocidos en un entorno no estructurado, las pruebas se centran en evaluar el novedoso método Xtru3D de reconstrucción propuesto usando una sola vista RGB-D. Una vez que el objeto se reconstruye se almacena en una base de datos, pero no se evalúan los algoritmos de reconocimiento ya que trabajos previos a esta tesis demostraron su eficacia [78]. A modo representativo se puede apreciar en la Fig. 4.20 resultados exitosos de reconocimiento usando los descriptores 2D y 3D propuestos. La Fig. 4.21 muestra la integración de los algoritmos de la estrategia propuesta para poder abarcar todos casos en un entorno de interiores no estructurado.

Para evaluar la eficacia de los algoritmos propuestos de percepción visual para un escenario de agarre y manipulación se han planteado dos casos de aplicación:

- Agarre: se valida la eficacia de los modelos 3D y la posición 6D obtenida para agarrar objetos desconocidos en una plataforma robótica integrando los algoritmos en el sistema de control y planificación de movimientos.

- Manipulación: se ha evaluado su integración en un sistema multi-modal de percepción visual y táctil desarrollado en el contexto del proyecto HANDLE. Los algoritmos de visión para reconocimiento, adquisición de un modelo 3D usando una sola vista así como el seguimiento del objeto una vez agarrado han sido validados junto a los algoritmos táctiles de identificación de posición mediante contacto desarrollados por KCL.

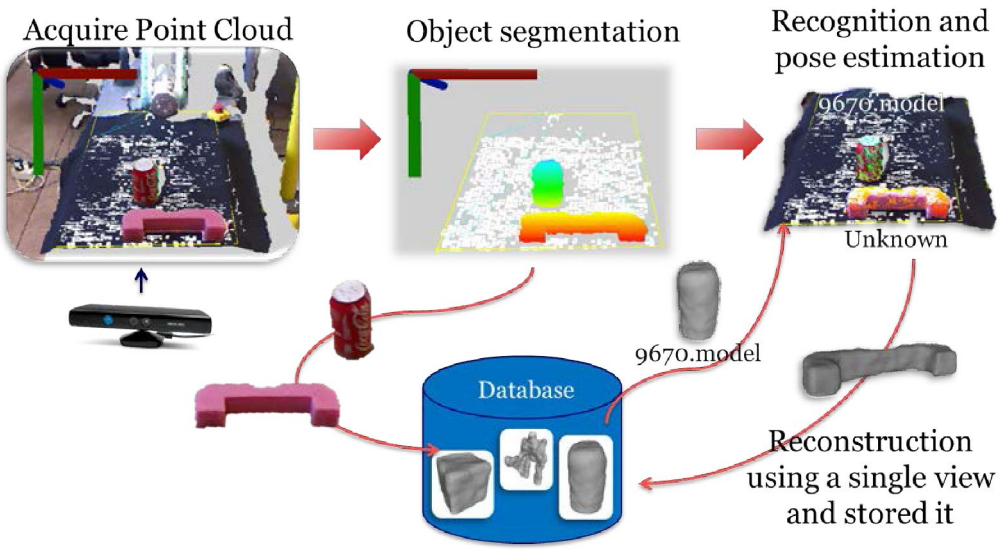


Figura 4.21: *Estrategia global de percepción 3D propuesta usando una sola vista para interiores levemente estructurados.*

A continuación se explicará en detalle cada una de las pruebas realizadas para la validación del sistema propuesto.

4.4.1. Evaluación del Método de Reconstrucción

4.4.1.1. Precisión del Modelo de Malla Reconstruido

El algoritmo propuesto de reconstrucción de una sola vista ha sido evaluado sobre un conjunto de 12 objetos reales con tamaños y formas diferentes, mostrados en la Fig. 4.22. Para cada objeto, se adquieren entre 5 y 9 modelos usando el algoritmo propuesto en el escenario mostrado en la Fig. 4.15, donde los objetos se encuentran sobre una mesa, colocados en diferentes posiciones y orientaciones. De esta forma, se han obtenido 72 modelos 3D a partir de una sola vista de la cámara Kinect. Para la evaluación, se ha calculado la diferencia geométrica entre el modelo de referencia y el reconstruido usando el algoritmo propuesto. Los modelos de referencia han sido adquiridos con un escáner láser comercial.

El tiempo de procesamiento del algoritmo es inferior a 2 segundos en un ordenador de 2Ghz para una nube de puntos de menos de 30000 puntos, significativamente mejor que el logrado en [68] para un número de puntos similar. Aunque este tiempo de computación es adecuado para la aplicación actual, su optimización es considerada como trabajo futuro.

Un software libre de procesamiento de modelos 3D, MeshLab [101], ha sido

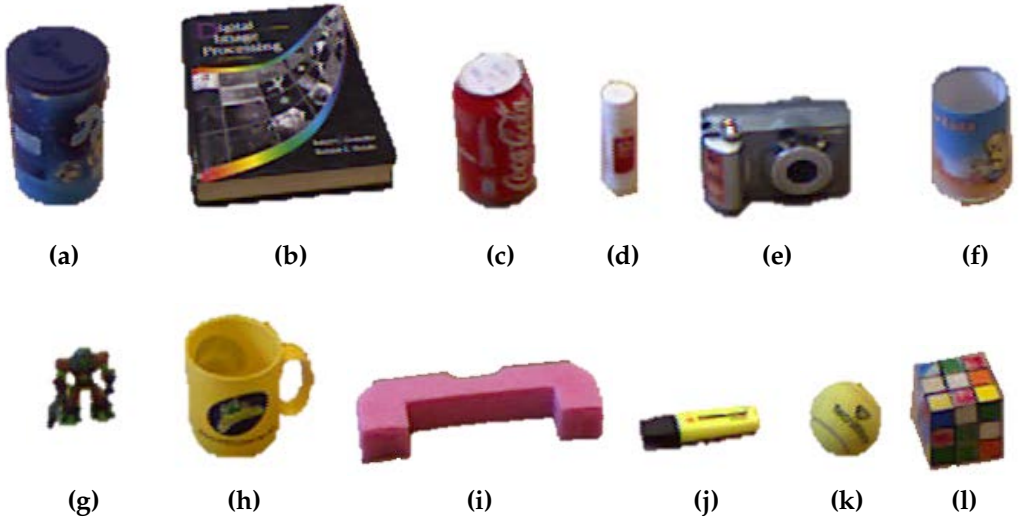


Figura 4.22: *Los 12 objetos reales de la base de datos: (a) bote, (b) libro, (c) lata, (d)pegamento, (e) cámara, (f) bote de lápices, (g) muñeco, (h) taza, (i) agarrador rosa, (j) subrayador, (k) pelota de tenis, (l) cubo de Rubik. Para cada objeto, se han adquirido al menos 5 imágenes del objeto en diferentes posiciones y orientaciones encima de la mesa.*

usado para computar la diferencia entre el modelo de referencia y el reconstruido estando ambos bien alineados en el mismo espacio. Un algoritmo iterativo de puntos cercanos, *Iterated Closed Point* (ICP) es usado para alinear las mallas de los modelos y posteriormente, se utiliza la distancia Hausdorff para medir la distancia geométrica entre ellos.

La Fig. 4.23 muestra la media y la desviación estándar entre los modelos de referencia y reconstruidos de todos los objetos. El error medio de todos los modelos es 3,87 mm y la desviación estándar de 0,96 mm. Teniendo en cuenta que los objetos son similares al conjunto usado en [68], nuestro algoritmo de extrusión proporciona una efectividad similar en comparación con métodos de simetría anteriores y una mejora significante para objetos grandes. Con Xtru3D, la media del error es inferior a 5 mm en todos los objetos, independientemente de su tamaño mientras que en [68], la media de error es inferior a 7 mm y para objetos grandes, inferior a 20 mm.

Es importante destacar que las medidas experimentales mostradas son estadísticamente completas en el sentido que cada modelo se obtuvo de una imagen capturada en diferentes localizaciones y orientaciones del objeto sobre una mesa, como se muestra en la Fig. 4.24. Este conjunto tiene en cuenta la mayor parte de posibles fuentes de errores, como detalles geométricos ocultos, reflexiones u otros efectos ópticos, que afectan a los resultados obtenidos e incrementan el error.

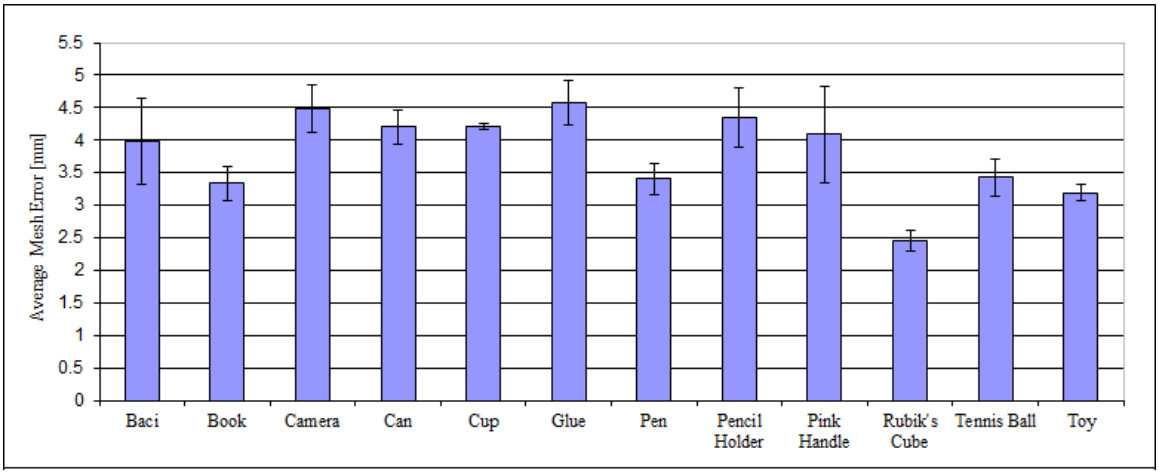


Figura 4.23: *Evaluación de la media y desviación estándar del error entre el modelo de referencia y el reconstruido de todos los objetos de la base de datos. La media de error es inferior a 5 mm en todos los objetos, siendo el error promedio inferior a 4 mm.*

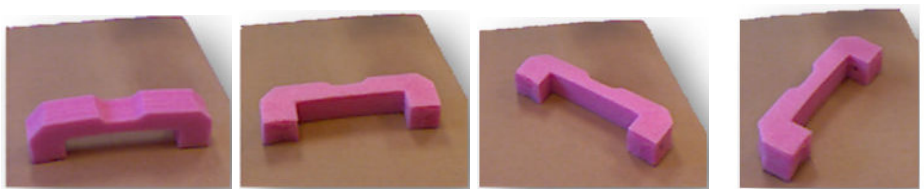
La Fig. 4.25 muestra como, debido a la posición del objeto, el error es inferior cuando las partes visibles proporcionan suficiente información para aproximar su geometría extruyendo la cara superior (caso 2 y 3), pero el error se incrementa cuando la parte superior no aporta demasiada información sobre la forma del objeto (caso 6).

Teniendo en cuenta que el tamaño de voxel escogido es 3 mm (Sección 4.2.6.1.1), el error mínimo con el que se pueden obtener reconstrucciones es en torno al valor mencionado de tamaño de voxel. Esto se puede observar en la tabla de errores medios de diferentes objetos, donde sólo el cubo de Rubik tiene un error medio en torno a 2,5 mm mientras que el resto es en torno a 3 mm. Si se desea más precisión en los modelos reconstruidos, entonces habría que sacrificar velocidad en el procesamiento teniendo que procesar más voxels de un tamaño inferior. Esto podría ser necesario para una determinada tarea, por lo que el algoritmo está preparado para soportar diversos tamaños de vóxeles.

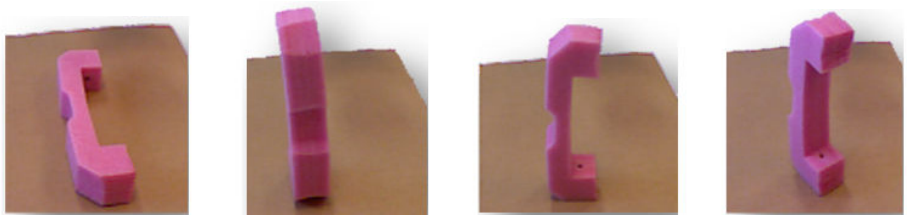
4.4.1.2. Modelos de Mallas 3D Reconstruidos

La Fig. 4.26 muestra algunos de los modelos adquiridos usando el algoritmo propuesto para el conjunto de evaluación formado por 12 objetos reales.

La Fig. 4.27, (a)-(c), muestra un ejemplo de buenos modelos 3D a pesar de la escasa información procedente de la nube de puntos. Como se puede apreciar, los



(a) *Caso 1.* **(b)** *Caso 2.* **(c)** *Caso 3.* **(d)** *Caso 4.*



(e) *Caso 5.* **(f)** *Caso 6.* **(g)** *Caso 7.* **(h)** *Caso 8.*

Figura 4.24: *El objeto denominado agarrador rosa situado sobre la mesa en las 8 orientaciones evaluadas.*

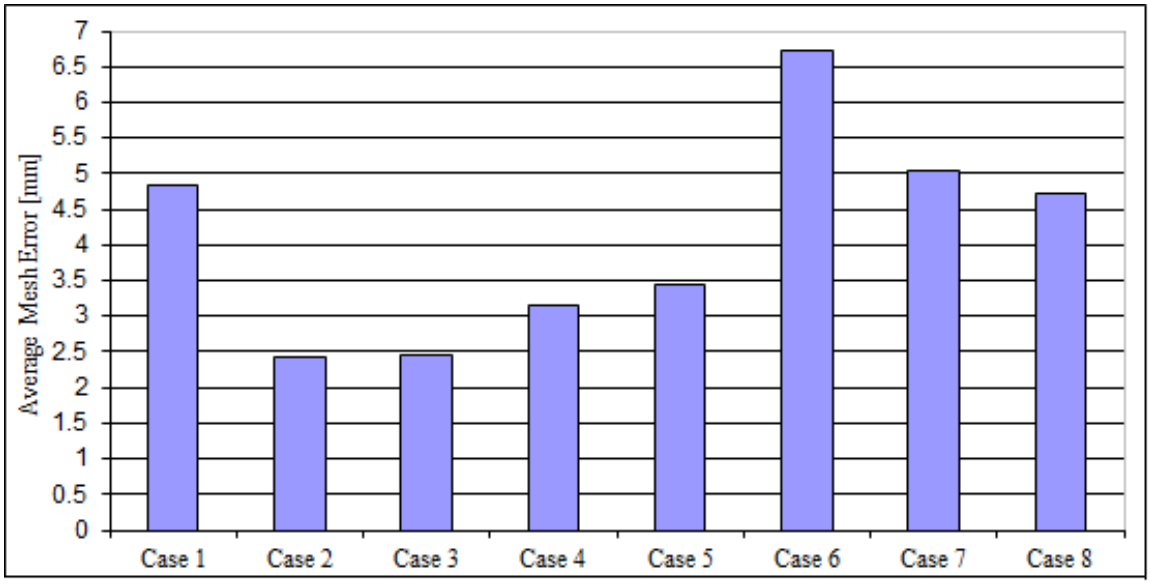


Figura 4.25: *Evaluación del error del modelo reconstruido para las 8 orientaciones del objeto denominado agarrador rosa situado sobre la mesa. Comparando con el modelo de referencia, el error medio es 4,09 mm y la desviación estándar es 1,49 mm.*

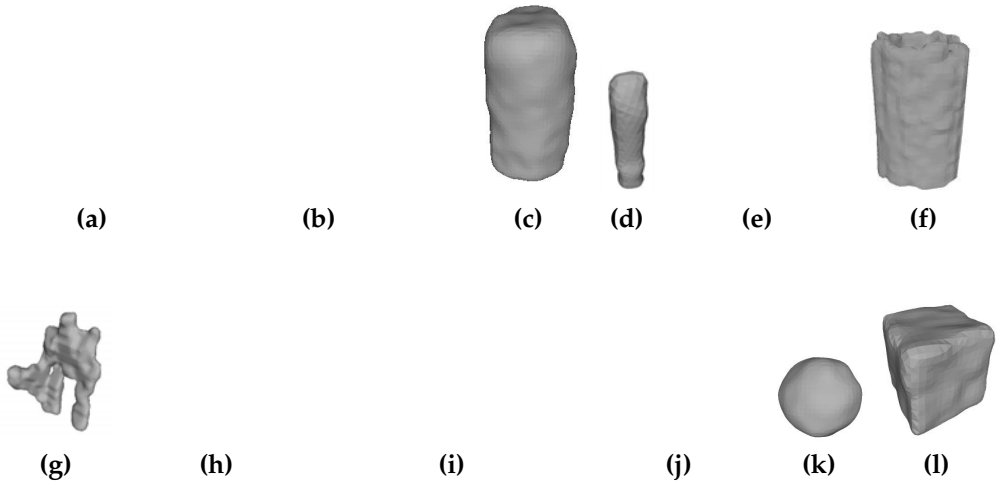


Figura 4.26: *Resultado de la reconstrucción a partir de una sola vista de 12 objetos reales de la base de datos mostrada Fig. 4.22: (a) bote, (b) libro, (c) lata, (d) pegamento, (e) cámara, (f) bote de lápices, (g) muñeco, (h) taza, (i) agarrador rosa, (j) subrayador, (k) pelota de tenis, (l) cubo de Rubik.*

modelos son más precisos cuanta mayor información proporcione la parte superior del objeto. Además, los pasos posteriores de segmentación y relleno de profundidad hacen más robusto el algoritmo y por tanto, los modelos resultantes, permitiendo reconstruir incluso en algunos casos donde apenas se tiene información procedente de la imagen original RGB-D. La Fig. 4.27,(d) y (e), muestra ejemplos de objetos cuya geometría no es la más adecuada para aproximar su forma mediante la extrusión de su parte superior. Cabe destacar que aunque los modelos obtenidos no son muy precisos, se obtienen estimaciones lo suficientemente útiles para el agarre, contexto de esta estrategia. La adición de otra cámara con un punto de vista diferente sería suficiente para obtener buenos modelos en estos casos.

4.4.2. Caso de Aplicación I: Agarre Mediante una Mano Robótica

La adecuación para el agarre de los modelos reconstruidos de objetos desconocidos ha sido evaluada para un solo objeto a modo representativo en la plataforma de la UC3M (Fig. 4.15), integrándolo con su sistema de agarre [102].

Los experimentos de planificación y agarre han sido realizados en el simulador de OpenRAVE [103]. La planificación se ha realizado mediante el algoritmo *Bi-directional Rapidly-Exploring Random Tree* (BiRRT) [104], diseñado para buscar de manera eficiente espacios no convexos. El planificador simula la trayectoria, evitando los puntos de colisión proporcionados por el sistema visual. Una vez que la posición y la orientación del objeto de interés son conocidos, se simulan agarres en

(a) (b) (c)

(d) (e)

Figura 4.27: *Modelos 3D reconstruidos con el algoritmo propuesto de los siguientes objetos: (a) bote de lapiceros, (b) un agarrador rosa, (c) una cámara, (d) una cámara en un trípode (e) una pelota de tenis. Izquierda: nube de puntos inicial. Derecha: modelo final reconstruido usando Poisson ((e) vista de lado y de frente).*

diversas posiciones para determinar aquellos que son estables para el objeto evaluado, como se muestra en la Fig. 4.28. Esta tabla de agarres estables se asocia al objeto y se utiliza para la planificación del agarre, una vez que el objeto es reconocido y su posición estimada. De esta forma, se calcula online el recorrido a realizar por el brazo robótico hasta llegar a la posición de agarre más estable en función de la posición y orientación del objeto en la escena.

La Fig. 4.29 recoge la secuencia de la trayectoria en simulación y en la plataforma robótica real, mostrando cómo el modelo reconstruido con Xtru3D es adecuado para el agarre. Se considera como trabajo futuro una evaluación más exhaustiva de agarres usando una sola vista en simulación y sobre la plataforma real robótica bajo estudio.

4.4.3. Caso de Aplicación II: Manipulación Robótica Usando Información Táctil

Los experimentos realizados en esta sección se llevaron a cabo en la plataforma de la UPMC, demostrador final del proyecto europeo HANDLE (Fig. 4.30).

La plataforma hardware dispone de un sistema robótico con un brazo de Shadow y una mano de la misma compañía que dispone de sensores táctiles ATI

Figura 4.28: *Cinco agarres de la tabla de agarres generada por OpenRAVE para el objeto "agarrador" rosa cuyo modelo ha sido generado usando el algoritmo propuesto.*

Nano17 (Sección 4.3.2). Además, cuenta con una cámara Microsoft Kinect montada verticalmente en el brazo robot. Se dispone de la arquitectura necesaria para controlar el hardware y el software:

- La arquitectura hardware se compone de cinco ordenadores con diferentes capacidades de computación, conectados por una red ethernet Gigabit y sincronizados usando el protocolo *Network Time Protocol* (NTP). Cada ordenador se dedica a las diferentes tareas de controla: mano, brazo, visión, manipulación y base de datos. La arquitectura de control está distribuida.

- La arquitectura software se compone de varios *stacks* de la versión Electric de ROS, cada pila contiene una agrupación de paquetes complementarios entre ellos. Un resumen del diagrama se muestra en la Fig. 4.31.

El sistema visual propuesto se engloba en la categoría de segmentación, reconocimiento y reconstrucción del diagrama de la Fig. 4.31. La adecuación del sistema visual propuesto ha sido validada en el demostrador final de HANDLE integrándolo con el resto de funcionalidades, concretamente con los algoritmos táctiles implementados por KCL. El diagrama de la base de datos global del proyecto HANDLE se muestra en Fig. 4.32.

Por tanto, el sistema visual se integra en un sistema de percepción que engloba algoritmos de visión y táctiles que se fusionan para implementar la manipulación robótica de forma eficiente [105]:

- El sistema de visión provee el modelo 3D del objeto así como su posición 6D (rotación 3D y traslación 3D).

- Los algoritmos táctiles proporcionan la posición de contacto, fuerzas normales y tangenciales, además del par de fuerzas local en la yema de los dedos de la mano robótica. Dada una posición inicial por el sistema de visión y la

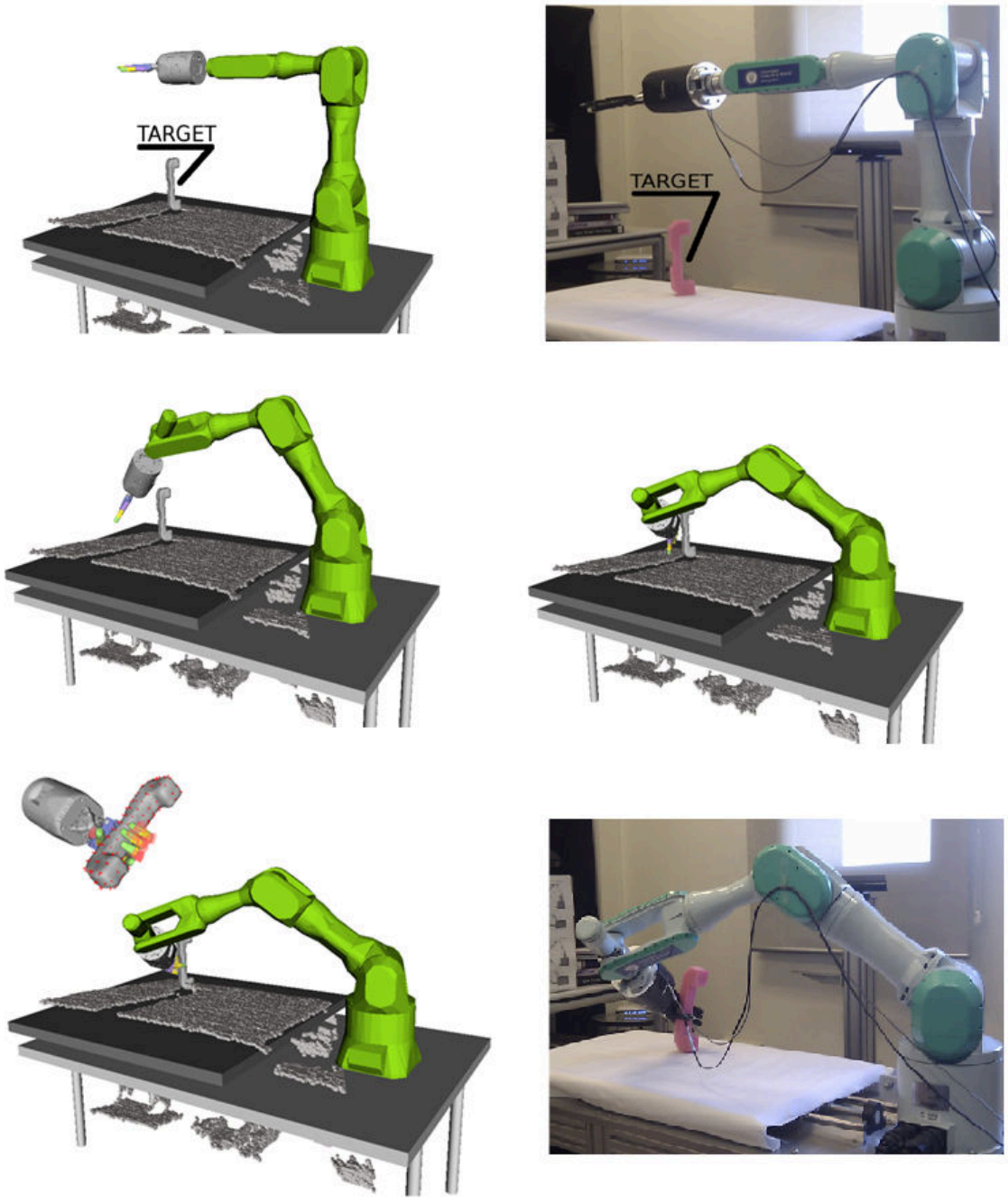


Figura 4.29: *Secuencia real y simulada de la trayectoria hacía la posición seleccionada de agarre, que ha sido calculada previamente off-line. Tanto la planificación de la trayectoria como el agarre ha sido calculado con OpenRAVE.*

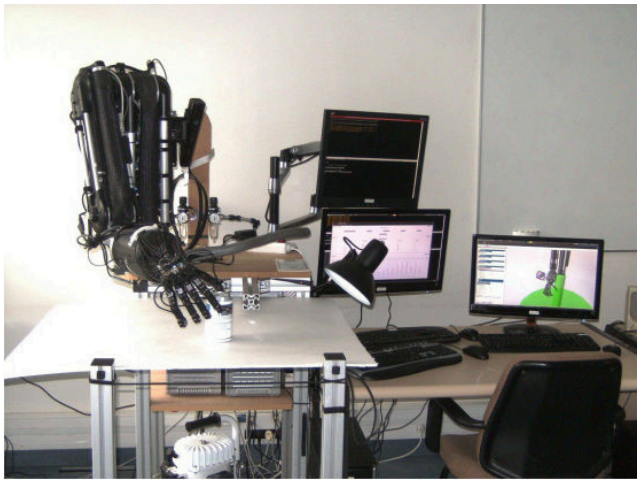


Figura 4.30: *Demostrador final en la UPMC.*

localización de contacto con las yemas, se calcula la posición del objeto mientras se manipula con un proceso iterativo que optimiza su posición mediante transformadas.

El método para verificar la efectividad del sistema propuesto de fusión en condiciones normales consiste en agarrar y manipular un objeto que se encuentra encima de una mesa. Se ha probado tanto con un objeto conocido (lata de refresco) como desconocido (bote de té). Para integrar los resultados visuales, táctiles y posiciones de la plataforma robótica, se ha replicado la configuración del robot en un visualizador en ROS.

Primeramente, se ha calibrado el sistema para conocer la posición de la cámara respecto a la base de la plataforma robótica como se muestra en la Fig. 4.33. Para ello se ha colocado un tablero de ajedrez 9x6 con cuadrados de 25mm en una posición conocida respecto a la base:

- Pose [x, y, z] (en metros): $[0,0675, 0,3975, -0,015]$
- Orientación (cuaternión): $[1, 0, 0, 0]$

Posteriormente, se ejecuta la segmentación y reconocimiento. En el caso de la lata, al ser un objeto conocido, se devuelve el modelo disponible en la base de datos y su posición. La nube de puntos del objeto se añade en el visualizador y se puede comprobar primeramente que existe un ligero desplazamiento debido a errores conocidos de calibración. Se realizan los movimientos pertinentes para colocar la mano en la posición de agarre y se van cerrando los dedos hasta detectar

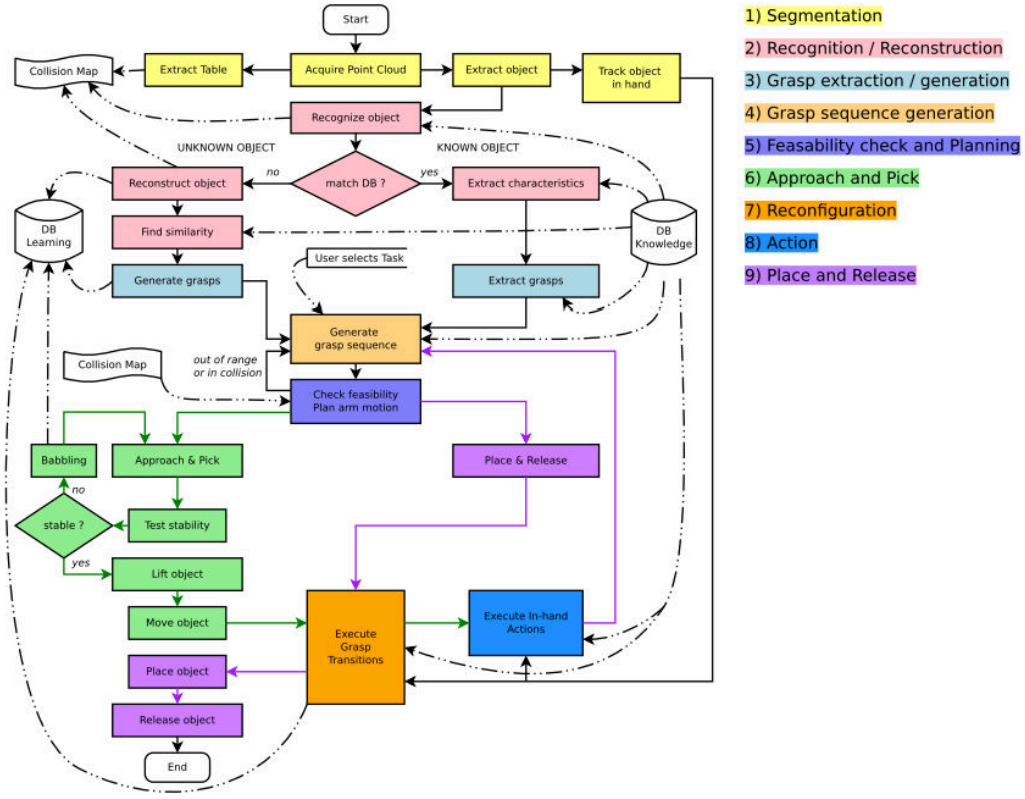


Figura 4.31: *Diagrama de la arquitectura software de la plataforma de la UPMC. Imagen tomada del entregable del informe final del proyecto HANDLE.*

contacto con la lata. Cuando el objeto está siendo agarrado con al menos dos dedos, la estimación dada por el sistema de visión se refina teniendo en cuenta la información táctil. La pose 6D visual se obtiene aplicando el descriptor SIFT (*Scale Invariant Feature Transform*) usado para reconocimiento mediante la evaluación de las características transformadas invariantes a escala [78]. Esa información visual se combina con la información táctil utilizando el algoritmo de optimización Levenberg-Marquardt que encuentra una transformación (vector de traslación y cuaternión de rotación) de la pose estimada por visión [106]. El algoritmo se ejecuta 10 iteraciones hasta que converge a una posición acorde entre la transformada de la superficie del objeto y las localizaciones de contacto de la mano robótica. Esta posición minimiza tanto la distancia entre los puntos de contacto a la superficie del objeto como el ángulo entre la dirección de la fuerza normal y la superficie normal.

La Fig. 4.34 muestra en gris la lata en la posición inicial dada por el sistema visual, y en rosa la posición rectificada con el algoritmo de fusión una vez agarrado. Hay que tener en cuenta que al agarrar el objeto, éste se mueve ligeramente

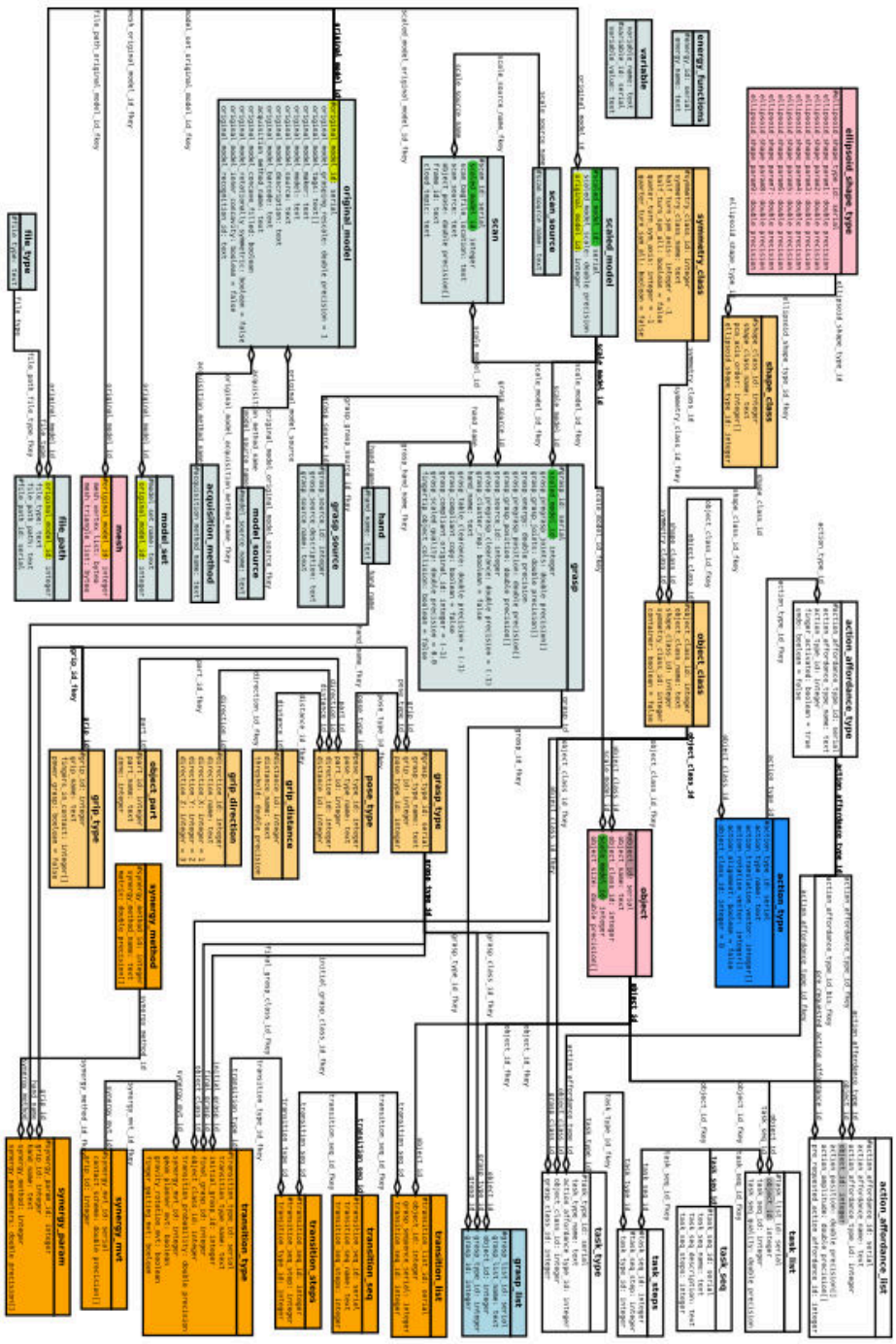


Figura 4.32: *Diagrama de la base de datos del proyecto HANDLE.*

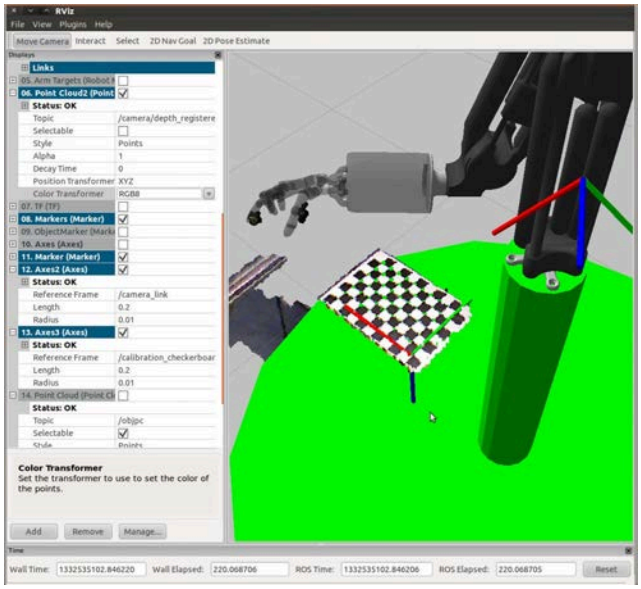


Figura 4.33: *Detalle en el visualizar 3D de ROS del resultado de la calibración del sistema.*

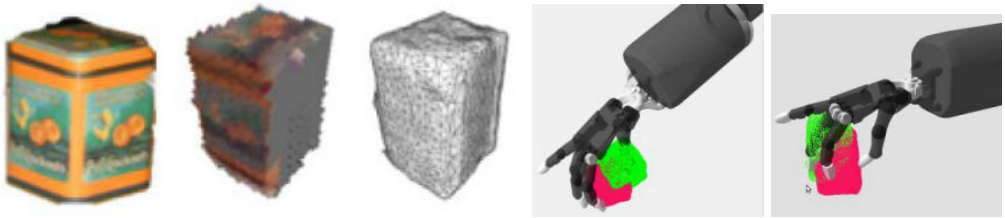
de la posición inicial y posteriormente las oclusiones no permiten hacer una buena manipulación con sólo la información visual. La distancia media entre el objeto y los puntos de contacto de los dedos es de 4,9 cm en la primera estimación, consiguiendo reducirse a 5 mm después de aplicar la corrección de pose con el algoritmo iterativo de fusión con información táctil. En simulación se había logrado un error de 2,5 mm [106]. La precisión del algoritmo viene limitada por la propia de la cámara de rango. En cuanto al rendimiento computacional, en un sistema real la duración media del algoritmo fue de 350 ms.

En el caso del bote de té, al ser un objeto desconocido se aplica el método de reconstrucción propuesto usando una sola vista como se aprecia en las imágenes (a)-(c) de la Fig. 4.35. Una vez que el objeto es reconstruido, se realiza un seguimiento de su posición 6D al cogerlo y manipularlo de la forma descrita anteriormente pero usando el modelo reconstruido del objeto [107]. Se ha probado a manipular el bote dentro de la mano y el método propuesto de corrección de pose mejora el error de la estimación de la visión con oclusiones en un 70 %, realizándolo en tiempos inferiores a 0,2 segundos. Esto permite el seguimiento en tiempo real de un objeto mientras se manipula, incluso cuando el objeto está ocluido o fuera del campo de visión de la cámara como se puede apreciar en las imágenes (d)-(e) de la Fig. 4.35.

Por tanto, se han logrado medidas precisas de los objetos agarrados a pesar de las limitaciones y posibles imprecisiones del sistema de calibración (errores provenientes de la precisión de la cámara, el uso de una sola imagen o incluso desviaciones del modelo cinemático del robot).

(a) **(b)** **(c)** **(d)**

Figura 4.34: *Manipulación de objeto conocido mediante la fusión de datos del sistema visual y táctil: (a) visualización en ROS de la fase previa al agarre tras obtener la posición de la lata visualmente (modelo en gris), (b) imagen real del agarre de la lata en la plataforma, (c)-(d) visualización en ROS de un momento del agarre, mostrando en gris el modelo en la posición dada inicialmente y en rosa el modelo en la posición final rectificada.*



(a) **(b)** **(c)** **(d)** **(e)**

Figura 4.35: *Manipulación de objeto desconocido mediante la fusión de datos del sistema visual (a)-(c) y táctil (d)-(e): (a) imagen del bote de té, (b) modelo voxelizado tras la extrusión (los voxels grises corresponden a las zonas no visibles por la cámara), (c) modelo final del bote tras la reconstrucción, (d)-(e) visualización del modelo del robot mientras agarra el bote, mostrándose en verde la pose del objeto dada por el sistema visual de seguimiento y en rosa la posición corregida.*

4.5. Conclusiones

En este capítulo se ha propuesto una estrategia visual para un escenario interior en un entorno semiestructurado en el ámbito del uso de robots para tareas domésticas. El sistema de percepción visual propuesto permite al robot hacer frente de forma autónoma a entornos poco estructurados de agarre y manipulación utilizando una sola vista RGB-D. Al no ser un entorno totalmente estructurado, para abordar la existencia de objetos desconocidos se ha desarrollado un novedoso algoritmo de reconstrucción 3D con una sola vista RGB-D al que hemos llamado Xtru3D. Los puntos clave y principales contribuciones de la estrategia visual propuesta para el escenario bajo estudio son:

1. Se ha desarrollado un sistema visual completo que incluye calibración, segmentación, detección, reconocimiento de objetos sobre una mesa y reconstrucción 3D de los objetos desconocidos, utilizando una sola vista proveniente de una cámara RGB-D tipo Kinect.

2. El enfoque se adapta perfectamente a cualquier escenario interior no estructurado donde constantemente se tiene que hacer frente a nuevos objetos que descansan sobre una mesa plana. Este escenario es común en robótica, pero es extensible también a otros ámbitos como la realidad aumentada.

3. En cuanto a la calibración, segmentación y reconocimiento de objetos se han integrado con éxito diversos métodos tanto de información 2D como de 3D. De esta forma, el sistema se adapta a la escena visualizada, pudiendo dar la posición 6D de los objetos reconocidos estando incluso ocluidos.

4. Para el caso de objetos desconocidos, se ha propuesto una nueva técnica denominada Xtru3D que utiliza los datos procedentes de una sola imagen RGB-D. Este método permite extruir una nube de puntos dispersa obtenida tras aplicar un detector de objetos sobre una mesa *tabletop object detector*:

   a) La sencillez, eficacia y rapidez del método radica en basarse en la propiedad de que la normal al plano de la mesa proporciona un eje de extrusión natural de los puntos correspondientes a la parte superior del objeto. La vista de esa zona de los objetos se logra simplemente orientando la cámara hacia la mesa y con una cierta altura, replicando la posición habitual de un adulto frente a una mesa o superficie plana con la que interactuar.

   b) El modelo 3D obtenido se perfecciona con una técnica de refinamiento desarrollada aprovechando la complementariedad de las imágenes de

color y profundidad. En este paso, es clave la inicialización elaborada cuidadosamente con los datos de profundidad una segmentación *graph-cut* basada en color.

c) Se ha logrado obtener el modelo 3D de objetos desconocidos con una sola vista con éxito. Para mostrar la eficacia del método propuesto, se ha evaluado de forma cuantitativa la exactitud de las mallas reconstruidas sobre un conjunto de 12 objetos reales de uso común:

- La media del error de todos los modelos es de menos de 4 mm y la desviación estándar de menos de 1 mm.
- Además, comparando los resultados con métodos anteriores [68], el algoritmo propuesto proporciona modelos 3D con precisión similar pero mejorando los tiempos significativamente. La mejora es incluso más significativa, tanto en precisión como en tiempo de ejecución en el caso de objetos grandes.
- Presenta algunas limitaciones cuando los objetos tienen zonas muy estrechas, o con objetos cuya parte superior no aporta suficiente información sobre la forma del objeto. Sin embargo, gracias a la generalidad del algoritmo propuesto, esto puede ser compensado añadiendo tantas cámaras como sean necesarias, aplicando la misma técnica a cada vista y finalmente juntando los voxels resultantes, similar a una extrusión en dos etapas [108]. Además, simetría y extrusión pueden complementarse, como en [109], donde se determina la forma extruída mediante la detección de simetrías planas de reflexión en una nube de puntos parcial.

5. La realización de los algoritmos bajo el framework ROS permite integrar fácilmente la estrategia propuesta en diversas plataformas de manipulación robótica para escenarios interiores poco estructurados:

- Se ha comprobado su idoneidad en dos plataformas robóticas diferentes, no solo en cuanto al robot en sí sino además en cuanto a la sala, objetos y localización de la mesa frente al robot. Los experimentos llevados a cabo para diferentes objetos muestran que los modelos son lo suficientemente precisos para poder realizar el cálculo de puntos de agarre viables en un entorno robótico interior poco estructurado.
- Por otra parte, se ha comprobado que el sistema de visión propuesto fusionado con el táctil desarrollado por KCL en el marco del proyecto HANDLE permite medidas precisas de los objetos agarrados.
- Por tanto, el sistema de percepción es fácil de integrar en escenarios

robóticos interiores poco estructurados, así como efectivo en tareas de agarre y manipulación de objetos, ya sean conocidos o no.

Por tanto, la estrategia de visión propuesta es adecuada en entornos interiores semiestructurados donde una cámara RGB-D tipo Kinect proporciona la información suficiente. En entornos mixtos, interiores y exteriores, este tipo de cámaras no son válidas debido a las condiciones de luminosidad por lo que hay que usar otro tipo de dispositivos de rango. Debido a ello, es necesario adaptar la estrategia al nuevo escenario como se verá en el próximo capítulo de la tesis.

Capítulo 5

Entorno Mixto Semiestructurado

5.1. Escenario: Sistema de Percepción 3D para Tareas de Agarre en un Sistema Robótico Móvil

La percepción 3D en el ámbito de agarre y manipulación robótica móvil en entornos complejos implica el modelado del escenario así como de los objetos que se encuentran a su alrededor. En este caso, los objetos de la escena son obstáculos y elementos clave para la tarea a realizar. Esta temática es de creciente interés por lo que numerosos grupos de investigadores trabajan en este tipo de sistemas de manipulación móvil en diversos entornos [94, 110, 111, 112, 113], siendo un contexto habitual el de robots de servicios.

Los robots industriales se centran en labores de fabricación mientras que los de servicios se utilizan en diversas áreas domésticas y profesionales, como sectores de limpieza del hogar, agricultura, sanidad, seguridad y defensa, entre otros. De estos sectores, los utilizados como respuesta en desastres así como en salud tienen un gran potencial de crecimiento futuro. El impulso dado para el desarrollo de robots dedicados a tareas de búsqueda y rescate en entornos urbanos degradados es debido fundamentalmente a la necesidad detectada durante el desastre nuclear de Fukushima [114], ocurrido en 2011.

En general, los robots de servicios son móviles por lo que tienen la capacidad de desplazarse en diversos ambientes, siendo habitual que tengan que hacer frente a situaciones en interiores y exteriores con cierto grado de incertidumbre. El sistema de percepción, debido a la falta de estructuración y al dinamismo de los entornos del mundo real, tiene que permitir al robot entender el entorno interpretando la información adquirida. En el caso concreto de interacción robótica con objetos en entornos semiestructurados se suele requerir que el robot responda de forma

rápida ante un mundo cambiante, por lo que la adquisición de imágenes y su procesamiento debe ser realizado acorde al tiempo requerido de ejecución. Además, es necesario realizar un agarre de objetos cuyas características no se conocen totalmente *a priori*, ya sea el modelo completo del objeto o su posición.

En este capítulo se propone una estrategia visual que aborda estas limitaciones para tareas de agarre en ambientes tanto interiores como exteriores, donde el conocimiento del mundo que le rodea y del objeto con el que interaccionar es reducido. Una vez identificado el objeto, debe realizar el agarre para la tarea a desarrollar. En este tipo de entornos semiestructurados se suelen hacer ciertas suposiciones para reducir la complejidad del espacio para el agarre y manipulación de objetos, ya que incluso en entornos no estructurados existe información significativa que puede ser explotada por el robot. La estrategia propuesta en este escenario tiene como premisa este criterio basándose en técnicas que explotan cualquier conocimiento *a priori* del objeto en ambientes humanos y se centran en características perceptuales de la tarea específica [60]. Por ello, se abordan ciertos enfoques característicos en tareas de agarre de robots móviles para simplificar este escenario poco estructurado.

En la mayoría de los enfoques de percepción se explota una restricción natural del espacio: los objetos están generalmente sobre planos, ya sean horizontales o verticales, lo que se conoce en inglés con el término *tabletop*. Hay situaciones en las que la escena no contiene un plano fundamental y en su lugar está compuesta por múltiples planos en diferentes orientaciones. La tarea de interaccionar con objetos que se encuentran sobre superficies planas es bastante común para robots móviles [113], por lo que la determinación de estructuras planas en la escena es una fase fundamental de la estrategia propuesta. Este planteamiento es bastante común a otros entornos con nube de puntos 3D [75, 115]. Por ello, en este escenario se contempla la interacción con objetos sobre superficies planas pero al ser un robot móvil no se conoce previamente dónde se encuentran estas superficies respecto al robot.

En cuanto al objeto, el escenario bajo estudio considera que ciertas características del objeto con el que interaccionar son conocidas pero no contempla el conocimiento previo del modelo ni su posición encima de la superficie plana. Para solventar esta falta de conocimiento *a priori* la estrategia propuesta se basa en extracción de características de color y forma. En este último caso buscando formas primitivas que componen el modelo del objeto. Esta premisa se ha elegido teniendo en cuenta la efectividad demostrada de este tipo de enfoques en entornos no estructurados para el agarre de objetos [116].

Por otra parte, en entornos semiestructurados y dinámicos la percepción tiene que hacer frente a una gran cantidad de información adquirida por diversos sensores. La existencia de múltiples sensores es común debido a que el robot es móvil,

por lo que el sistema hardware de visión debe funcionar en ambientes interiores y exteriores, además de facilitar la información necesaria tanto para navegación como para tareas de agarre. Se ha comprobado que el uso de cámaras estéreo junto con escáner láser 2D giratorio es adecuado para percepción 3D en este tipo de escenarios [117, 118]. Debido a las características de cada sensor se ha comprobado que para obtener un sistema robusto y eficaz es recomendable la fusión de los datos para combinar las ventajas de ambos sensores [119]: la precisión en la medida de la distancia y la información de color. Por tanto, el sistema propuesto de percepción 3D explotará al máximo las características de ambos sensores, integrando la información procedente de ellos.

Las fases de la estrategia propuesta de percepción combinando información procedente del láser 2D y la cámara estéreo se explican en detalle en la Sección 5.2. En este bloque se explicará cómo controlar las incertidumbres en la posición del objeto y la orientación dentro del espacio de trabajo del robot para que pueda ser agarrado de forma fiable. Para validar la estrategia propuesta se ha escogido un contexto robótico de rescate con un nivel de complejidad elevado debido tanto al entorno como a los requisitos de tiempos de ejecución por enmarcarse en la competición mundial *DARPA Robotics Challenge* (DRC). Dicho contexto se detallará en la Sección 5.3 y se validará mediante pruebas experimentales relacionadas con la tarea de manipulación de la competición en el entorno de simulación. Finalmente, tras evaluar los resultados, en la Sección 5.5 se recopilarán las conclusiones sobre la estrategia visual propuesta para el escenario bajo estudio de entornos interiores y exteriores poco estructurados.

5.2. Estrategia Propuesta

La estrategia propuesta de percepción visual para acciones de agarre en este entorno tiene en cuenta que un robot móvil se desplaza por un ambiente interior o exterior hasta llegar a la zona donde tiene que realizar la tarea de interaccionar con un objeto. Para lograr dicha acción de manera autónoma y sin colisionar con obstáculos, el robot debe ser capaz de identificar y localizar los diferentes objetos relevantes en el espacio de trabajo.

Como se comentó en la sección anterior, este tipo de robots habitualmente disponen de cámara estéreo y un escáner láser para poder abarcar la percepción 3D tanto para navegación como agarre y manipulación. El enfoque propuesto explota las capacidades de cada sensor por separado para buscar en conjunto el mejor resultado. Para ello es necesario que la nube de puntos procedente de la cámara estéreo y de los escaneos del láser estén calibrados para minimizar la diferencia en las zonas superpuestas y que la información redundante no genere ambigüedad.

En este caso, se parte del supuesto que ya están calibrados debido a que la posición de ambos dispositivos en el robot es conocida teniendo en cuenta los datos de la Unidad de Medida Inercial, IMU, (*Inertial Measurement Unit*) integrada en el robot.

Debido a la complejidad estructural del entorno, la estrategia empleada se basa en las pautas establecidas para reducir la dificultad en entornos poco estructurados. Concretamente explota las características conocidas *a priori* del objeto en ambientes humanos y se centra en obtener aspectos claves del objeto así como de la tarea específica a realizar [60]. Por tanto, se especifican una serie de premisas para abordar el entorno. La primera es que la estrategia propuesta se centra en tareas de agarre de objetos sobre planos, concretamente sobre una mesa o una pared. Esto es debido a que la tarea de interaccionar con objetos que se encuentran sobre superficies planas es bastante común en estos entornos, como se vio en la sección anterior. En segundo lugar, de entre las posibles formas básicas, se focaliza en objetos cilíndricos ya que son bastante comunes en objetos realizados por el hombre [120].

Teniendo en cuenta estas premisas, el sistema propuesto de percepción consta de varios pasos (Fig. 5.1).

1. **Preprocesamiento**: la nube de puntos del escáner láser contiene información del entorno hasta varias decenas de metros para disponer de la información 3D necesaria para desplazarse por el espacio libre evitando los obstáculos. Sin embargo, para tareas de agarre sólo se necesitan los puntos de la escena que se encuentran a corto alcance, por tanto, se filtran los escaneos eliminado los puntos distantes innecesarios.

2. **Segmentación** (remarcado en amarillo en Fig. 5.1): la nube de puntos 3D filtrada se analiza para encontrar los puntos que conforman un plano 3D. La mesa se considera el plano horizontal dominante sobre el suelo y la pared es el plano vertical más cercano a la mesa. Para calcular la trayectoria de agarre libre de colisiones, se añaden tanto la mesa como la pared al mapa de colisiones.

3. **Detección** (remarcado en verde en Fig. 5.1): se analizan los datos de profundidad mediante un detector de objetos sobre un plano, considerando tanto la mesa como la pared. De esta forma, se eliminan los puntos que están fuera del área de proyección de los planos y los restantes se agrupan en objetos individuales denominados *clusters*. Cada *cluster* se encierra en una caja delimitadora o *bounding box* que contendrán objetos candidatos con los que interaccionar.

4. **Reconocimiento** (remarcado en rosa en Fig. 5.1): en esta fase se integran los datos recibidos de la cámara estéreo debido a que la información de color es

clave. Para ello, se filtra la nube de puntos de color y profundidad (RGB-D) generada por la cámara estéreo teniendo en cuenta las cajas delimitadoras del bloque anterior. Debido a que las características conocidas del objeto de interés son reducidas se explotan características intrínsecas del mismo, como la forma y el color, proporcionando suficiente información para reconocer el objeto de interés de la tarea específica.

5. **Estimación de posición 6D** (remarcado en naranja en Fig. 5.1): una vez que el objeto ha sido reconocido, sus coordenadas se calculan con respecto al frame de las coordenadas globales del robot.

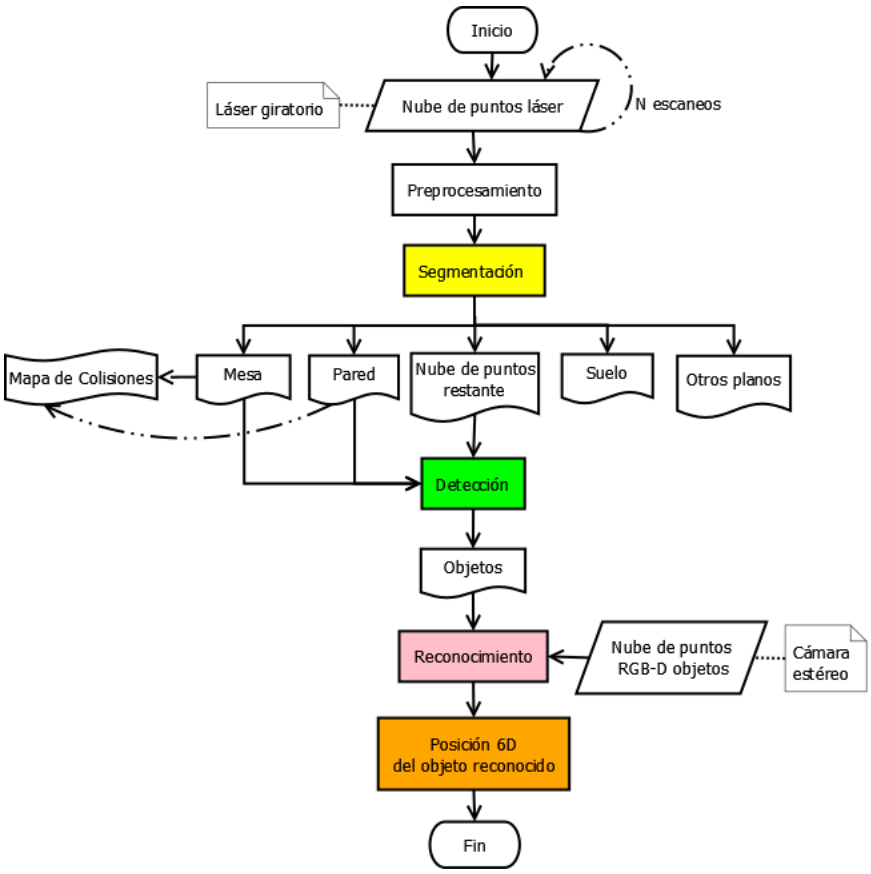


Figura 5.1: *Diagrama general del sistema de percepción propuesto para el entorno mixto semiestructurado bajo estudio.*

En las próximas subsecciones se explicará detalladamente cada bloque.

5.2.1. Preprocesamiento: Filtrado

El robot móvil debe poder desplazarse sin colisiones hasta el área de trabajo para la tarea de agarre. Para resolver los problemas de localización se suele emplear el enfoque Simultáneo de Localización y Mapeo, *Simultaneous Localization and Mapping*, (SLAM) [121] u otros métodos de discretización que mejoran el rendimiento computacional, como los basados en un modelo probabilístico tridimensional usando árboles de decisión [122]. En estos casos habitualmente se utilizan escáneres láser para obtener la información 3D de la escena debido a la alta precisión y fiabilidad de las mediciones. Se pueden utilizar tanto escáneres láser 3D como 2D. Sin embargo, debido al alto coste de los primeros, suelen emplearse escáneres láser 2D giratorios para poder obtener mediciones en todos los planos en lugar de sólo donde está posicionado [123]. De esta forma, se obtiene la medición 3D de los obstáculos a diferentes alturas, pudiendo detectar por ejemplo tanto las patas como la superficie de una mesa.

Teniendo en cuenta esta información de la escena, el robot se desplaza hasta el área de trabajo. Una vez que alcanza la superficie plana donde se encuentra el objeto con el que interaccionar, el robot se para. En este punto es donde comienza la estrategia propuesta, ya que para recibir información 3D consistente para el agarre se considera que el robot no está en movimiento.

En esta primera fase, la técnica propuesta recoge durante un determinado tiempo varios escaneos del láser giratorio para combinarlos en una sola nube de puntos. Posteriormente, se filtra esa nube teniendo en cuenta el espacio de trabajo requerido por el robot, eliminando los puntos distantes de dicha área. Esto es debido a que para agarrar y manipular objetos sólo se necesitan los puntos de las zonas que se encuentran al alcance del brazo robótico. Esta distancia debe poder ser configurada teniendo en cuenta el rango de movimientos del brazo. Además, este bloque puede adaptarse fácilmente a un escáner láser 3D debido a que sólo tendría que suprimirse la fase de combinar escaneos.

La gran ventaja de este bloque es que permite trabajar en las siguientes fases con la información 3D exclusivamente necesaria para la tarea de agarre, disminuyendo drásticamente los tiempos de procesamiento al trabajar con un reducido número de puntos.

5.2.2. Segmentación

Debido a los buenos resultados obtenidos en el escenario del capítulo anterior, la nube de puntos 3D filtrada, procedente del escáner láser, se segmenta usando *RANdom Sampling Consensus* (RANSAC) [73]. El método empleado se basa en la misma técnica, encontrando los puntos de la nube que conforman un plano 3D y calculando el modelo de regresión planar que mejor encaje con el subconjunto.

Posteriormente, se selecciona el soporte plano horizontal 3D dominante sobre el suelo, cuya envolvente convexa de puntos se identificará como la mesa en la que se apoyan los objetos [75]. Además, el plano vertical más cercano a ella se selecciona como pared. Para calcular la trayectoria de agarre sin chocarse con otros objetos, se añaden tanto la mesa como la pared al mapa de colisiones.

En esta etapa no es posible utilizar la información procedente de la cámara estéreo debido a que no se puede suponer que la pared y la mesa tengan suficiente textura. Por esto, para entornos con poca textura la información 3D dada por la cámara estéreo es poco robusta debido a la dificultad de encontrar entre dos imágenes de color zonas de correspondencia sin ambigüedad.

5.2.3. Detección

En este bloque de detección de objetos candidatos, al igual que en el escenario del capítulo anterior, se analizan los datos de profundidad mediante un detector de objetos sobre una superficie plana, denominado en inglés *table top object detector*. El concepto es similar al desarrollado por Rusu et al. [16] en el que extrae agrupaciones de puntos sobre una mesa y elimina aquellos que están fuera de la caja que engloba la mesa. En este caso, la superficie plana a considerar es tanto vertical como horizontal por lo que el algoritmo de detección propuesto engloba:

- Eliminación de información fuera de la mesa y de la pared: los puntos que están fuera de un prisma en torno al plano son eliminados.

- Clusterización: usando distancias euclídeas con un determinado umbral, los puntos restantes cuyas proyecciones caen dentro de la envolvente convexa correspondiente a la mesa o a la pared se agrupan en objetos individuales denominados *clusters*. Las agrupaciones de puntos que son demasiado pequeñas o no tocan ni la mesa ni la pared se eliminan.

- Cálculo de la caja delimitadora mínima: cada *cluster* se encierra en una caja delimitadora o *bounding box*, conteniendo objetos candidatos para la tarea de agarre. Para ello se calcula el centro de gravedad del *cluster* tomando el promedio de todos los puntos y teniendo en cuenta una serie de valores mínimos y máximos en las coordenadas XYZ.

Esta fase también considera la posibilidad de que la mesa esté pegada a la pared. De esta forma, se consideran que son objetos encima de la mesa o de la pared según su proyección y la localización más cercana del *cluster* al plano horizontal o vertical.

5.2.4. Reconocimiento

Tras la detección de varios objetos candidatos es necesario reconocer el objeto de interés para interaccionar con él. Debido a que no se conoce el modelo completo del objeto ni su tamaño, se plantea una técnica que utilice un descriptor de objeto que combine color y forma. Se escogen estas propiedades intrínsecas ya que no varían con el tamaño y peso del objeto. En este último caso se buscan formas básicas en lugar de un modelo completo del objeto, debido a que el nivel de complejidad del entorno semiestructurado se reduce por ser más fácil obtener información sobre formas básicas. La ventaja de esta integración es que se amplía el abanico de posibles objetos a reconocer ya que se pueden buscar tanto objetos de una forma básica con diferente tamaño, de un determinado color, entre dos zonas de colores diferentes, etc.

Como ejemplo de objeto cotidiano se ha utilizado el cilindro como forma básica a reconocer ya que puede relacionarse con una amplia lista de objetos comunes en el ámbito humano [120]. Cabe destacar que esta fase podría ser extensible a otras formas básicas como cubos, esferas y prismas, como han abordado otros autores con métodos similares basados en muestreo aleatorio [124].

En este bloque es necesario explotar la información de color por lo que se trabaja con la nube de puntos generada por la cámara estéreo. Para ello se toma de la fase anterior la caja delimitadora de cada *cluster* candidato y se seleccionan aquellos puntos de color y profundidad (RGB-D) que están dentro de los límites XYZ de cada caja. En este paso es crucial que la información 3D del láser y la cámara estéreo estén correctamente calibradas, y que se esté evaluando una zona común del campo de visión de ambos sensores. De esta forma, se integran en el proceso las dos nubes de puntos, utilizando el escaneo láser filtrado para estimar planos y objetos candidatos, cuya caja delimitadora se emplea para filtrar la nube de puntos RGB-D generada por la cámara estéreo.

Una vez que se tiene la nube de puntos RGB-D de los candidatos a ser objeto de interés se obtienen sus características de color y forma según el siguiente proceso:

- **Color**: es necesario realizar un paso previo de segmentación por crecimiento de regiones basada en color (*color-based region growing segmentation*) [125] sobre cada conjunto de nube de puntos RGB-D filtrada. De esta forma se puede comprobar si en un *cluster* existen varias zonas de colores diferentes. Esta fase permite encontrar un objeto de un determinado color o situado entre dos zonas de colores diferentes. El proceso de crecimiento de regiones (*Region Growing Process*) es similar al euclídeo de la clusterización. En este caso, en lugar de comparar las normales y la curvatura de los puntos, se compara el color. Para buscar los vecinos más cercanos de cada punto RGB-D también se utiliza un árbol kd (k-dimensional). Este tipo de árboles son una estructura

de datos que organiza los puntos en un espacio de k dimensiones. En este algoritmo, el primer paso es que las regiones con una pequeña diferencia en media entre los colores se fusionan, teniendo en cuenta un umbral. En segundo lugar, las agrupaciones que son más pequeñas que un tamaño mínimo se fusionan con sus vecinos.

- **Forma**: se realiza una segmentación y detección de una forma cilíndrica sobre la nube de puntos bajo evaluación. Para ello, se calcula el modelo del cilindro que mejor encaje con el subconjunto mediante RANSAC [73], aplicando la búsqueda de los puntos de la nube que conforman un cilindro. El objeto quedará definido por el eje de orientación, un punto del eje y el radio del cilindro. Este ajuste es bastante común para detectar objetos con estructuras geométricas comunes, por ejemplo, detectando una taza mediante el ajuste de un modelo de cilindro. Para dar mayor flexibilidad a la estrategia, no se ponen restricciones respecto a cómo se encuentra el eje del cilindro respecto al plano.

El descriptor obtenido con información de color y forma proporciona la información suficiente para reconocer al objeto de interés, buscando de entre los candidatos aquella nube de puntos cuyo valor de similitud sea mayor. La función de comparación de coincidencias se implementa evaluando color y forma e integrando los resultados, teniendo como entrada la información conocida del objeto a agarrar. Estas características conocidas se deben especificar junto con la tarea a realizar.

5.2.5. Estimación de posición 6D

Una vez que el objeto con el que interaccionar ha sido detectado, se calcula su posición 6D (vector de traslación y cuaternión de rotación) con respecto a las coordenadas globales del robot. Para ello se tiene en cuenta la característica extraída del bloque anterior en relación a su forma. Como se comentó anteriormente, se consideran objetos que se pueden aproximar a una forma básica, como el caso de un cilindro, por lo que su posición y su eje se calculan mediante el modelo de la forma básica que más se aproxime a la nube de puntos. Para ello se utiliza una técnica probabilística de concordancia entre la nube de puntos y el cilindro ideal como se ha explicado anteriormente. Teniendo en cuenta que la orientación puede verse influenciada por la simetría, para el agarre se considera fundamentalmente la dirección de los ejes.

Debido a que no se conoce previamente el peso del objeto, el centro de gravedad se considera adecuado como punto inicial de agarre. De hecho, para tareas de

agarre suele ser habitual tener una serie de agarres predefinidos para formas primitivas (cubo, esfera, prisma triangular y cilindro). Por este motivo, el punto que se proporcionará para el agarre es el centro de gravedad, la normal respecto a la superficie plana y la orientación dada por la dirección del eje del cilindro.

Una vez explicados los distintos bloques de la estrategia, a continuación se validará en un contexto experimental representativo. Para ello se selecciona un entorno donde se tengan que realizar tareas, tanto en interiores como en exteriores, de interacción con objetos de los que apenas se tiene información *a priori*.

5.3. Contexto Experimental

El sistema de percepción desarrollado en este capítulo se enmarca en el contexto de una competición robótica celebrada a nivel mundial denominada *DARPA Robotics Challenge* (DRC) [126]. Las tareas de rescate en los desastres, ya sean naturales o provocados por el ser humano, conllevan grandes riesgos para la salud de los operarios, poniendo en peligro su vida por salvar las de otros. El objetivo del DRC consiste en desarrollar robots terrestres capaces de ayudar a los seres humanos en la realización de tareas peligrosas, manipulando objetos y escombros en los escenarios de desastre. Para interactuar con el entorno humano en operaciones de respuesta a desastres, los robots requieren múltiples capas de software y herramientas que cada equipo tiene que desarrollar. Para evaluar su desarrollo, la competición DRC tenía tres fases:

- El *Virtual Robotics Challenge* (VRC) en Junio 2013: los mejores equipos de esta competición virtual pasan a las siguientes fases en entornos reales.
- La competición DRC preliminar en Diciembre 2013.
- La competición DRC final en Junio 2015.

El VRC fue la primera fase de esta competición y consistía en controlar en simulación un robot a través de una carrera de obstáculos realizando un conjunto de tareas complejas relacionadas con locomoción, percepción y manipulación en un entorno virtual. El equipo SARBOT (Search And Rescue roBOT) [1], compuesto por la Universidad Carlos III de Madrid, la Universidad Politécnica de Madrid, la Universidad de Alcalá y el Centro de Automatización y Robótica UPM-CSIC, participó en dicha competición.

Los escenarios del VRC requieren tareas de agarre y manipulación en entornos poco estructurados, tanto en interiores como en exteriores, por lo que el método

[1] http://www.sarbot-team.es/

propuesto proporcionará la información necesaria para ejecutar las tareas. Por tanto, es un contexto ideal para validar el sistema de percepción visual tridimensional desarrollado.

5.3.1. Ámbito de Aplicación

La estrategia propuesta se validará en el entorno de simulación de la competición DRC. El Reglamento del VRC [127] indica que un robot humanoide debe ser capaz de resolver tres tareas representativas de los retos a superar en los entornos reales. Cada una de las tareas se desarrolla en un determinado escenario simulado, como se puede ver en la Fig. 5.2.:

1. Conducción de un vehículo por una carretera salvando los obstáculos que encuentre: el robot deberá caminar una corta distancia, subirse al vehículo, conducir a una velocidad inferior a 16 km/h, bajarse del vehículo y caminar hacia la meta.

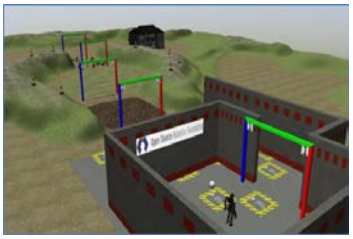
2. Locomoción por un terreno irregular de diversa complejidad: el robot debe salir de la zona inicial y seguir un determinado circuito que contempla superficies planas, con cuestas, lodo y escombros.

3. Conexión y manejo de una manguera: el robot debe salir de la zona inicial y dirigirse hacia una mesa donde se encuentra la manguera. Esta tarea se puede descomponer en cuatro subtareas:

   - Agarrar y levantar la manguera.
   - Transportarla hasta la tubería de la pared.
   - Conectar la manguera a la tubería.
   - Girar la válvula de agua situada en la pared.

Para cada tarea hay 5 rondas de un máximo de duración de 30 minutos, dando lugar a 15 pruebas en el VRC. En cada una de las rondas las condiciones de estos escenarios varían: fricción, luz, posición, dimensión de los objetos, etc. Durante la competición, la simulación se produce en la nube por lo que existe una latencia de 500 ms y para cada ronda se limitan de forma más o menos restrictiva los bits de subida y bajada.

La evaluación de la estrategia desarrollada en este capítulo se va a centrar en el ámbito la primera subtarea de la tarea 3, con el objetivo de dotar al robot de la percepción visual necesaria para agarrar la manguera interpretando el escenario con los sensores disponibles. Por tanto, la primera meta a alcanzar es la identificación y localización de los objetos de interés del espacio de trabajo, y en particular



(a) **(b)** **(c)**

Figura 5.2: *Detalle de los escenarios de las tres tareas VRC. (a) Tarea 1; (b) Tarea 2; (c) Tarea 3.*

la manguera y su conector. Según las especificaciones, la manguera se modela con un máximo de diez segmentos rígidos. Sin embargo, su peso, su rigidez, el tamaño de la rosca, y el diámetro del mango del conector de la manguera pueden variar en las diferentes rondas de la competición, lo que significa que son parámetros que no se conocen *a priori*.

En la tercera tarea del VRC (Fig. 5.2c), la manguera se encuentra sobre una mesa, mientras que el tubo vertical y la válvula están en una pared. Por lo tanto, se cumple la restricción física mencionada de la ubicación de los objetos en entornos humanos sobre planos.

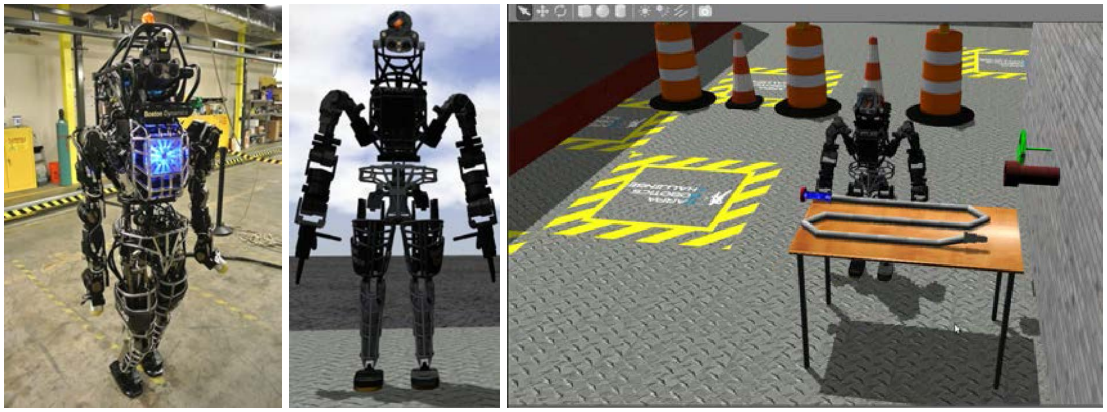
A continuación se detallará más en detalle la configuración empleada para la competición.

5.3.2. Configuración

Todos los equipos compiten en simulación con el mismo robot denominado Atlas (Fig. 5.3), desarrollado por Boston Dynamics, Inc. para este desafío, y basado en un robot humanoide anterior denominado Petman. El robot tiene 2 brazos con manos de la compañía Sandia National Laboratories, dos piernas y un torso. En total tiene 28 grados de libertad, *Degrees Of Freedom* DoF, (6 en cada extremidad, 3 en el torso y 1 en el cuello), y 4 DoF en cada mano. El robot real mide 2,25 metros y pesa 150 kg.

El Atlas está equipado de varios sensores: una cámara estéreo y un láser de rango en su cabeza, una unidad de medida inercial (*Inertial Measurement Unit*, IMU) de 3 ejes para la odometría, sensores de fuerza en los pies y muñecas, de esfuerzo en cada unión y sensores táctiles en las manos. La cabeza del sensor del robot es el modelo MultiSense-SL [28] (Fig. 5.4) de la compañía llamada Carnegie Robots, siendo sus características las siguientes:

- Cámara estéreo: rango desde 0.4 metros (m) hasta 18 m. Resolución de profundidad: ± 0.31 milímetro (mm) a 1 m; ± 30.00 mm a 10 m.



(a) (b) (c)

Figura 5.3: *Robot Atlas. (a) Robot real (cortesía de Boston Dynamics); (b) Modelo del robot Atlas en Gazebo; (c) Robot Atlas en simulación en el escenario de la tarea 3.*

(a) (b)

Figura 5.4: *Cabeza sensora del robot Atlas. (a) Real; (b) Simulada en Gazebo.*

- Láser: rango desde 0.1 m hasta 30 m. Precisión: ± 30 mm@<10 m; ± 50 mm@>10 m.

El robot Atlas y los escenarios están simulados en en el *DARPA Robotics Challenge Simulator* (DRCSim), una plataforma de código libre (*open-source*) hecha sobre el entorno de Gazebo [2] bajo el *middleware Robot Operating System* (ROS) [100]. El software de simulación de los sensores de la cabeza está disponible en DRCSim. El terreno y los objetos que constituyen el entorno están disponibles como archivos *SDF world*.

[2] http://gazebosim.org/

La Fig. 5.5 muestra el entorno de computación del VRC, donde el DRCSim se ejecuta en la nube desde un ordenador dedicado. Durante cada ronda se utilizan cuatro ordenadores en la nube: uno para el simulador, dos para el código y otro para el tráfico de datos [128]. Cada equipo tiene una Unidad de Control del Operador, *Operator Control Unit* (OCU) que sirve como estación de control de los comandos del robot.

Figura 5.5: *Diagrama de bloques del entorno de computación del VRC [128].*

Los algoritmos se han desarrollado basándose en los implementados en la librería *Point Cloud Library* (PCL) [76]. Estos algoritmos propuestos han sido encapsulados en varios nodos ROS procesando las nubes de puntos procedentes del láser y la cámara estéreo en simulación, usando la librería PCL en ROS dentro del DRCSim. Como herramienta de visualización 3D del escenario se empleará RViz, disponible para ROS. En este trabajo, se han utilizado las siguientes versiones en Ubuntu 12.04 (64 bits): ROS fuerte, DRCSim 2.6, Gazebo 1.8 y PCL 1.7.

5.4. Resultados Experimentales

El sistema de percepción propuesto está enfocado a tareas de agarre en entornos complejos dinámicos por lo que se van a validar los algoritmos en el entorno previsto de la tercera tarea de clasificación del VRC. Como se comentó en la anterior sección, la cabeza sensora del robot Atlas está equipada con un láser y una cámara estéreo que proporcionan la nube de puntos 3D a procesar.

La evaluación de la eficacia de los algoritmos se realiza en dos fases, evaluando primeramente la estrategia de percepción visual adaptada a la tarea específica y posteriormente para dos casos de aplicación: agarre de la manguera y en la propia competición del VRC.

5.4.1. Evaluación de la Estrategia de Percepción Visual Adaptada al VRC

Para validar la estrategia propuesta y reducir el nivel de desconocimiento se adapta el sistema de percepción a la tarea concreta a desarrollar. Para ello, aunque no se conoce el modelo completo de los objetos durante la competición, se tiene en cuenta todo el conocimiento previo disponible del espacio de trabajo (Fig. 5.6a) [127]:

- La manguera descansa sobre una mesa, estando formada por varios segmentos rígidos blancos, un conector de color azul terminando en una boquilla roja de rosca. Todo ello con forma cilíndrica.
- El número de elementos rígidos de la manguera puede variar hasta un máximo de 10. Además, el radio del cilindro puede también ser distinto a lo largo de la competición.
- El color entre el conector y el resto de la manguera siempre será diferente, aunque puede variar entre las diferentes rondas de la tarea.
- La manguera puede estar colocada sobre la mesa de una forma diferente a la mostrada.
- La tubería y la válvula se sitúan sobre la pared. Los colores de ambas son rojo y verde, respectivamente.

Por tanto, la estrategia de percepción 3D desarrollada se evalúa comprobando la eficacia en la detección del conector de la manguera así como su posición, proporcionando un mapa de colisiones para asegurar en los casos de aplicación que el robot puede realizar el movimiento por una zona libre de obstáculos. Para ello, también se dará la posición de la válvula y la tubería en el caso de estar en el campo de visión.

Para probar el algoritmo de percepción para agarre, el robot se inicializa en una posición predefinida en la que se garantiza que puede alcanzar el área de trabajo. La cabeza del sensor se posiciona, mediante los actuadores del cuello, para inspeccionar la mesa. El sistema de coordenadas de referencia será el dado por la pelvis del robot, disponiendo de todas las transformadas necesarias al origen.

(a) **(b)**

Figura 5.6: *El espacio de trabajo de la tercera tarea del VRC. (a) Captura de pantalla del mundo de Gazebo desde diferentes puntos de vista; (b) Nube de puntos del láser de rango girando durante 10 segundos, visualizado en RViz.*

Teniendo en cuenta estas premisas, a continuación se detallará cada una de las fases de la estrategia propuesta adaptada al contexto experimental, evaluando el resultado obtenido.

El láser comienza a girar una vez que el robot está en la posición prefijada, y con la cabeza orientada hacia el espacio de trabajo (Fig. 5.6a). Para obtener una nube de puntos cartesiana de coordenadas 3D (XYZ) se ensamblan los escaneos del láser girando durante 10 segundos (Fig. 5.6b).

Los escaneos del láser contienen puntos de la escena hasta distancias de 30 m. Sin embargo, para tareas de manipulación sólo se necesitan los puntos de la escena que se encuentran a corto alcance. Por tanto, se filtran los escaneos eliminado los puntos distantes innecesarios (Fig. 5.7a). La nube de puntos procedente de la cámara estéreo (Fig. 5.7a) y los escaneos filtrados del láser (Fig. 5.7c) están calibrados, minimizando la diferencia entre ellos en las zonas superpuestas (Fig. 5.7d). Como cabía esperar, existen deficiencias de la nube de puntos RGB-D en la zona de la mesa (Fig. 5.7b) debido a su poca textura. Sin embargo, en el caso del láser la nube de puntos es más escasa en algunas zonas por proporcionar un campo más amplio de visión.

La información filtrada de los escaneos del láser se utiliza para la segmentación de planos (Fig. 5.8a). La mesa se considera el plano horizontal dominante por encima del suelo y la pared se considera el plano vertical más cercano. El resultado se puede apreciar en la Fig. 5.8b donde los puntos correspondientes a la mesa se destacan en verde y los de la pared en azul.

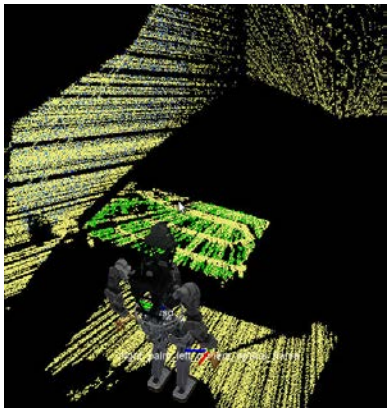
(a) **(b)** **(c)** **(d)**

Figura 5.7: *Visualización de las nubes de puntos 3D procedentes por del láser de rango y la cámara estéreo, mostrado en RViz. (a) La nube de puntos filtrada se muestra en naranja, y el escaneo láser completo en blanco; (b)nube de puntos RGB-D de la cámara estéreo; (c) nube de puntos filtrada del láser; (d) comparación entre la nube de puntos del láser y la cámara estéreo.*

Una vez que la mesa y la pared son detectadas, los objetos de interés se consideran aquellos que se encuentran encima del plano. Siguiendo la estrategia propuesta, se detectan los puntos sobre el plano (puntos blancos de la Fig. 5.8c) y se agrupan en *clusters* considerados como objetos candidatos. Para terminar la fase de detección se calculan las cajas delimitadoras para todos los objetos detectados en el espacio de trabajo. En la Fig. 5.9 se muestran dos cajas que contienen las nubes de puntos correspondientes a la válvula y a la manguera. Como se puede apreciar, la manguera se encierra por completo en una caja delimitadora alineada con el eje en la parte superior de la mesa. En este punto se pasa a la fase de reconocimiento del conector.

La manguera está formada por segmentos rígidos y el conector. Aunque sus propiedades pueden variar durante la competición, el color entre el conector y el resto de la manguera será siempre diferente. El enfoque adoptado en la estrategia propuesta permite reconocer el conector de la manguera en la nube de puntos RGB-D proporcionada por la cámara estéreo utilizando segmentación de crecimiento de regiones basada en color (*color-based region growing segmentation*).

En esta fase se integran los datos recibidos de la cámara estéreo debido a que la información de color es clave. Para ello, se filtra la nube de puntos de color y profundidad generada por la cámara estéreo teniendo en cuenta los límites de las cajas delimitadoras del bloque anterior (Fig. 5.10a). El sistema propuesto utiliza una nube de puntos filtrada para reducir el tiempo de funcionamiento debido a que la eficiencia de tiempo de cómputo es un requisito fundamental para la competición del VRC.



(a) **(b)** **(c)**

Figura 5.8: *Visión general de la detección de planos: la nube de puntos filtrada se muestra en amarillo, la mesa en verde, la pared en azul y los clusters de interés en blanco. (a) Detalle de la nube de puntos filtrada, mesa detectada y pared; (b) Detalle de la mesa y la pared; (c) Detalle de los clusters de interés sobre la mesa y la pared.*

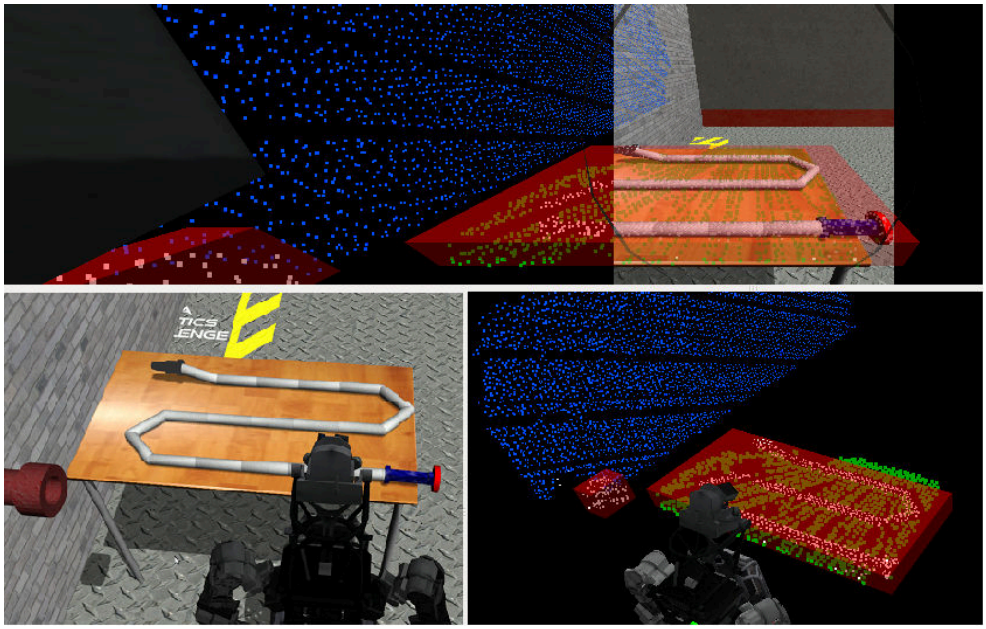


Figura 5.9: *Estimación de la caja o bounding box que contiene los objetos relevantes del espacio de trabajo.*

Debido a las características de la cámara estéreo, la nube de puntos RGB-D contiene lagunas y menos información, existiendo zonas de la manguera sin información 3D. Por esta razón, la manguera ya no aparece completa como en la nube

de puntos del láser sino que los puntos se agrupan en tres unidades individuales (Fig. 5.10b). Debido a que se tiene que trabajar con la información de color, se aplica segmentación de regiones coherentes basada en color a la nube RGB-D de cada unidad. Como resultado de ello, la manguera se segmenta en tres áreas: elementos rígidos, conector y la boquilla de rosca del conector (Fig. 5.10c).

(a) **(b)** **(c)**

Figura 5.10: *Detalle de la detección de la manguera. La nube de puntos verde corresponde a los puntos pertenecientes a la mesa procedentes de la información 3D del láser. (a) Caja delimitadora según el láser, conteniendo la nube de puntos RGB-D de la cámara estéreo; (b) Cajas delimitadoras según la nube RGB-D; (c) Resultado de la segmentación basada en color a la nube de puntos RGB-D: mesa (puntos rojos), elementos rígidos de la manguera (verde oscuro), conector (beige) y rosca (rosa).*

El conector de la manguera se detecta teniendo en cuenta que es la región azul que se encuentra entre dos zonas de diferente color (Fig. 5.11a). En este escenario, el sistema propuesto se ha ejecutado 50 veces, en 48 de las cuales se ha realizado con éxito la detección del conector de la manguera.

Por último, se estima la posición del conector. Debido a que la forma 3D de la manguera puede aproximarse por un cilindro, su posición y su eje se calculan mediante el modelo del cilindro que más se ajuste a la nube de puntos del conector de la manguera (Fig. 5.11b). Su eje hace referencia al frame de coordenadas de la pelvis. Para comprobar la eficacia del enfoque propuesto se ejecutaron 50 ensayos, logrando 48 detecciones con éxito del conector de la manguera. Uno de esos casos experimentales correspondiente a la tasa del 96 % de consecución se muestra en la Fig. 5.12. En dicha figura se puede apreciar en verde el detalle del modelo del cilindro que más se ajusta a la nube de puntos del conector de la manguera.

Además, para comprobar la robustez de la estrategia propuesta se han hecho múltiples pruebas cambiando la posición del robot respecto a la mesa (Fig. 5.13a) y

se ha llevado a situaciones extremas, tirando parte de la manguera al suelo. Como se observa en la Fig. 5.13b, a pesar de no estar el conector apoyado sobre la mesa, es detectado correctamente.

(a) **(b)**

Figura 5.11: *Detalle de la estimación de posición del conector. (a) Detección del cilindro que más se aproxima al conector de la manguera. El modelo del cilindro calculado se muestra con puntos rojos; (b) El eje y la posición del centroide del conector respecto al frame de la pelvis del robot.*

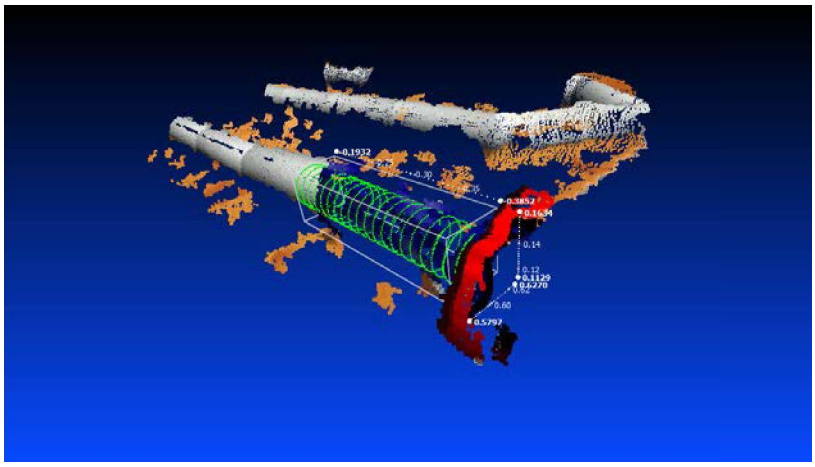
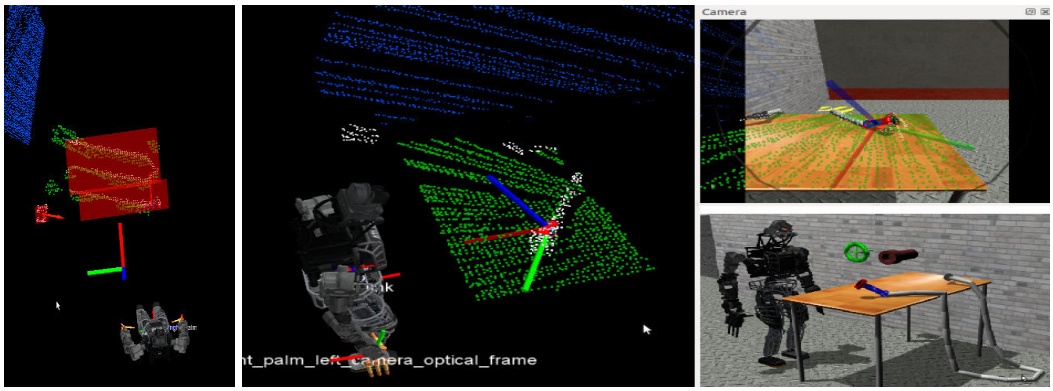


Figura 5.12: *Detalle de la nube de puntos RGB-D de la manguera y en verde, el modelo del cilindro que más se ajusta el conector.*



(a) **(b)**

Figura 5.13: *Detalle de la detección correcta del conector en diversas situaciones. (a) Posición del robot no perpendicular frente a la mesa; (b) Situación extrema en la que parte de la maguera está caída de la mesa y el conector no se encuentra directamente apoyado sobre la mesa.*

5.4.2. Caso de Aplicación I: Agarre Mediante una Mano Robótica

Para demostrar la fiabilidad de la posición estimada del conector de la manguera se realiza de forma representativa un agarre en el escenario simulado de la tarea 3 del VRC. Como se comentó anteriormente, el robot debe situarse en una posición en la que se garantice que puede alcanzar el área de trabajo.

El algoritmo propuesto proporciona la posición del centroide del conector de la manguera y su eje principal en el plano horizontal. Esta orientación se utiliza para planificar el agarre por la parte superior del conector a lo largo de dicho eje (Fig. 5.14). El agarre se realiza en tres pasos [129, 130]:

1. **Pre-agarre**: considerando la forma cilíndrica del objeto, se define una posición de pre-agarre para la mano derecha. Las características de esta posición son: situación a varios centímetros por encima de la posición de agarre, con la orientación según el eje del cilindro y planificación de movimiento libre de colisiones (Fig. 5.14a).

2. **Movimiento reactivo**: se deshabilitan las colisiones y el robot mueve el brazo hacía el centroide del conector. Durante el movimiento se monitorizan los sensores táctiles de la mano para detectar el contacto con el objeto.

3. **Agarre**: cuando se detecta contacto, el robot agarra el conector. La mano está configurada para agarres cilíndricos y los sensores táctiles validan la robustez del agarre por lo que si se detecta que no es correcto, se abre la mano y se vuelve a la primera fase de pre-agarre.

Como se puede apreciar, se realiza con éxito el agarre por la parte superior del eje principal del objeto (Fig. 5.14c). En el siguiente caso de aplicación se validará la estrategia propuesta directamente en la competición VRC, integrando los algoritmos en el sistema completo del equipo SARBOT.

5.4.3. Caso de Aplicación II: Competición en el Virtual Robotics Challenge (VRC)

Debido a que el contexto experimental se enmarca dentro de la competición DRC y concretamente en la fase virtual VRC, se evalúa la efectividad de la estrategia propuesta en la competición. Los entornos donde se compite son variantes tanto en apariencia como en tiempos de ejecución máximos y límite de datos transferidos, como se explicó en el ámbito de aplicación. Por tanto, el nivel de complejidad es muy elevado.

SARBOT formaba parte de los más de 100 equipos no financiados inscritos para acceder al VRC, etiquetados como Track C (Fig. 5.15). Estos equipos del Track C

(a) **(b)** **(c)**

Figura 5.14: *Ejemplo de un agarre por la parte superior en torno al eje principal del conector de la manguera. (a) Posición de pre-agarre; (b) Posición de agarre; (c) Agarre final de la manguera.*

procedían de centros de investigación y universidades ubicadas la mayor parte en Estados Unidos pero también en Japón, Reino Unido, Brasil, México, Polonia y España. Considerando el Track B y C sólo el 20 % de los equipos inscritos lograron acceder a la competición virtual. En la fase del VRC compitieron 26 equipos de todo el mundo, siendo SARBOT el único equipo español de los 3 equipos europeos que se clasificaron.

El equipo SARBOT compitió durante tres días de Junio de 2013 en el VRC a través de una herramienta de simulación de entornos degradados en la nube. El sistema autónomo supervisado integraba, para la toma de decisiones, métodos de percepción complejos junto con algoritmos de control de movimiento. En este sentido, la estrategia de percepción visual para agarre propuesta en este capítulo se integró con el resto de algoritmos desarrollados por los otros grupos de investigación de SARBOT [131]:

- Métodos de percepción del entorno y guiado para navegación mediante la reducción de un ambiente 3D a un mapa 2D [132].
- Algoritmos reactivos de control de locomoción para terrenos complejos y ocluidos [133].
- Técnicas de agarre y manipulación remotas [134], teniendo en cuenta el sistema de percepción desarrollado en este capítulo [130].

La ejecución de las tareas se realizó mediante una Unidad de Control del Operador (OCU) [129]. Esta unidad está centralizada en un ordenador que permite

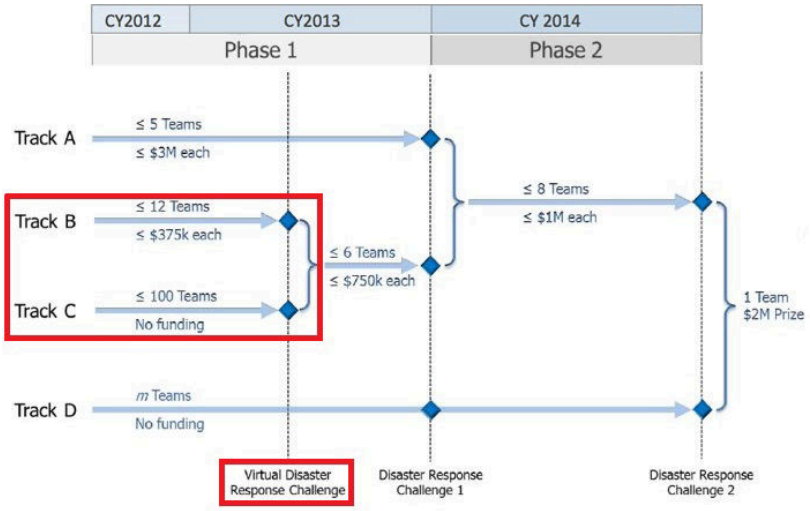


Figura 5.15: *Tabla resumen de la competición, destacando la fase del VRC. Imagen obtenida del reglamento de DRC [126].*

enviar y verificar la correcta ejecución de los comandos teniendo en cuenta el límite de datos de subida y bajada impuestos por el reglamento de la competición. El esquema seguido de cliente-servidor se muestra en la Fig. 5.16, donde se dispone de herramientas ROS tanto en la nube del VRC como en la unidad OCU. La comunicación se realiza mediante sockets ZeroMQ, que es librería de mensajes que permite crear complejos sistemas de comunicación de forma sencilla y rápida.

Para la tarea de agarre y manipulación, el robot se mueve por el escenario hasta llegar a una zona cercana a la mesa. Como se puede ver en la parte inferior derecha de la Fig. 5.16, existe una interfaz de navegación muy sencilla que permite al controlador seleccionar una posición objetivo con el ratón, destacada con una flecha verde en dicha figura. El robot genera automáticamente una trayectoria y se mueve dando realimentación visual cuando es requerido. Esta estrategia es de gran utilidad para percepción a corto alcance, y específicamente para la tarea 3, ya que permite acercarse a la mesa en una determinada posición tras conocer dónde se encuentran los objetos sobre la pared y la mesa. El tamaño visual del mapa se dibuja con un tamaño de cuadrícula de 0,2 metros, indicando en negro las áreas con obstáculos (pared y borde de la mesa), en gris las desconocidas y en blanco las no ocupadas.

Una vez que el robot está cerca de la mesa se dispone de una interfaz de manipulación (parte inferior izquierda de la Fig. 5.16), sólo con propósitos de visualización para validar la información generada automáticamente. Como se observa en la Fig. 5.17, la estrategia propuesta en este capítulo, adaptada a este caso de

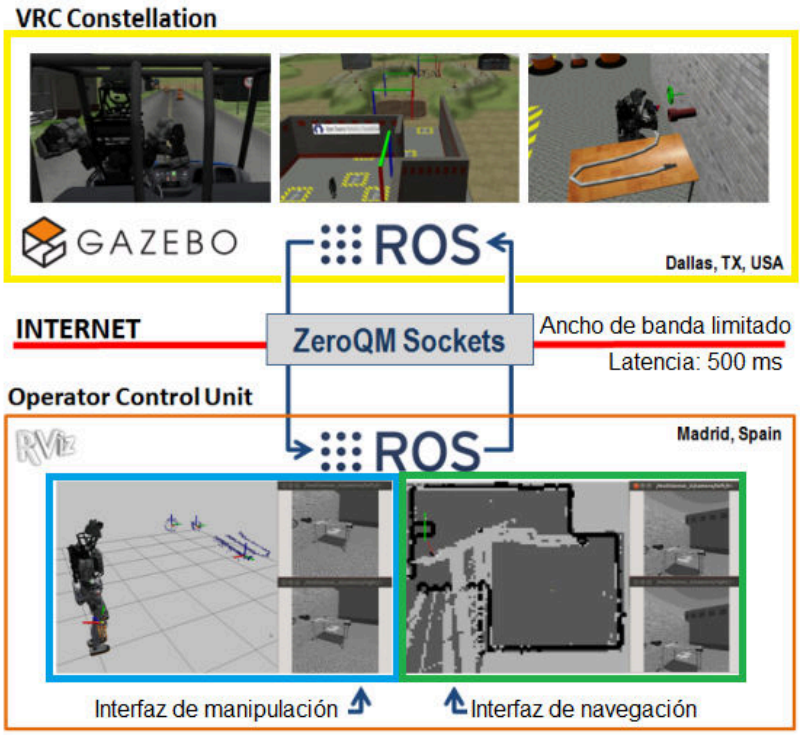


Figura 5.16: *Esquema del cliente-servidor utilizado para la competición del VRC por el equipo SARBOT. Imagen adaptada de la Tesis doctoral de Francisco Suárez [129].*

aplicación, proporciona de forma exitosa y automáticamente la posición de los objetos de interés (conector de la manguera, tubería y llave de paso) así como la nube de puntos correspondiente a cada objeto. Tanto la pared como la mesa se añaden al mapa de colisiones. La interfaz puede mostrar la posición de los objetos, sus correspondientes nubes de puntos, la posición del robot y su mano, así como las imágenes de color de la cámara estéreo con una comprensión del 5 % en jpeg.

Teniendo en cuenta la posición del conector y el rango de alcance del robot, el controlador decide acercar el robot o ejecutar directamente el agarre de la forma indicada en el caso de aplicación anterior. Los mensajes recibidos en el OCU para el control supervisado ocupan en total una media de 20 Kilobytes, siendo el tamaño del dato de posición en torno a 240 bytes y las imágenes en torno a 4795 bytes. Por tanto, se tiene un control exhaustivo de qué información se solicita en cada uno de los intentos de la tarea para evitar superar el límite impuesto de tráfico de subida y bajada.

Durante la competición, cada tarea tiene 5 escenarios diferentes en los que las condiciones ambientales, posición y características de los objetos varían. Además,

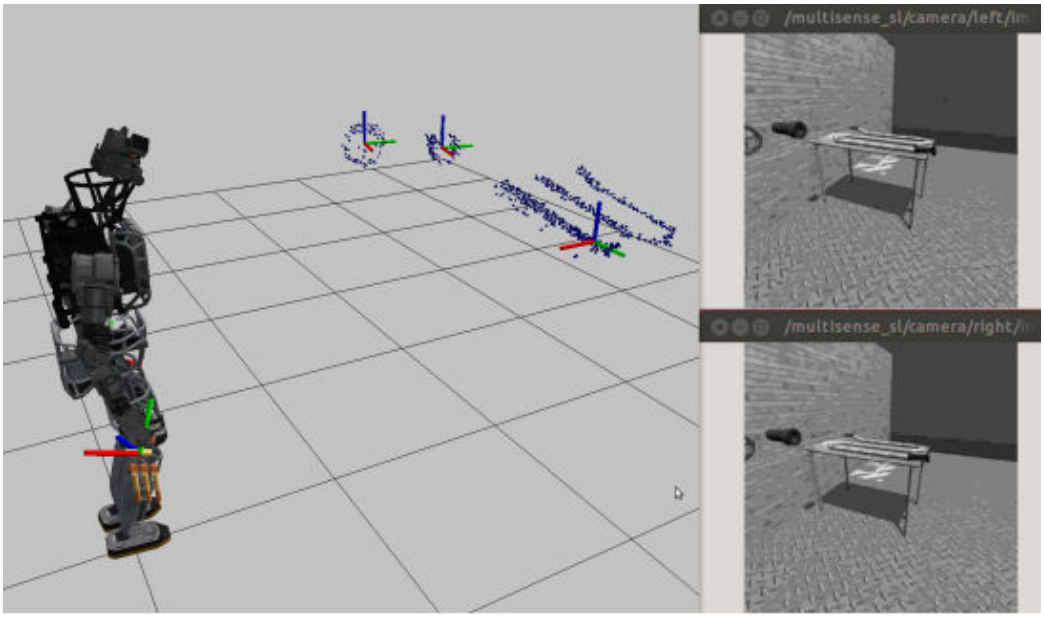


Figura 5.17: *Interfaz de manipulación del cliente corriendo en el OCU, mostrando todos los datos que se pueden recibir durante la tarea 3 del VRC. Imagen obtenida de la Tesis doctoral de Francisco Suárez [129].*

se modifican las restricciones del tráfico de subida y bajada. Para la tarea 3 se consumió entre el 2 % y el 55,8 % de la cantidad disponible de tráfico de subida, por lo que nunca se consumieron todos los bytes disponibles. Hay que tener en cuenta que en el caso peor el límite era de 14 Kilobytes. Para el tráfico de bajada se consumió para la tarea 3 entre un 2 % y 26,6 % del total disponible en los 5 intentos. El caso peor corresponde al límite de 7 MB. Por tanto, se demuestra que la estrategia de percepción propuesta para el agarre permite proporcionar la información requerida de forma automática y sin tener que recibir apenas realimentación visual, por lo que la robustez de la técnica es elevada.

Tras la competición, de los 26 equipos que participaron en el VRC sólo 7 de ellos se clasificaron para la siguiente fase del DRC en entornos reales. Los resultados oficiales de la competición VRC revelaron que las tareas fueron tan complejas que 4 de los 26 equipos clasificados fueron incapaces de anotar un solo punto durante los tres días. El equipo SARBOT alcanzó el puesto 17, quedando por delante de equipos con financiación directa de DARPA. Por tanto, se demostró que, a pesar de no haber quedado entre los 7 finalistas, se alcanzó una solución robótica adecuada para situaciones de rescate y fácilmente adaptable, integrando algoritmos robustos de manipulación, locomoción y percepción, como el propuesto en este capítulo para tareas de agarre.

5.5. Conclusiones

En este capítulo se ha propuesto una estrategia de percepción 3D válida tanto para escenarios poco estructurados en interiores como exteriores. Un contexto habitual de este tipo de entornos es el de robótica humanoide, ya que mientras están moviéndose tienen que hacer frente a escenarios complejos y de condiciones ambientales cambiantes. La estrategia propuesta se ha validado en un escenario de rescate debido a que, actualmente es un gran reto que un robot humanoide realice tareas complejas de manera autónoma durante un desastre. La estrategia de percepción 3D desarrollada permite que un robot móvil sea capaz de hacer frente a tareas de agarre en estos entornos complejos, donde apenas se tiene información del objeto con el que interaccionar ni la escena global.

Para abordar este tipo de escenarios donde apenas se tiene información previa se ha explotado cualquier conocimiento *a priori* del objeto en ambientes humanos, centrándose en las características perceptuales de la tarea específica [60]. A continuación se detallan por puntos las conclusiones y principales contribuciones de la estrategia desarrollada para este tipo de entornos:

- Los algoritmos propuestos trabajan con información de profundidad procedente de sensores que permiten el correcto funcionamiento del sistema robótico tanto en interiores como en exteriores. La solución explota las ventajas de cada uno de los sensores: la precisión en la medida de la distancia del láser y la información de color de la cámara estéreo.

- Para reducir la complejidad del escenario, se consideran dos premisas comunes en el entorno humano: los objetos de interés se encuentran sobre superficies planas y se pueden extraer características específicas del objeto a reconocer.

- En cuanto a la extracción de características del objeto con el que interaccionar, el enfoque se basa en propiedades intrínsecas del objeto ya que se considera que el tamaño puede variar. Para ello se utiliza la integración de descriptores de color y forma.

    - Respecto al color, se ha explotado al máximo la nube RGB-D generada por la cámara estéreo para realizar una segmentación de crecimiento de regiones basada en color sobre la nube de puntos. De esta forma, se permite distinguir zonas con colores diferentes que servirán para el reconocimiento cuando se especifique el objeto al que agarrar. Este enfoque es pertinente tanto para tareas con un objeto de un determinado color como para aquel cuyo color es diferente a los de alrededor.

  - En cuanto a la forma, se trabaja detectando cilindros ya que es una forma básica que puede relacionarse con una amplia lista de objetos comunes en el ámbito humano [120]. Este bloque se ha hecho lo suficientemente global para adaptarse fácilmente a otras formas básicas como cuadrados o esferas.

- En cuanto a la validación, se ha comprobado su idoneidad en un contexto experimental de gran complejidad, como es la competición VRC organizada por DARPA a nivel mundial. Los resultados experimentales demuestran la eficiencia del método propuesto de percepción 3D en un tiempo reducido, requisito indispensable para la competición.

  - La estrategia se ha adaptado con éxito al ámbito de aplicación de la tarea 3 del entorno de simulación del VRC, consistente en la detección del conector de una manguera para conectarlo a una tubería y girar la válvula de agua. La evaluación se ha centrado en la primera subtarea de dicha tarea, que es localizar el conector de la manguera situada sobre una mesa.
  - El algoritmo desarrollado de percepción permite segmentar con éxito los diferentes planos de la escena, detectar la mesa y la pared, así como los objetos relevantes en el espacio de trabajo. Además, se genera el mapa de colisiones de la forma correcta.
  - Analizando la información conocida *a priori* del objeto de interés, se ha establecido como criterio de reconocimiento del conector que, de los objetos encima de la mesa, es aquel que se encuentra entre dos zonas de diferente color y cuya forma es cilíndrica.
  - Se ha detectado el conector con éxito en un porcentaje superior al 95 % de las pruebas con el robot en una posición inicializada cerca de la mesa. Por tanto, los criterios de detección se consideran adecuados.
  - La robustez de la detección ha sido validada tanto en la competición como con pruebas en situaciones extremas donde el conector apenas estaba apoyado sobre la mesa.
  - Mediante casos experimentales, y en la propia competición del VRC, se ha demostrado que la posición estimada de la manguera es adecuada para realizar un agarre fiable.
  - Durante la competición se ha comprobado que el conector se detecta automáticamente y de forma correcta, utilizando la mínima cantidad de datos para su validación previa a la ejecución del agarre.

- Teniendo en cuenta la visualización de la información desde la interfaz de manipulación en el OCU, el filtro aplicado a la información proporcionada es suficiente y adecuado ya que en ninguna de las rondas se ha sobrepasado la tasa máxima de datos de subida y bajada. Por tanto, se consiguen buenos resultados con la mínima información posible, clave durante la competición.
- Como último punto cabe destacar la integración exitosa de la estrategia en el sistema global del equipo SARBOT, funcionando correctamente con los algoritmos de locomoción, agarre y navegación.

Tras haber tratado en este capítulo y en el anterior estrategias para entornos mixtos e interiores, respectivamente, en el próximo capítulo se abordarán los problemas que plantean los exteriores levemente estructurados para las cámaras de rango. Debido a ello, se tendrá que hacer un planteamiento completamente diferente para proponer una estrategia visual que permita detectar objetos en este tipo de escenarios exteriores.

Capítulo 6

Entorno Exterior Levemente Estructurado

6.1. Escenario: Sistema de Percepción con Cámara de Tiempo de Vuelo para Detección de Objetos en Exteriores

Los dispositivos de rango son ampliamente usados en entornos interiores pero en exteriores, muy pocos son válidos. Como se vio en el Capítulo 2, existen algunas cámaras de Tiempo de Vuelo, *Time-of-Flight* (ToF) en inglés, que funcionan en exteriores, sin embargo su efectividad en estos escenarios lleva años bajo estudio [43]. En cambio, en interiores han sido aplicadas con éxito para percepción en diversos ámbitos como aplicaciones en tiempo real, robótica, interacción hombre máquina [30], detección de obstáculos para vehículos inteligentes [40], etc.. Esto es debido a que estos dispositivos solventan algunos problemas de las técnicas clásicas de visión como la complejidad en calcular un mapa de disparidad de forma robusta de sistemas estéreos, o soportar cambios de iluminación.

La limitación en exteriores es debida a varios aspectos derivados fundamentalmente del uso habitual de una señal infrarroja para medir distancias. Algunas de las limitaciones son: la influencia de la luz del sol en las medidas y el máximo rango sin ambigüedad de un sensor, *non-ambiguity range* (NAR) en inglés, limitado por la modulación de la longitud de onda y la potencia de iluminación. Para abordar y superar estas limitaciones, este capítulo se centra en la evaluación del uso de cámaras ToF en exteriores poco estructurados. Además, se aborda el planteamiento de una estrategia de detección de objetos superando todas las limitaciones que

el entorno bajo estudio presenta. La cámara utilizada es la PMD CamCube 3.0 debido a su idoneidad para abordar el escenario y cuyas características se explicaron en la Sección 2.1.2.2 del Capítulo 2.

La detección de objetos se basa en localizar objetos de interés dentro de una escena [135]. Los algoritmos de detección suelen utilizar características extraídas y algoritmos de aprendizaje para reconocer instancias de una categoría de objetos. El proceso se complica cuando los objetos o eventos a detectar son múltiples, tienen formas variables o poco conocidas [136] como ocurre en los entornos semiestructurados. Los enfoques basados en sensores de rango han tenido un gran éxito en los últimos años debido a que se puede adquirir la información 3D en tiempo real, proporcionando directamente imágenes de rango o nubes de puntos [137]. Sin embargo, en exteriores las imágenes suelen ser ruidosas y el entorno es variable en cuanto a luminosidad [138] por lo que los sistemas de percepción visuales deben adaptarse a estas condiciones tan adversas y complejas.

Este Capítulo recoge en la Sección 6.2 el método general y un razonamiento adaptado para la detección de un objeto en exteriores poco estructurados usando una cámara de ToF. Para obtener un criterio de decisión automático se propone un análisis estadístico. En cuanto a la validación del sistema propuesto para el escenario bajo estudio, se ha escogido un contexto experimental aéreo de repostaje en vuelo cuyas características se explican en la Sección 6.3. Por último, en la Sección 6.4 se evalúa previamente la influencia de la potencia transmitida, la frecuencia de modulación y la reflectividad de los objetos para un entorno exterior semiestructurado. Posteriormente, se evalúan y validan los métodos propuestos mediante pruebas experimentales realizadas en un entorno que simula las condiciones bajo estudio. Por último, en la Sección 6.5 se resumen las principales conclusiones del sistema propuesto y de su validación en un entorno exterior poco estructurado.

6.2. Estrategia Propuesta

Las imágenes de rango adquiridas en exteriores por la cámara PMD muestran que los valores de profundidad de las regiones de fondo describen típicas configuraciones con propiedades estadísticas, similar a ruido uniforme. Por otra parte, el objeto a detectar se percibe como una desviación de un modelo de aleatoriedad completa. Teniendo en cuenta esta apreciación y que en el entorno semiestructurado se desconoce la forma del objeto a detectar, se propone para este escenario exterior un sistema basado en la estimación de umbrales de detección de regiones homogéneas con aprendizaje basado en imágenes de ruido [139].

Dado un conjunto de regiones, el algoritmo de segmentación decide para cada región si es parte de un objeto o si pertenece al fondo. Esta técnica permite eliminar

todas las regiones que son significativamente similares a un modelo estadístico de ruido. Para la detección del objeto, con el fin de medir el grado de relevancia de un evento se introduce una medida directamente relacionada con el número de ocurrencias bajo un único criterio: el Número de Falsas Alarmas (NFA).

El criterio de decisión *a-contrario* para la detección de objetos en exteriores permite especificar un modelo estadístico para el fondo en lugar de modelar el objeto. Este hecho ha permitido que el método *a-contrario* haya sido aplicado con éxito para detección en numerosos campos como el reconocimiento de iris [140], asistencia en conducción [141], cambio de subpíxeles en imágenes de satélite [142] y detección de manchas en fondos texturizados [143], entre otros. En el caso del enfoque *a-contrario* utilizando el criterio NFA cabe destacar que ha tenido muy buenos resultados en diversos problemas de detección de objetos en imágenes [144] y más recientemente en secuencias de vídeo [136].

El método de detección propuesto engloba primeramente la fase de segmentación en regiones homogéneas con aprendizaje basado en imágenes de ruido (fondo), teniendo en cuenta múltiples parámetros: tamaño de regiones, distribución global de los píxeles, diferencia de niveles de grises, diferencia de varianzas y contraste de la frontera. Por otra parte, la detección automática de objeto basado en la estimación de un umbral para el NFA.

Por tanto, en esta sección se va a explicar la teoría *a-contrario* aplicada para la detección automática de objetos aéreos en imágenes captadas por una cámara ToF, ámbito de un escenario exterior semiestructurado. En primer lugar, en la Subsección 6.2.1.1 se describe el método general de *a-contrario*. Posteriormente, en la Subsección 6.2.1.2 se propone la adaptación al escenario bajo estudio. En las siguientes subsecciones, se evalúa exhaustivamente cada etapa de la estrategia *a-contrario* propuesta para el entorno exterior.

6.2.1. Método *A-Contrario*

6.2.1.1. Principio General de Detección *A-Contrario*

La detección *a-contrario* se basa en el principio de agrupación perceptual denominado el principio de Helmholtz [144]. Dicho principio explica que los objetos perceptualmente relevantes se definen como grandes desviaciones de una situación aleatoria. Por lo tanto, el entorno *a-contrario* requiere un modelo de fondo aproximado y la incorporación de una o varias medidas relacionadas con el objeto a detectar. Estos parámetros se combinan para obtener el NFA, que corresponde con la expectación de un número de ocurrencias de un determinado evento bajo el modelo de fondo.

La función de NFA es una medida indirecta de importancia: un evento es menos probable que sea fondo a menor NFA, y por lo tanto es más probable que

corresponda al objeto de interés. Para tomar dicha decisión, se requiere sólo un parámetro relacionado con el número medio de falsas alarmas toleradas, un umbral ϵ. Este parámetro tiene una fuerte interpretación estadística, y resulta ser bastante insensible en la práctica ya que la mayoría de funciones de probabilidad tienen un comportamiento exponencial.

Sea $NFA(E)$ la expectación del número de falsas alarmas de un evento E bajo el modelo de fondo, E es ϵ-significativo si

$$NFA(E) < \epsilon \tag{6.1}$$

De esta forma, un evento ϵ-significativo es un evento que se espera que ocurra un número inferior a ϵ veces de media bajo un modelo de fondo.

6.2.1.2. Estrategia Propuesta de Detección Automática *A-Contrario*

Teniendo en cuenta los puntos clave del método general *a-contrario* explicados en la Subsección 6.2.1.1, el sistema propuesto de percepción con cámara ToF para detección de objetos en exteriores engloba dos etapas: segmentación y detección automática utilizando el criterio NFA. El entorno *a-contrario* requiere un modelo de fondo aproximado y la incorporación de una o varias medidas relacionadas con el objeto a detectar, por lo que es necesario realizar previamente un estudio del modelo de fondo y del objeto para obtener los parámetros idóneos y validar la estrategia propuesta. Por tanto, en la estrategia propuesta de percepción se llevan a cabo los siguientes procesos (Fig. 6.1):

- **Segmentación** (remarcado en amarillo en Fig. 6.1): obtención del modelo de fondo para posteriormente comparar su distribución con la del objeto usando diversas medidas (centrales y de dispersión). Si se pueden considerar distribuciones suficientemente diferentes entre las zonas de los píxeles de fondo y del objeto se podría confirmar que la técnica propuesta de segmentación *a-contrario* es válida para esta aplicación por lo que se pasaría a la siguiente etapa.

- **Detección** (remarcado en verde en Fig. 6.1): para lograr una detección automática utilizando el criterio NFA es necesario buscar el umbral ϵ más adecuado. Para su elección, se evalúan dos criterios de detección para buscar aquel que sea más robusto y eficaz en el escenario.

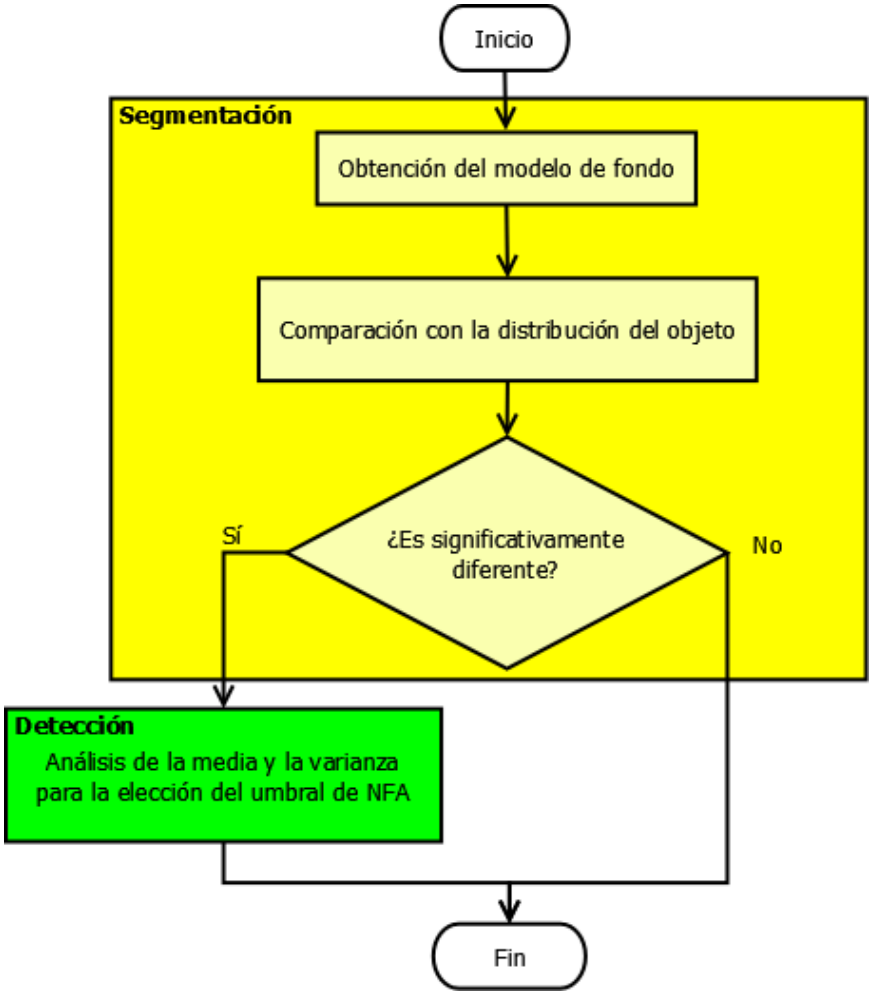


Figura 6.1: *Diagrama general de la estrategia propuesta a-contrario.*

6.2.2. Segmentación

6.2.2.1. Obtención del Modelo de Fondo

Para la segmentación *a-contrario* es necesario elegir un modelo de fondo. Para ello, se compara la distribución de un conjunto de imágenes de fondo con distribuciones de ruido comunes (blanco, gaussiano...). En la Fig. 6.2 se puede ver el proceso de obtención de un modelo de fondo. Como se aprecia, para comparar las distribuciones se utilizan las siguientes funciones:

- La Función de Densidad de Probabilidad (*Probability Density Function*), PDF. Para distribuciones discretas, PDF es la probabilidad de observar un valor particular y tiene dos propiedades teóricas:

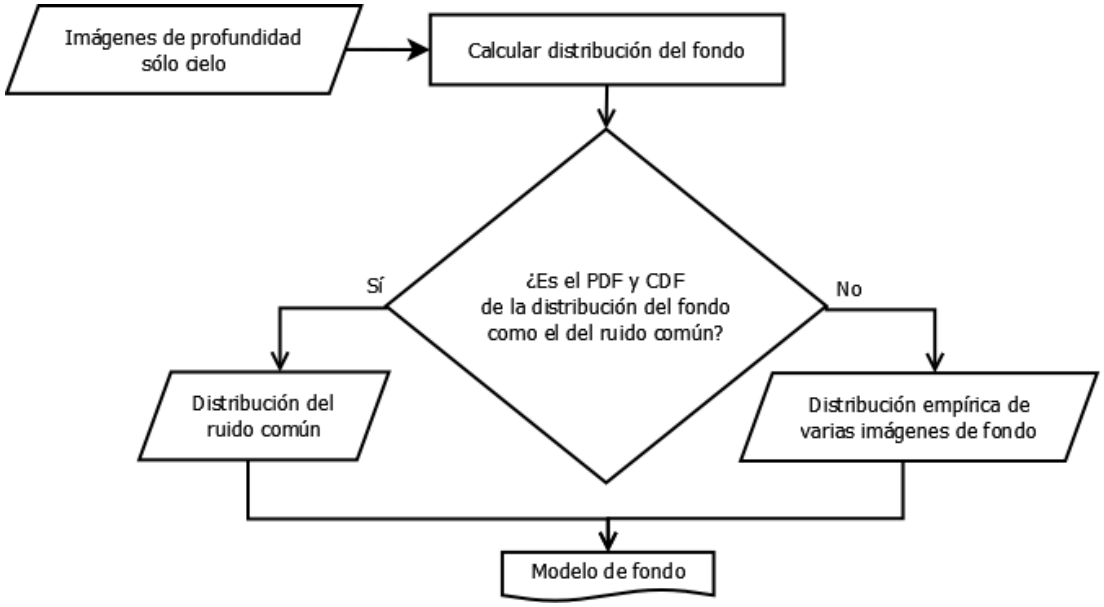


Figura 6.2: *Diagrama general de la obtención del modelo de fondo.*

- La función es positiva o cero para cualquier resultado.
- La integral de una función PDF sobre su rango completo de valores es uno.

- La Función de Distribución Acumulativa (*Cumulative Distribution Function*), CDF. Si f es una función de densidad de probabilidad para una variable aleatoria X, CDF de un valor x, es decir $F(x)$, es la probabilidad de observar algún resultado que sea menor o igual que x:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \tag{6.2}$$

  Una Función de Distribución Acumulativa tiene las siguientes propiedades:

  - El rango de la función es de 0 hasta 1.
  - Si $y > x$, entonces el $CDFy \geq CDFx$.

6.2.2.2. Comparación con la Distribución del Objeto

Las imágenes ToF en exteriores suelen caracterizarse por ser bastante ruidosas (Fig. 6.3). En estos casos, los objetos corresponden a regiones con una profundidad

más regular en la imagen (parte central de la Fig. 6.3). Para medir esta característica de homogeneidad se suele utilizar la media y la mediana de gradiente así como la varianza. Por tanto, para la distribución del objeto se estudian las siguientes medidas características:

- **Medidas centrales**: sirven para medir los valores centrales, siendo las más representativas son la media y la mediana.
  - **Media**: se calcula utilizando todos los valores de la variable, teniendo un resultado único. El problema es que los efectos que sobre ella producen los valores extremos, muchas veces son poco significativos.
  - **Mediana**: se utiliza especialmente cuando los valores extremos son excepcionales y cuando los datos están agrupados en clases y las clases extremas son (al menos una de ellas) abiertas.
- **Medidas de dispersión**: proporcionan valores sobre la dispersión de los datos respecto de los valores centrales.
  - **Varianza**: la media aritmética de los cuadrados de las desviaciones respecto a la media. Es una de las más representativas, siendo su valor mayor cuanto más grande sea la dispersión, y en consecuencia, menos representativos son los valores centrales.

La varianza se aplica como medida característica de la distribución en la imagen de profundidad ya que proporciona directamente información de contraste. Por otra parte, la mediana y media se aplican para la imagen de magnitud de gradiente ya que se encuentran cambios de la mediana o media evaluando descensos por gradiente. Esto es debido a que las medidas centrales de una distribución dan información sobre ella a partir de los valores medios, pudiendo haber dos distribuciones muy distintas que tengan valores medios similares. Por lo que es necesario saber en qué medida los datos numéricos están agrupados o no alrededor de los valores centrales.

Teniendo estos puntos en cuenta, las distribuciones a comparar corresponden a:

- **Fondo**: se espera una distribución similar a ruido blanco. Se evaluará en detalle en la Subsección 6.4.2.1.
- **Objeto a detectar**: Para poder realizar una adecuada detección a partir de la segmentación *a-contrario* se emplea la Probabilidad de Falsas Alarmas (PFA) para la media, mediana y varianza.

Figura 6.3: *Imagen de profundidad adquirida por una cámara ToF en exteriores. Se puede apreciar un objeto en la parte central de la imagen ruidosa.*

Si la comparación entre la distribución del fondo y la del objeto es significativamente diferente, la segmentación *a-contrario* basada en regiones se considera válida por lo que se puede pasar a la siguiente fase de detección.

Antes de pasar a la siguiente etapa, se va a explicar a qué corresponde la imagen de magnitud de gradiente de la de profundidad, utilizada para las medidas de mediana y media.

6.2.2.2.1 Imagen de Magnitud de Gradiente

Matemáticamente, el gradiente de una función de dos variables (en este caso, la función de intensidad de la imagen) es un vector bidimensional para cada punto, cuyos componentes están dados por las primeras derivadas de las direcciones verticales y horizontales. Para cada punto de la imagen, el gradiente del vector apunta en dirección del incremento máximo posible de intensidad, y la magnitud del gradiente del vector corresponde a la cantidad de cambio de intensidad en esa dirección. Por este motivo, la imagen de magnitud de gradiente muestra en cada punto analizado cómo de abrupto es el cambio, mostrando a su vez si un punto representa un borde y la orientación a la que tiende ese borde.

La imagen de magnitud de gradiente ha sido calculada usando operadores locales basados en distintas aproximaciones discretas de la primera y segunda derivada de los niveles de grises de la imagen I. Una vez calculado el vector gradiente

∇J, se calcula el módulo del mismo obtenido en cada píxel de la imagen por lo que los valores grandes corresponden a píxeles del borde o a ruido:

$$\nabla J = (J_x, J_y) = \left[\frac{\partial J}{\partial x}, \frac{\partial J}{\partial y}\right] \Rightarrow \parallel \nabla J \parallel = \sqrt{J_x^2 + J_y^2} \tag{6.3}$$

siendo J_x y J_y:

$$J_x = I \otimes \left(\frac{\partial}{\partial x} \otimes G\right) = I \otimes G_x;\ J_y = I \otimes \left(\frac{\partial}{\partial y} \otimes G\right) = I \otimes G_y \tag{6.4}$$

Para implementar estas operaciones existen máscaras u operadores. En un entorno de vecindad 2x2 se usa el operador de Roberts y en el caso de vecindad 3x3 se aproximan por los operadores de Prewitt, Sobel y el de Fre-Chen. En este caso, la imagen de magnitud de gradiente ha sido calculada usando operadores locales basados en distintas aproximaciones discretas:

- **Vecindad considerada 2 x 2**: El operador Robert proporciona una imagen de magnitud de gradiente calculada usando operadores locales basados en distintas aproximaciones discretas de la primera y segunda derivada de los niveles de grises de la imagen. El gradiente de cada fila y columna puede aproximarse por la diferencia de píxeles adyacentes de la misma.

$$G_x = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\ G_y = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{6.5}$$

- **Vecindad considerada 3 x 3**: El operador Sobel representa una primera aproximación imprecisa del gradiente de la imagen, pero es rápido, efectivo y de calidad suficiente para ser de uso práctico en muchas aplicaciones. Es ampliamente usado para detección de bordes ya que para cada punto muestra cuánto de probable es que represente un borde así como la orientación a la que tiende. Éste operador utiliza sólo valores de intensidad en una región de 3x3 alrededor de cada punto analizado para calcular el gradiente correspondiente, y además utiliza sólo números enteros para los coeficientes que indican la aproximación del gradiente:

$$G_x = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{pmatrix}\ G_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} \tag{6.6}$$

6.2.3. Detección

Teniendo en cuenta la teoría de la Subsección 6.2.1.1, los parámetros relacionados con el fondo y el objeto para cada región se combinan para obtener el NFA. Dicho parámetro corresponde con la expectación de un número de ocurrencias de un determinado evento bajo el modelo de fondo.

Para elegir un método de detección *a-contrario* para el contexto aéreo usando cámaras ToF, se van a estudiar los criterios para la media y varianza. Estos métodos han demostrado que proporcionan buenos resultados de segmentación de regiones coherentes con aprendizaje basado en ruido de fondo [139] y son los más estudiados en entornos similares [145].

Estos métodos de detección se basan en observar a un grupo de píxeles adyacentes en lugar de un único píxel. De esta forma, estos métodos comparan una región de una imagen (una ventana de tamaño w) con el cielo, mediante el análisis de la media o varianza de cada distribución. En el primer caso, el modelo se adapta a las propiedades de la imagen de magnitud de gradiente de la de profundidad y en el segundo caso, a las propiedades de la imagen de profundidad en sí misma.

Por tanto, se sigue el mismo principio de segmentación que la sección anterior pero se busca un determinado umbral que permita la detección automática. De esta forma, se propone que una región de la imagen es detectada como objeto de interés si existe una ventana de tamaño w cuyo Número de Falsas Alarmas cumple:

$$\text{Análisis de la media: } NFA(\mu_w) < \epsilon \tag{6.7}$$

$$\text{Análisis de la varianza: } NFA(\sigma_w^2) < \epsilon \tag{6.8}$$

En la Fig. 6.4 se puede ver el proceso de detección, analizando la media o la varianza de cada región, habiendo calculado previamente un umbral ϵ. Por tanto, en la fase de evaluación que se verá en la Subsección 6.4.3 se profundizará en la obtención del umbral necesario para automatizar el proceso. En este caso se obtendrá por experimentos debido a que por la novedad de esta aplicación no existen unos umbrales previos que sirvan de referencia.

A continuación se desarrolla en más detalle y de forma teórica, esta propuesta de detección y elección del umbral analizando la media y la varianza de cada región de la imagen evaluada.

6.2.3.1. Análisis de la Media de la Imagen de Magnitud de Gradiente

Para el análisis de la media, consideramos que el valor del gradiente medio es menos susceptible a valores atípicos [145]. De esta forma, el análisis está hecho basado en dos criterios de magnitud de gradiente a probar, con vecindad 2x2 o 3x3.

Figura 6.4: *Diagrama general del proceso de detección de objeto y fondo en regiones (ventanas) de la imagen.*

Teniendo en cuenta que la *Probability Density Function* y la *Cumulative Distribution Function* de una distribución normal es:

$$PDF = f(x, \mu, \theta^2) = \frac{1}{\sqrt{2\pi\theta^2}} \exp^{-\frac{(x-\mu)^2}{2\theta^2}} \tag{6.9}$$

$$\begin{aligned} CDF = \int_{-\infty}^{x} PDF \mathrm{d}t &= \frac{1}{\sqrt{2\pi\theta^2}} \int_{-\infty}^{x} \exp^{-\frac{(t-\mu)^2}{2\theta^2}} \mathrm{d}t \\ &= \phi\left(\frac{x-\mu}{\theta}\right) = \frac{1}{2} erfc\left(-\frac{x-\mu}{\theta\sqrt{2}}\right) \end{aligned} \tag{6.10}$$

Siendo $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp^{-\frac{t^2}{2}} \mathrm{d}t$ y $erfc(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} \exp^{-t^2} \mathrm{d}t$.

Entonces, debido a que en este caso lo que se quiere medir es cuánto de bajo es el gradiente, el PFA se calcula teniendo en cuenta la probabilidad de que una observación en una distribución normal con media μ_{sky} y desviación típica θ, en el intervalo $(-\infty, x]$. Debido a que el modelo de fondo corresponde a la distribución empírica de la magnitud de gradiente para varias imágenes del cielo, la media de la magnitud de gradiente es μ_{sky} y la desviación estándar es σ_{sky}. Todas las ventanas de tamaño w con media μ_w, y un número de píxeles independientes e idénticamente distribuidos N_{indp} tienen una desviación típica $\theta_{indp} = \frac{\theta_{sky}}{\sqrt{N_{indp}}}$. La probabilidad de observar una ventana por casualidad es:

$$PFA = P_{ac}(\mu \leq \mu_w) \equiv N_{\leq}\left(\mu_w, \mu_{sky}, \theta^2_{indp}\right) PFA = \int_{-\infty}^{\mu_w} PDF \mathrm{d}x \tag{6.11}$$

Por tanto, el Número de Falsas Alarmas para la media del gradiente de una ventana, $NFA(\mu_w)$, es la esperanza del número de ventanas que tienen una media más baja que el umbral en una imagen de ruido puro. Se calcula multiplicando esta probabilidad por el número de ventanas evaluadas en la imagen M_w. De esta forma, $NFA(\mu_w)$ se define de la siguiente forma:

$$NFA(\mu_w) = M_w \times P_{ac}(\mu \leq \mu_w) \tag{6.12}$$

Teniendo en cuenta que el teorema del valor central de la teoría de la probabilidad, considerando un gran número de píxeles aleatorios independientes de cada ventana evaluada de la imagen de profundidad, la media de sus magnitudes de gradiente debería ser aproximadamente de distribución normal. Entonces, $NFA(\mu_w)$ se define por:

$$NFA(\mu_w) = M_w \times N_{\leq}\left(\mu_w, \mu_{sky}, \frac{\sigma_{sky}}{\sqrt{N_{indp}}}\right) \tag{6.13}$$

Donde:

- M_w es el número de ventanas.
- $N_{\leq}(x, \mu, \sigma)$ es CDF de una distribución normal con media μ y la desviación σ se aplica a x.
- N_{indp} es el número de píxeles independientes de cada ventana. De esta forma, si se eligen ventanas de 20x40, por ejemplo, el número de píxeles por ventana N es 800. De esos 800, los independientes se calculan teniendo en cuenta la vecindad de la imagen de magnitud de gradiente $N_{indp} = \frac{N}{vecindad}$:

$$\text{Imagen de gradiente 2x2: } N_{indp} = \frac{N}{4} \tag{6.14}$$

$$\text{Imagen de gradiente 3x3: } N_{indp} = \frac{N}{9} \tag{6.15}$$

Por tanto, $NFA(\mu_w)$ es el número esperado de ventanas con media μ_w inferior al umbral ϵ considerando como modelo de fondo del cielo, por lo que la detección se producirá cuando:

$$NFA(\mu_w) < \epsilon \tag{6.16}$$

Si se escoge un umbral de $0,01$, quiere decir, que en media, ocurre una falsa alarma de cada 100 imágenes.

Si los valores de PFA son muy pequeños, se utiliza el logaritmo neperiano. Esto es debido a que su interpretación gráfica es muy clara, por lo que la elección del umbral es mucho más sencilla:

$$ln(NFA(\mu_w)) < ln(\epsilon) \rightarrow ln(M_w) + ln(PFA(\mu_w)) < ln(\epsilon) \tag{6.17}$$

De esta forma, las ventanas que contienen un objeto deben corresponder a valores negativos del logaritmo de NFA.

6.2.3.2. Análisis de la Varianza de la Imagen de Profundidad

Para el análisis de varianza [139], consideramos la varianza de cada ventana en la imagen de profundidad. El modelo de fondo considerado corresponde a la distribución empírica de varias imágenes de profundidad del cielo, por lo que se considera el cuarto momento μ_4 y la varianza σ^2_{sky}.

El cuarto momento se relaciona con la kurtosis, que representa la elevación o achatamiento de una distribución, comparada con la distribución normal.

Para todas las ventanas de tamaño w con varianza σ^2_w , y un número de píxeles independientes e idénticamente distribuidos N_{indp},la probabilidad de observar una ventana por casualidad es:

$$PFA(\sigma^2_w) = P_{ac}(\theta^2 \leq \sigma^2_w) \equiv N_{\leq}\left(\sigma^2_w, \sigma^2_{sky}, \frac{\sqrt{\mu_4 - (\sigma^2_{sky})^2}}{\sqrt{N_{indp}}}\right) \tag{6.18}$$

De esta forma, el Número de Falsas Alarmas $NFA(\sigma^2_w)$ para el valor de la varianza de una ventana se obtiene multiplicando esta probabilidad por el número de ventanas evaluadas en la imagen M_w. Así $NFA(\sigma^2_w)$ es:

$$NFA(\sigma^2_w) = M_w \times P_{ac}(\theta^2 \leq \sigma^2_w) \tag{6.19}$$

Teniendo en cuenta el teorema central del límite de la teoría de la probabilidad, la varianza de los píxeles de cada ventana evaluada debería ser aproximadamente normalmente distribuida. Entonces, $NFA(\sigma^2_w)$ se define por:

$$\begin{aligned} NFA(\sigma^2_w) &= M_w \times N_{\leq}\left(\sigma^2_w, \sigma^2_{sky}, \frac{\sqrt{\mu_4 - \left(\sigma^2_{sky}\right)^2}}{\sqrt{N_{indp}}}\right) \\ &= M_w \times N_{\leq}\left(\sigma^2_w, \sigma^2_{sky}, \sigma^2_{sky}\frac{\sqrt{K_{sky} - 1}}{\sqrt{N_{indp}}}\right) \end{aligned} \tag{6.20}$$

Donde:

- M_w es el número de ventanas.
- $N_{\leq}(x, \mu, \sigma)$ es la CDF de una distribución normal con media μ y desviación σ aplicada a x.
- μ_4 es el cuarto momento de la distribución por lo que $\mu_4 = K_{sky}(\sigma^2)^2$.
- N_{indp} es el número de píxeles independientes de cada ventana, en este caso N.

Por tanto, $NFA(\sigma_w^2)$ es el número de ventanas esperado que tiene una varianza σ_w^2 inferior a ϵ en el modelo del fondo del cielo, por lo que la detección se producirá cuando:

$$NFA(\sigma_w^2) < \epsilon \tag{6.21}$$

Si los valores de PFA son muy pequeños, se utiliza el logaritmo neperiano como se explicó en el apartado anterior.

6.3. Contexto Experimental

El sistema de percepción que se evalúa en este capítulo se enmarca en el proyecto titulado PROSAVE2 (PROyecto de investigación en Sistemas Avanzados para aViones más Eco-Eficientes), del Programa Nacional Español INGENIO 2010, referencia CENIT-2010-1039. PROSAVE2 tiene como objetivo investigar y desarrollar nuevas tecnologías en el ámbito aeronáutico a nivel nacional a través de sistemas ambientales avanzados. El proyecto está específicamente dirigido al desarrollo de tecnologías de futuro para su aplicación en sistemas de actuación avanzada, trenes de aterrizaje, sistemas de reabastecimiento en vuelo, sistemas de generación, purificación de gases y reaprovechamiento energético.

El escenario bajo estudio permite validar el sistema de percepción propuesto para detección de objetos en exteriores, ya que en un entorno real todo estará lo suficientemente lejos por lo que el fondo tendrá una distribución similar a ruido excepto el objeto a detectar. El método propuesto busca explotar los mapas densos de profundidad proporcionados por las cámaras ToF en tiempo real para el reabastecimiento en vuelo, *Aerial Refueling* (AR) en inglés. De esta forma, el sistema permitirá obtener información relevante sobre la posición del receptor en las operaciones de repostaje con pértiga.

6.3.1. Ámbito de Aplicación

El repostaje en vuelo, en general, se puede clasificar en dos categorías: sonda y cesta, *probe-and-drogue* en inglés, y pértiga, *boom-and-receptacle* en inglés. El contexto de este capítulo es el segundo método de repostaje (Fig. 6.5), donde un avión cisterna transfiere combustible a otro (receptor) extendiendo un dispositivo denominado pértiga hacía el receptáculo. La pértiga es un tubo telescópico rígido localizado en la parte inferior del avión cisterna, mientras que el receptáculo es una apertura situada en la parte superior del avión receptor.

Hay tres fases durante la operación de repostaje usando pértiga:

1. Acercamiento (Fig. 6.5a). El avión cisterna vuela a velocidad y altitud constante, en dirección recta, mientras que el receptor se coloca debajo y por detrás del avión cisterna, guardando la distancia de seguridad. El operador del cisterna, extiende la pértiga dirigiéndolo hacía el receptáculo del receptor.

2. repostaje (Fig. 6.5b). Una vez que la conexión se ha realizado, la válvula de la pértiga y la del receptáculo se abren hidraúlicamente, y el combustible se empieza a transferir mediante su propulsión por bombas.

3. Separación. Cuando el combustible necesario ha sido transferido completamente, las válvulas se cierran y la pértiga se desconecta y se retrae, deshaciendo el receptor la formación.

Para un posicionamiento preciso de los aviones en un sistema de repostaje en vuelo, varios métodos han sido considerados [146] como sensores basados en radar, sistemas de visión tanto pasivos como activos y Sistema de Posicionamiento Global, *Global Positioning System* (GPS). Sin embargo, tienen ciertas limitaciones debido a su poca fiabilidad en ciertos ambientes (como por ejemplo, baja cobertura o condiciones de visibilidad adversas). Además, las tecnologías basadas en visión pierden información debido al mapeo de los puntos 3D a planos de imagen 2D. Por ello, debido a las limitaciones que presentan otros sistemas, este escenario es adecuado para evaluar y validar la estrategia propuesta para detección de objetos en exteriores.

6.3.2. Configuración

Teniendo en cuenta las características técnicas de las cámaras de profundidad ToF del Capítulo 2, las más adecuadas para las aplicaciones exteriores son PMD CamCube [35] [44] y 3D Flash LIDAR [147]. La tecnología 3D Flash Lidar era de reciente aparición durante el contexto de PROSAVE2 y demasiado específica, por lo que la cámara ToF utilizada para validar el sistema de percepción propuesto es

(a) **(b)**

Figura 6.5: *Ejemplo de repostaje mediante pértiga. (a) Fase de acercamiento. El avión receptor vuela por detrás del avión cisterna (foto de U.S. Air Force tomada por Tech. Sgt. Mark R. W. Orders-Woempner). (b) Fase de repostaje. Un avión cisterna durante el proceso de respotaje a otro avión (foto tomada por Tech. Sgt. Jacob N. Bailey, U.S. Air Force/Released).*

la cámara PMD CamCube 3.0. Como se comentó en la Sección 2.1.2.2.3, el rango máximo sin ambigüedad del sensor está limitado a la mitad de la longitud de onda de modulación, λ_{mod}, debido a que la señal es periódica. Además, hay que tener en cuenta que la energía de la luz disminuye con el cuadrado de la distancia al objeto. La cámara PMD CamCube por defecto, está preparada para abarcar como máximo 7 metros usando dos módulos de iluminación en interiores (Fig. 6.6a), pero tiene una gran ventaja que es su fácil adaptación a las especificaciones requeridas según su uso.

En el escenario bajo contexto se requiere abarcar distancias de hasta 30 metros pero se comprobó que la cámara PMD CamCube en exteriores no permitía apreciar objetos a distancias superiores a 6 metros. Por tanto, para el ámbito de aplicación es necesario bajar la frecuencia de modulación y además, incrementar el número de fuentes de iluminación a la configuración básica. La compañía PMDTechnologies GmbH proporcionó el software necesario para poder modificar la frecuencia de modulación de la cámara desde 40 hasta 0,1 MHz, pero en fase experimental. Para la elección del número de módulos adicionales se tuvo en cuenta que los desarrolladores de PMDTec alcanzaron distancias superiores a 35 metros con 8W (8 módulos de iluminación) en una aplicación con vehículos en exteriores [40].

Por tanto, para poder abarcar el rango de distancias requerido en exteriores, se ha utilizado una cámara PMD CamCube 3.0 con los siguientes componentes adicionales, que se muestran en la Fig. 6.6b:

(a) **(b)**

Figura 6.6: *La cámara PMD CamCube 3.0: (a) Configuración básica con dos módulos de iluminación. (b) Configuración usada en este escenario, con el hardware adicional añadido teniendo en cuenta el contexto bajo evaluación.*

- **Módulos de iluminación**: 7 módulos de iluminación en total (7W de potencia de iluminación). Para la integración de estos módulos se ha añadido un *splitter* que proporciona una señal diferencial de bajo voltaje apropiada para cada unidad de iluminación.
- **Filtro**: se ha añadido delante del sensor un filtro que sólo permite pasar las longitudes de onda comprendidas entre 800 y 900 nm. De esta forma se evita que las medidas se vean afectadas por radiaciones IR que no corresponden a la longitud de onda de la señal IR emitida de 870 nm.

6.4. Resultados Experimentales

En esta sección, se evalúa el sistema de percepción propuesto a través de resultados experimentales. Como se mencionó en el apartado anterior, debido a las características del contexto se tuvo que adaptar la cámara PMD CamCube añadiéndole hardware adicional.

Debido a la novedad del uso de este tipo de cámaras para exteriores, y su configuración experimental, para poder validar el sistema propuesto se ha evaluado en dos fases la segmentación. Primeramente se ha comprobado cómo afectan los parámetros técnicos en la práctica a las medidas de rango y por tanto, a la segmentación. Posteriormente, en la Subsección 6.4.2 se evalúa la distribución de los

píxeles de fondo y del objeto para validar la técnica de segmentación *a-contrario* propuesta. Por último, se evalúa el algoritmo de detección automático para verificar la robustez del sistema propuesto de percepción en exteriores.

6.4.1. Fase I de Evaluación de Segmentación: Influencia de Parámetros Técnicos sobre las Imágenes de Rango

Para abordar la estrategia propuesta para detección de objetos en exteriores, en la Subsección 6.4.1.1 se analizan primeramente en detalle los datos que proporciona la cámara en exteriores mediante diversos casos prácticos. Posteriormente, teniendo en cuenta la influencia de los parámetros técnicos en las imágenes de rango tomadas en exteriores, en la Subsección 6.4.1.2 se evalúan los parámetros más limitantes para un entorno similar al escenario bajo contexto.

En ambos casos se evaluará el algoritmo de segmentación común para este tipo de cámaras, basado en filtrado por amplitud, cuyos resultados se estudiarán en detalle teniendo en cuenta los diversos parámetros que influyen. Por último, se realizará una comparativa preliminar entre una técnica de segmentación clásica y *a-contrario*.

6.4.1.1. Casos Prácticos en Exteriores

El entorno donde se han desarrollado las pruebas de esta subsección ha sido la azotea del edificio Betancourt del campus de Leganés de la Universidad Carlos III de Madrid (UC3M). La zona ofrece un entorno exterior sin obstáculos, y expuesto a la luz directa del sol, como se puede apreciar en la Fig. 6.7, por lo que es un lugar adecuado para realizar las pruebas requeridas.

En esta fase se han realizado los ensayos que se especifican a continuación, cuya explicación detallada se muestra en los próximos apartados:

- Influencia del material utilizado en el alcance máximo del sistema.
- Evaluación del alcance máximo en función de la potencia de iluminación, desde 1W hasta 7W.
- Influencia del ángulo de inclinación del objeto a evaluar.
- Análisis de los datos de profundidad para proponer el método más adecuado de detección automática.

Para estas pruebas se ha usado la cámara PMD CamCube 3.0 con la configuración comentada en la Sección 6.3.2 (7 módulos de iluminación y un filtro paso de

Figura 6.7: *Imagen de la zona de ensayos en la azotea de la UC3M.*

banda delante del sensor), cuya configuración se muestra en la Fig. 6.8. Las imágenes de 200x200 píxeles han sido adquiridas a 40 Frames Por Segundo (FPS) y a un tiempo de integración de 2500 μs o 500 μs, según el caso. El resto de características empleadas han sido las de por defecto de la cámara, mostradas en la Tabla 2.3.

6.4.1.1.1 Influencia de la Reflectividad en el Alcance Máximo de Detección

En la Sección 2.1.2.2.7 se explicó que la reflectividad de los objetos influye en la cantidad de luz recibida en el sensor, y por tanto, en la distancia máxima a la que pueden ser vistos por una cámara ToF. Hay estudios previos [49] sobre la influencia de la reflectividad de los objetos en la medida de distancias con las cámaras ToF, donde se ve que ciertos materiales provocan problemas de saturación o de luminosidad, así como exceso de ruido, siendo la información de medida más débil.

Para comprobar la influencia del material en el alcance, en este ensayo se han utilizado tres placas rectangulares de materiales con reflectividad diferente (de menor a mayor: cartón, poliestireno y metal). Se han utilizado todos los módulos de iluminación (7W) y se ha alejado cada objeto hasta la distancia máxima donde eran detectados por la cámara ToF, conservando las condiciones de trabajo entre experimentos. En la Fig. 6.9 se muestran los resultados, y como cabía esperar, a mayor reflectividad mayor distancia máxima de detección:

$$\rho_{carton} < \rho_{poliestireno} < \rho_{metal}$$

Figura 6.8: *La cámara PMD CamCube 3.0 usada en este escenario, con el hardware adicional añadido teniendo en cuenta el contexto bajo evaluación.*

$$d_{dmax\,carton} < d_{dmax\,poliestireno} < d_{dmax\,metal}$$

$$d_{dmax\,carton} = 19m;\ d_{dmax\,poliestireno} = 22m;\ d_{dmax\,metal} = 24,5m$$

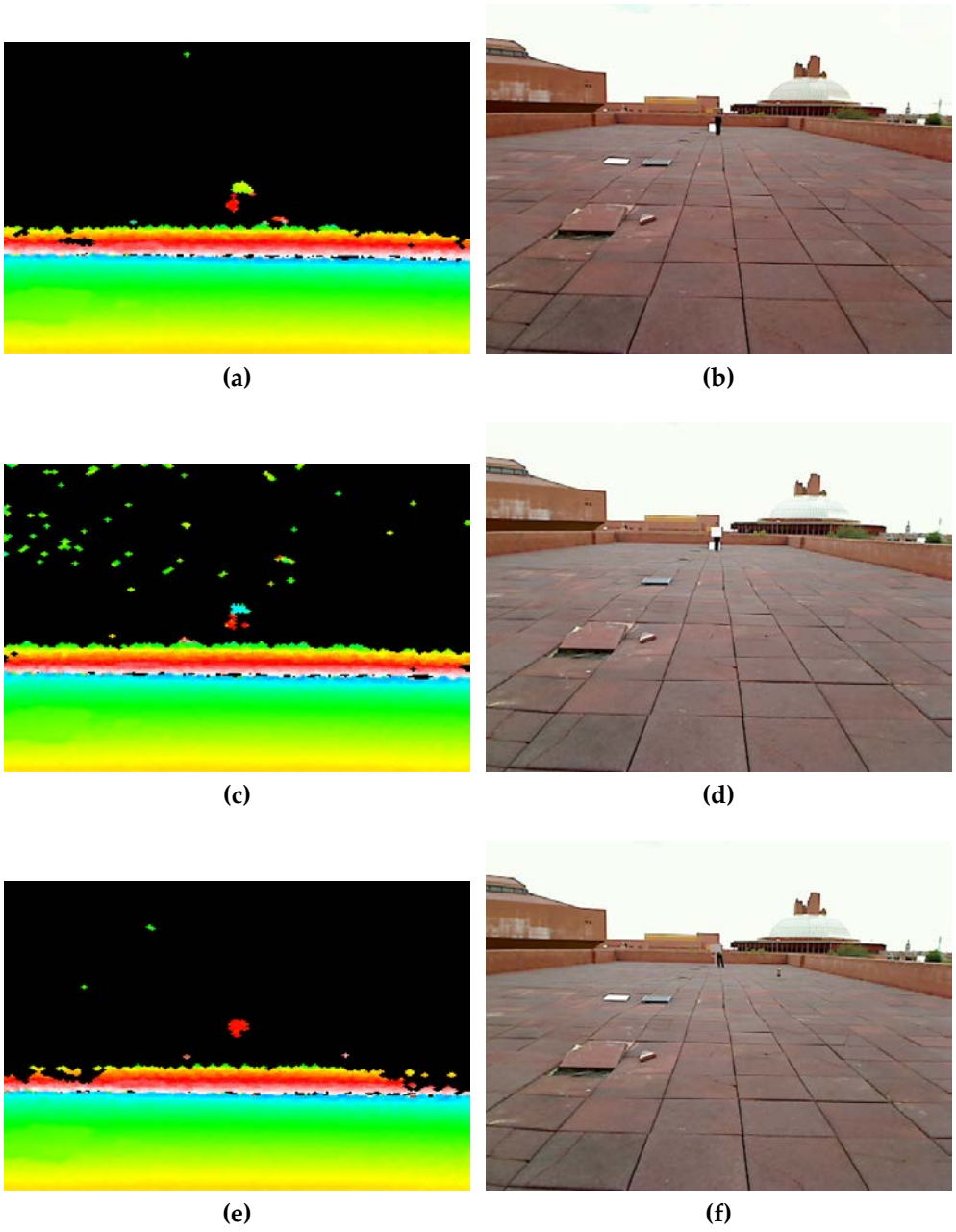
Las imágenes de profundidad que se muestran a la izquierda de la Fig. 6.9, son el resultado de aplicar un filtrado semiautomático tras la captura. La secuencia de filtros utilizada es un umbral sobre la amplitud (elegido manualmente), un filtro de mediana, y un filtro morfológico de apertura 5x5 para eliminar las estructuras demasiado pequeñas y aquellos valores espúreos.

6.4.1.1.2 Alcance Máximo Frente a Potencia de Iluminación

Para poder abarcar las distancias requeridas en el contexto bajo estudio se han añadido módulos de iluminación a la configuración básica de la cámara PMD CamCube, como se comentó previamente. En esta sección se evalúa la distancia máxima a la que se puede detectar un objeto según la potencia de la fuente emisora de luz.

Para comprobar el máximo alcanzable en función de la potencia de la iluminación se ha considerado como caso de referencia un poste de aluminio. Este objeto está constituido por un material dotado de gran reflectividad a 870 nm, como se observa en la Fig. 2.13b. Para otros materiales con peor reflectividad se ha comprobado en la sección anterior que la distancia máxima de alcance se reduce.

Se ha partido de 1W de potencia iluminadora (un módulo de iluminación) y se ha ido aumentando la potencia de uno en uno hasta 7W, es decir, activando módulos de iluminación hasta que todos estuviesen funcionando. Para cada potencia, el



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figura 6.9: *Distancia máxima de detección de tres objetos con distinta reflectivida, a la izquierda se muestran las imágenes de profundidad y a la derecha las de color: (a)-(b) caja de cartón; (c)-(d) placa de poliestireno; (d)-(e) placa metálica.*

Figura 6.10: *Gráfica de potencia de iluminación frente a la distancia máxima a la que se detecta el poste de aluminio. Se ha considerado el uso desde 0 hasta 7 módulos de iluminación.*

objeto se ha ido alejando de la cámara y se han grabado los datos en la posición más alejada en la cual el objeto es detectado. El poste de aluminio se ha podido apreciar con todos los módulos activos hasta una distancia de 24,5 metros.

En la Fig. 6.10 se muestra la gráfica que representa la relación entre potencia de iluminación y distancia máxima de detección del objeto a evaluar, en este caso un poste de aluminio.

Para hallar la función que expresa la relación entre potencia de iluminación y distancia se ha recurrido a la función *polyfit* de Matlab [148]. Teniendo en cuenta que existen 8 datos (considerando también el valor para 0W), el polinomio de mínimos cuadrados de grado 7 que mejor se ajusta a los datos es:

$$p(x) = 0,0008x^7 - 0,0238x^6 + 0,3095^5 - 2,1276x^4 + 8,2331x^3 - 18,1586x^2 + 24,2667x \tag{6.22}$$

Cada uno de los casos comentados en la Fig. 6.10 se pueden apreciar en la Fig. 6.11-6.13 donde se muestran las imágenes de amplitud y profundidad tras aplicar un filtro en función de un umbral de amplitud (elegido manualmente).

6.4.1.1.3 Influencia del Ángulo de Inclinación

En la Sección 2.1.2.2.7 se explicó que la reflectividad de los objetos influye en la cantidad de luz recibida en la dirección del sensor, y por tanto, en el alcance máximo al que pueden ser vistos por una cámara ToF. El rayo que incide sobre el objeto se reflejará con un ángulo igual al incidente de tal forma que si el objeto está paralelo al plano del sensor, la onda se refleja en la dirección del sensor por lo que llega todo lo reflejado. Pero si el objeto no está paralelo al plano del sensor, la onda reflejada puede no llegar al sensor (en función del ángulo y de la distancia del objeto). Este efecto es determinante en objetos con reflexión especular (por ejemplo, espejos) ya que en objetos con reflexión difusa, la onda se refleja en varias direcciones por lo que parte de la luz puede llegar al sensor.

Se puede observar este efecto en las imágenes de la Fig. 6.14, donde se muestra una placa metálica a unos 14,5 metros colocada paralela al sensor, con un ángulo de inclinación de 10º y de 20º, aproximadamente. A pesar de tener una reflectividad alta, se observa como a medida que se va girando se pierden los datos de profundidad, debido a la reflexión especular con un ángulo de incidencia distinto de 0º. Si se mantiene la placa metálica perfectamente paralela al plano del sensor, se han conseguido abarcar distancias de hasta 31 metros, como se puede observar en la Fig. 6.15 con diversos tiempos de integración.

Las imágenes de profundidad de las Fig. 6.14 y Fig. 6.15 son el resultado de aplicar un filtrado semiautomático tras la captura. La secuencia de filtros utilizada es un umbral sobre la amplitud (elegido manualmente), un filtro de mediana, y un filtro morfológico de apertura 5x5 para eliminar las estructuras demasiado pequeñas y aquellos valores espúreos.

6.4.1.1.4 Segmentación Basada en Filtrado de Amplitud

Cada píxel del sensor detecta tanto intensidad como distancia por lo que la cámara proporciona, como se comentó en la Subsección 2.1.2.2.2, una imagen de intensidad, amplitud y rango en tiempo real. La precisión de la distancia para cada píxel depende de diversos factores externos (luz del sol y/o condiciones climáticas extremas) y propiedades de la escena (reflexiones múltiples, ángulo de inclinación de la luz sobre el objeto a evaluar, su acabado y composición). Estas fluctuaciones causan ruido, como se ha podido observar en las secciones anteriores donde en exteriores la cámara produce imágenes de profundidad brutas que son difíciles de

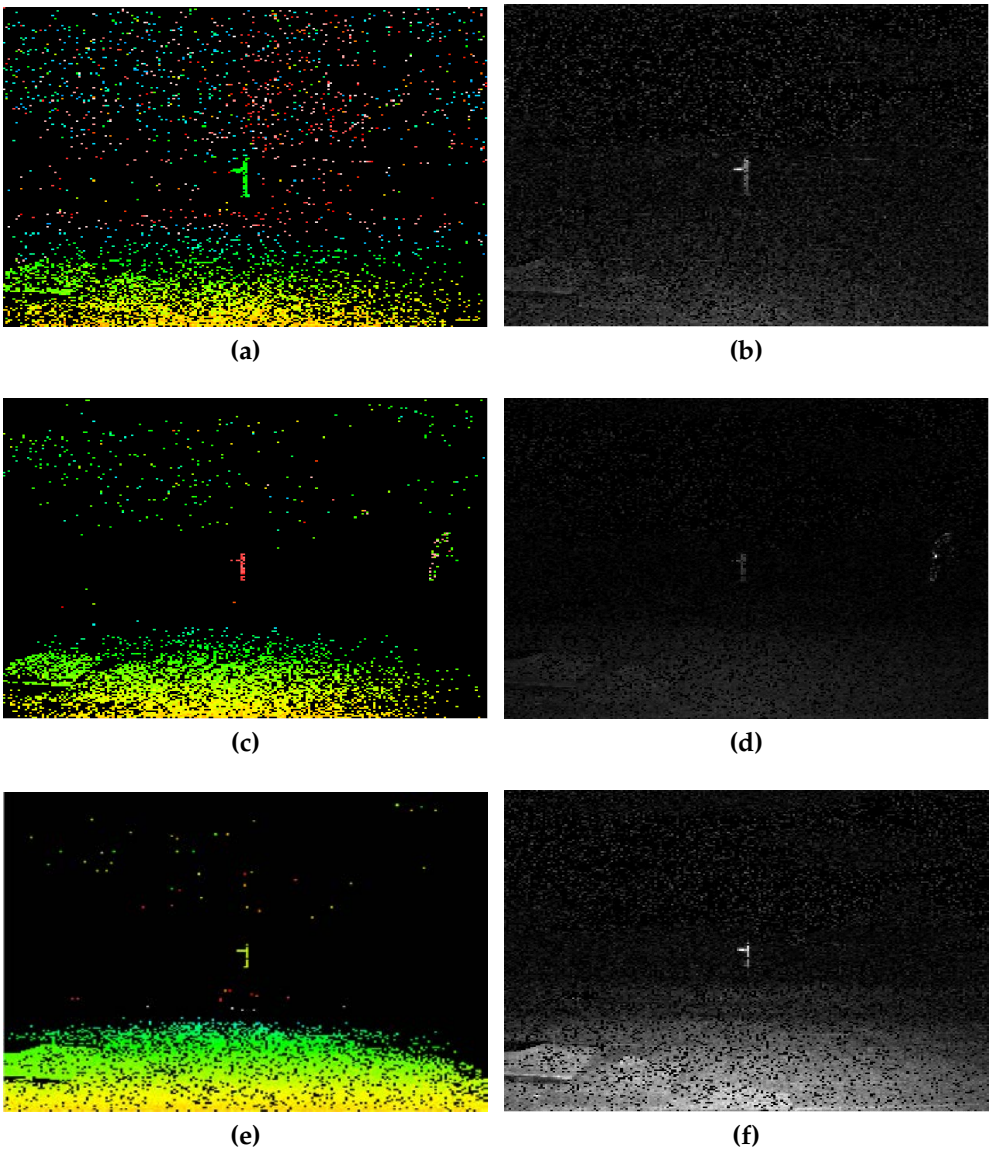


Figura 6.11: *Imágenes de evaluación de un poste de aluminio, a la izquierda se muestran las imágenes de profundidad en función de un umbral de amplitud y a la derecha las de amplitud: (a)-(b) 1 W y 429 de amplitud; (c)-(d) 2W y 649 de amplitud; (d)-(e) 3W y 809 de amplitud.*

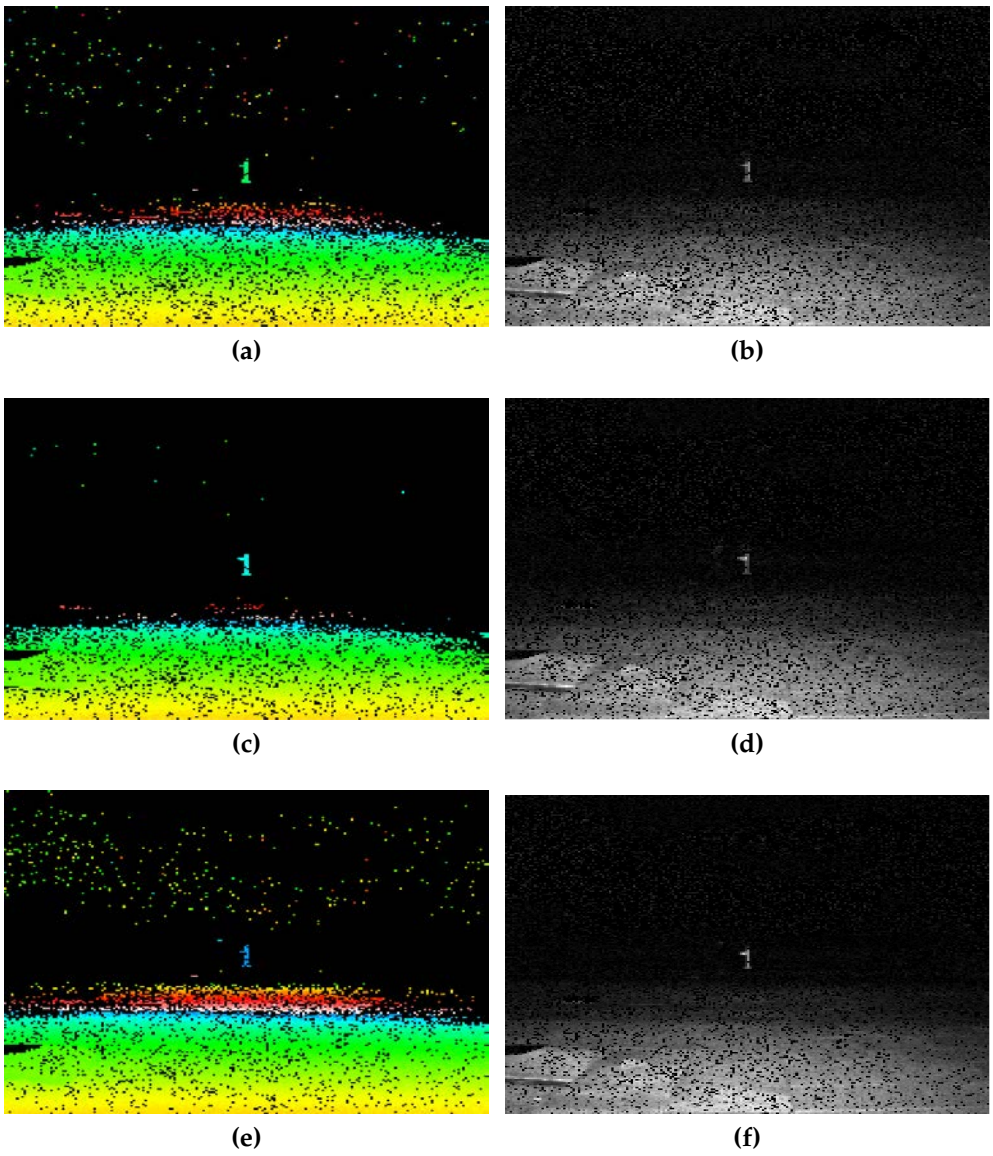
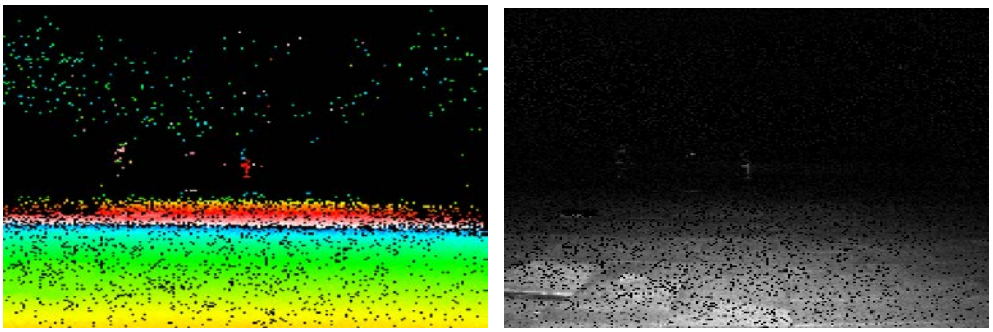


Figura 6.12: *Imágenes de evaluación de un poste de aluminio, a la izquierda se muestran las imágenes de profundidad en función de un umbral de amplitud y a la derecha las de amplitud: (a)-(b) 4 W y 629 de amplitud; (c)-(d) 5W y 920 de amplitud; (e)-(f) 6W y 650 de amplitud*



(a) **(b)**

Figura 6.13: *Imágenes de evaluación de un poste de aluminio, a la izquierda se muestran las imágenes de profundidad en función de un umbral de amplitud y a la derecha las de amplitud: (a)-(b) 7W y 567 de amplitud.*

usar directamente. Por este motivo, primeramente se ha evaluado la segmentación de un objeto eliminando los píxeles con amplitud baja, ya que como se dijo en la Subsección 2.1.2.2.2 del Capítulo 2, la imagen de amplitud indica la fiabilidad de los valores de distancia.

Por lo tanto, para minimizar el ruido en una imagen de profundidad adquirida en un entorno al aire libre, el filtrado estándar se basa en ignorar aquellos píxeles cuyo valor está por debajo de un determinado umbral de amplitud, y la adición de un filtro de mediana [149]. De esta forma, los objetos con buena reflectividad o cercanos producirán valores de amplitud altos, mientras que los objetos lejanos (es decir, el cielo) o con baja reflexión se verán afectados por el ruido dando lugar a regiones de profundidad no uniformes y con baja amplitud.

En las imágenes de exterior grabadas durante los ensayos (días diferentes y con condiciones de iluminación distintas), se observó que un umbral único permite eliminar la mayoría de los píxeles de fondo en todos los casos, como se puede comprobar en las imágenes de la Fig. 6.16. En ambos casos, un umbral de 500 permite eliminar la mayoría de los píxeles de fondo. Si este resultado se combina con un filtro mediano de tamaño 3x3, el cielo y los edificios del fondo son eliminados en ambos casos.

Sin embargo, se detectan dos limitaciones en este método:

1. No es demostrable, al menos con los ensayos realizados en la azotea, que un solo umbral sea suficiente para todos los casos que se pueden presentar en un entorno real.

2. Las zonas del objeto a evaluar que tienen un material poco reflectante, un

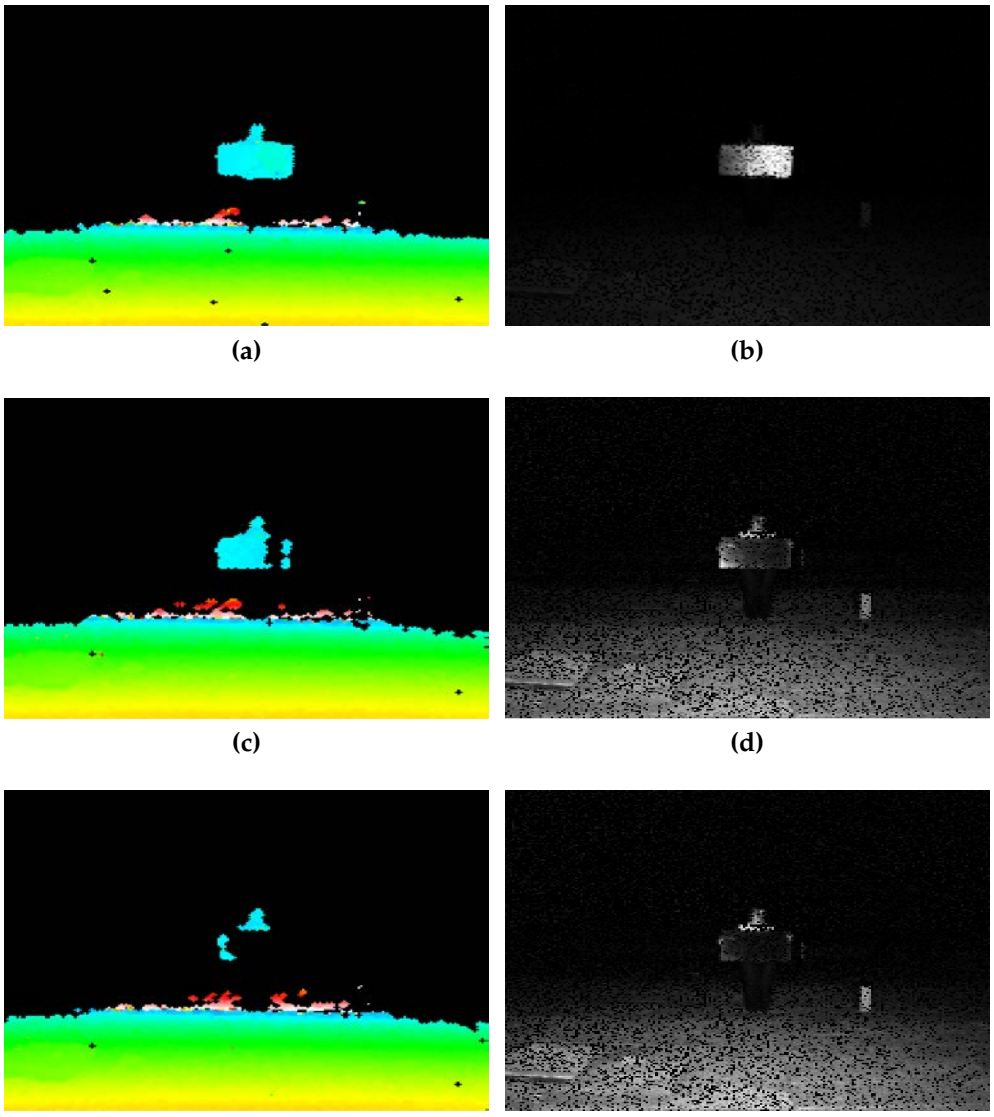
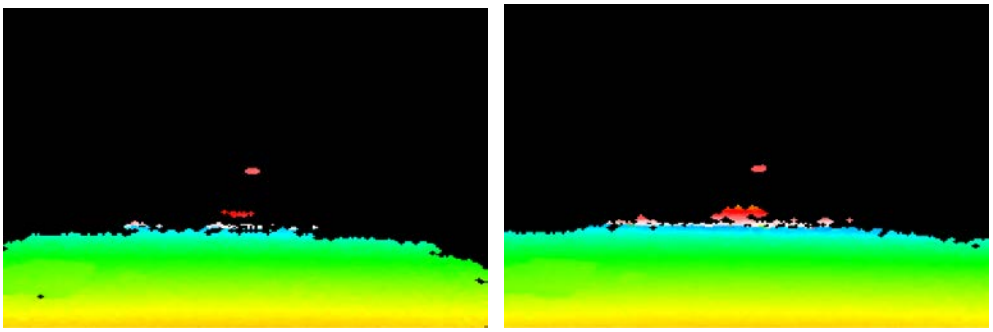


Figura 6.14: *Evaluación de los datos de profundidad en relación al ángulo de inclinación de una placa metálica respecto al sensor. A la izquierda se muestran las imágenes de profundidad con un filtrado semiautomático y a la derecha las de amplitud: (a)-(b) ángulo de 0º; (c)-(d) ángulo de 10º; (d)-(e) ángulo de 20º.*



(a) **(b)**

Figura 6.15: *Imágenes de profundidad con filtrado automático de la placa metálica a 31 metros y perfectamente paralela al plano del sensor: (a) tiempo de integración de 500 μs y 299 de amplitud; (b) tiempo de integración de 2500 μs y 950 de amplitud.*

ángulo de incidencia demasiado grande, o simplemente lejanas, pueden también tener una amplitud baja, comparable con la del fondo. Esto se puede observar en el suelo de la Fig. 6.16a-6.16c.

6.4.1.2. Escenario Bajo Contexto

En la Subsección anterior se ha comprobado que con las modificaciones incorporadas a la cámara PMD CamCube 3.0 se consigue abarcar distancias de hasta 31 metros, pero el alcance máximo de detección depende del acabado y la composición del objeto a detectar, así como el ángulo de inclinación de la luz sobre el objeto a evaluar.

Teniendo en cuenta estos resultados, para validar la tecnología ToF en un contexto aéreo es necesario realizar ensayos en condiciones similares a las reales. Debido a que este tipo de entorno experimental no existe en la universidad se recogieron imágenes para la detección automática en las instalaciones de Airbus Military - EADS en Getafe (Madrid, España). En sus instalaciones se encuentra una plataforma diseñada para la evaluación de todos los parámetros y posibles escenarios durante el proceso de repostaje en vuelo con pértiga. El edificio está equipado con una maqueta a tamaño real de una pértiga así como de una plataforma hidráulica sobre la que se sitúa el receptáculo (Fig. 6.17).

La cámara estaba montada en el primer piso del edificio, orientada para observar el receptáculo desde la plataforma. En la Fig. 6.18 se puede ver una vista a color del objeto aéreo bajo estudio así como las imágenes de profundidad, intensidad y amplitud de la escena adquiridas por la cámara en una sola captura.

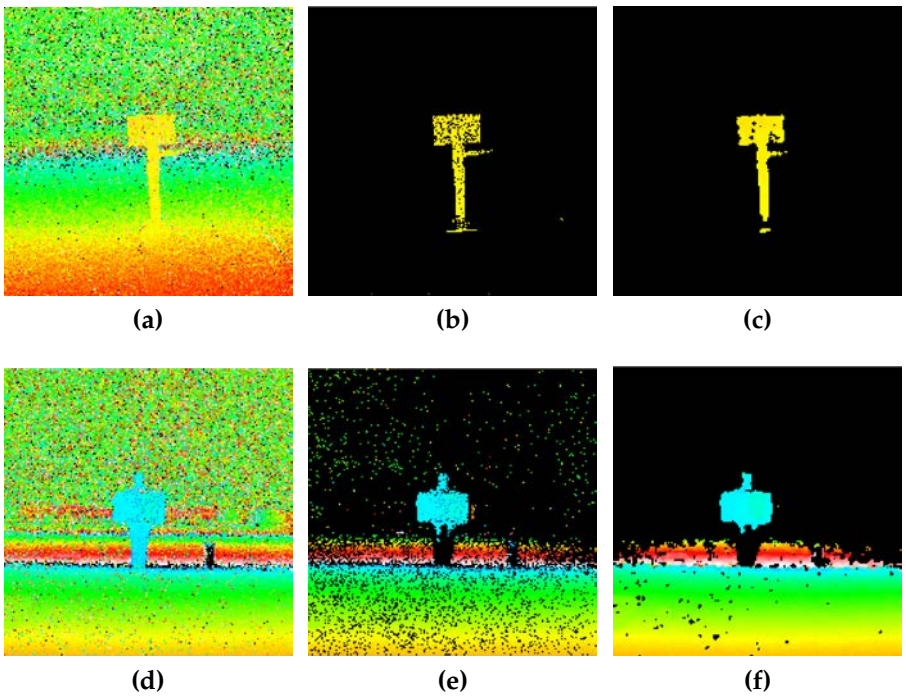


Figura 6.16: *Evaluación del filtrado en amplitud en imágenes de profundidad tomadas en la terraza: (a)-(c) día soleado; (d)-(f) día nublado. Imagen de profundidad bruta (izquierda), filtrada con un umbral sobre la amplitud (centro), y con un filtro mediano adicional (derecha).*

Las imágenes del receptáculo en movimiento fueron adquiridas en condiciones de luz variables (luz solar directa con sol y nubes). El objeto reproducía movimientos verticales y horizontales, variando su orientación para simular las maniobras de repostaje en vuelo mediante una pértiga.

6.4.1.2.1 Alcance Máximo

Debido a que el material del receptáculo tiene menor reflectividad que la placa metálica de la sección anterior, cabe esperar que la distancia máxima alcanzable sea inferior a los 31 metros abarcados. Tras diversas pruebas en exteriores y moviendo el receptáculo a diversas posiciones se pudo comprobar que el receptáculo se puede apreciar hasta una distancia aproximada de 15 metros con 7 módulos de iluminación activos. A distancias inferiores a 11 metros, el receptáculo, incluso en situaciones adversas de luminosidad, es claramente identificable en la mayor parte de las imágenes de profundidad (Fig. 6.19) y amplitud (Fig. 6.20)).

El aumento del número de módulos de iluminación activos amplía considerablemente el rango de trabajo y el margen de maniobras, ya que con dos módulos de

Figura 6.17: *Escenario de pruebas: (izquierda) Entorno experimental de Airbus Military - EADS en Getafe (Madrid) [150]; (derecha) Detalle del receptáculo bajo estudio.*

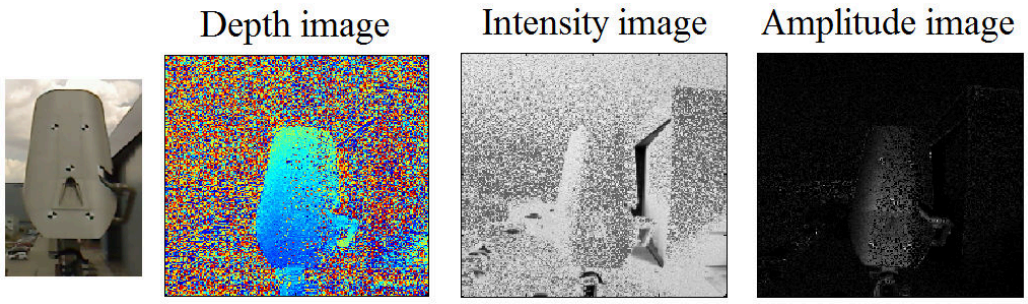


Figura 6.18: *Objeto aéreo bajo estudio. De izquierda a derecha se muestra una vista a color, y posteriormente las imágenes adquiridas de la escena con la PMD CamCube: profundidad, intensidad y amplitud.*

Figura 6.19: *Imágenes de profundidad del receptáculo en movimiento a unos 11 metros de distancia, desde una posición perpendicular al suelo a una paralela a él.*

iluminación sólo se podían abarcar distancias inferiores a 6 metros en las mejoras condiciones climatológicas.

6.4.1.2.2 Segmentación Basada en Filtrado de Amplitud

Como se observó en el apartado anterior, en las imágenes de exterior grabadas en la terraza durante los ensayos, los píxeles de fondo (correspondientes al cielo y a edificios) se eliminaban utilizando un solo umbral sobre la imagen de amplitud y un filtro mediano de tamaño 3x3. Sin embargo, se detectaban dos limitaciones en cuanto a la elección de un umbral único general y el bajo valor que tenían algunos píxeles de materiales poco reflectantes.

En el caso de las imágenes adquiridas del receptáculo, aplicando un umbral

Figura 6.20: *Imágenes de amplitud del receptáculo en movimiento a unos 11 metros de distancia, desde una posición perpendicular al suelo a una paralela a él.*

único de amplitud y un filtro mediano de tamaño 3x3, quedan patentes las dos limitaciones comentadas en la sección anterior. Este efecto se puede comprobar observando la Fig. 6.21, correspondiente a una secuencia de imágenes del receptáculo parado e incidiendo el sol sobre la zona derecha:

- Fig. 6.21e: Se eliminan ciertas partes del receptáculo, mientras que se ve una gran parte del fondo.
- Fig. 6.21f: Se consiguen más datos del receptáculo que en el caso anterior pero sigue habiendo información correspondiente al fondo.
- Fig. 6.21g: Todos los píxeles del fondo han sido eliminados, pero se han perdido datos del receptáculo en comparación al caso anterior.
- Fig. 6.21h: En comparación al resto de casos, se han eliminado una gran cantidad de datos del receptáculo, mientras que aún siguen existiendo píxeles

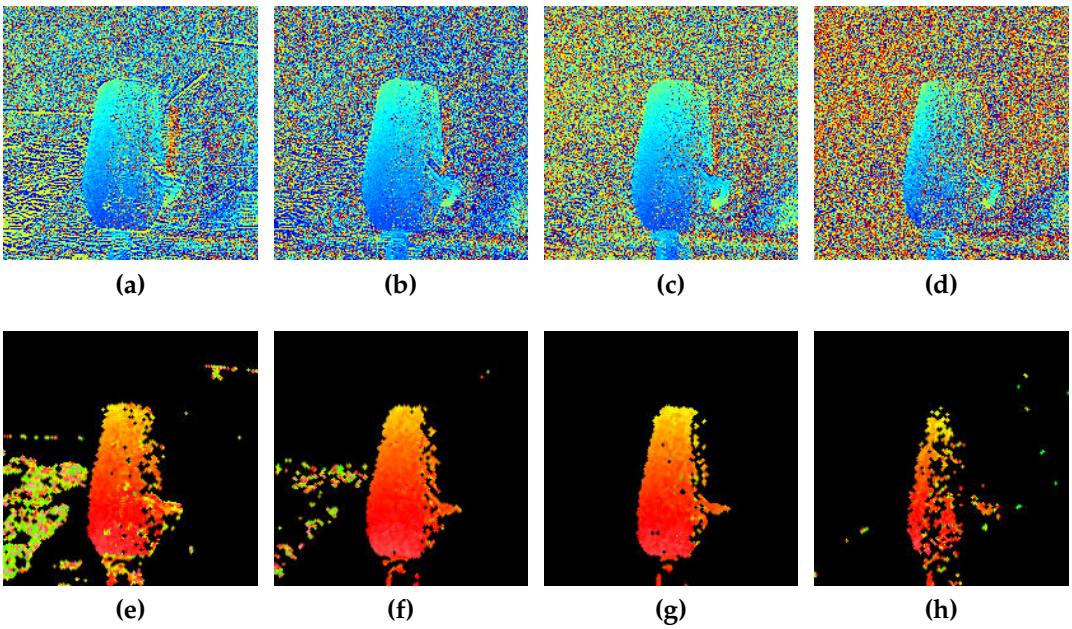


Figura 6.21: *Evaluación del filtrado en amplitud en imágenes de profundidad del receptáculo. (a)-(d) Imagen de profundidad bruta; (e)-(h) Imagen de profundidad filtrada con un umbral de 419 sobre la amplitud y con un filtro mediano adicional.*

del fondo.

Por tanto, a la vista de los resultados de los ensayos con el receptáculo, se descarta esta opción para su segmentación, debido a que:

- El receptáculo está hecho de un material poco reflectante, por lo que hay zonas en las que debido al ángulo de incidencia, luz solar, o simplemente zonas más lejanas, pueden también tener una amplitud baja, comparable con la del fondo.
- Un solo umbral no es suficiente para obtener resultados óptimos de detección en todos los casos que se pueden presentar en un entorno real, donde las condiciones climatológicas y de iluminación son variantes.

Estos efectos se observan en la Fig. 6.22, donde se muestran varias imágenes del objeto bajo estudio a una distancia de 11 metros y en movimiento. Se puede observar cómo hay píxeles no válidos que pertenecen al cielo, objetos del fondo e incluso al objeto bajo estudio. El uso de un umbral de amplitud constante como filtrado de datos no es satisfactorio debido al ruido variable en el área correspondiente al

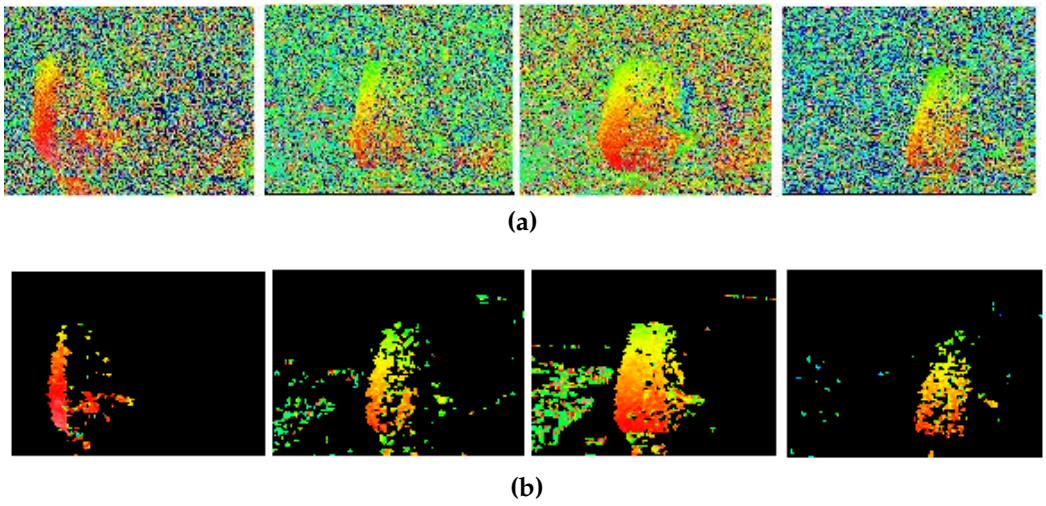


Figura 6.22: *De izquierda a derecha, imágenes del objeto bajo estudio a una distancia de 11 metros, en diferentes posiciones: (a) Imágenes de profundidad; (b) Imágenes de profundidad filtradas aplicando un umbral a la amplitud, y un filtro de mediana 3x3.*

receptáculo. Esta variación es causada por la baja reflectancia del objeto, la luz incidente y la luz solar, provocando que en algunas ocasiones el objeto apenas se vea al aplicar un filtrado de amplitud constante (Fig. 6.22b).

6.4.1.3. Comparativa de Técnicas de Segmentación

Si se observan las características de las regiones de fondo en las imágenes de rango, se aprecia que los valores de profundidad describen configuraciones con propiedades estadísticas similares a ruido uniforme. Las técnicas de segmentación en regiones pueden solventar los problemas mencionados previamente con el filtrado de amplitud, permitiendo separar las zonas que tienen valores coherentes de las zonas con valores aleatorios de profundidad.

El sistema propuesto tiene en cuenta esta apreciación, proponiendo estimar umbrales de detección de regiones homogéneas con aprendizaje basado en imágenes de ruido puro [139]. Esta técnica permite eliminar todas las regiones que son significativamente similares a un modelo estadístico de ruido, en este caso un ruido uniforme. Este razonamiento es particularmente pertinente para esta aplicación, ya que las técnicas clásicas de segmentación [151] detectan muchas regiones en las zonas de ruido como se aprecia en la Fig. 6.23.

En dicha Figura se muestra una comparativa de la segmentación realizada con técnicas clásicas y *a-contrario* para mostrar el potencial y eficacia que tiene esta última técnica para el caso bajo estudio. Para asegurar la compatibilidad con el algoritmo de segmentación se ha trabajado con las imágenes de profundidad en

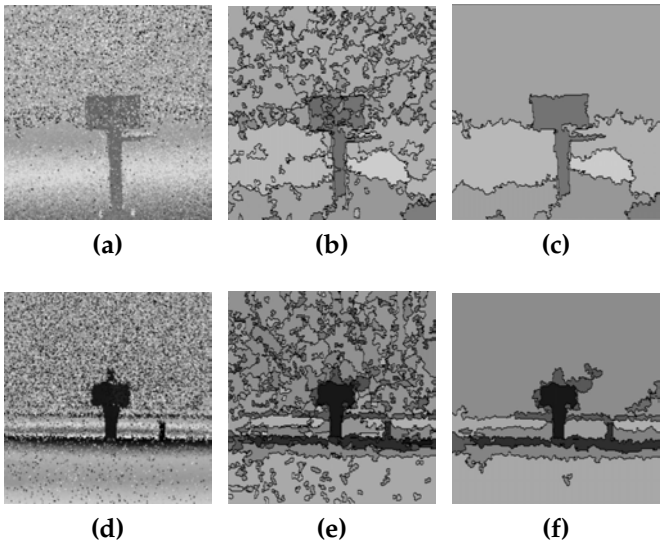


Figura 6.23: *Segmentación de imágenes de profundidad: (a), (d) Imágenes de profundidad en escala de grises; (b), (e) Segmentación basado en grafos; (c), (f) Segmentación basada en regiones coherentes.*

escala de grises (Fig. 6.23a y Fig. 6.23d).

- En el centro, se observa el resultado de un método de segmentación basado en grafos [151], donde se detectan muchas regiones del fondo.
- Con el método de filtrado *a-contrario* propuesto [139], estas regiones pueden ser eliminadas, y sólo se conservan las regiones significativamente homogéneas (Fig. 6.23b y Fig. 6.23e).

En el siguiente apartado se va a hacer una evaluación más detallada del método de segmentación propuesto para lograr un sistema de detección automático para el escenario bajo contexto.

6.4.2. Fase II de Evaluación de Segmentación: Método *A-Contrario*

Para la evaluación del método propuesto de segmentación *a-contrario* se van a seguir los siguientes pasos para validar que la técnica propuesta es adecuada para esta aplicación:

- **Evaluación del modelo de fondo**: caracterización del modelo de fondo para comprobar si sigue una distribución similar a ruido. Para ello se evalúa el histograma del cielo, del ruido blanco uniforme y del gaussiano de media 0,5

y varianza 0,1. Además, se compara el PDF y CDF del cielo y las distribuciones de tipo normal, valor extremo y uniforme que mejor se aproximan a los datos.

- **Comparativa de las distribuciones de fondo y objeto de interés**: se evalúa región por región de la imagen su distribución. Para ello se comparan los valores normalizados de las varianzas, medias y medianas de cada ventana de la imagen. Si los valores de las medidas características de las regiones de cielo y objeto son significativamente diferentes significa que la técnica de segmentación propuesta es válida.

6.4.2.1. Evaluación del Modelo de Fondo

En esta sección se evalúa y caracteriza la distribución que sigue el cielo o cualquier elemento a larga distancia en la imagen de profundidad. Para este estudio, el entorno donde se captaron las imágenes es el mismo que el de la Subsección 6.4.1.1.

Las tres imágenes utilizadas para el análisis han sido capturadas consecutivamente, con una diferencia aproximada de un segundo entre cada captura. Para comprobar si el cielo sigue una distribución de ruido blanco, en la Fig. 6.24, Fig. 6.25 y Fig. 6.26 se muestran respectivamente los histogramas normalizados de la región del cielo de cada una de las tres imágenes. Además de una imagen de ruido blanco uniforme y otra de ruido gaussiano de media 0,5 y varianza 0,1, generado mediante *rand* y *wgn* de Matlab [148], respectivamente.

Para comprobar la similitud de la distribución del cielo respecto a otras funciones se ha realizado un *fitting* con diferentes funciones de distribución proporcionadas por Matlab, y mediante el *toolbox dfittol*. Los resultados de Función de Distribución Acumulativa y Función de Densidad de Probabilidad para el cielo de la primera imagen se pueden ver en las Fig. 6.27a y 6.27b; para la segunda imagen en las Fig. 6.27c y 6.27d y para la tercera imagen en las Fig. 6.27e y 6.27f. En todas ellas se puede observar que las dos funciones que más se aproximan son la normal y valor extremo.

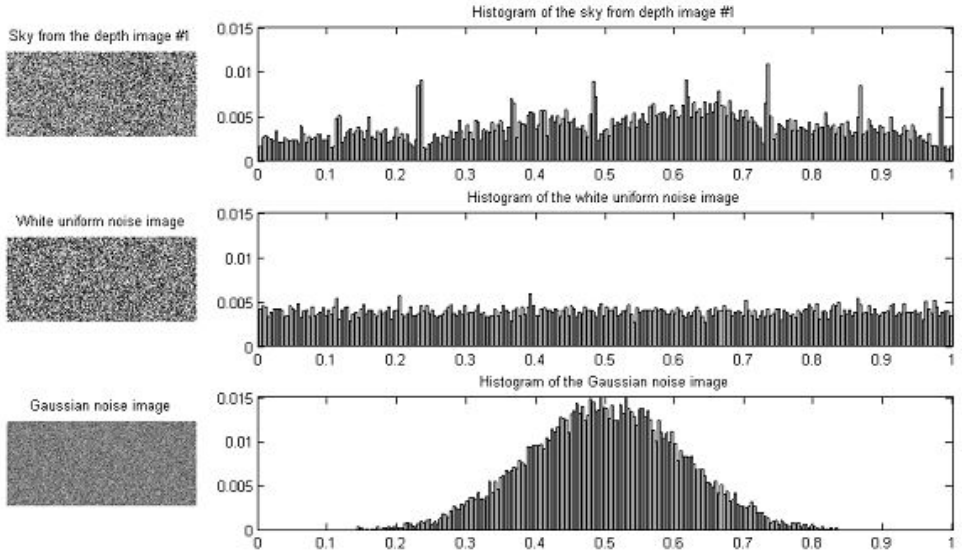


Figura 6.24: *Histograma de diversas imágenes: (arriba) cielo de la primera imagen adquirida; (centro) ruido blanco uniforme; (abajo) ruido gaussiano de media 0,5 y varianza 0,1.*

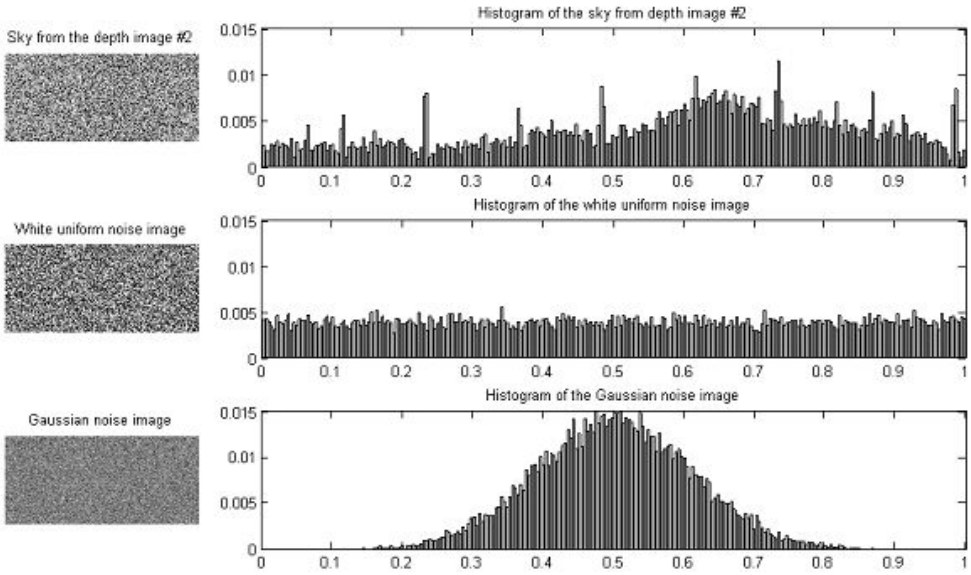


Figura 6.25: *Histograma de diversas imágenes: (arriba) cielo de la segunda imagen consecutiva adquirida; (centro) ruido blanco uniforme; (abajo) ruido gaussiano de media 0,5 y varianza 0,1.*

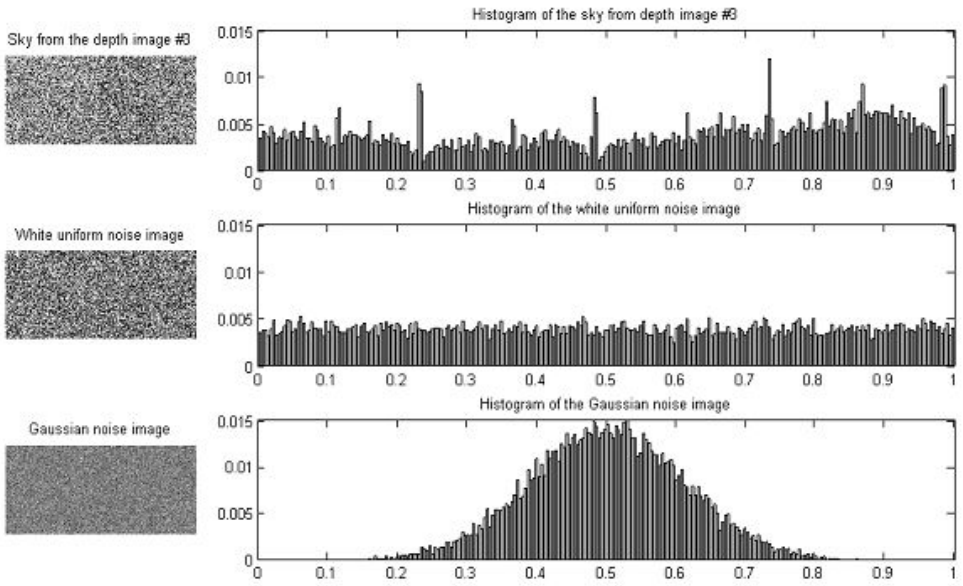


Figura 6.26: *Histograma de diversas imágenes: (arriba) cielo de la tercera imagen consecutiva adquirida; (centro) ruido blanco uniforme; (abajo) ruido gaussiano de media 0,5 y varianza 0,1.*

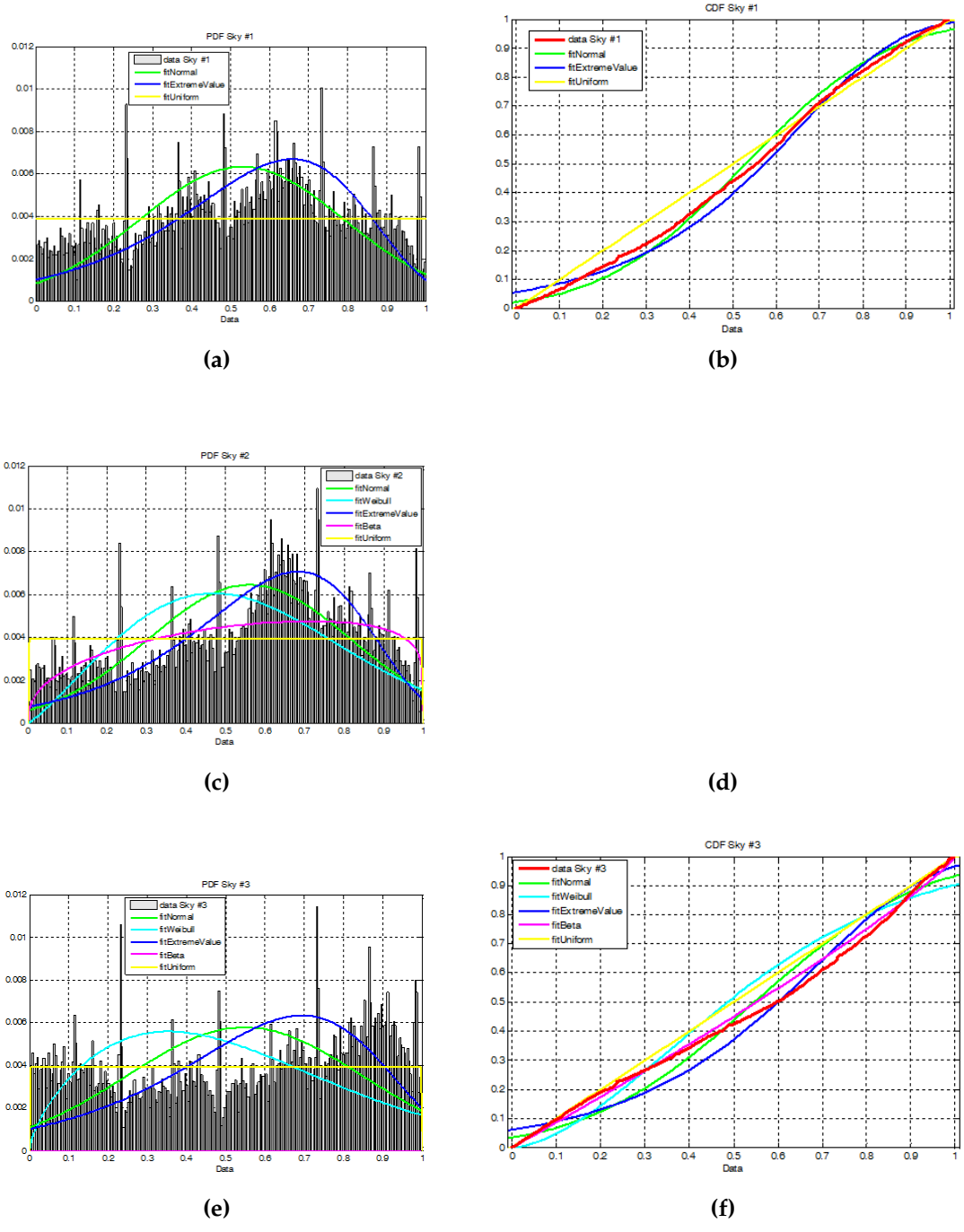


Figura 6.27: *Evaluación de la PDF y CDF de una secuencia de tres imágenes (en orden, de arriba hacia abajo) y las funciones de tipo normal, valor extremo y uniforme, que mejor se aproximan a los datos. (Izquierda) PDF de la imagen y las funciones; (Derecha) CDF de la imagen y las funciones.*

Figura 6.28: *Resultados por columnas de la evaluación de la distribución del cielo: (arriba) Imágenes de profundidad del cielo; (centro) PDF; (abajo) CDF para el cielo, la función normal, uniforme y valor extremo más similares.*

En la Fig. 6.28 se muestra un resumen de las funciones que más se aproximan a los datos del cielo de cada imagen: normal, uniforme y valor extremo.

Los parámetros característicos de la función que más se aproxima son los que se muestran a continuación (cuanto más cercano a 0 sea la probabilidad, más se parecen las distribuciones):

- Primera imagen del cielo:
  - Función normal: media, $\mu = 0,5289$, y varianza, $\sigma = 0,2618$. $Probabilidad = 1,626458e^{-165}$ (usando *Chi-square goodness-of-fit test*).
  - Función valor extremo: *location parameter*, $\mu = 0,6570$ y *scale parameter*, $\sigma = 0,2315$. $Probabilidad = 6,761430e^{-233}$ (usando *Chi-square goodness-of-fit test*).

- Función uniforme: $Probabilidad = 6{,}114156e^{-225}$ (usando *Chi-square goodness-of-fit test*.

- Segunda imagen del cielo:
  - Función normal: media, $\mu = 0,5610$, y varianza, $\sigma = 0,2570$. $Probabilidad = 1,976263e^{-323}$ (usando *Chi-square goodness-of-fit test*).
  - Función valor extremo: *location parameter*, $\mu = 0,6837$ y *scale parameter*, $\sigma = 0,2154$. $Probabilidad = 2,327126e^{-156}$ (usando *Chi-square goodness-of-fit test*).
  - Función uniforme: $Probabilidad = 0$ (usando *Chi-square goodness-of-fit test*.

- Tercera imagen del cielo:
  - Función normal: media, $\mu = 0,5479$, y varianza, $\sigma = 0,3004$. $Probabilidad = 0$ (usando *Chi-square goodness-of-fit test*).
  - Función valor extremo: *location parameter*, $\mu = 0,6931$ y *scale parameter*, $\sigma = 0,2511$. $Probabilidad = 0$ (usando *Chi-square goodness-of-fit test*).
  - Función uniforme: $Probabilidad = 3{,}725821e^{-199}$ (usando *Chi-square goodness-of-fit test*.

Se puede apreciar, tanto en las gráficas como observando los valores de probabilidad, que la función de distribución de las dos primeras imágenes se puede aproximar por una normal o valor extremo. En cambio, la imagen 3 tiene mayor similitud con una distribución uniforme. Por tanto, la similitud entre el cielo y otras distribuciones no es muy notable en todos los casos.

Por todo ello el modelo de fondo se considera como la distribución empírica de una serie de imágenes de profundidad consecutivas del cielo.

6.4.2.2. Comparativa de las Distribuciones de Fondo y Objeto de Interés

Como se ha visto anteriormente, las técnicas de segmentación *a-contrario* permiten separar las zonas que tienen valores coherentes de las zonas con valores aleatorios. Por tanto, es necesario identificar el tipo de distribución que siguen los píxeles de fondo y realizar una comparativa entre las distribuciones de zonas de la imagen de profundidad correspondientes a píxeles de fondo y del receptáculo. Si se pueden considerar distribuciones suficientemente diferentes entre ambas zonas se podría confirmar que la técnica propuesta de detección de un objeto mediante segmentación *a-contrario* es válida para esta aplicación.

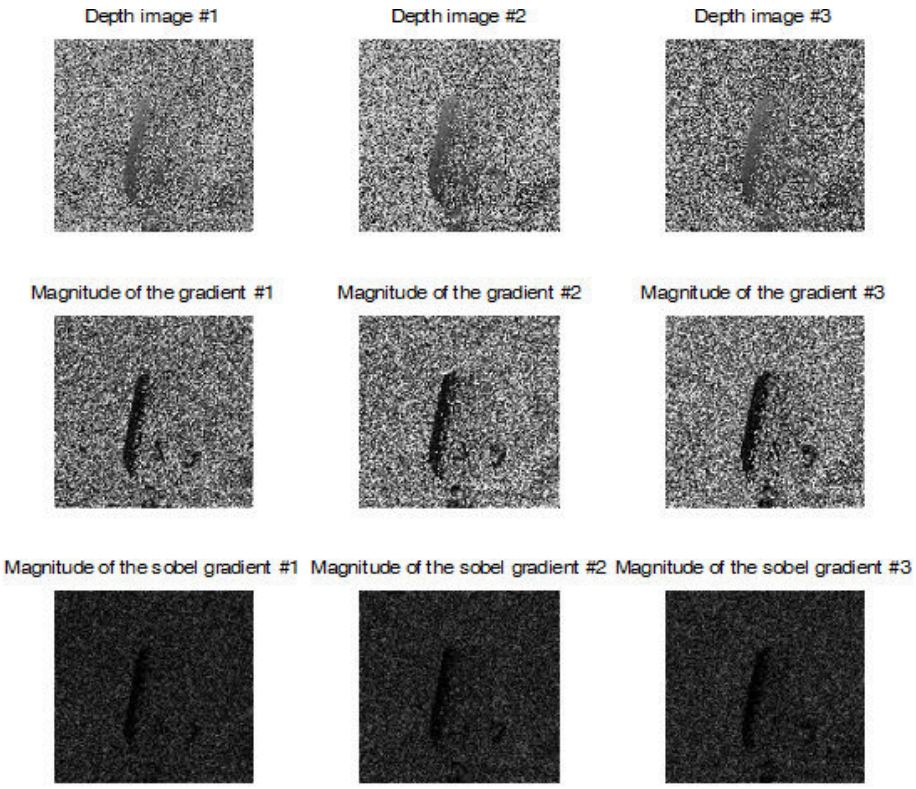


Figura 6.29: *Imágenes bajo estudio, siendo las de referencias las superiores: (arriba) Imágenes de profundidad 1, 2 y 3; (centro) Imágenes correspondientes de magnitud de gradiente con vecindad 2x2; (abajo) Imágenes correspondientes de magnitud de gradiente de Sobel.*

Por tanto, como en la subsección anterior se ha caracterizado la distribución del fondo, en esta subsección se analiza la del objeto de interés y ambas se comparan. Para este caso se utilizan tres imágenes obtenidas en el escenario de la Sección 6.4.1.2, con la PMD CamCube 3.0 y 7 módulos de iluminación observando el receptáculo en movimiento desde la torre de las instalaciones de Airbus Military. En la Fig. 6.29 se muestran las tres imágenes bajo estudio, así como sus correspondientes imágenes de magnitud de gradiente con vecindad 2x2 y 3x3 (Sobel).

Se han calculado medidas características de las distribuciones de 50 ventanas de 20x40 píxeles (ancho x alto) de imágenes 200x200 píxeles, asignando a cada una un número identificativo (de izquierda a derecha y de arriba a abajo). Se ha escogido esta dimensión de ventana considerando el tamaño del receptáculo en la imagen. De esta forma, como se puede observar en la Fig. 6.30, se recogen todos los casos posibles de distribuciones a estudiar en las ventanas:

- Fondo: como es el caso de las ventanas 1 a 14.

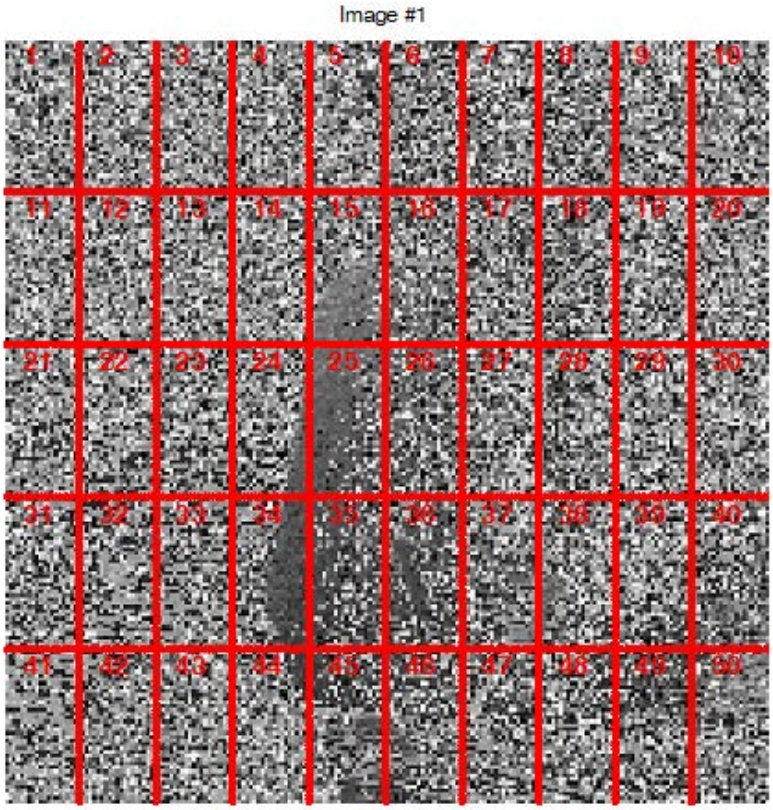


Figura 6.30: *Detalle de las 50 ventanas numeradas, de tamaño 20x40 píxeles (ancho x alto) de una imagen de profundidad 200x200 píxeles.*

- Receptáculo: como por ejemplo la ventana 25.
- Fondo y receptáculo: por ejemplo, la ventana número 34.

Para evaluar la distribución de cada ventana, y de esta forma comprobar cuán diferentes son las funciones de zonas de fondo respecto a las correspondientes a receptáculo, se han escogido como medidas características de las distribuciones las siguientes:

- Varianza de las 50 ventanas de la imagen de profundidad (Fig. 6.31).
- Mediana (Fig. 6.32) y media (Fig. 6.33) de cada una de las 50 ventanas en la imagen de magnitud de gradiente con vecindad 2x2.
- Mediana (Fig. 6.34) y media (Fig. 6.35) de cada una de las 50 ventanas en la imagen de magnitud de gradiente con vecindad 3x3 (Sobel).

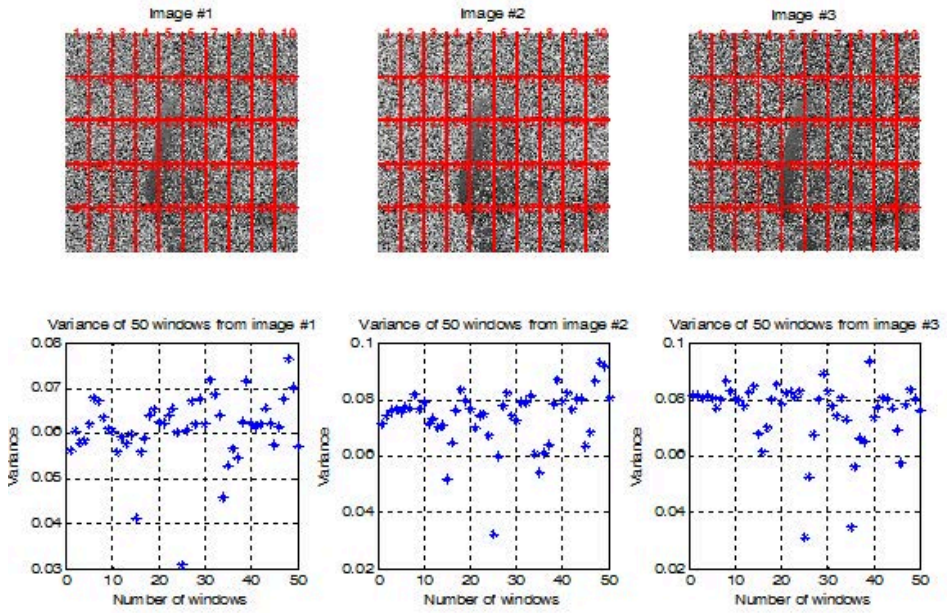


Figura 6.31: *Imágenes bajo estudio: (arriba) Imágenes de profundidad, mostrando en rojo las 50 ventanas numeradas evaluadas; (abajo) Valores de la varianza de cada una de las ventanas.*

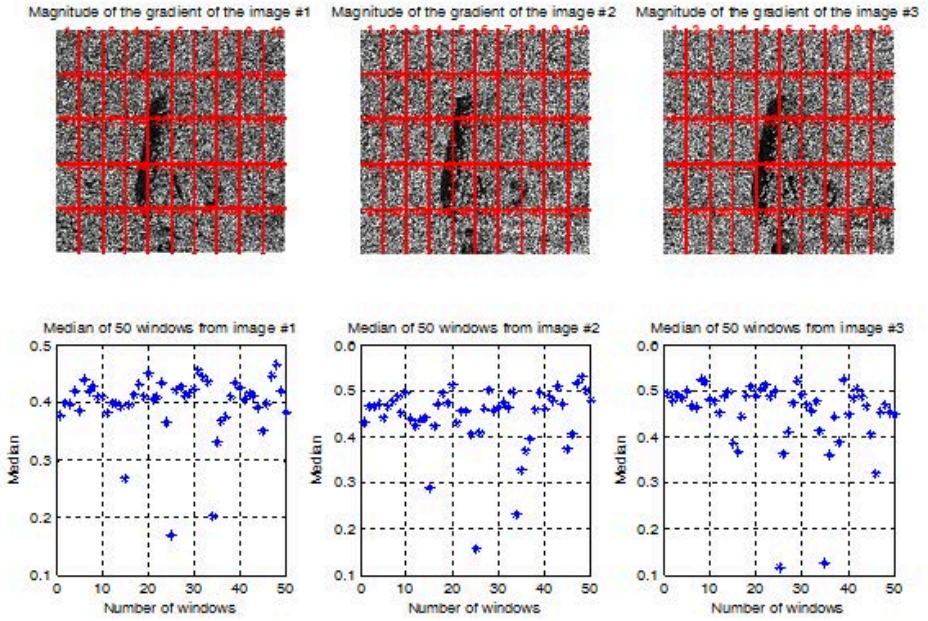


Figura 6.32: *Imágenes bajo estudio: (arriba) Magnitud de gradiente 2x2, mostrando en rojo las 50 ventanas numeradas evaluadas; (abajo) Valores de la mediana de cada una de las ventanas.*

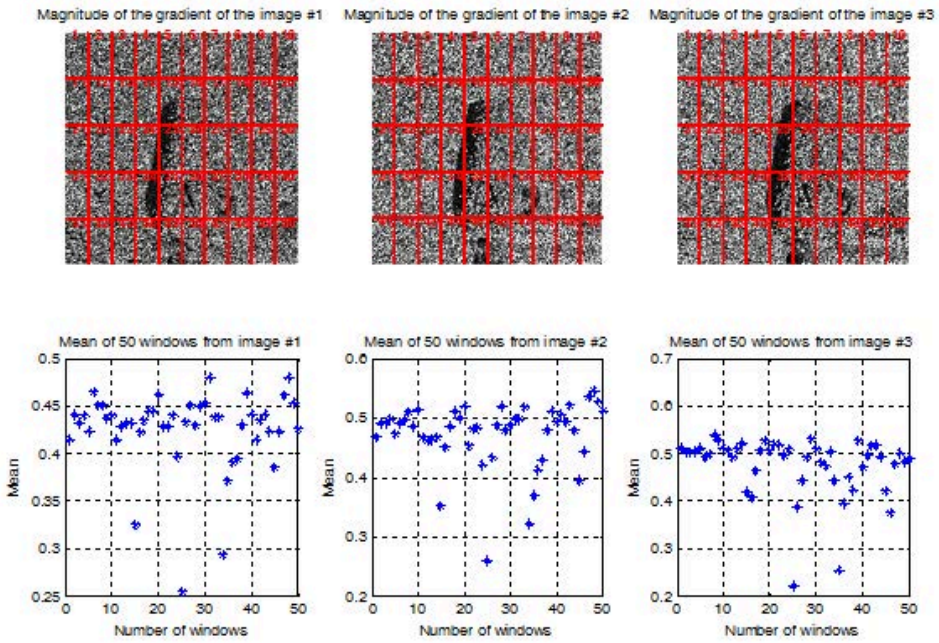


Figura 6.33: *Imágenes bajo estudio: (arriba) Magnitud de gradiente 2x2, mostrando en rojo las 50 ventanas numeradas evaluadas; (abajo) Valores de la media de cada una de las ventanas.*

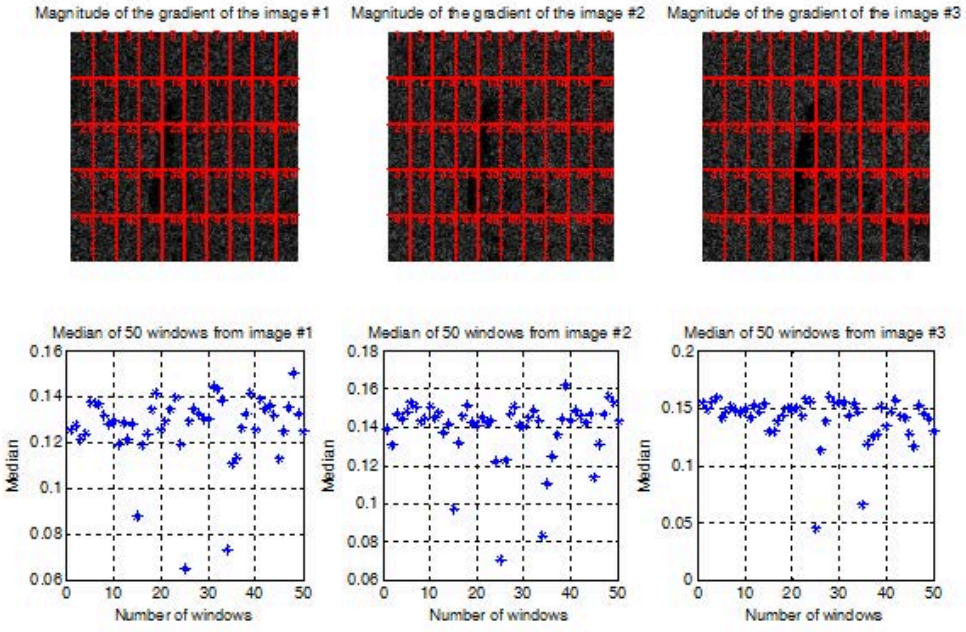


Figura 6.34: *Imágenes bajo estudio: (arriba) Magnitud de gradiente de Sobel, mostrando en rojo las 50 ventanas numeradas evaluadas; (abajo) Valores de la mediana de cada una de las ventanas.*

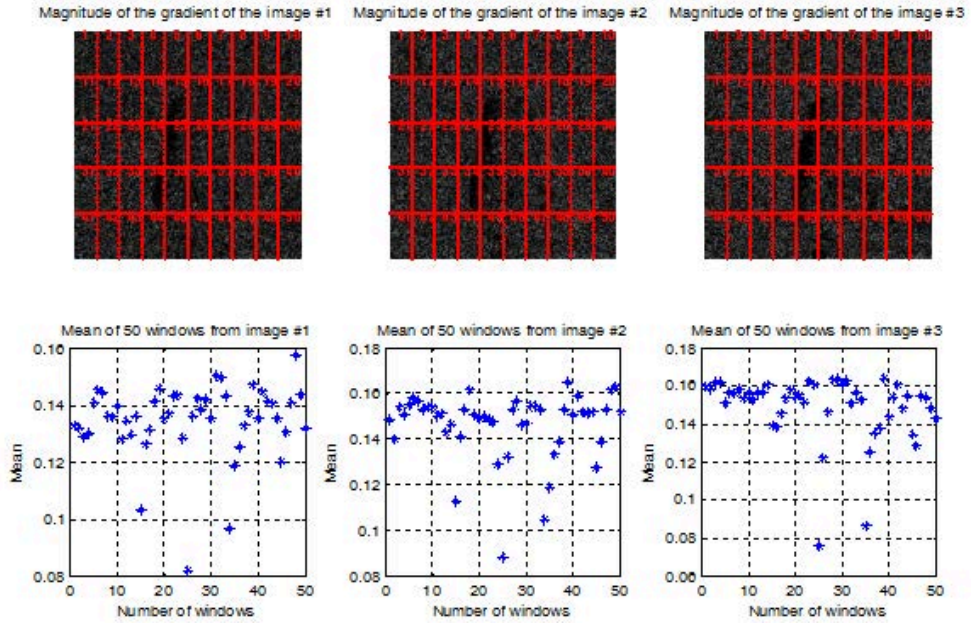


Figura 6.35: *Imágenes bajo estudio: (arriba) Magnitud de gradiente de Sobel, mostrando en rojo las 50 ventanas numeradas evaluadas; (abajo) Valores de la media de cada una de las ventanas.*

Figura 6.36: *Valores de varianza de las 50 ventanas evaluadas de la imagen de profundidad, así como valores de la mediana y media de la imagen de magnitud del gradiente 2x2 y 3x3: (arriba) Imagen 1; (centro) Imagen 2; (abajo) Imagen 3.*

Se puede ver el resumen normalizado de las distribuciones anteriores en la Fig. 6.36, donde se puede apreciar que la imagen de magnitud de gradiente con Sobel no proporciona sustanciales mejoras. Por tanto, para facilitar la visualización de los resultados, en la Fig. 6.37 se muestran los valores normalizados de las varianzas de las ventanas de la imagen de profundidad, la mediana y la media de las de la imagen de magnitud de gradiente de vecindad 2x2.

Teniendo en cuenta los valores de las medidas características de las distribuciones de cada ventana, en la Fig. 6.37 se pueden observar los siguientes aspectos:

- Las ventanas que engloban píxeles correspondientes a cielo tienen unas medidas características similares entre sí, siendo cercanas al valor 1 cuando se normalizan.

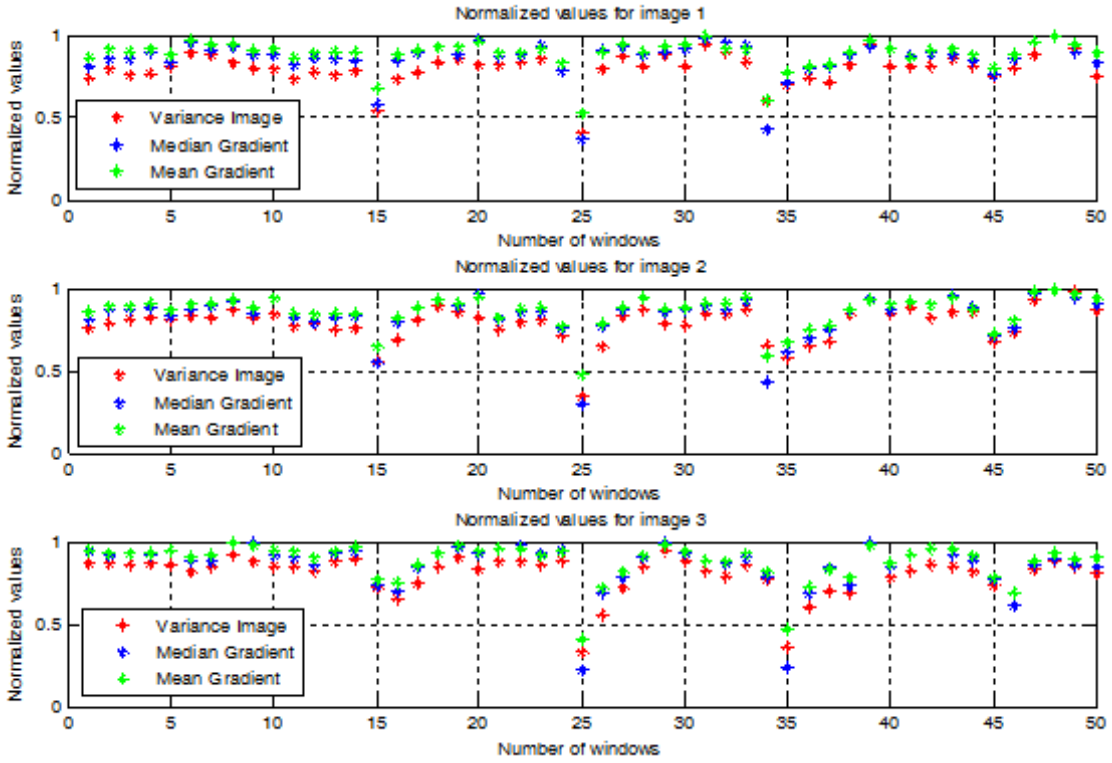


Figura 6.37: *Valores de varianza de las 50 ventanas evaluadas de la imagen de profundidad, así como los valores de la mediana y media de la imagen de magnitud del gradiente 2x2: (arriba) Imagen 1; (centro) Imagen 2; (abajo) Imagen 3.*

- En las ventanas que engloban al receptáculo se aprecia que los valores de las medidas características son inferiores a los del caso anterior, siendo los valores normalizados cercanos a cero.
- La distribución entre ventanas que contienen fondo o receptáculo es diferente entre sí.

Tras evaluar los resultados, se ha podido comprobar que el tipo de distribución que siguen los píxeles de fondo es significativamente diferente a las distribuciones de las zonas que corresponden a receptáculo. Por tanto, la técnica de segmentación *a-contrario* propuesta para esta aplicación se considera válida y podría proporcionar resultados satisfactorios. Sin embargo, es necesario evaluar un algoritmo de detección automática, cuya comparativa utilizando la media y varianza se muestran en la siguiente sección.

6.4.3. Evaluación del Método *A-Contrario* de Detección Automática

Los resultados de la Subsección anterior han proporcionado un modelo de fondo aproximado basado en la distribución empírica de una serie de imágenes de profundidad consecutivas del cielo. Además, se ha comprobado que su distribución es suficientemente diferente a la de las zonas correspondientes al objeto. Por tanto, una vez validada la técnica de segmentación *a-contrario*, se procede a determinar de forma experimental el tamaño y forma de la región a explorar más idónea para establecer el método *a-contrario* de detección automático:

- **Evaluación del método más detección más eficaz**: La comparativa de los resultados obtenidos sobre la detección correcta o errónea de las regiones del receptáculo se realiza mediante el análisis de las curvas ROC (acrónimo de *Receiver Operating Characteristic*, o Característica Operativa del Receptor). La elección de estas curvas se debe a su idoneidad para comprobar la eficacia de métodos de detección *a-contrario* validada para localizar cambios en imágenes de radar [152]. La elección del criterio se basará en seleccionar aquel con mayor área bajo la curva, siendo el más discriminatorio y por tanto, más cercano al caso ideal de $(0, 1)$.
- **Evaluación del umbral de detección automática**: Teniendo en cuenta las pruebas realizadas con las curvas ROC, se elige un umbral ϵ que permita obtener una tasa de detección fallida mínima. Se buscará para ello un compromiso entre la tasa de detección y el número de falsas alarmas. Se realizan pruebas experimentales con dicho umbral sobre tres imágenes de referencia

para mostrar la idoneidad del criterio escogido. Esta tasa se fija de acuerdo a un análisis de los resultados debido a que es una forma general de validar el método en contexto experimental y no se puede fijar de manera intuitiva [153].

Las curvas ROC constituyen una herramienta muy adecuada para la visualización del comportamiento de clasificadores [154]. Estas curvas muestran la Tasa de Verdaderos Positivos o *True Positive Rate* (TPR) frente a la Tasa de Falsos Positivos o *False Positive Rate* (FPR) para varios umbrales, de forma acumulada. Si la curva se encuentra cerca de la diagonal principal de $(0, 0)$ a $(1, 1)$ significa que el método no es adecuado debido a que la proporción de falsos positivos es igual a la de verdaderos positivos. Por el contrario, una curva ROC con un punto en $(0, 1)$ es el modelo ideal ya que significa que el método propuesto separa perfectamente positivos y negativos. El área bajo la curva (AUC, de Area Under the Curve) resume el comportamiento global. De hecho, el método ideal sería aquel cuyo AUC tomara el valor 1 en todo el rango, considerándose ya como razonable a partir de 0,5 [155].

Teniendo en cuenta este criterio, la imagen de 200x200 píxeles se divide en 100 ventanas de 20x20 píxeles. Se parte de ventanas más pequeñas para evaluar un mayor número en este caso. Observando la imagen de la izquierda de la Fig. 6.38, 26 ventanas contienen objeto por lo que 74 no, siendo estos valores los utilizados para calcular la tasa de TPR y FPR, respectivamente. Se toma un rango de 101 valores de umbrales, entre 0 y 15 espaciados linealmente, para el cálculo de las curvas ROC. Una vez obtenidas estas curvas se calcula su envolvente convexa para poder obtener el AUC [155]. Los resultados relacionados con las imágenes de magnitud de gradiente son los más cercanos al caso ideal de $(0, 1)$ como se puede apreciar a la derecha en la Fig. 6.38. De los dos criterios, el más discriminatorio es el de la media de la imagen de gradiente 2x2 ya que su AUC es mayor, siendo cercano al 95 %, y disminuyendo la FPR casi a cero mucho antes que la TPR empiece a decrementar.

Se repite la misma prueba con 50 ventanas de 20x40 píxeles (ancho x alto). En este caso también el criterio de decisión de la media usando la imagen de magnitud de gradiente 2x2 es más robusto, como se aprecia en la Fig. 6.39. La tasa de falsos positivos es del 17 % para una tasa de detección del 100 %, y considerando aceptable una baja de tasa de detección fallida, la tasa de falsas alarmas se reduce al 3 % para una tasa de detección del 95 %.

Debido a que las ventanas 20x40 píxeles son más adecuadas para el tamaño del receptáculo, como se mencionó anteriormente, se realiza la elección del umbral sobre este caso ya que, además, la validación será más consistente porque muestra resultados más restrictivos en las curvas ROC. En las Figuras 6.40 y 6.41 se muestran respectivamente los valores de la media de cada una de las ventanas bajo estudio, μ_w, para las imágenes de magnitud de gradiente obtenidas con vecindad 2x2 y 3x3 de la imagen de profundidad. Se puede apreciar cómo determinadas

Figura 6.38: *Envolventes convexas de las curvas ROC del método a-contrario evaluado con ventanas de 20x20 sobre la imagen de la izquierda.*

Figura 6.39: *Las curvas ROC del método a-contrario evaluado con ventanas de 20x40, mostrando en naranja su envolvente convexa.*

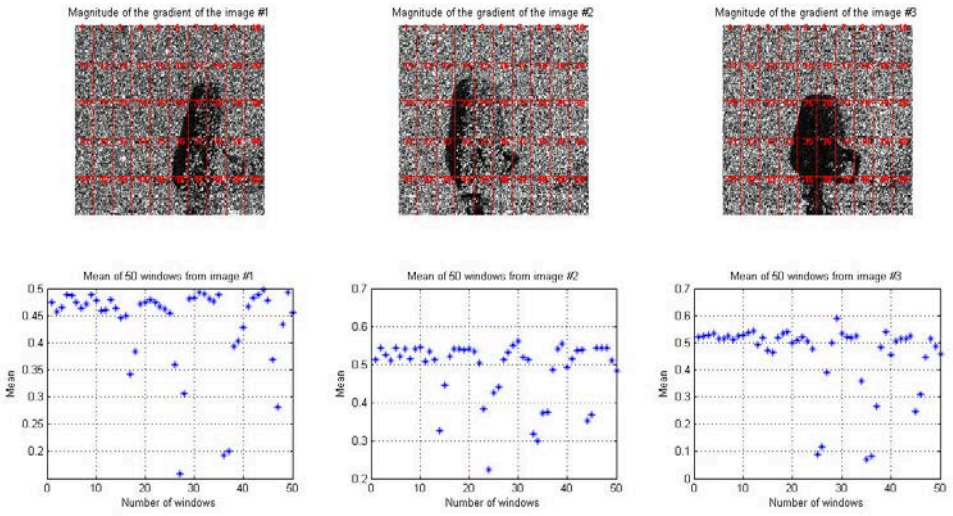


Figura 6.40: *Media de las ventanas 20x40 píxeles de la imagen de magnitud de gradiente 2x2 de un receptáculo en tres posiciones y orientaciones diferentes.*

ventanas tienen una media muy diferente por lo que es necesario establecer un umbral sobre el NFA para la detección automática.

Teniendo en cuenta las pruebas realizadas con las curvas ROC, se elije $0,11$ como umbral ϵ para el $NFA(\mu_w)$ usando la medida de la media de la imagen de magnitud de gradiente 2x2 para la detección automática *a-contrario*. Este valor se ha escogido teniendo en cuenta el rango de umbrales bajo evaluación en las curvas ROC y llegando a un compromiso de baja tasa de detección fallida frente a pocos falsos positivos. Se ha aplicado el umbral propuesto sobre tres imágenes de referencia para su validación.

La Fig. 6.42 muestra la exitosa detección automática del receptáculo en diferentes posiciones y orientaciones usando el método de detección *a-contrario* con un umbral de $0,11$ para el $NFA(\mu_w)$ sobre la imagen de magnitud de gradiente de la de profundidad. Teniendo en cuenta que los valores de PFA son muy pequeños, se ha utilizado el logaritmo neperiano, como se mencionó en la Sección 6.2.1.2.

Se puede comprobar en la Fig. 6.43 que, al intentar aplicar la detección automática del receptáculo con un umbral prefijado sobre la imagen de magnitud de gradiente 3x3, no se obtienen resultados exitosos en todos los casos. Esta conclusión es reflejo de la mayor lejanía de la curva ROC del método al caso ideal mostrada en la Fig. 6.39.

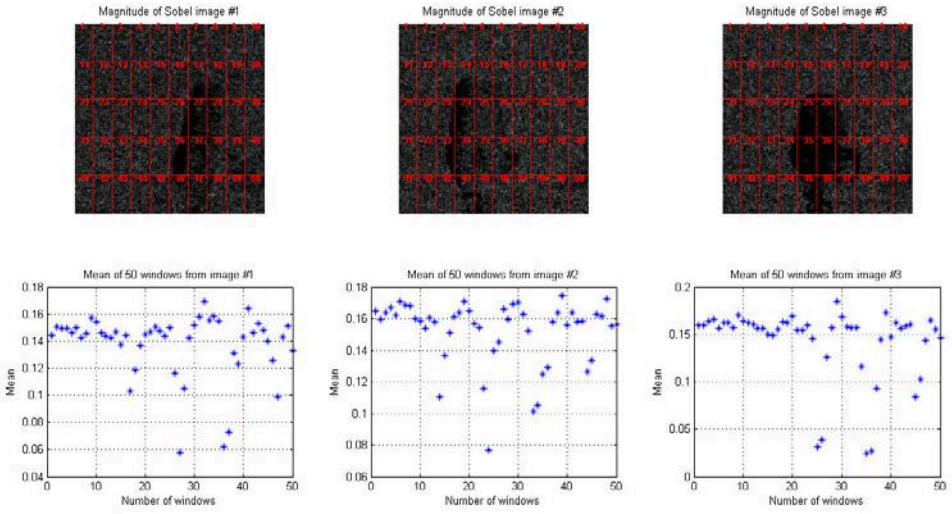


Figura 6.41: *Media de las ventanas 20x40 píxeles de la imagen de magnitud de gradiente 3x3 de un receptáculo en tres posiciones y orientaciones diferentes.*

Figura 6.42: *Resultados de detección automática con un umbral* ϵ *de* $0,11$ *para valores de media de la imagen de magnitud de gradiente 2x2 de un receptáculo en tres posiciones y orientaciones diferentes. En rojo se muestra el* $ln(0,11)$.

Figura 6.43: *Resultados de detección automática con un umbral ϵ de $0,11$ para valores de media de la imagen de magnitud de gradiente 2x2 de un receptáculo en tres posiciones y orientaciones diferentes. En rojo se muestra el $ln(0,11)$.*

6.5. Conclusiones

Con este último escenario se han abordado todos los entornos a evaluar en esta tesis. En este caso, se ha desarrollado un sistema de percepción para un entorno exterior semiestructurado proponiendo una estrategia para la detección de objetos con una cámara de rango ToF. De entre los pocos dispositivos ToF que funcionan en exteriores se ha seleccionado la cámara PMD CamCube 3.0, elegida debido a su idoneidad para el entorno bajo estudio.

En este tipo de escenarios exteriores, las imágenes de profundidad de una cámara ToF habitualmente están influenciadas por ruido debido a la luz del sol, reflectividad de los objetos, condiciones de iluminación, entre otros. Por este motivo, se ha propuesto un razonamiento *a-contrario* para la detección automática de objetos en exteriores usando imágenes de rango. Este método permite separar las zonas que tienen valores coherentes de las zonas que tienen valores de profundidad aleatorias. El sistema de detección *a-contrario* propuesto se caracteriza por las siguientes características:

- El modelo de fondo corresponde a la distribución empírica de una serie de imágenes de profundidad consecutivas del cielo.
- En ausencia de cualquier suposición acerca de la forma del objeto, el método de detección se basa en observar a un grupo de píxeles adyacentes en lugar de un único píxel. Esto es debido a que el objeto ocupará más de un píxel.
- El sistema compara una región de una imagen con el modelo de fondo, mediante el análisis de la media o varianza de cada distribución.
- Cada detección está asociada a un nivel de confianza estableciendo un valor de similitud. El umbral de detección automática se establece evaluando varias imágenes de referencia.

Estos parámetros permiten adaptar fácilmente el método propuesto a numerosas aplicaciones haciendo un uso eficaz de la técnica de detección *a-contrario* propuesta para exteriores.

Debido a la novedad de esta aplicación, se realizó además una exhaustiva evaluación de las imágenes de profundidad. Para ello se caracterizó la influencia de la potencia transmitida, la frecuencia de modulación y la reflectividad de los objetos mediante el análisis de la información relevante de las imágenes así como su fiabilidad.

- Evaluación de la cámara PMD CamCube mediante un análisis técnico y experimental de la influencia de la potencia transmitida, la frecuencia de modulación y la reflectividad de los objetos en la distancia máxima alcanzable.

- Comparativa entre la versión 2.0 y la 3.0 con el hardware adicional: el número de módulos de iluminación y el filtro de paso banda 800-900 nm han permitido reducir enormemente el efecto negativo del sol, ya que sólo existe ruido en las zonas del receptáculo donde incide directamente el sol, teniendo en las zonas de sombra datos de profundidad correctos. La cámara ToF PMDCamCube 3.0 con el hardware adicional permite:
  - Apreciar el receptáculo hasta una distancia aproximada de 15 metros, cumpliendo el objetivo inicial.
  - A distancias inferiores a 11 metros, el receptáculo, incluso en situaciones adversas de luminosidad, es claramente identificable en la mayor parte de las imágenes de profundidad y amplitud.

Estos resultados revelaron que:

- No es posible usar un filtrado de amplitud sobre las medidas de profundidad obtenidas y, principalmente cuando el objeto a detectar tiene baja reflectividad. Esto es debido a que al aplicar un cierto umbral de amplitud para el filtrado, los valores de distancia tanto del cielo como del objeto son ignorados debido a que apenas reflejan la luz emitida.
- Se ha apreciado que los valores de las regiones de fondo en las imágenes de rango describen típicas configuraciones con propiedades estadísticas, similar a ruido uniforme. Por otra parte, el objeto a detectar se percibe como una desviación de un modelo de aleatoriedad completa.
- Es eficaz el enfoque de modelar las propiedades del fondo debido a que, al no tener conocimiento *a priori* del objeto a detectar no es posible modelarlo.

Estas conclusiones afianzan el sistema propuesto de percepción ya que es eficaz frente a las configuraciones similares a ruido y no necesita conocer previamente la forma del objeto.

Para validar la estrategia propuesta se evalúa el sistema en un contexto aéreo de repostaje en vuelo. Este escenario es idóneo ya que el cielo ofrece amplitudes cercanas a cero y además, el fuselaje del avión tiene una baja reflectancia a infrarrojos, por lo que se producen mediciones ruidosas de distancia y valores bajos de amplitud. Los puntos clave de las diferentes fases de la estrategia propuesta evaluada son:

- Evaluación de los píxeles de fondo: se ha identificado que la distribución de los píxeles del cielo es similar a ruido, siendo claramente diferente a la distribución de las zonas donde existen datos de profundidad procedentes

del receptáculo. Se concluye que el fondo debe ser definido empíricamente a partir de una serie de imágenes de profundidad del cielo.

- Comparación entre una región del receptáculo y el cielo: una vez que el fondo se ha definido, se han propuesto dos criterios *a-contrario* para detectar al objeto mediante el análisis de la media o la varianza de las distribuciones de las regiones a evaluar. Para la media, se utilizan las imágenes de magnitud de gradiente con vecindad 2x2 y 3x3 calculadas a partir de la imagen de profundidad. Por otro lado, para la varianza se ha utilizado directamente la imagen de profundidad.

- Elección del umbral: las curvas ROC revelan que el enfoque más discriminativo es el de detección *a-contrario* utilizando la imagen de magnitud de gradiente con vecindad 2x2. Este método realiza una tasa de detección del 100 % con un 17 % de falsos positivos, que se pueden reducir al 3 % si se acepta reducir la tasa de detección al 95 %. Las pruebas se han llevado a cabo con 50 ventanas de 20x40 píxeles sobre tres imágenes del receptáculo en diferentes posiciones y orientaciones. Para otros tamaños de ventana la tasa es similar, y el criterio de decisión de la media usando la imagen de magnitud de gradiente 2x2 es siempre más robusta.

- Para el contexto bajo evaluación se concluye que el método de detección *a-contrario* automático propuesto se caracteriza por un umbral de $\epsilon = 0,11$ para el $NFA(\mu_w)$. Para ello se utiliza la media de la imagen de magnitud de gradiente con vecindad 2x2. De esta forma, se propone que una región de la imagen es detectada como objeto de interés si existe una ventana de tamaño w cuyo Número de Falsas Alarmas cumple $NFA(\mu_w) < 0,11$.

Este umbral se ha fijado de acuerdo a un análisis posterior de los resultados de detección, una forma general de validar el método en un contexto experimental [153]. Los resultados sobre las imágenes reales muestran que la detección automática se realiza de forma exitosa con la estrategia propuesta, mostrando su robustez frente a falsas alarmas. Sin embargo, debido a que, en teoría, el NFA corresponde a una cantidad intuitiva, al indicar que el umbral es $0,11$ significa que de media se debe observar una falsa alarma cada 10 imágenes. En la práctica, en los experimentos con el objeto a diferentes distancias existen mas falsas alarmas que lo que predice el modelo. Por tanto, cabe esperar que los píxeles no son todos independientes en la imagen de profundidad y el tamaño de ventana w debería adaptarse al objeto a buscar. Se concluye que, a pesar de la eficacia para la detección de objetos en imágenes ruidosas, para una generalización completa del método automático es necesario llevar más estudios a cabo con otros conjuntos de datos.

Una explicación más detallada de este trabajo futuro así como las conclusiones generales de las estrategias propuestas para los escenarios poco estructurados evaluados se expondrán en el siguiente capítulo de la tesis.

Capítulo 7

Conclusiones y Trabajo Futuro

En esta tesis se han presentado estrategias de percepción 3D robustas y eficaces para entornos poco estructurados. Para ello, partiendo de los datos adquiridos por una cámara de rango se ha desarrollado una estrategia global mediante la definición de diversos bloques que apoyan un diseño unificado e integrado de sistema visual.

Debido a la complejidad de estos entornos, la tesis describe un enfoque que permite reducir la incertidumbre centrándose en un determinado contexto para plantear la estrategia a seguir, reduciendo el espacio de estados sin afectar a la tarea a desempeñar. De esta forma, se ha planteado una forma eficaz de seleccionar los ambientes complejos según niveles de estructuración del entorno, analizando las condiciones claves.

Se ha demostrado que, considerando condiciones tecnológicas, ambientales y de incertidumbre en torno a la identidad y localización del objeto es posible establecer unos contextos experimentales donde validar una estrategia visual que aborde los desafíos de una forma eficaz. Aunque los requisitos del entorno repercuten en la elección del hardware y el software del sistema, cabe destacar que los métodos se han desarrollado de la forma más general posible para su fácil adaptación a otros entornos de escenarios similares.

El trabajo ha demostrado bajo un contexto experimental que una misma estrategia visual se puede integrar en escenarios de percepción de objetos tanto en interiores como en exteriores, con diferentes grados de incertidumbre. Para ello se ha adaptado cada bloque de la estrategia global al contexto para explotar al máximo las características de la escena y de los datos adquiridos. Por último, cabe destacar que en estos escenarios poco estructurados se han abordado con éxito y de manera explícita varios desafíos críticos en la percepción visual 3D de objetos.

A continuación se detallan las contribuciones relacionadas con estos desafíos

abordados. Posteriormente, se mencionarán las líneas futuras de investigación así como las publicaciones más relevantes relacionadas con esta tesis.

7.1. Resumen de Contribuciones

Esta tesis contribuye en líneas generales al desarrollo de estrategias de percepción 3D robustas y eficaces para entornos poco estructurados. Para ello se han explotado al máximo las características de los entornos y las ventajas de cada cámara de rango empleada. En línea con los desafíos identificados en el Capítulo 1, este trabajo comprende las siguientes contribuciones:

- **Método de reconstrucción 3D de objetos desconocidos partiendo de información parcial de la escena**: la estrategia descrita en el Capítulo 4 permite obtener el modelo 3D de un objeto cuando la cámara sólo puede ver la escena desde un único punto de vista, siendo robusto incluso cuando los datos disponibles de un objeto son escasos. Los puntos clave de este método son tres. En primer lugar, se propone una nueva técnica para la extrusión de una nube de puntos dispersa basándose en la extrusión de los correspondientes a la parte superior de los objetos. En segundo lugar, se propone una refinamiento que se aprovecha de la complementariedad de las imágenes de profundidad y color inicializando con los datos de profundidad una segmentación a color basada en un algoritmo de cortes de grafos. Finalmente, se realiza una evaluación cuantitativa de la exactitud de las mallas reconstruidas demostrando que la eficacia del método propuesto es comparable en precisión a un enfoque utilizando simetrías, logrando mejores tiempos de ejecución.

- **Combinación de técnicas de color y profundidad para percibir objetos independientemente del grado de textura**: en el Capítulo 4 se han integrado con éxito diversos métodos tanto de información 2D como de 3D para obtener un sistema visual completo que incluye calibración, segmentación, detección y reconocimiento de objetos sobre una mesa. En este caso el sistema se adapta a la escena visualizada, dando la posición 6D de los objetos reconocidos incluso estando ocluidos.

- **Conjunción de características fotométricas y geométricas para lograr una percepción eficaz y más fácilmente adaptable a objetos diferentes**: se ha desarrollado un enfoque basado en propiedades intrínsecas del objeto para poder reconocerlo explotando las propiedades que se conocen *a priori* de él. Para ello, en el Capítulo 5, se utiliza la integración de descriptores de color y forma. La segmentación de crecimiento de regiones basada en color sobre la

nube de puntos permite distinguir tanto un objeto de un determinado color como aquel cuyo color es diferente a los de alrededor. Por otra parte, en cuanto a forma, se propone una detección basada en formas básicas que puede relacionarse con una amplia lista de objetos comunes en el ámbito humano. En este caso el enfoque se centra en cilindros, pero este bloque se ha hecho lo suficientemente global para adaptarse fácilmente a otras formas básicas como cuadrados o esferas.

- **Detección y estimación de posición de objetos eficaz para tareas de agarre y manipulación**: los experimentos llevados a cabo en los Capítulos 4 y 5 para diferentes objetos muestran que tanto los modelos como la posición estimada de los objetos son lo suficientemente precisos para realizar un agarre fiable. Se ha comprobado la idoneidad de las estrategias propuestas en diferentes plataformas robóticas acordes al entorno bajo evaluación. Además, se ha comprobado que el sistema de visión propuesto fusionado con otro táctil permite medidas precisas de los objetos agarrados para su manipulación.

- **Algoritmos de percepción visual fácilmente integrables**: se ha comprobado su integración con éxito tanto para fusión de información sensórica para agarres y manipulación robótica (Capítulo 4) como en sistemas globales de navegación y locomoción (Capítulo 5). En estos casos, la realización de los algoritmos bajo el framework Sistema Operativo Robótico, ROS, ha permitido integrar fácilmente la estrategia propuesta en diversas plataformas robóticas.

- **Métodos temporalmente eficientes que permiten la ejecución de tareas sin tiempos de espera innecesarios por procesamiento**: los algoritmos implementados en los contextos robóticos de los Capítulos 4 y 5 muestran cómo la estrategia de percepción propuesta permite interaccionar con los objetos en los tiempos previstos para la tarea a realizar, incluso cuando los objetos son desconocidos. Además, se ha logrado en el Capítulo 5 una reducción eficaz a la mínima cantidad de datos intercambiados para la detección automática en entornos implementados en la nube.

- **Técnica basada en razonamiento *a-contrario* como método de detección de un objeto en imágenes con ruido**: en este tipo de escenarios se comprobó que es posible percibir al objeto a detectar como una desviación de un modelo de aleatoriedad completa. Este enfoque solventa las dificultades que presentan las detecciones basadas en aplicar un umbral único de amplitud o técnicas clásicas de segmentación de las regiones coherentes. Este planteamiento fue validado exitosamente en un entorno de exteriores partiendo de los datos obtenidos de una cámara de Tiempo de Vuelo, ToF, (Capítulo 6). La generalización del método propuesto es sencilla ya que las aplicaciones de detección

a-contrario requieren la definición de un modelo de fondo así como la incorporación de una o varias medidas relacionadas con el objeto a detectar. Para extender el modelo sólo es necesario una fase de evaluación de la distribución del fondo, del objeto y un análisis del Número de Falsas Alarmas en regiones para encontrar el umbral óptimo.

Aunque se han logrado importantes contribuciones para lograr una percepción 3D de objetos robusta y eficaz en entornos poco estructurados, aún quedan retos por abarcar en el futuro, como se verá a continuación.

7.2. Trabajo Futuro

Las futuras líneas de investigación en este trabajo incluyen:

- **Robustez del método de reconstrucción 3D de objetos desconocidos a partir de una sola imagen de color y profundidad**: se identifican dos vías para aumentar la robustez de la técnica:
  - **Generalización del método de reconstrucción 3D**: para ello se explotará la combinación del método de extrusión con simetrías de rotación adoptando técnicas de estimación de formas como la descrita en [69].
  - **Mejora del refinamiento del modelo reconstruido**: se considera el estudio de curvas y superficies NURBS (B-splines racionales no uniformes) [156], determinadas por puntos de control, como método para refinamiento adaptativo de la malla triangular [157, 158]. Además, para aplicaciones de manipulación, una vez que el objeto se agarra, se considerarán diversas técnicas para el refinamiento del modelo obtenido. Por un lado, se estudiará la integración de técnicas incrementales de refinamiento considerando más vistas e información de la plataforma en sí [159, 160]. Por otro lado, centrándose en los avances de fusión de información de sensores para la manipulación de objetos desconocidos, el refinamiento del modelo podría desarrollarse fusionando información visual y táctil [105].
- **Segmentación de escenas desordenadas con muchos objetos juntos**: los algoritmos propuestos en 3D abordan enfoques en los que los objetos están ligeramente separados entre ellos, por lo que se pretende abordar en el futuro escenas complejas con objetos ocluidos y desordenados colocados indistintamente sobre una mesa o sobre otros objetos. Para ello se estudiarán técnicas de combinación de planos y NURBS junto con máquinas de vectores soporte (SVM) para la toma de decisiones [161]. Además, se evaluarán

filtros de partículas y descriptores de características de color de puntos por pares, CPPF, (*Color Point Pair Feature*) [18, 162] para mejorar la efectividad de la posición en estas situaciones.

- **Combinación de técnicas generales de extracción de formas**: se ha realizado de forma exitosa la extracción de cilindros a partir de nubes de puntos incompletas pero se busca abordar métodos más eficientes mediante parámetros descriptivos [163] y directamente extensibles a otras formas básicas como cubos, esferas y prismas [164].

- **Clasificación de objetos por categoría**: en esta tesis se ha planteado un algoritmo de reconocimiento de objetos individuales, pero se considera una vía de gran interés abordar en el futuro la generalización del reconocimiento al nivel de clasificación por categoría de una forma escalable. Se plantea el estudio de diversos métodos como la categorización basada en grafos con aprendizaje de formas no supervisado [165] así como basados en redes neuronales convolucionales pre-entrenadas para la categorización de imágenes para la obtención de características semánticas [166].

- **Mayor integración en un sistema multi-modal de información visual y táctil para tareas de agarre y manipulación**: este factor es clave en las situaciones en las que el objeto está prácticamente ocluido por la mano robótica por lo que el uso exclusivo de información visual no es suficiente para interaccionar con el objeto. En este caso, es fundamental fusionar la información disponible de la plataforma. El algoritmo de fusión evaluado experimentalmente en el Capítulo 4 presenta ciertas limitaciones relacionadas con objetos simétricos ya que las localizaciones de contacto pueden coincidir con diferentes partes del modelo proporcionando varias soluciones posibles siendo sólo una válida. El trabajo futuro en el algoritmo de fusión se centra en solventar estas limitaciones buscando otras fuentes de información. Se plantea el estudio de normales de contacto así como añadir una segunda cámara con otro punto de vista para poder estimar mejor la posición del objeto a pesar de las oclusiones. Además, con el fin de validar este método muti-modal de forma más exhaustiva, se llevarán a cabo experimentos para obtener resultados estadísticos de la tasa de éxito en tareas de agarre y manipulación con planificadores online.

- **Mayor investigación en un sistema multi-modal locomoción, agarre y navegación**: a pesar de la exitosa integración vista en el Capítulo 5, aún quedan desafíos abiertos en este ámbito por abordar. De hecho, en la competición final del Darpa Robotics Challenge, celebrada en Junio 2015, se comprobó que una tarea aparentemente sencilla como era abrir una puerta, fue imposible

para la mayor parte de robots porque se requería una mejor integración de percepción y control [167], entre otras cosas, para manipular el pomo y evitar chocar con el marco de la puerta al entrar.

- **Robustez de la técnica *a-contrario* para detección en imágenes con ruido**: en este caso se han considerado dos vías de estudio futuro:

  - **Generalización de la técnica**: la técnica de detección *a-contrario* propuesta para esta aplicación ha dado resultados satisfactorios pero debido a que se han realizado sólo una serie de experimentos preliminares, es necesario un estudio más exhaustivo como trabajo futuro. Para dicho estudio será necesario evaluar objetos de diferentes formas y tamaños con el objetivo de generalizar este sistema de percepción propuesto a cualquier aplicación de exteriores con cámaras ToF. Por otra parte, para una generalización completa del método automático, se deberían llevar más estudios a cabo con otros conjuntos de datos para establecer la tasa de falsas alarmas.
  - **Incremento de la adaptabilidad**: se incorporarán técnicas de ventanas adaptativas de tamaño variable para aportar mayor robustez y flexibilidad en tiempo real al método *a-contrario* propuesto con imágenes de magnitud de gradiente 2x2 de la imagen de profundidad.

Todas estas líneas futuras se enmarcan en el objetivo de lograr el desarrollo de estrategias de percepción 3D de objetos para poder interaccionar con ellos de forma autónoma en cualquier escenario y en los tiempos de ejecución requeridos.

7.3. Publicaciones Relevantes

Las publicaciones y trabajos más relevantes relacionadas con esta tesis se enumeran a continuación:

1. S. Rodríguez-Jiménez, N. Burrus y M. Abderrahim. "Xtru3D: Single-View 3D Object Reconstruction from Color and Depth". Computer Vision, Imaging and Computer Graphics – Theory and Applications Journal, pp 163-178, 2014.

2. S. Rodríguez-Jiménez y M. Abderrahim. "3-Dimensional Object Perception for Manipulation Tasks Using the Atlas Robot". ROBOT2013: First Iberian Robotics Conference. Advances in Intelligent Systems and Computing, Springer, vol. 253, pp. 359-368, 2014.

3. E. García, M. Ocaña, L. Bergasa, M. Ferre, M. Abderrahim, J. C. Arévalo, D. Sanz-Merodio, E. Molinos, N. Hernández, A. Llamazares, F. Suárez y S. Rodríguez-Jiménez. "Competing in the DARPA Virtual Robotics Challenge as the SARBOT Team". ROBOT2013: First Iberian Robotics Conference. Advances in Intelligent Systems and Computing, Springer, vol. 253, pp. 381-396, 2014.

4. E. García, J. C. Arévalo, D. Sanz-Merodio, L. Bergasa, M. Ocaña, E. Molinos, N. Hernández, A. Llamazares, M. Abderrahim, S. Rodríguez-Jiménez, M. Ferre y F. Suárez. "PROYECTO SARBOT: Introducción de robots humanoides en tareas de búsqueda y rescate en entornos urbanos degradados". Congreso de I+D en Defensa y Seguridad (DESEi+d 2013), Madrid, Noviembre, 2013.

5. J. Gonzalez-Quijano, M. Abderrahim, C. Bensalah y S. Rodríguez-Jiménez. "RoMPLA: An efficient robot motion and planning learning architecture". IEEE International Conference on Intelligent Robots and Systems (IROS). Tokyo, Japón, Noviembre, 2013.

6. J. Bimbo, S. Rodríguez-Jiménez, H. Liu, N. Burrus, L. Senerivatne, M. Abderrahim y K. Althoefer. "Fusing Visual and Tactile Sensing for Manipulation of Unknown Objects". Workshop on Interactive Perception - ICRA 2013 Mobile Manipulation. Karlsruhe, Alemania, Mayo, 2013.

7. S. Rodríguez-Jiménez. "3D Object Perception Using Depth Cameras for Indoor and Outdoor Applications". 2nd Workshop Robotics Lab Spring 2013, Leganés, España, Abril, 2013.

8. S. Rodríguez-Jiménez, N. Burrus and M. Abderrahim. "3D Object Reconstruction with a Single RGB-Depth Image". International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 155–163, Barcelona, España, Febrero, 2013.

9. S. Rodríguez-Jiménez. "Visual Perception System within HANDLE EU project using ROS". IEEE International Conference on Intelligent Robots and Systems (IROS) Tutorial: Handling ROS - Introductory tutorial to ROS and its use for robot in-hand manipulation. Vila Moura, Portugal, Octubre, 2012.

10. S. Rodríguez-Jiménez, N. Burrus and M. Abderrahim. "A-Contrario Detection of Aerial Target Using a Time-of-Flight Camera". Proceedings of Sensor Signal Processing for Defence Conference (SSPD) 2012, pp. 1–5, Londres, Reino Unido, Septiembre, 2012.

11. J. Bimbo, S. Rodríguez-Jiménez, H. Liu, X. Song, N. Burrus, L. Senerivatne, M. Abderrahim and K. Althoefer. “Object Pose Estimation and Tracking by Fusing Visual and Tactile Information”. Proceedings of the 2012 IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). pp. 65–70. Hamburg, Alemania, Septiembre, 2012.

12. S. Rodríguez-Jiménez, N. Burrus, J. Muñoz, J. González-Quijano, C. Bensalah, A. Al-kaff and M. Abderrahim. ”3D Visual Perception System for Grasping with an Anthropomorphic Hand”. 5th International Conference on Cognitive Systems, COGSYS (POSTER). Viena, Febrero, 2012.

13. J. González-Quijano, M. Abderrahim, C. Bensalah, S. Rodríguez-Jiménez, N. Burrus, A. Al-Kaff and A. Villoslada. “RoMPLA: Robot Motion Planning and Learning Architecture”. 5th International Conference on Cognitive Systems, COGSYS (POSTER). Viena, Febrero, 2012.

14. S. Rodríguez-Jiménez, N. Burrus y M. Abderrahim. ”Vision Algorithms Integration with a Robotic System Control”. HANDLE Workshop Benicassim 2012 (POSTER). Benicassim, España, Febrero, 2012.

Bibliografía

[1] A. M. Research, "Global 3d camera market (type, technology, application and geography) - size, share, global trends, company profiles, demand, insights, analysis, research, report, opportunities, segmentation and forecast 2013 - 2020," tech. rep., 2015.

[2] L. Shao, J. Han, P. Kohli, and Z. Zhang, *Computer Vision and Machine Learning with RGB-D Sensors*. Springer Publishing Company, Incorporated, 2014.

[3] T. Kurke, "Lazeeeye: 3d capture device phone add-on." `http://3dsolver.com/lazeeeye-3d-capture-device-phone-add-on/`, 2014. [Online].

[4] Occipital, "Kickstarter: Structure sensor - capture the world in 3d." `https://www.kickstarter.com/projects/occipital/structure-sensor-capture-the-world-in-3d?lang=es`, 2013. [Online].

[5] Google, "Project tango." `https://www.google.com/atap/projecttango/`, 2014. [Online].

[6] B. X. Chen, "For hints at apple plans, read its shopping list. the new york times." `http://nyti.ms/1pgvPRT`, 2014. [Online].

[7] L. G. Roberts, *Machine Perception of Three-Dimensional Solids*. Optical and Electrooptical Information Processing, Massachusetts Institute of Technology Press, 1965.

[8] A. Guzman, "Analysis of curved line drawings using context and global information," in *Machine Intelligence (6)* (B. Meltzer and D. Mitchie, eds.), pp. 325–376, Edinburgh University Press, 1971.

[9] T. O. Binford, "Visual perception by computer," in *Proceedings of the IEEE Conference on Systems and Control*, 1971.

[10] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-d objects from appearance," *International Journal of Computer Vision*, vol. 14, no. 1, pp. 5–24, 1995.

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision*, vol. 2, pp. 1150–1157, 1999.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding (CVIU)*, vol. 110, pp. 346–359, 2008.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[15] W. Wang, L. Chen, Z. Liu, K. Kahnlenz, and D. Burschka, "Textured/textureless object recognition and pose estimation using rgb-d image," *Journal of Real-Time Image Processing*, pp. 1–16, 2013.

[16] R. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3d recognition and pose using the viewpoint feature histogram," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2155–2162, IEEE, 2010.

[17] R. Rusu and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[18] C. Choi, *Visual Object Perception in Unstructured Environments*. PhD thesis, Georgia Institute of Technology, 2014.

[19] B. Curless, "From range scans to 3d models," *SIGGRAPH Computer Graphics*, vol. 33, pp. 38–41, November 1999.

[20] M. Abdul-Rani, K. K. Chong, A. F. M. Hani, Y. B. Yap, and A. Jamil, "Analytical studies on volume determination of leg ulcer using structured light and laser triangulation data acquisition techniques," vol. 5, no. 10, pp. 785–790, 2011.

[21] R. Schwarte, "Principles of 3d imaging techniques," *Handbook of Computer Vision and Applications*, 1999.

[22] S. Hovanessian, *Introduction to Sensor Systemss*. Artech House Communication and Electronic Defense Library, 1988.

[23] G. Sansoni, M. Trebeschi, and F. Docchio, "State-of-the-art and applications of 3d imaging sensors in industry, cultural heritage, medicine, and criminal investigation," *Sensors*, vol. 9, no. 1, pp. 568–601, 2009.

[24] R. Lange, *3D Time-of-Flight Distance Measurement with Custom Solid-State Image Sensors in CMOS/CCD-Technology*. PhD thesis, University of Siegen, 2000.

[25] M. Frade, J. M. Enguita, I. Álvarez, and S. Rodríguez-Jiménez, "Fringe pattern characterization by opd analysis in a lateral shearing interferometric profilometer," in *Proceedings of SPIE*, vol. 8082, 2011.

[26] Infaimon, "Ensenso n20 800." `http://www.infaimon.com/es/ensenso-n20-800`, 2015. [Online].

[27] G. Bradsky and A. Kaehler, *Learning OpenCV*. O'Reilly, 2008.

[28] C. Robotics, "MultiSense-SL Datasheet." `http://www.theroboticschallenge.org/local/documents/MultiSense_SL.pdf/`, 2013. [Online].

[29] A. Jongenelen, *Development of a Compact, Configurable, Real-Time Range Imaging System*. PhD thesis, Victoria University of Wellington, 2011.

[30] A. Kolb, E. Barth, R. Koch, and R. Larsen, "Time-of-Flight Cameras in Computer Graphics," *Computer Graphics Forum*, vol. 29, no. 1, pp. 141–159, 2010.

[31] A. Jongenelen, *Development of a Compact, Configurable, Real-Time Range Imaging System*. PhD thesis, Victoria University of Wellington, May 2011.

[32] G. J. Iddan and G. Yahav, "3D Imaging in the Studio (and elsewhere ...)," *Proc. of SPIE*, vol. 4298, pp. 48–55, 2001.

[33] Canesta, "Canesta 101: Introduction to 3D Vision in CMOS," tech. rep., 2008.

[34] A. S. Concepts, "3d flash lidar." `http://www.advancedscientificconcepts.com/`, 2015. [Online].

[35] PMDtec, "PMD[vision] CamCube 2.0." `http://www.pmdtec.com`, 2010. [Online].

[36] Mesa Imaging, "Swiss Ranger SR4000 Overview." http://www.mesa-imaging.ch, 2010. [Online].

[37] IEE, "3D MLI Sensor." http://www.iee.lu, 2010. [Online].

[38] Softkinetic-Optrima, "OptriCam 3D time-of-flight imager using DepthSense CMOS sensor." http://www.softkinetic-optrima.com, 2010. [Online].

[39] S. B. Gokturk, H. Yalcin, and C. Bamji, "A Time-Of-Flight Depth Sensor - System Description, Issues and Solutions." http://canesta.com, 2010. [Online].

[40] S. Hussmann, T. Ringbeck, and B. Hagebeuker, "A performance review of 3D ToF vision systems in comparison to stereo vision systems," *Stereo Vision (Online book publication)*, pp. 103–120, 2008.

[41] T. Ringbeck, "A 3d time of flight camera for object detection," *Measurement*, vol. 9, no. 2, pp. 1–16, 2007.

[42] T. Ringbeck, T. Möller, and B. Hagebeuker, "Multidimensional measurement by using 3-d pmd sensors," *Advances in Radio Science*, vol. 5, pp. 135–146, 2007.

[43] D. Piatti and F. Rinaudo, "Sr-4000 and camcube3.0 time of flight (tof) cameras: Tests and comparison," *Remote Sensing*, vol. 4, no. 4, pp. 1069–1089, 2012.

[44] P. Technologies, "PMD[vision] CamCube 3.0." http://www.pmdtec.com, 2011. [Online].

[45] S. B. Gokturk, H. Yalcin, and C. Bamji, "A Time-Of-Flight Depth Sensor, System Description, Issues and Solutions," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, pp. 35–45, 2004.

[46] M. Hebert and E. Krotkov, "3-d measurements from imaging laser radars: How good are they?," *International Journal of Image and Vision Computing*, vol. 10, pp. 170–178, 1992.

[47] W. Rohsenow and H. Choi, *Heat Mass and Momentum Transfer*. Prentice Hall, 1961.

[48] D. Anderson, H. Herman, and A. Kelly, "Experimental characterization of commercial flash ladar devices," in *Proceedings of the International Conference of Sensing and Technology*, 2005.

[49] D. Piatti, *Time-of-Flight cameras: tests, calibration and multi-frame registration for automatic 3D object reconstruction*. PhD thesis, Politecnico Di Torino, 2010.

[50] L. Zhang, B. Curless, and S. M. Seitz, "Rapid shape acquisition using color structured light and multi-pass dynamic programming," in *In The 1st IEEE International Symposium on 3D Data Processing, Visualization, and Transmission*, pp. 24–36, 2002.

[51] Microsoft, "Kinect for Xbox 360." `http://www.xbox.com/en-US/kinect/`, 2010. [Online].

[52] ASUS, "Wavi-xtion camera." `http://event.asus.com/wavi/product/WAVI_Xtion.aspx`, 2015. [Online].

[53] PrimeSense, "The PrimeSensor Technology." `http://www.primesense.com`, 2010. [Online].

[54] R. El-laithy, J. Huang, and M. Yeh, "Study on the use of microsoft kinect for robotics applications," in *Position Location and Navigation Symposium (PLANS), 2012 IEEE/ION*, pp. 1280–1288, 2012.

[55] Microsoft, "The Natal Project: Introducing Controller-Free Gaming and Entertainment." `http://www.xbox.com/en-US/live/projectnatal/`, 2010. [Online].

[56] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[57] J. Garcia, *Reconstruction and Recognition of Confusable Models using Three-Dimensional Perception*. PhD thesis, Universidad Carlos III de Madrid, 2014.

[58] D. Silver, *Learning Preference Models for Autonomous Mobile Robots in Complex Domains*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, December 2010.

[59] D. Kragic and M. Björkman, "Strategies for object manipulation using foveal and peripheral vision," in *International Conference on Computer Vision Systems (ICVS), New York, USA*, pp. 50–60, 2006.

[60] D. Katz, J. Kenney, and O. Brock, "How can robots succed in unstructured environments," in *Workshop on Robot Manipulation: Intelligence in Human Environments at Robotics*, Science and Systems, June 2008.

[61] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Seven guiding scenarios for information visualization evaluation," tech. rep., Department of Computer Science, University of Calgary, 2011.

[62] M. A. Neerincx, A. Bos, A. Olmedo-Soler, U. Brauer, L. Breebaart, N. Smets, J. Lindenberg, T. Grant, and M. Wolff, "The mission execution crew assistant: Improving human-machine team resilience for long duration missions," in *Proceedings of the 59th International Astronautical Congress, IAC*, 2008.

[63] J. Kuehnle, Z. Xue, M. Stotz, J. Zoellner, A. Verl, and R. Dillmann, "Grasping in depth maps of time-of-flight cameras," in *International Workshop on Robotic and Sensors Environments (ROSE)*, pp. 132–137, 2008.

[64] A. Miller and P. Allen, "Graspit! a versatile simulator for robotic grasping," *Robotics Automation Magazine, IEEE*, vol. 11, pp. 110–122, December 2004.

[65] M. Sun, S. S. Kumar, G. Bradski, and S. Savarese, "Toward automatic 3d generic object modeling from one single image," in *3DIMPVT*, (Hangzhou, China), May 2011.

[66] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, and L. Van Gool, "Depth-from-recognition: Inferring meta-data by cognitive feedback," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

[67] S. Thrun and B. Wegbreit, "Shape from symmetry," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, pp. 1824–1831, IEEE, 2005.

[68] J. Bohg, M. Johnson-Roberson, B. León, J. Felip, X. Gratal, N. Bergstrom, D. Kragic, and A. Morales, "Mind the gap-robotic grasping under incomplete observation," in *2011 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 686–693, IEEE, 2011.

[69] Z. Marton, D. Pangercic, N. Blodow, J. Kleinehellefort, and M. Beetz, "General 3d modelling of novel objects from a single view," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3700–3705, IEEE, 2010.

[70] P. Benko, R. R. Martin, and T. Varady, "Algorithms for reverse engineering boundary representation models," *Computer-Aided Design*, vol. 33, pp. 839–851, 2001.

[71] G. Bradski, "The opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.

[72] J. Y. Bouguet, "Matlab calibration tool." `http://www.vision.caltech.edu/bouguetj/calib_doc/`, 2013. [Online].

[73] M. Fischler and R. 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.

[74] M. Y. Yang and W. Förstner, "Plane detection in point cloud data," in *Department of Photogrammetry, University of Bonn*, 2010.

[75] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Close-range scene segmentation and reconstruction of 3d point cloud maps for mobile manipulation in human environments," in *The 22nd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (St. Louis, MO, USA), October 2009.

[76] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *International Conference on Robotics and Automation*, (Shanghai, China), 2011.

[77] L. A. Alexandre, "3d descriptors for object and category recognition: a comparative evaluation," in *in Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[78] N. Burrus, M. Abderrahim, J. Garcia, and L. Moreno, "Object reconstruction and recognition leveraging an rgb-d camera," in *Machine Vision Applications (MVA), the 12th IAPR Conference on*, June 2011.

[79] R. Pérula, "Estado del arte e implementación de un clasificador de objetos de uso cotidiano," Master's thesis, Universidad Carlos III de Madrid, March 2013.

[80] L. Martínez, P. Loncomilla, and J. Ruiz-del Solar, "Object recognition for manipulation tasks in real domestic settings: A comparative study," in *RoboCup 2014: Robot World Cup XVIII*, vol. 8992 of *Lecture Notes in Computer Science*, pp. 207–219, Springer International Publishing, 2015.

[81] W. E. L. Grimson and T. Lozano-Perez, "Localizing overlapping parts by searching the interpretation tree," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, pp. 469–482, 1987.

[82] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols," *CoRR*, vol. abs/1502.03143, 2015.

[83] W. Garage, "Household objects database." `http://www.ros.org/wiki/household_objects//`, 2013. [Online].

[84] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

[85] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Symposium on Geometry Processing*, pp. 61–70, 2006.

[86] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary amp; region segmentation of objects in n-d images," in *Proceedings of Eighth IEEE International Conference on Computer Vision (ICCV)*, vol. 1, pp. 105–112, 2001.

[87] H. Lombaert, Y. Sun, L. Grady, and C. Xu, "A multilevel banded graph cuts method for fast image segmentation," in *In IEEE International Conference on Computer Vision (ICCV)*, pp. 259–265, 2005.

[88] C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut": interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, pp. 309–314, 2004.

[89] K. Vaiapury, A. Aksay, and E. Izquierdo, "Grabcutd: improved grabcut using depth information," in *Proceedings of the 2010 ACM workshop on Surreal media and virtual cloning*, SMVC '10, (New York, NY, USA), pp. 57–62, ACM, 2010.

[90] W. Chiu, U. Blanke, and M. Fritz, "Improving the kinect by cross-modal stereo.," in *22nd British Machine Vision Conference (BMVC)*, (Dundee, UK), 2011.

[91] K.-J. Oh, S. Yea, and Y.-S. Ho, "Hole filling method using depth based inpainting for view synthesis in free viewpoint television and 3-d video," in *Picture Coding Symposium, 2009. PCS 2009*, pp. 1–4, May 2009.

[92] A. Telea, "An Image Inpainting Technique Based on the Fast Marching Method," *Journal of graphics, gpu, and game tools*, vol. 9, no. 1, pp. 23–34, 2004.

[93] A. E. Ichim, *RGB-D Handheld Mapping and Modeling*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2013.

[94] J. A. Bagnell, F. Cavalcanti, L. Cui, T. Galluzzo, M. Hebert, M. Kazemi, M. Klingensmith, J. Libby, T. Y. Liu, N. S. Pollard, M. Pivtoraiko, J.-S. Valois, and R. Zhu, "An integrated system for autonomous robotics manipulation," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 2955–2962, 2012.

[95] S. Rodríguez-Jiménez, N. Burrus, and M. Abderrahim, "3d object reconstruction with a single rgb-depth image," in *2013 International Conference on Computer Vision Theory and Applications, VISAPP*, vol. 2, pp. 155–163, 2013.

[96] S. R. Company, "Shadow Dexterous Hand." `http://www.shadowrobot.com/products/dexterous-hand/`, 2014. [Online].

[97] A. Company, "ATI Nano17 sensors." `http://www.ati-ia.com//`, 2014. [Online].

[98] H. Liu, X. Song, J. Bimbo, L. Seneviratne, and K. Althoefer, "Surface material recognition through haptic exploration using an intelligent contact sensing finger," in *Proceedings of the 2012 IEEE International Conference on Intelligent Robots and Systems, IROS*, pp. 52–57, 2012.

[99] L. Mitsubishi Heavy Industries, "PA-10 Instruction Manual for Instalation, Maintenance and Safety," 2010.

[100] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, 2009.

[101] MeshLab, "Visual Computing Lab-ISTI-CNR.." `http://meshlab.sourceforge.net/`, 2011. [Online].

[102] J. Muñoz, "3d perception integration with robotic manipulation," Master's thesis, Universidad Carlos III de Madrid, October 2011.

[103] R. Diankov, *Automated Construction of Robotic Manipulation Programs*. PhD thesis, Carnegie Mellon University, Robotics Institute, August 2010.

[104] J. J. K. Jr. and S. M. LaValle, "Rrt-connect: An efficient approach to single-query path planning," in *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA*, pp. 995–1001, 2000.

[105] J. Bimbo, S. Rodríguez-Jiménez, H. Liu, N. Burrus, I. Seneviratne, M. Abderrahim, and K. Althoefer, "Fusing visual and tactile sensing for manipulation of unknown objects," in *Proceedings of the ICRA 2013 Mobile Manipulation Workshop on Interactive Perception*, 2013.

[106] J. Bimbo, S. Rodríguez-Jiménez, H. Liu, S. Xiaojing, N. Burrus, I. Seneviratne, M. Abderrahim, and K. Althoefer, "Object pose estimation and tracking by fusing visual and tactile information," in *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2012.

[107] J. Bimbo, H. Liu, I. Seneviratne, , and K. Althoefer, "Combining touch and vision for the estimation of an object's pose during manipulation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013.

[108] S. Shum, W. Lau, M. Yuen, and K. Yu, "Solid reconstruction from orthographic views using 2-stage extrusion," *Computer-Aided Design*, vol. 33, no. 1, pp. 91 – 102, 2001.

[109] O. Kroemer, H. Ben Amor, M. Ewerton, and J. Peters, "Point cloud completion using symmetries and extrusions," in *Proceedings of the International Conference on Humanoid Robots (HUMANOIDS)*, 2012.

[110] S. Srinivasa, D. Ferguson, C. Helfrich, D. Berenson, A. C. Romea, R. Diankov, G. Gallagher, G. Hollinger, J. Kuffner, and J. M. Vandeweghe, "Herb: a home exploring robotic butler," *Autonomous Robots*, vol. 28, pp. 5–20, January 2010.

[111] S. Srinivasa, D. Ferguson, J. M. Vandeweghe, R. Diankov, D. Berenson, C. Helfrich, and K. Strasdat, "The robotic busboy: Steps towards developing a mobile robotic home assistant," in *International Conference on Intelligent Autonomous Systems*, July 2008.

[112] M. Kojima, K. Okada, and M. Inaba, "Manipulation and recognition of objects incorporating joints by a humanoid robot for daily assistive tasks," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 1564–1569, 2008.

[113] J. Stückler, R. Steffens, D. Holz, and S. Behnke, "Efficient 3d object perception and grasp planning for mobile manipulation in domestic environments," *Robotics and Autonomous Systems*, 2012.

[114] T. D. Mechanical System Group and A. Department, "2014 white paper on robotization of industry, business and our life," tech. rep., New Energy and Industrial Technology Development Organization, NEDO, 2014.

[115] D. Holz, R. Schnabel, D. Droeschel, J. Stückler, and S. Behnke, "Robocup 2010," ch. Towards semantic scene analysis with time-of-flight cameras, pp. 121–132, Berlin, Heidelberg: Springer-Verlag, 2011.

[116] M. Nieuwenhuisen, J. Stückler, A. Berner, R. Klein, and S. Behnke, "Shape-primitive based object recognition and grasping," in *Proceedings of ROBOTIK*, VDE-Verlag, 2012.

[117] K.-H. Lin, C.-H. Chang, A. Dopfer, and C.-C. Wang, "Mapping and localization in 3d environments using a 2d laser scanner and a stereo camera," *Journal of Information Science and Engineering*, vol. 28, pp. 131–144, 2012.

[118] B. Schafer, M. Proetzsch, and K. Berns, "Extension approach for the behaviour-based control system of the outdoor robot ravon," pp. 149–155, 2006.

[119] D. Klimentjew, N. Hendrich, and J. Zhang, "Multi sensor fusion of camera and 3d laser range finder for object recognition," in *2010 IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 236–241, 2010.

[120] C. Beder and W. Förstner, "Direct solutions for computing cylinders from minimal sets of 3d points," in *European Conference on Computer Vision (ECCV)* (A. Leonardis, H. Bischof, and A. Pinz, eds.), vol. 3951 of *Lecture Notes in Computer Science*, pp. 135–146, Springer Berlin Heidelberg, 2006.

[121] H. Durrant-Whyte and T. Bailey, "Simultaneous localisation and mapping (slam): Part i the essential algorithms," *IEEE Robotics and Automation Magazine*, vol. 2, pp. 99–108, 2006.

[122] A. Hornung, K. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: an efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[123] M. Montemerlo and S. Thrun, "Large-Scale Robotic 3-D Mapping of Urban Structures," in *Experimental Robotics IX* (M. H. Ang and O. Khatib, eds.), vol. 21 of *Springer Tracts in Advanced Robotics*, ch. 14, pp. 141–150, Springer-Verlag, 2006.

[124] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," *Computer Graphics Forum*, vol. 26, no. 2, pp. 214–226, 2007.

[125] Q. Zhan, Y. Liang, and Y. Xiao, "Color-based segmentation of point clouds," in *ISPRS Laser Scanning Workshop*, vol. XXXVII, pp. 248–252, 2009.

[126] DARPA, "Drc.." `http://www.darpa.mil/Our_Work/TTO/Programs/DARPA_Robotics_Challenge.aspxl/`, 2012. [Online].

[127] DARPA, "VRC Technical Guide." `http://www.theroboticschallenge.org/`, 2013. [Online].

[128] B. G. C. Aguero and E. Krotkov, "Technical guide, virtual robotics challenge, case 21251, release 2," tech. rep., DISTAR, 2013.

[129] F. Suárez-Ruiz, *Human-Robot Interaction for Telemanipulation in Large Workspaces*. PhD thesis, Universidad Politécnica de Madrid, December 2014.

[130] S. Rodríguez-Jiménez and M. Abderrahim, "3-dimensional object perception for manipulation task using the atlas robot," in *ROBOT2013: First Iberian Robotics Conference*, vol. 253, pp. 359–368, Springer, 2014.

[131] E. García, M. Ocaña, L. M. Bergasa, M. Ferre, M. Abderrahim, J. Arevalo, D. Sanz-Merodio, E. Molinos, N. Hernandez, A. Llamazares, F. Suárez-Ruiz, and S. Rodríguez-Jiménez, "Competing in the darpa virtual robotics challenge as the sarbot team," in *ROBOT2013: First Iberian Robotics Conference*, vol. 253, pp. 381–396, Springer, 2014.

[132] E. Molinos, Á. Llamazares, N. Hernández, R. Arroyo, A. Cela, J. J. Yebes, M. Ocaña, and L. M. Bergasa, "Perception and navigation in unknown environments: The darpa robotics challenge," in *ROBOT2013: First Iberian Robotics Conference* (M. A. Armada, A. Sanfeliu, and M. Ferre, eds.), vol. 253 of *Advances in Intelligent Systems and Computing*, pp. 321–329, Springer International Publishing, 2014.

[133] J. Arevalo, D. Sanz-Merodio, and E. Garcia, "Reactive humanoid walking algorithm for occluded terrain," in *ROBOT2013: First Iberian Robotics Conference* (M. A. Armada, A. Sanfeliu, and M. Ferre, eds.), vol. 253 of *Advances in Intelligent Systems and Computing*, pp. 397–410, Springer International Publishing, 2014.

[134] F. Suárez-Ruiz, A. Owen-Hill, and M. Ferre, "Internet-Based Supervisory Teleoperation of a Virtual Humanoid Robot," in *ROBOT2013: First Iberian Robotics Conference*, vol. 253 of *Advances in Intelligent Systems and Computing*, pp. 345–358, Springer, 2014.

[135] H. Chen and B. Bhanu, "3d free-form object recognition in range images using local surface patches," *Pattern Recogn. Lett.*, vol. 28, no. 10, pp. 1252–1262, 2007.

[136] M. Ammar and S. Le Hégarat-Mascle, "An a-contrario approach for object detection in video sequence.," *International Journal of Pure and Applied Mathematics*, vol. 89, no. 2, pp. 173–201, 2013.

[137] S. Hyojoo, K. Changwan, and C. Kwangnam, "Rapid 3d object detection and modeling using range data from 3d range imaging camera for heavy equipment operation," *Automation in Construction*, vol. 19, pp. 898–906, 2010.

[138] B. Langmann, K. Hartmann, and O. Loffeld, "Depth camera technology comparison and performance evaluation," in *ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pp. 438–444, 2012.

[139] N. Burrus, B. Thierry, and J.-M. Jolion, "Image segmentation by a contrario simulation," *Pattern Recognition*, vol. 42, no. 7, pp. 1520–1532, 2009.

[140] M. Mottalli, M. Tepper, and M. Mejail, "A contrario detection of false matches in iris recognition," in *Proceedings of the 15th Iberoamerican congress conference on Progress in pattern recognition, image analysis, computer vision, and applications*, CIARP'10, (Berlin, Heidelberg), pp. 442–449, Springer-Verlag, 2010.

[141] S. Le Hegarat Mascle, A. Robin, and R. Reynaud, "Simultaneous localization and object detection using an a-contrario approach," in *Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 440–447, 2010.

[142] A. Robin, L. Moisan, and S. Le Hégarat-Mascle, "An a-contrario approach for subpixel change detection in satellite imagery.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1977–1993, 2010.

[143] B. Grosjean and L. Moisan, "A-contrario detectability of spots in textured backgrounds," *J. Math. Imaging Vis.*, vol. 33, no. 3, pp. 313–337, 2009.

[144] A. Desolneux, L. Moisan, and J.-M. Morel, "A grouping principle and four applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 4, pp. 508–513, 2003.

[145] N. Burrus and T. M. Bernard, "Adaptive Vision Leveraging Digital Retinas: Extracting Meaningful Segments," *Advanced Concepts for Intelligent Vision Systems, 8th International Conference*, pp. 220–231, 2006.

[146] S. Khanafseh, *GPS Navigation Algorithms for Autonomous Airborne Refueling of Unmanned Air Vehicles*. PhD thesis, Illinois Institute of Technology, 2008.

[147] A. S. Concepts, "3D Flash LIDAR." `http://www.advancedscientificconcepts.com/`, 2013. [Online].

[148] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[149] M. Schulze, "3d-camera based navigation of a mobile robot in an agricultural environment," in *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXVIII, Part 5, pp. 538–542, 2010.

[150] WebRigEADS, "Rig building at EADS CASA in Getafe (Spain)." `http://webs.ono.com/iconnect/Imagenes/Secprofesional/Titulos/EdificioRig.jpg`, 2012. [Online].

[151] P. Felzenszwalb and D. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[152] A. Robin, G. Mercier, G. Moser, and S. Serpico, "An a-contrario approach for unsupervised change detection in radar images," in *Geoscience and Remote Sensing Symposium,2009 IEEE International,IGARSS*, vol. 4, pp. 240–243, 2009.

[153] T. Veit, F. Cao, and P. Bouthemy, "An a contrario Decision Framework for Region-Based Motion Detection," *International Journal of Computer Vision*, vol. 68, no. 2, pp. 163–178, 2006.

[154] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[155] A. Slaby, "Roc analysis with matlab," in *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*, pp. 191–196, 2007.

[156] L. Piegl and W. Tiller, *The NURBS book*. 2nd edition, Springer-Verlag, 1996.

[157] V. Estellers, M. Scott, K. Tew, and S. Soatto, "Robust poisson surface reconstruction," in *Scale-Space Variational Methods (SSVM)*, May 2015.

[158] K. Yin, H. Huang, H. Zhang, M. Gong, D. Cohen-Or, and B. Chen, "Morfit: Interactive surface reconstruction from incomplete point clouds with curve-driven topology and geometry control," *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia*, vol. 33, no. 6, pp. 202:1–202:12, 2014.

[159] M. Krainin, P. Henry, X. Ren, and D. Fox, "Manipulator and object tracking for in hand model acquisition," in *Proc. of the Workshop on Best Practice in 3D Perception and Modeling for Mobile Manipulation at the Int. Conf. on Robotics Automation (ICRA)*, 2010.

[160] M. Krainin, B. Curless, and D. Fox, "Autonomous generation of complete 3d object models using next best view manipulation planning," in *IEEE International conference on Robotics and Automation (ICRA)*, 2011.

[161] A. Richtsfeld, T. Morwald, J. Prankl, M. Zillich, and M. Vincze, "Segmentation of unknown objects in indoor environments," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 4791–4796, IEEE, 2012.

[162] C. Choi and H. I. Christensen, "3d pose estimation of daily objects using an rgb-d camera," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 3342–3349, IEEE, 2012.

[163] T. Tran, V. Cao, and D. Laurendeau, "Extraction of cylinders and estimation of their parameters from point clouds," *Computers & Graphics*, vol. 46, pp. 345–357, 2015.

[164] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," *Computer Graphics Forum*, vol. 26, no. 2, pp. 214–226, 2007.

[165] C. Mueller, K. Pathak, and A. Birk, "Object recognition in rgbd images of cluttered environments using graph-based categorization with unsupervised learning of shape parts," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 2248–2255, 2013.

[166] H. S. Max Schwarz and S. Behnke, "Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[167] E. Ackerman and E. Guizzo, "Darpa robotics challenge: Amazing moments, lessons learned, and what is next." `http://spectrum.ieee.org/robotics/humanoids`, 2015. [Online].