Escuela Superior Politécnica del Litoral

Facultad de Ingeniería en Electricidad y Computación

Generación de nubes de puntos a través de un sistema de estereovisión para aplicaciones industriales y de investigación

TECH-378

Proyecto Integrador

Previo la obtención del Título de:

Ingeniero/a en Ciencias de la Computación

Presentado por:

Kevin Elihan Muñoz Calva

Loberlly Noelia Salazar Aspiazu

Guayaquil - Ecuador

Año: 2024

Dedicatoria

Dedico este proyecto a mis padres y a mi abuela, quienes siempre han estado a mi lado, brindándome su apoyo incondicional a lo largo de este extenso camino académico. Su confianza en mí y en mis decisiones ha sido fundamental para poder avanzar, respaldándome en cada paso que he dado. Finalmente, quiero reconocerme a mí mismo, ya que creo profundamente que he tomado decisiones acertadas las mismas que me han formado en la persona que soy.

Kevin Elihan Muñoz Calva

Dedicatoria

Les dedico este proyecto con todo mi amor y eterna gratitud a mi familia, en especial a mi madre, mi hermana y mi abuela, quienes han sido el pilar fundamental en mi vida. Gracias por su amor incondicional, por cada palabra de aliento y por estar a mi lado en cada paso de este camino. Su apoyo constante durante toda mi vida me dio la fuerza para seguir adelante y superar cada obstáculo. Este logro no habría sido posible sin ustedes.

Loberlly Noelia Salazar Aspiazu

Agradecimientos

Nos gustaría expresar nuestro más sincero y profundo agradecimiento al Ph.D. Boris Vintimilla y al Ing. Steven Araujo por brindarnos la invaluable oportunidad de formar parte de este proyecto. Su confianza en nuestras capacidades y su disposición para guiarnos a lo largo de este proceso han sido fundamentales para el desarrollo de nuestro trabajo. Agradecemos no solo la oportunidad, sino también el apoyo constante, el asesoramiento experto y la dedicación que han demostrado hacia nosotros.

Declaración Expresa

Nosotros Kevin Elihan Muñoz Calva y Loberlly Noelia Salazar Aspiazu acordamos y reconocemos que:

La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique a los autores que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 20 de mayo del 2024.

Kevin Elihan Muñoz Calva

Loberlly Noelia Salazar Aspiazu

Evaluadores

Boris Xavier Vintimilla Burgos Ph.D

Profesor de Materia

Boris Xavier Vintimilla Burgos Ph.D

Tutor de proyecto

Resumen

La estereovisión, una técnica que genera información tridimensional a partir de imágenes 2D capturadas por dos cámaras, destaca por su capacidad de proporcionar representaciones 3D detalladas, aunque enfrenta desafíos técnicos como la calibración de cámaras, la rectificación de imágenes y la generación de mapas de disparidad para producir nubes de puntos con bajo ruido. Este proyecto se enfoca en desarrollar un sistema de estereovisión que combina métodos tradicionales y avanzados para generar mapas de disparidad y nubes de puntos 3D, validando y analizando su precisión mediante la medición de la altura de personas en la escena. Se implementaron técnicas tradicionales como SGBM y enfoques modernos basados en machine learning, como RAFT-Stereo y Selective-IGEV. El desarrollo también incluyó procesos de normalización y corrección dimensional para garantizar la coherencia y comparabilidad de las nubes de puntos generadas. Los resultados mostraron que RAFT-Stereo y Selective-IGEV tienden a crear una mejor representación visual de escenarios tridimensionales mientras que SGBM acompañado del filtro WLS, destaca por su precisión uniforme. Finalmente, se concluye que RAFT-Stereo y Selective-IGEV pueden competir en precisión con métodos avanzados basados en coincidencia de bloques como SGBM, y en ciertas aplicaciones, podrían superarlos, generando representaciones visuales tridimensionales superiores.

Palabras Clave: Machine Learning, Mapas De Disparidad, Visión Por Computador, Análisis de Precisión.

Abstract

Stereovision, a technique that generates three-dimensional information from 2D images captured by two cameras, stands out for its ability to provide detailed 3D representations. However, it faces technical challenges such as camera calibration, image rectification, and disparity map generation to produce low-noise point clouds. This project focuses on developing a stereovision system that combines traditional and advanced methods to generate disparity maps and 3D point clouds, validating and analyzing their accuracy by measuring the height of people in the scene. Traditional techniques such as SGBM and modern approaches based on machine learning, like RAFT-Stereo and Selective-IGEV, were implemented. The development also included normalization and dimensional correction processes to ensure consistency and comparability of the generated point clouds. The results showed that RAFT-Stereo and Selective-IGEV tend to create a better visual representation of three-dimensional scenarios, while SGBM accompanied by the WLS filter stands out for its uniform accuracy. Finally, it is concluded that RAFT-Stereo and Selective-IGEV can compete in accuracy with advanced block-matching methods like SGBM, and in certain applications, may even surpass them by generating superior three-dimensional visual representations.

***Keywords**: Machine Learning, Disparity Maps, Computer Vision, Accuracy Analysis.*

Índice General

Resumen I
Abstract II
Índice General III
Abreviaturas V
Simbología VI
Índice de Figuras VII
Capítulo 1 1
1.1 Introducción 2
1.2 Descripción del Problema 2
1.3 Justificación del Problema 4
1.4 Objetivos 5
1.4.1 Objetivo General 5
1.4.2 Objetivos Específicos 5
1.5 Marco Teórico 5
1.5.1 Generación de Nube de Puntos 6
1.5.2 Estereovisión: Conceptos Fundamentales 7
1.5.3 Block Matching en la Estereovisión 9
1.5.4 Machine Learning en la Estereovisión 12
1.5.5 Proyección al Espacio Tridimensional 15
1.5.6 Solución Propuesta para el Proyecto 18
Capítulo 2 20
2.1 Metodología. 21
2.1.1 Generación de Bases de Datos 22
2.1.2 Calibración de las Cámaras Estereoscópicas 23
2.1.3 Rectificación de Imágenes Estereoscópicas 25
2.1.4 Generación de Nubes de Puntos Densa 28
2.1.5 Generación de Nube de Puntos No Densa 32
2.1.6 Cálculo de la Profundidad y Altura Estimada 34
2.1.7 Corrección Dimensional y de Profundidad 36
Capítulo 3 41
3.1 Resultados y Análisis 41
3.1.1 Generación de Nube de Puntos 42
3.1.2 Estimación de Profundidad 46
3.1.3 Análisis de la Altura en Función de la Profundidad 49

3.1.4 Análisis de la Altura en Función de una Profundidad Constante 52
3.1.5 Análisis de la Precisión en Estimación de Alturas a Profundidades Variables ... 55
Capítulo 4 58
4.1 Conclusiones y Recomendaciones 59
4.1.1 Conclusiones 59
4.1.2 Recomendaciones 60
Referencias 63

Abreviaturas

2D	Dos Dimensiones
3D	Tres Dimensiones
API	Application Programming Interface
BM	Block Matching
CIDIS	Centro de Investigación, Desarrollo e Innovación de Sistemas Computacionales
CNN	Convolutional Neural Network
CSA	Contextual Spatial Attention
ESPOL	Escuela Superior Politécnica del Litoral
FOV	Fiel of View
FPS	Frames per Second
GRU	Gated Recurrent Unit
JSON	JavaScript Object Notation
KD	K-Dimensional
Lidar	Light detection and Ranging
ML	Machine Learning
ROI	Region of Interest
SGBM	Semi-Global Block Matching
SGM	Semi-Global Matching
SRU	Selective Recurrent Unit
WSL	Weighted Least Squares
XML	Extensible Markup Language

Simbología

B	Línea Base
Q	Matriz de reproyección
Z	Profundidad
c	*Centro óptico de la cámara*
cm	Centímetro
d	Disparidad
f	Distancia focal
h	*Altura*
m	*Rango optimo*

Índice de Figuras

Figura 1 8
Figura 2 11
Figura 3 16
Figura 4 17
Figura 5 18
Figura 6 21
Figura 7 22
Figura 8 23
Figura 9 24
Figura 10 26
Figura 11 28
Figura 12 33
Figura 13 38
Figura 14 39
Figura 15 40
Figura 16 43
Figura 17 44
Figura 18 47
Figura 19 51
Figura 20 53
Figura 21 55

Capítulo 1

1.1 Introducción

Los sistemas de visión por computador son esenciales en aplicaciones industriales, científicas y de entretenimiento, ya que permiten a las máquinas percibir y analizar su entorno con precisión. La estereovisión, una técnica que genera información tridimensional a partir de imágenes de dos dimensiones (2D) capturadas por dos cámaras, se destaca por su capacidad de proporcionar representaciones de tres dimensiones (3D) detalladas. Sin embargo, enfrenta desafíos técnicos como la calibración de cámaras, la rectificación de imágenes, generación de mapas de disparidad para finalmente generar de nubes de puntos con bajo ruido.

El proyecto descrito en este documento aborda estos desafíos mediante el desarrollo de un sistema de estereovisión que no solo genera nubes de puntos 3D, sino que también utiliza algoritmos avanzados para segmentar objetivos y extraer características específicas. Con esto, se desea resolver problemas técnicos y extender los beneficios de la estereovisión a diversas aplicaciones industriales.

En este contexto, se ha llevado a cabo el desarrollo de un sistema la estereovisión, analizando y evaluando la evolución de las diferentes técnicas para la generación de mapas de disparidad y la posterior creación de nubes de puntos 3D. Las técnicas estudiadas incluyen el método tradicional de 2005 Semi-Global Block Matching (SGBM) (Hirschmuller, 2005) para la generación de mapas de disparidad y métodos basados en *Machine Learning* (ML) considerados modernos posteriores al año 2020, como RAFT-Stereo y Selective-IGEV (Lipson et al., 2021; Wang et al., 2024). Este análisis busca identificar los pros y contras de cada técnica, con el fin de proporcionar una visión integral de las herramientas disponibles y sus aplicaciones potenciales en distintos campos.

1.2 Descripción del Problema

La problemática principal en la generación de una nube de puntos mediante sistemas de estereovisión radica en las diferentes técnicas existentes para generar mapas de disparidad y la robustez del proceso de reconstrucción 3D. La estereovisión implica el uso de dos cámaras para

capturar imágenes de un objeto desde diferentes ángulos, simulando la visión humana, lo que permite calcular la profundidad y generar una representación 3D del objeto mediante la triangulación.

Sin embargo, existen varios desafíos técnicos asociados con este proceso:

- Calibración de cámaras: Las cámaras deben estar correctamente calibradas para garantizar la precisión en la correspondencia entre las imágenes capturadas desde diferentes puntos de vista.
- Correspondencia estéreo: La identificación precisa de puntos comunes en las imágenes estéreo es crucial para calcular la disparidad y, por lo tanto, la profundidad. Este proceso puede ser complicado debido a factores como la oclusión, la iluminación variable y el ruido en las imágenes.
- Consistencia de escala: Las diferentes técnicas de generación de disparidad pueden producir nubes de puntos con escalas inconsistentes. Esto puede dificultar la integración de datos de múltiples fuentes o la comparación de resultados
- Eficiencia computacional: El proceso de generación de nubes de puntos puede ser intensivo en términos computacionales, especialmente al trabajar con grandes cantidades de datos de imagen.

Abordar estas problemáticas es fundamental para garantizar la precisión y la fiabilidad de la generación de nubes de puntos mediante sistemas de estereovisión, lo que a su vez tiene aplicaciones importantes en diversos campos como la robótica, la visión por computador y la realidad aumentada.

El propósito fundamental de este proyecto es desarrollar un sistema de estéreo visión compuesto por dos cámaras usando técnicas tradicionales y actuales para generar la nube de puntos correspondiente a su campo de visión (FOV). Este sistema será validado mediante la cuantificación de la altura de personas que estén dentro del FOV de la nube de puntos generada. Adicionalmente, en un futuro se espera montar este sistema de estereovisión a los robots móviles

que el Centro de Investigación, Desarrollo e Innovación de Sistemas Computacionales (CIDIS) de la ESPOL tiene para sus trabajos de investigación y de desarrollo.

1.3 Justificación del Problema

El proyecto propuesto para la generación de nubes de puntos tridimensionales utilizando sistemas de estereovisión responde a una necesidad crucial en múltiples áreas de aplicación, donde es esencial una percepción precisa del entorno. El análisis de las diferentes técnicas de generación de mapas de disparidad para la generación de representaciones tridimensionales exactas del entorno mediante estereovisión no solo es relevante desde el punto de vista técnico, sino que también tiene implicaciones significativas en la investigación y la industria.

La estereovisión es una tecnología accesible y económica en comparación con otras técnicas de reconstrucción tridimensionales, lo que la hace ideal para aplicaciones que requieren precisión moderada sin incurrir en altos costos. Esto es especialmente importante para instituciones educativas y centros de investigación con presupuestos limitados o para el desarrollo de prototipos de bajo costo.

En robótica, la estereovisión mejora la capacidad de los robots para detectar obstáculos y calcular distancias, esencial para la navegación autónoma. La integración de este sistema en los robots móviles del CIDIS fomentará la innovación en el campo.

Además, la estereovisión tiene aplicaciones importantes en la visión por computador, la medicina, la biología, la seguridad, la realidad virtual y aumentada. Permite la reconstrucción tridimensional de objetos y la detección de movimiento, facilita diagnósticos médicos y estudios biológicos, mejora la detección de personas y objetos para seguridad, y contribuye a crear experiencias inmersivas en realidad virtual y aumentada.

1.4 Objetivos

1.4.1 Objetivo General

Desarrollar una solución tecnológica para la generación de nubes de puntos tridimensional mediante un sistema de estereovisión con dos cámaras analizando técnicas tradicionales y actuales para la obtención de mapas de disparidad, validando su desempeño a través de la medición de la altura de personas dentro del campo de visión del sistema.

1.4.2 Objetivos Específicos

1. Identificar con precisión los puntos comunes en las imágenes estéreo utilizando imágenes adecuadas para el cálculo de la disparidad entre los pixeles de ambas imágenes.
2. Investigar la precisión de diferentes métodos de generación de mapas de disparidad para analizar la evolución de las técnicas existentes tanto actuales como tradicionales.
3. Generar una representación tridimensional estandarizada de la escena mediante el ajuste de parámetros y técnicas de procesamiento para asegurar uniformidad y precisión en la visualización de las geometrías espaciales.
4. Validar la precisión de la nube de puntos generada a través de la medición de la profundidad y la altura de personas que están presentes en la escena.
5. Definir una plataforma escalable que incluya los módulos de generación de nubes de puntos, que además pueda acoger otros módulos de visión por computador para futuros proyectos de investigación aplicables a robots móviles.

1.5 Marco Teórico

La estereovisión es una técnica esencial en el ámbito de la visión por computador, utilizada para extraer información tridimensional a partir de dos imágenes bidimensionales. Un sistema de estereovisión binocular emplea dos cámaras situadas en diferentes posiciones para simular la visión humana, proporcionando una percepción de profundidad y la capacidad de reconstruir la escena en 3D.

1.5.1 Generación de Nube de Puntos

En el ámbito de la generación de nubes de puntos, existen diversas metodologías y tecnologías que permiten la obtención de datos espaciales, cada una con sus propias ventajas y desventajas.

Para comenzar, se encuentra el método Lidar (*Light Detection and Ranging*), desarrollado en la década de 1960 (Wandinger, 2005), el cual utiliza pulsos de luz láser para medir distancias con gran precisión, creando nubes de puntos detalladas de la superficie escaneada. Sus ventajas incluyen una alta precisión y resolución, y la independencia de las condiciones de iluminación, lo que permite su uso en una amplia variedad de entornos. No obstante, los sistemas Lidar pueden ser significativamente más costosos que otros métodos debido al *hardware* especializado requerido, y su alcance puede ser limitado, con una densidad de puntos que disminuye con la distancia (W. Zhang, 2010).

El método de estereovisión binocular, utilizado ampliamente en computación desde los años 1970 (Gregory et al., 1997; Marr et al., 1997; Ogle, 1952), es un método ampliamente utilizado debido a su capacidad para capturar la geometría 3D de una escena utilizando dos cámaras. Entre sus ventajas se encuentran el costo relativamente bajo en comparación con otros métodos como el Lidar, y su utilidad en aplicaciones dinámicas, como la navegación autónoma y la robótica móvil, debido a su capacidad para proporcionar datos 3D de alta frecuencia (Rusu & Cousins, 2011). Sin embargo, este método es sensible a las variaciones en la iluminación y texturas homogéneas que dificultan el cálculo de correspondencias, y los algoritmos de correspondencia y triangulación pueden ser computacionalmente intensivos (Sankowski et al., 2017; Scharstein et al., 2001).

La fotogrametría, desarrollada en el siglo XIX (Bache, 1892; Deville & Survey, 1895) y modernizada significativamente en el siglo XX (Lehmann, 1975), es otro método que utiliza múltiples imágenes capturadas desde diferentes ángulos para reconstruir modelos 3D de alta densidad. Sus ventajas incluyen la capacidad de generar nubes de puntos muy detalladas,

dependiendo de la distancia en la que se usa, y su versatilidad, ya que puede aplicarse utilizando drones para capturar grandes áreas y estructuras complejas desde múltiples ángulos. Sin embargo, la reconstrucción 3D a partir de imágenes puede ser computacionalmente intensiva y requerir mucho tiempo para procesar, la precisión de los modelos generados depende en gran medida de la calibración precisa de las múltiples cámaras y la calidad de las imágenes capturadas (Remondino & El-Hakim, 2006).

Finalmente, el método de luz estructurada, desarrollado en las últimas décadas del siglo XX (Gonzalez Blanco, 1998; Wolf, 1996), proyecta un patrón conocido sobre una superficie y captura la deformación del patrón para calcular la geometría 3D (Chen & Kak, 1987). Este método es muy preciso en entornos controlados y se utiliza comúnmente para escaneos en interiores y objetos pequeños, siendo menos dependiente de las texturas de la superficie. Sin embargo, la luz ambiental puede interferir con los patrones proyectados, limitando su uso en exteriores, y requiere proyectores adicionales, aumentando la complejidad y el costo del sistema (Geng, 2011).

La generación de nubes de puntos es fundamental para obtener representaciones tridimensionales detalladas a partir de datos espaciales. Los métodos como Lidar, estereovisión binocular, fotogrametría y luz estructurada han demostrado ser herramientas eficaces, cada uno con sus ventajas y limitaciones específicas. La evolución de estas técnicas ha permitido avanzar en campo de la generación de nubes de puntos tridimensionales. En la siguiente sección, se analizará la evolución de las técnicas para la generación de mapas de disparidad, que son cruciales para la creación de nubes de puntos tridimensionales precisas.

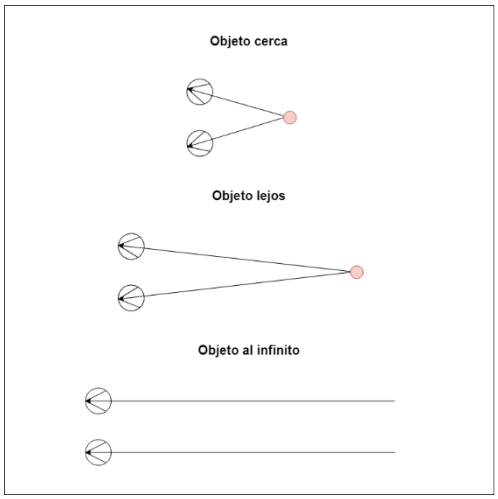
1.5.2 Estereovisión: Conceptos Fundamentales

La estereovisión se basa en el principio de la disparidad binocular, que es la diferencia de posición de un objeto en las dos imágenes capturadas por las cámaras (Scharstein et al., 2001). Esta técnica permite convertir la disparidad en información de profundidad mediante triangulación (Sombekke, s. f.).

La disparidad es un concepto crucial en estereovisión, ya que se refiere a la diferencia en la posición horizontal de un punto observado en dos imágenes tomadas desde perspectivas ligeramente diferentes. Esta diferencia se debe a la separación entre las dos cámaras, conocida como línea base (Sombekke, s. f.). La disparidad se utiliza para calcular la profundidad de los objetos en la escena. Como se muestra en la Figura 1, cuanto mayor es la disparidad, más cerca está el objeto de las cámaras, y viceversa. La precisión en el cálculo de la disparidad es esencial para obtener mediciones de profundidad exactas y fiables (Scharstein et al., 2001).

Figura 1

Disparidad relativa a diferentes distancias



Nota: La Figura 1 muestra la modificación de la disparidad de los objetos a diferentes distancias.

Para calcular la disparidad de manera precisa, es fundamental contar con cámaras bien calibradas. Los parámetros intrínsecos y extrínsecos de las cámaras juegan un papel vital en este proceso (Z. Zhang, 2000). Los parámetros intrínsecos incluyen características propias de cada cámara, como la distancia focal, el tamaño del sensor y las distorsiones de la lente. La distancia focal, en particular, es la distancia entre el centro óptico de la lente y el sensor de la cámara, y determina el campo de visión y el nivel de magnificación de las imágenes capturadas. Estos parámetros determinan cómo se proyecta una escena 3D en una imagen 2D y son necesarios para corregir las distorsiones y obtener una representación precisa de la escena (Sombekke, s. f.).

Finalmente, la disparidad d se calcula utilizando la diferencia entre las coordenadas horizontales de un punto en las dos imágenes (Barnard & Thompson, 1980), tal y como se muestra en la Fórmula 1 a continuación:

$$\boldsymbol{d = x_{izquierda} - x_{derecha}} \quad (1)$$

Por otro lado, los parámetros extrínsecos describen la posición y la orientación de las cámaras en el espacio tridimensional, incluyendo la traslación y rotación de una cámara respecto a la otra. La correcta estimación de estos parámetros permite alinear las imágenes capturadas por las dos cámaras y es fundamental para el proceso de rectificación de imágenes (Z. Zhang, 2000).

Una vez calibradas las cámaras, el siguiente paso es la rectificación de imágenes. Este proceso transforma las imágenes de las cámaras para que las líneas epipolares sean horizontales y coincidan entre las dos imágenes (Fusiello et al., 2000; Sombekke, s. f.). Las líneas epipolares son líneas proyectadas en la segunda imagen a partir de un punto en la primera imagen, teniendo en cuenta la posición relativa de las dos cámaras (Fusiello et al., 2000; Z. Zhang, 2000). En un sistema de cámaras calibradas, las líneas epipolares son horizontales y paralelas, lo que simplifica el proceso de búsqueda de puntos correspondientes entre las dos imágenes (Fusiello et al., 2000). La rectificación de imágenes ajusta las imágenes de tal manera que las líneas epipolares sean horizontales, restringiendo la búsqueda de correspondencias a una dimensión y facilitando el cálculo de disparidad.

1.5.3 Block Matching en la Estereovisión

El cálculo de disparidad es un paso crucial en la estereovisión, ya que permite la percepción de la profundidad al identificar diferencias de posición entre puntos correspondientes en las dos imágenes estéreo. Para lograr esto, se utilizan algoritmos de correspondencia como el Block Matching (BM), desarrollado entre las décadas de 1980 y 1990 (Koschan et al., 1996; Koschan & Rodehorst, 1995), y el Semi-Global Block Matching (SGBM), introducido en 2005 (Hirschmuller, 2005). Estos métodos comparan bloques de píxeles en las dos imágenes

rectificadas y buscan la mejor correspondencia para minimizar las diferencias de intensidad, lo que resulta en un mapa de disparidad preciso (Scharstein et al., 2001).

Detalles Técnicos del Método Block Matching

El *Block Matching* es un método que divide la imagen de referencia en pequeños bloques de píxeles (habitualmente cuadrados) y busca el bloque más similar en la imagen secundaria dentro de una región de búsqueda definida. La similitud entre bloques se mide comúnmente mediante métricas como la suma de diferencias absolutas (SAD) o la suma de cuadrados de diferencias (SSD). Este proceso se repite para cada bloque en la imagen de referencia, lo que da como resultado un mapa de disparidad que indica la posición relativa de los objetos en las imágenes.

Una de las principales limitaciones del BM es su sensibilidad a las variaciones de iluminación y texturas homogéneas, donde los bloques pueden parecer similares, pero no representar la misma parte de la escena. Además, BM es computacionalmente intensivo, ya que implica la búsqueda exhaustiva en una ventana de búsqueda para cada bloque, lo que lo hace menos eficiente en aplicaciones en tiempo real (Sankowski et al., 2017).

Mejoras Introducidas por Semi-Global Block Matching

El *Semi-Global Block Matching* (SGBM) es una mejora del BM tradicional que busca optimizar la correspondencia de bloques al incorporar la información de múltiples direcciones de búsqueda en lugar de solo una. SGBM minimiza una función de energía que considera no solo la similitud entre bloques de píxeles, sino también la coherencia de la disparidad en toda la imagen. Esto se logra al realizar la correspondencia a lo largo de múltiples direcciones (por ejemplo, horizontales, verticales y diagonales) y luego combinar estas evaluaciones para obtener un mapa de disparidad más robusto y suave (Hirschmüller, 2008).

SGBM aborda varios de los desafíos inherentes al BM, como las oclusiones, las texturas homogéneas y las variaciones de iluminación, que pueden dificultar la identificación precisa de

puntos correspondientes. Al considerar la coherencia global en la imagen, SGBM es capaz de producir mapas de disparidad más precisos y menos propensos a errores en áreas problemáticas.

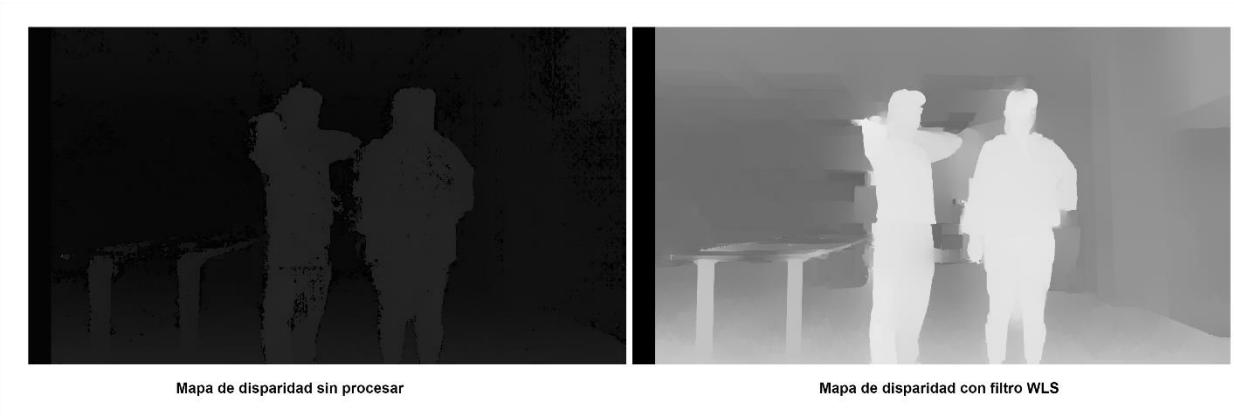
Post-procesamiento para la Mejora de Mapas de Disparidad

El resultado de utilizar BM o SGBM es un mapa de disparidad, que es una representación en escala de grises de la disparidad de cada píxel en las imágenes rectificadas. En este mapa, los tonos más claros indican una mayor disparidad (objetos más cercanos) y los tonos más oscuros indican una menor disparidad (objetos más lejanos) (Jang & Ho, 2011). Sin embargo, para mejorar la calidad de los mapas de disparidad, es común aplicar técnicas de post-procesamiento.

Uno de los métodos de post-procesamiento más efectivos es el uso de filtros de paso bajo, como el filtro de mínimos cuadrados ponderados (Weighted Least Squares, WLS). Este filtro ha demostrado ser eficaz en la eliminación de ruido *speckle* y en la mejora de la suavidad del mapa de disparidad (Choi et al., 2019). El WLS pondera las diferencias entre píxeles vecinos basándose en su similitud, lo que permite preservar los bordes mientras suaviza las áreas homogéneas, resultando en un mapa de disparidad más limpio y utilizable para aplicaciones de triangulación y cálculo de profundidad.

Figura 2

Mapa sin filtro vs. Mapa con filtro WLS



Nota: La Figura 2 muestra la diferencia entre un mapa de disparidad SGBM básico y un mapa de disparidad una vez aplicado el filtro WLS.

1.5.4 Machine Learning en la Estereovisión

Si bien, el uso de modelos ML para la generación de mapas de disparidad comenzó a ganar popularidad a partir de mediados de la década de 2000 (Montenegro et al., 2008; Venkatesh et al., 2004; Z. Zhang et al., 2011), no fue hasta 2016 que de Zbontar y LeCun (Žbontar & LeCun, 2016) realizaron uno de los trabajos más significativos en este campo al usar redes neuronales convolucionales (CNNs) para calcular la similitud entre parches de imágenes estéreo. Y es que, a partir de 2020 los avances en ML han revolucionado el campo de la estereovisión, permitiendo la creación de mapas de disparidad con una muy buena precisión y eficiencia. Modelos como RAFT-Stereo[1] y Selective-IGEV[2] han emergido como referentes en este ámbito, introduciendo enfoques arquitectónicos innovadores que abordan los desafíos inherentes al cálculo de la disparidad en imágenes estéreo.

Arquitectura de los Modelos

Los modelos de ML utilizados en estereovisión se basan en arquitecturas CNNs, optimizadas para la extracción y procesamiento de características visuales. RAFT-Stereo, por ejemplo, implementa una arquitectura híbrida que combina CNNs con Unidades Recurrentes Cerradas (GRUs), permitiendo un procesamiento iterativo que refina continuamente las estimaciones de disparidad.

RAFT-Stereo (Lipson et al., 2021): Inspirado en el modelo RAFT para flujo óptico, RAFT-Stereo introduce un volumen de costo de todos los pares dentro de una pirámide que mantiene alta resolución, asegurando la preservación de detalles finos. La arquitectura de RAFT-Stereo se distingue por el uso de GRUs multinivel que expanden el campo receptivo, mejorando

---

[1] Para más información sobre RAFT-Stereo, consultar el trabajo *"RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching"* (Lipson et al., 2021)

[2] Para más información sobre Selective-IGEV, consultar el trabajo *"Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching"* (Wang et al., s. f.)

la capacidad del modelo para capturar tanto información de alta frecuencia en bordes como de baja frecuencia en regiones homogéneas. Este enfoque iterativo permite mejorar progresivamente la precisión de los mapas de disparidad generados, abordando de manera efectiva los problemas de pérdida de detalles y correspondencias falsas en áreas de textura uniforme.

Selective-IGEV (Wang et al., 2024): En contraste, Selective-IGEV emplea una arquitectura CNN con un enfoque en la selección adaptativa de características relevantes a través de una Unidad Recurrente Selectiva (SRU). La SRU es un operador de actualización iterativa diseñado específicamente para el emparejamiento estereoscópico, que fusiona de manera adaptativa la información de disparidad en múltiples frecuencias. Esto es particularmente eficaz en la preservación de la precisión en áreas de bordes y regiones suaves. Además, el modelo incorpora un Módulo de Atención Espacial Contextual (CSA), que genera mapas de atención utilizados como pesos de fusión para mejorar la precisión en las correspondencias, mitigando la pérdida de información crítica durante los procesos iterativos.

Proceso de Generación de Mapas de Disparidad

El proceso de generación de mapas de disparidad con estos modelos sigue una serie de etapas clave:

1. **Extracción de Características**: Ambos modelos comienzan con la extracción de características visuales a partir de las imágenes estéreo utilizando CNNs.
2. **Cálculo de Correlación y Volumen de Costos**: RAFT-Stereo genera un volumen de costo a partir de las correlaciones locales, lo que facilita la evaluación exhaustiva de las correspondencias entre píxeles en las imágenes estéreo. Selective-IGEV, en cambio, emplea un enfoque más ligero para la agregación de costos, optimizando así la eficiencia computacional y reduciendo la necesidad de iteraciones prolongadas.
3. **Modulación de Costos y Refinamiento Iterativo**: En RAFT-Stereo, la disparidad se actualiza iterativamente mediante operadores basados en GRU, refinando la estimación en cada

ciclo. Selective-IGEV utiliza la SRU para ajustar adaptativamente las características de disparidad en diferentes frecuencias, lo cual es crucial para mantener la precisión en regiones con texturas homogéneas o bordes complejos.

4. **Generación del Mapa de Disparidad**: El resultado de este proceso es un mapa de disparidad que asigna un valor a cada píxel en las imágenes rectificadas. Estos mapas se emplean posteriormente para generar información de profundidad tridimensional. Tanto en RAFT-Stereo como en Selective-IGEV, se utilizan pérdidas fotométricas y geométricas durante el entrenamiento, garantizando la consistencia y precisión de los mapas de disparidad generados.

5. **Entrenamiento y Recursos Computacionales:** La capacitación de estos modelos no solo requiere arquitecturas complejas, sino también recursos computacionales significativos. Ambos modelos fueron entrenados en hardware especializado, como GPUs NVIDIA RTX 3090 (Selective-IGEV) y RTX 6000 (RAFT-Stereo), utilizando *frameworks* como PyTorch. Este entrenamiento se realizó con optimizadores avanzados, como AdamW, y con técnicas de ajuste como la programación de tasas de aprendizaje de un ciclo. Los modelos RAFT-Stereo y Selective-IGEV se entrenaron en múltiples iteraciones, lo que les permitió refinar sus capacidades de predicción de disparidad de manera iterativa.

El entrenamiento también se realizó utilizando *datasets* especializados que proporcionan un gran volumen de datos estereoscópicos y mapas de disparidad como *ground truth*. Entre estos *datasets* se encuentran "Middlebury"[3], "KITTI2015"[4], y "ETH3D"[5], que contienen miles o incluso millones de pares de imágenes estereoscópicas y sus correspondientes mapas de

---

[3] *"High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth",* (Scharstein et al., 2014)

[4] *"Object Scene Flow for Autonomous Vehicles",* (Menze & Geiger, 2015)

[5] *"A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos"*, (Schops et al., 2017)

disparidad. Cabe recalcar que estos *datasets* no son privados de cada modelo, sino que son creados por diversas instituciones de investigación y empresas tecnológicas con el objetivo de impulsar la investigación en visión por computador y proporcionar *benchmarks* comunes para evaluar el rendimiento de los algoritmos. Crear un *dataset* propio de estas características requiere una inversión significativa en hardware especializado, personal capacitado y un proceso meticuloso para generar *ground truth* de alta calidad, lo que lo convierte en una tarea compleja y costosa.

1.5.5 Proyección al Espacio Tridimensional

La proyección al espacio tridimensional es un paso crucial en la estereovisión, donde la información de disparidad obtenida de las imágenes estéreo se convierte en coordenadas 3D. Este proceso, conocido como triangulación, utiliza la geometría de las cámaras para calcular la profundidad de los puntos en la escena (Sombekke, s. f.). Tradicionalmente, este método se basa en la geometría de triángulos, donde la línea base conocida entre las cámaras y los ángulos de disparidad forman triángulos con los puntos correspondientes. Como se puede observar en la Figura 3, de manera teórica e ideal, la profundidad Z de un punto en la escena se calcula mediante la Fórmula 2 de triangulación, considerando la distancia entre las cámaras (línea base B) y la distancia focal f:

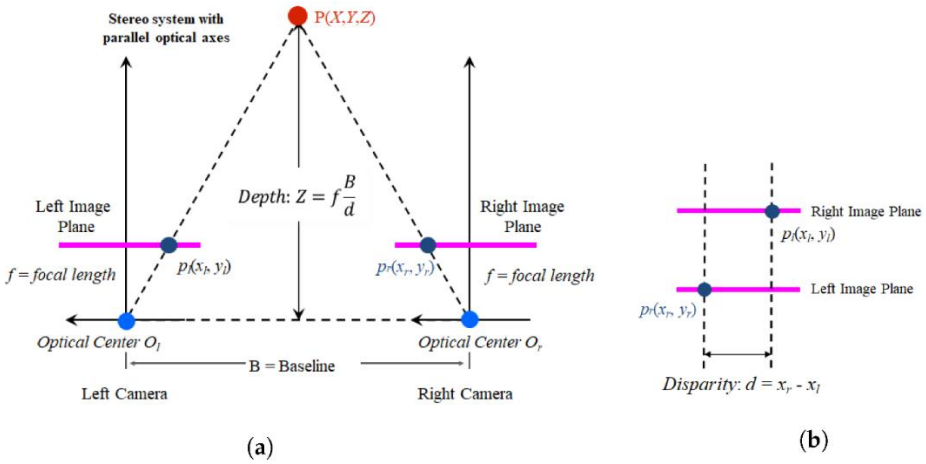
$$Z = \frac{f \cdot B}{d} \qquad (2)$$

Donde:

- $\boldsymbol{Z}$ es la profundidad del punto.
- $\boldsymbol{f}$ es la distancia focal de las cámaras.
- $\boldsymbol{B}$ es la línea base entre las dos cámaras.
- $\boldsymbol{d}$ es la disparidad.

Figura 3

Stereo matching: (a) configuración de la cámara estéreo y definición de profundidad, y (b) cálculo de disparidad



Nota: La Figura 3 obtenida de (Kim et al., 2022) muestra dos subfiguras etiquetadas como (a) y (b), que ilustran el proceso de coincidencia estéreo y el cálculo de profundidad.

Método de Reproyección Usando Matrices de Reproyección

Además del enfoque teórico, existe un método más práctico y robusto para proyectar puntos al espacio tridimensional utilizando matrices de reproyección. Estas matrices se obtienen a partir del proceso de calibración de las cámaras, que no solo ajusta las propiedades intrínsecas y extrínsecas de las cámaras, sino que también genera las matrices necesarias para reproyectar los puntos de las imágenes 2D al espacio 3D.

En este método, las coordenadas 3D (X, Y, Z) de un punto en la escena se calculan utilizando las matrices de reproyección, que son derivadas de las matrices de calibración de la cámara. Estas matrices transforman las coordenadas de imagen y la disparidad en coordenadas en el espacio tridimensional. La Fórmula 4, representa la fórmula general para la reproyección y se expresa como:

$$\begin{pmatrix} X \\ Y \\ Z \\ W \end{pmatrix} = Q \begin{pmatrix} x \\ y \\ d \\ 1 \end{pmatrix} \quad (3)$$

Donde:

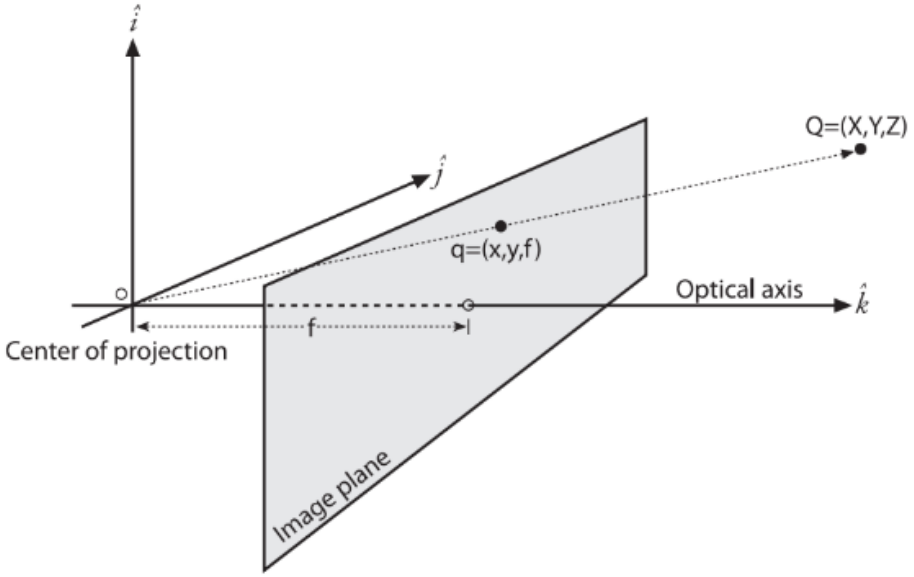
- $\boldsymbol{Q}$ es la matriz de reproyección.
- $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ son las coordenadas del píxel en la imagen.
- $\boldsymbol{d}$ es la disparidad en ese punto.
- $\boldsymbol{W}$ es el factor de escala homogéneo.

Como se muestra en la Figura 4, la matriz Q, que se obtiene durante el proceso de calibración, permite calcular las coordenadas tridimensionales de un punto en el espacio a partir de sus coordenadas de imagen y su disparidad. Las coordenadas 3D reales en el espacio físico se obtienen dividiendo X, Y, Z por W, según la Fórmula 5 descrita en (Bradski & Kaehler, 2011).

$$X_{real} = \frac{X}{W} \quad , \quad Y_{real} = \frac{Y}{W} \quad , \quad Z_{real} = \frac{Z}{W} \tag{4}$$

Figura 4

Reproyección de puntos tridimensionales utilizando Q



Nota: La Figura 4 obtenida de (Bradski & Kaehler, 2011) muestra el proceso de proyección de puntos tridimensionales utilizado la matriz Q.

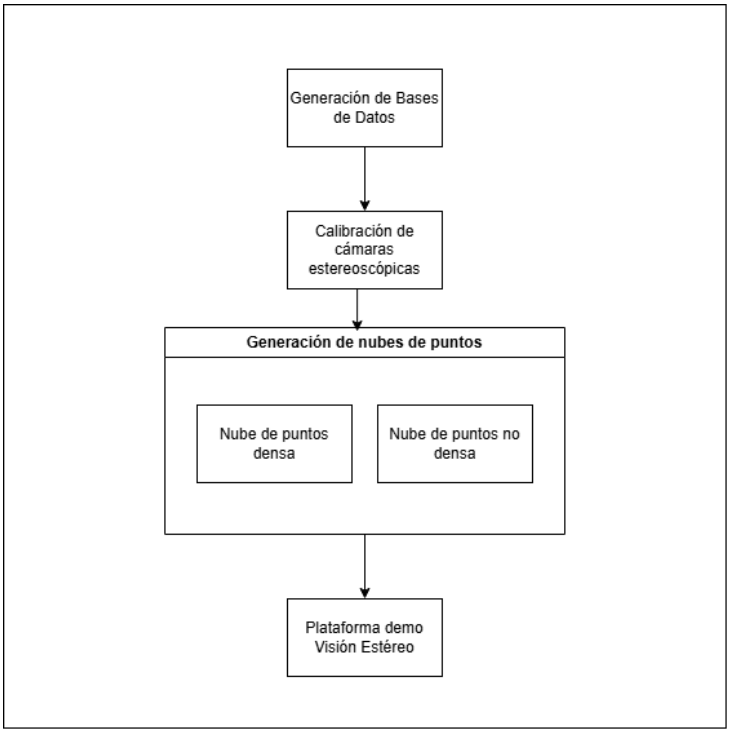
El uso de matrices de reproyección permite una transformación directa y eficiente de los puntos de las imágenes rectificadas al espacio 3D, teniendo en cuenta tanto la geometría de la cámara como las correcciones necesarias para eliminar distorsiones y errores. Este método es especialmente valioso en aplicaciones donde la precisión en la reconstrucción tridimensional es crucial, como en la visión por computador y la estereovisión avanzada.

1.5.6 Solución Propuesta para el Proyecto

Para la realización de este proyecto, se propone la implementación de varios módulos y submódulos interrelacionados como se aprecia en la Figura 5. En primer lugar, se llevará a cabo la creación de un *dataset* de verdad base "*Ground Truth"*. Posteriormente, se procederá a la calibración de las cámaras utilizando *MatLab*, seguido de la rectificación de la entrada. Una vez calibradas las cámaras y rectificadas las imágenes, se avanzará hacia la generación de la nube de puntos, que incluye la creación de mapas de disparidad utilizando SGBM, RAFT-Stereo y Selective-IGEV, seguido a esto se realizará un post-procesado utilizando el filtro WLS para el enfoque tradicional representado por SGBM. Luego, se generará una nube de puntos densa y dos tipos de nubes no densas identificando ROIs (*Region Of interest*) y *Keypoints* mediante un modelo de ML especializado en visión por computador. Finalmente, se realizarán pruebas y rectificaciones, evaluando el error en la profundidad de las diferentes nubes y corrigiéndolo mediante un proceso de normalización y regresión lineal. El proyecto concluirá con el cálculo de la altura de las personas en la escena acompañada de una plataforma de Visión Estéreo *demo* integrando los módulos desarrollados.

Figura 5

Flujo de trabajo propuesto



Nota: La Figura 5 muestra el flujo de trabajo propuesto a seguir para el desarrollo del proyecto.

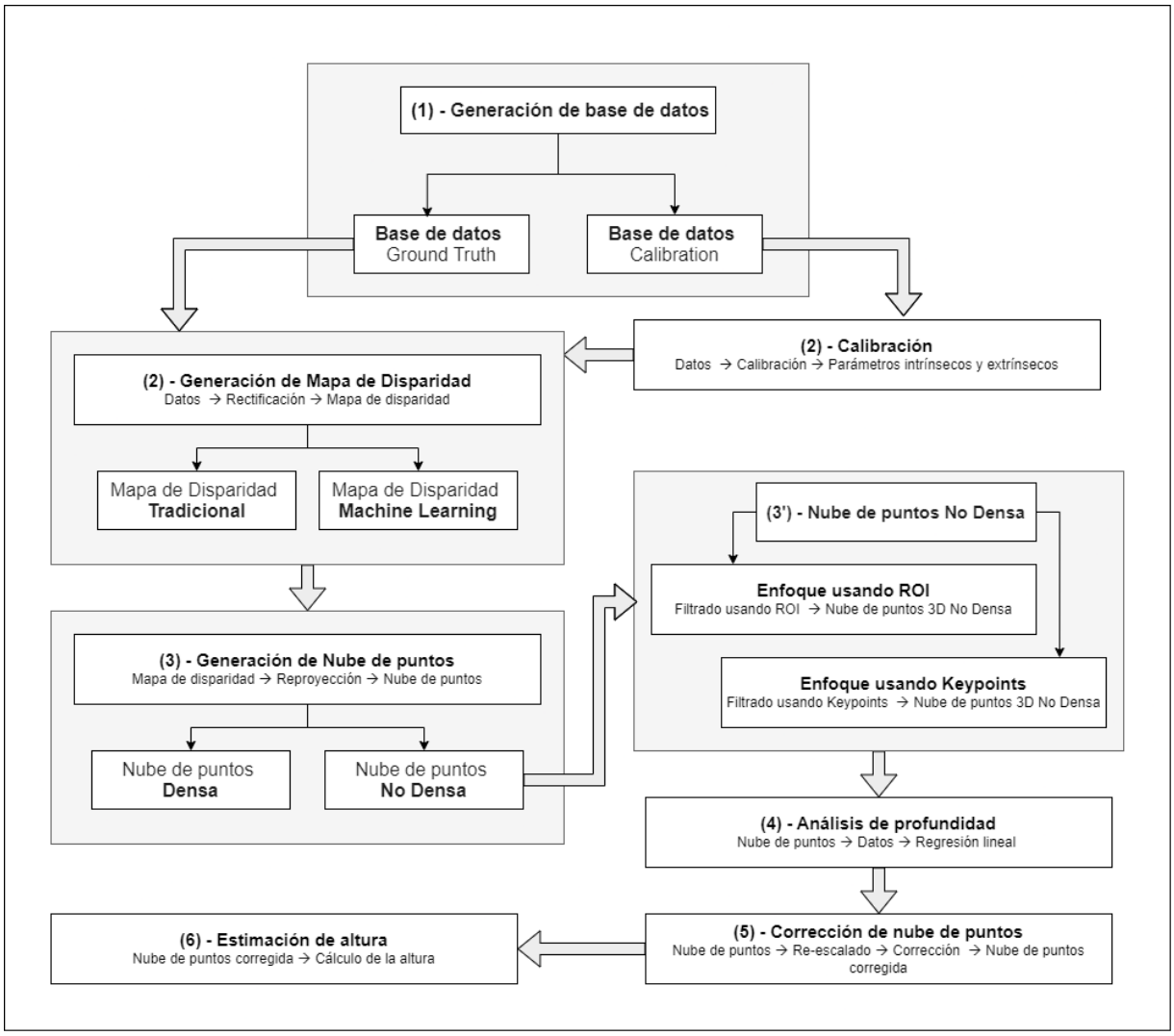
Capítulo 2

2.1 Metodología.

A partir del flujo de trabajo propuesto en la Figura 5, se diseñó una propuesta modular más detallada para la solución del proyecto, tal y como se muestra en la Figura 6.

Figura 6

Diagrama modular para la generación de nubes de puntos



Nota: La Figura 6 muestra un diagrama modular usado para la solución la generación de las nubes de puntos.

Primeramente, se ha abordado a detalle los conceptos y configuraciones relacionadas con la rectificación de imágenes, crucial para la generación de mapas de disparidad y nubes de puntos 3D en aplicaciones de visión por computadora utilizando la biblioteca OpenCV. Se

describen los métodos de interpolación y los modos de borde, proporcionando recomendaciones específicas para mejorar la precisión y la calidad de los resultados.

2.1.1 Generación de Bases de Datos

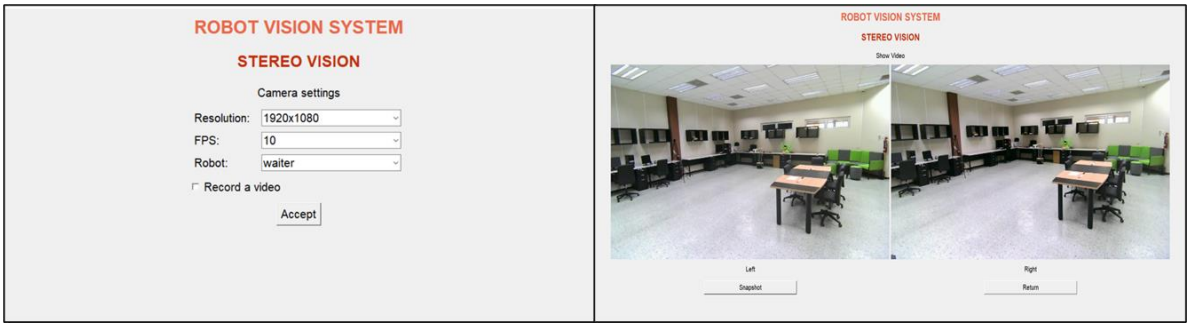
Se desarrollaron dos tipos principales de bases de datos: una destinada a la calibración de las cámaras y otra enfocada en la validación de la precisión mediante la comparación con datos de referencia.

Captura de Imágenes

Las imágenes se capturaron utilizando un software especializado desarrollado por el CIDIS, como se muestra en la Figura 7. Este software permite seleccionar diferentes resoluciones de la cámara, la cantidad de cuadros por segundo (FPS) y el robot asignado. Se tomaron imágenes y videos a una resolución de 1920x1080 píxeles y 30 FPS, los cuales fueron etiquetadas sistemáticamente en el formato <Date_time>_IMG_LEFT, <Date_time>_IMG_RIGHT y <Date_time>_IMG_ORIGINAL, permitiendo un seguimiento preciso de las imágenes y sus metadatos temporales.

Figura 7

Software de captura de imágenes



Nota: La Figura 7 muestra la interfaz de del software de captura de imágenes para sistemas de estereovisión del CIDIS.

Base de Datos de Calibración.

Para la calibración de las cámaras estereoscópicas, se empleó un conjunto de datos compuesto por 234 pares de imágenes que incluían un patrón de tablero de ajedrez con cuadros

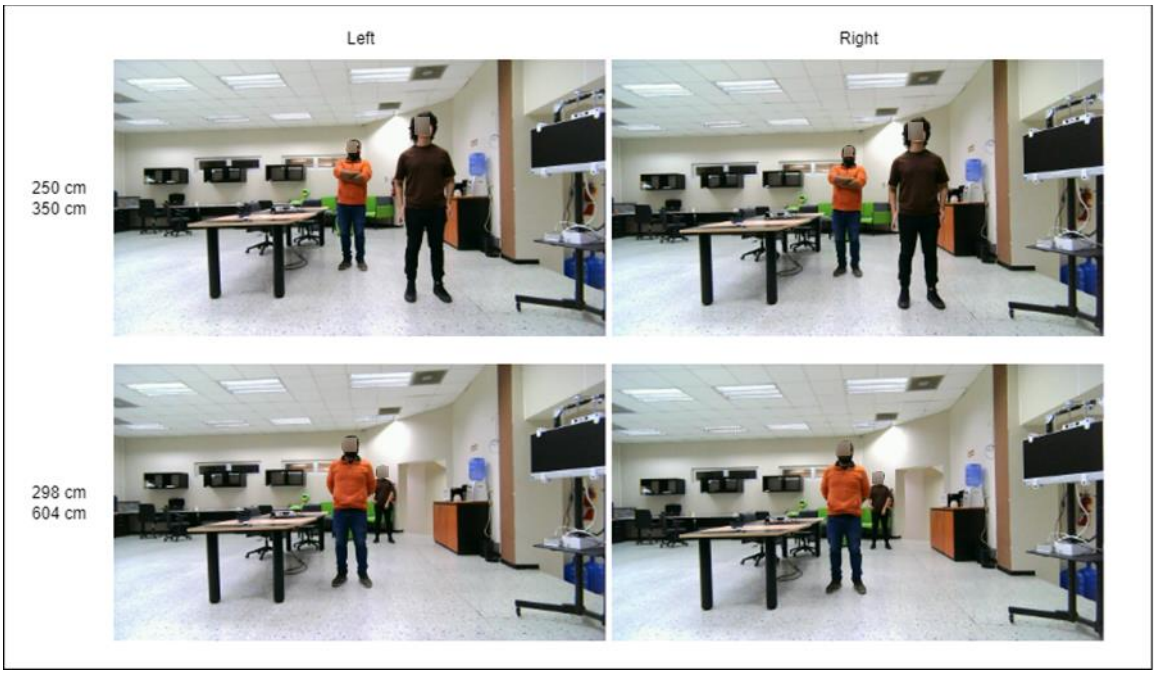
de 10 mm de lado. El patrón de tablero de ajedrez fue colocado en diversas posiciones y orientaciones dentro del campo de visión de las cámaras, asegurando una cobertura completa del campo visual.

Base de Datos *Ground Truth*

Para evaluar y validar la precisión de las nubes de puntos generadas, se utilizó un conjunto de datos compuesto por 138 pares de imágenes etiquetadas a diferentes distancias y alturas, como se puede observar en la Figura 8. Estas distancias abarcaban un rango entre 150 cm y 620 cm en intervalos de variables.

Figura 8

Imágenes del dataset Ground Truth



Nota: La Figura 8 muestra pares de imágenes estereoscópicas de personas capturadas a diferentes profundidades.

2.1.2 Calibración de las Cámaras Estereoscópicas

La calibración de las cámaras estereoscópicas consistió en un procedimiento que comprendió la detección y corrección de distorsiones, la alineación de las cámaras y la

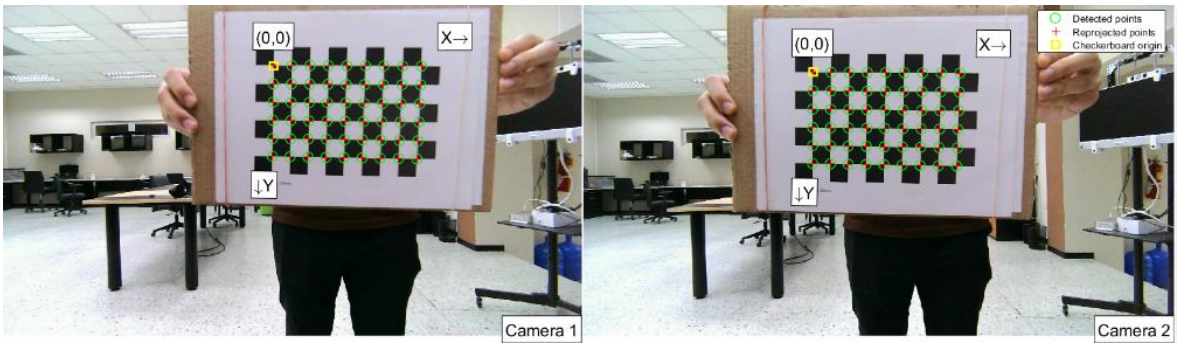
optimización global de los parámetros. A continuación, se describen los pasos y metodologías empleados en detalle.

Detección y correspondencia de Puntos

Como se muestra en la Figura 9, se detectaron los patrones en las esquinas del tablero de ajedrez en cada imagen utilizando algoritmos de procesamiento de imágenes, concretamente se utilizó MATLAB, lo que fue crucial para calcular los parámetros intrínsecos de las cámaras, como las matrices de calibración y los coeficientes de distorsión. Posteriormente, la correspondencia de estos puntos permitió estimar los parámetros extrínsecos, describiendo la posición y orientación relativas de las cámaras, asegurando su alineación precisa y la correcta medición de distancias en la visión estereoscópica. Los parámetros intrínsecos y extrínsecos calibrados se almacenaron en archivos JSON y XML para su uso en procesos posteriores de rectificación y triangulación.

Figura 9

Detección de esquinas de tablero de ajedrez



Nota: La Figura 9 muestra la detección de las esquinas en los tableros de ajedrez durante la calibración en MATLAB.

Métodos de Calibración

Se utilizó el *dataset Calibration* para la respectiva calibración de las cámaras empleadas. Para este fin, se utilizó la aplicación de escritorio de MATLAB debido a su capacidad de integración con otros módulos de análisis. El módulo usado, "*Stereo Camera Calibration*", permitió la carga y procesamiento de las imágenes de calibración, y la estimación precisa de los

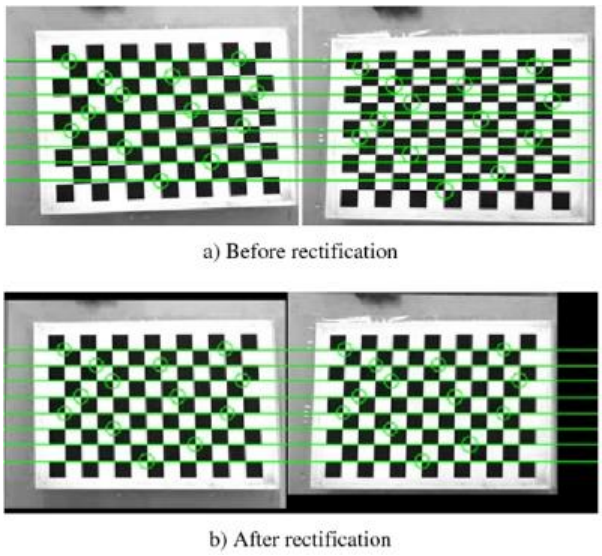
parámetros deseados, los cuales se almacenaron en archivos JSON para rectificación y triangulación posteriores.

2.1.3 Rectificación de Imágenes Estereoscópicas

Una vez completada la calibración de las cámaras estereoscópicas, el siguiente paso crucial fue la rectificación de las imágenes capturadas. La rectificación es un proceso esencial que permite alinear dos imágenes de manera que los pares de puntos correspondientes se ubiquen en la misma línea horizontal en ambas imágenes, como se puede apreciar en la Figura 10. Esto simplifica significativamente el cálculo de la disparidad y la generación de un mapa de disparidad preciso. A continuación, se detallan los procedimientos y técnicas utilizados para la rectificación de imágenes.

Figura 10

Rectificación de imágenes estéreo



Nota: La Figura 10 obtenida de (Su & He, 2011) muestra el antes y el después de un proceso de rectificación de un par de imágenes estereoscópicas visibilizando el cambio en sus líneas epipolares.

Proceso de Rectificación

El proceso de rectificación se llevó a cabo utilizando los parámetros intrínsecos y extrínsecos obtenidos durante la calibración. Los principales objetivos de este proceso fueron corregir las distorsiones presentes en las imágenes y alinear vertical y horizontalmente las vistas de las dos cámaras de manera precisa.

Obtención de Parámetros de Rectificación:

Utilizando los parámetros intrínsecos y extrínsecos, junto con los coeficientes de distorsión de las cámaras, se calcularon las matrices de rectificación y los mapas de reasignación necesarios para transformar las imágenes. Con *cv.stereoRectify* de OpenCV, se calcularon las matrices de transformación que alinean las imágenes estereoscópicas.

- **Matrices Intrínsecas:** Describen las propiedades internas de las cámaras.
- **Coeficientes de Distorsión:** Valores que describen las distorsiones de las lentes.

- **Matriz de Rotación y Vector de Traslación:** Transformaciones que relacionan las posiciones de las dos cámaras.

Posteriormente, *cv.initUndistortRectifyMap* generó los mapas de reasignación que corrigen las distorsiones y transforman las imágenes a una vista rectificada, eliminando los efectos causados por las lentes.

Aplicación de Mapas de Reasignación:

Con los mapas de reasignación obtenidos, se aplicaron transformaciones a las imágenes originales para obtener vistas rectificadas y asegurar que los pares de puntos correspondieran a líneas horizontales en ambas imágenes. Para esto, se utilizó la función *cv.remap* de OpenCV, que transforma cada píxel de la imagen original según los mapas de coordenadas, reubicándolos en su nueva posición para producir las imágenes rectificadas, tal y como se observa en la Figura 11.

La configuración utilizada incluyó:

- **Interpolación**: Se uso *cv.INTER_LANCZOS4,* el cual utiliza una función *sinc* para calcular el valor del píxel, considerando un vecindario de 8x8 píxeles. Ofrece muy alta calidad, preservando detalles finos, pero es la más lenta de todas.

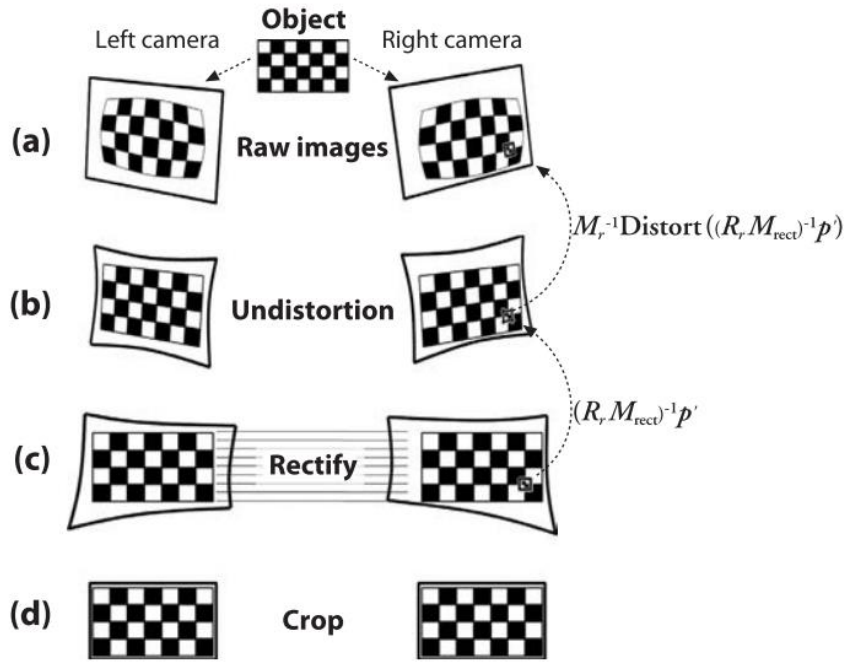
- **Borde**: Se usó *cv.BORDER_CONSTANT,* el cual rellena con un valor constante definido por el usuario (Para este proyecto se usó el valor de 0). Permite un control preciso del valor del borde, pero puede introducir artefactos si el valor no coincide con la imagen.

Finalmente, se obtuvo la matriz Q[6], la cual es fundamental para la reproyección de las disparidades de las imágenes estereoscópicas en coordenadas tridimensionales reales (andijakl, 2020; Y.-J. Zhang, 2021).

[6] Para una comprensión más detallada de la matriz Q, se sugiere revisar: *Learning OpenCV* (Bradski & Kaehler, 2011).

Figura 11

Proceso de rectificación



Nota: La Figura 11 obtenida de (Bradski & Kaehler, 2011) describe los pasos realizados en el proceso de rectificación de imágenes estéreo.

2.1.4 Generación de Nubes de Puntos Densa

Con las imágenes rectificadas obtenidas, el siguiente paso fue la generación de mapas de disparidad y, a partir de estos, la creación de nubes de puntos tridimensionales.

Generación de Mapas de Disparidad

Un mapa de disparidad es una representación de las diferencias de posición entre dos imágenes estereoscópicas. Cada valor en el mapa de disparidad corresponde a la diferencia de posición de un punto entre las dos imágenes. Como se mostró en la Figura 1, este valor es inversamente proporcional a la distancia del punto respecto a las cámaras.

Para calcular el mapa de disparidad, se utilizaron tres métodos diferentes: SGBM, un enfoque más tradicional, y SELECTIVE-IGEV y RAFT-STEREO, que son métodos avanzados basados en aprendizaje automático.

Mapa de disparidad usando SGBM

Para la obtención del mapa de disparidad de SGBM, se empleó la *función cv.StereoSGBM_create* de OpenCV. Este método, basado en Semi-Global Matching, dependiendo del *mode* en uso puede equilibrar precisión y/o eficiencia computacional. Los parámetros clave incluidos fueron *minDisparity*, el valor mínimo de disparidad; *numDisparities*, el rango de valores de disparidad a considerar; *blockSize*, el tamaño del bloque de comparación; y P1 y P2, que son parámetros de control de suavidad de la disparidad. Estos parámetros fueron ajustados para optimizar la calidad del mapa de disparidad, asegurando una buena detección de la profundidad y reduciendo el ruido.

Durante la generación del mapa de disparidad, se observó la presencia de ruido significativo, lo que podría afectar la precisión de la reconstrucción tridimensional. Para mitigar este problema, se aplicó el filtro de paso bajo WLS, que suaviza el mapa de disparidad preservando los bordes tal y como se puede apreciar en la Figura 2. Esto es crucial para mantener la integridad de los objetos en la escena mientras se reduce el ruido. El filtro WLS minimiza una función de costo que incluye términos de fidelidad y suavidad. El término de fidelidad asegura que el filtro no altere significativamente las disparidades en las regiones homogéneas, mientras que el término de suavidad regula la cantidad de suavización aplicada, especialmente en los bordes. Este filtro utiliza una imagen guía (normalmente la imagen proveniente de la cámara izquierda de las imágenes de entrada) para ajustar la cantidad de suavización en función de la estructura de la imagen, permitiendo un mejor manejo de los bordes y detalles finos (Farbman et al., s. f.).

Generación de Mapas de Disparidad usando RAFT-Stereo y Selective-IGEV

En esta sección se describen los métodos empleados para generar mapas de disparidad utilizando los modelos avanzados de ML RAFT-Stereo y Selective-IGEV. Para este proyecto se han usado modelos preentrenados utilizando el *dataset* Middlebury y se implementaron utilizando la configuración por defecto recomendada por los autores, asegurando así la

reproducibilidad y la eficacia de los resultados obtenidos. A continuación, se detallan los pasos metodológicos seguidos en la implementación de estos modelos.

Mapa de Disparidad usando RAFT-Stereo

El proceso comienza con la extracción de características visuales a partir de las imágenes estéreo. RAFT-Stereo utiliza una CNN para generar un volumen de costo tridimensional, que captura la similitud entre cada píxel de la imagen izquierda con sus posibles correspondencias en la imagen derecha. Este volumen de costo es esencial para identificar con precisión las disparidades entre las imágenes.

Una vez extraídas las características, se construye un volumen de costo piramidal. Este enfoque permite que el modelo mantenga una alta resolución en todas las escalas, lo cual es crucial para preservar detalles finos en el proceso de correspondencia de píxeles. La estructura piramidal facilita la captura de correlaciones a diferentes niveles de resolución. Posteriormente se aplica un proceso de refinamiento iterativo mediante GRU. En cada iteración, el modelo ajusta las disparidades calculadas previamente, refinando las estimaciones iniciales. Este enfoque iterativo permite mejorar la precisión del mapa de disparidad, particularmente en áreas de bordes finos y texturas homogéneas.

Mapa de Disparidad usando Selective-IGEV

Selective-IGEV también utiliza inicialmente una arquitectura de CNN diseñada para extraer características a múltiples niveles de resolución. Este enfoque permite que el modelo capture tanto los detalles finos como las estructuras globales presentes en las imágenes estéreo.

Un componente central de Selective-IGEV es el Módulo CSA, que genera mapas de atención. Estos mapas guían el proceso de correspondencia al asignar pesos específicos a diferentes regiones de la imagen, según su relevancia. Seguidamente, el modelo emplea SRU para fusionar de manera adaptativa la información de disparidad en múltiples frecuencias. Este proceso iterativo permite ajustar las correspondencias de manera más precisa, especialmente en áreas de baja textura o con cambios de iluminación.

Generación de Nubes de Puntos Densa

A partir del mapa de disparidad generado, se procede a la conversión de estos datos en nubes de puntos tridimensionales. Este proceso implica transformar las coordenadas de disparidad en coordenadas 3D, lo que permite visualizar la geometría de la escena capturada. Se utilizaron dos enfoques vinculados a cada método en particular.

Enfoque con SGBM

Para los mapas de disparidad generados utilizando el enfoque SGBM, se empleó la matriz Q calculada previamente durante la calibración de las cámaras. La función *reprojectImageTo3D* de OpenCV fue utilizada para realizar la conversión. Esta función toma como entradas el mapa de disparidad y la matriz Q, produciendo una matriz tridimensional donde cada elemento contiene las coordenadas 3D correspondientes. Cabe recalcar que se utilizó este tipo de triangulación debido al que es un método propio de OpenCV misma librería del método SGBM especifico usando en este proyecto.

Enfoque con RAFT-Stereo y Selective-IGEV

En contraste, para los mapas de disparidad obtenidos mediante RAFT-Stereo y Selective-IGEV, se utilizó un método de triangulación manual, ajustado según las recomendaciones de los autores en sus repositorios oficiales. La Fórmula 5 ha sido empleada para calcular la profundidad Z:

$$depth = \frac{f_x \cdot B}{-d + (c_{x2} + c_{x1})} \quad (5)$$

Donde:

- $\boldsymbol{f_x}$ es la distancia focal horizontal de la cámara en pixeles.
- $\boldsymbol{B}$ es la línea base entre las dos cámaras.
- $\boldsymbol{d}$ es la disparidad en ese punto.
- $\boldsymbol{c_{x1}, c_{x2}}$ son las coordenadas x de los centros ópticos de las cámaras.

Posteriormente, se genera una rejilla de puntos que representa las coordenadas X, Y

y Z en el espacio tridimensional usando la Fórmula 6 mostrada a continuación:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} \frac{(xx - c_{x1})}{f_x} \\ \frac{(yy - c_{y1})}{f_{y1}} \\ 1 \end{pmatrix} \cdot depth \quad (6)$$

Donde:

- $\boldsymbol{xx, yy}$ con las coordenadas de la imagen en el plano x y en el plano y.
- $\boldsymbol{c_{y1}}$ es la coordenada y del centro óptico de la cámara 1.
- $\boldsymbol{f_y}$ es la distancia focal vertical de la cámara en pixeles.

Se realizaron ajustes adicionales a este método, permitiendo estandarizar el proceso de reproyección para los diferentes enfoques utilizados (SGBM, RAFT-Stereo, y Selective-IGEV), facilitando la comparación y análisis de los resultados. Estos ajustes aseguran que las nubes de puntos generadas sean compatibles entre sí y puedan ser utilizadas en análisis y aplicaciones posteriores de manera coherente.

2.1.5 Generación de Nube de Puntos No Densa

Tras la obtención de la nube de puntos tridimensionales densa, se procedió a la generación de una nube de puntos no densa. Este proceso es relevante para aplicaciones donde una representación menos detallada de la escena es necesaria, suficiente y/o más eficiente en términos computacionales. Para reducir la densidad de la nube de puntos, se utilizó el modelo de ML, YOLOv8 (*You Only Look Once* versión 8) y las llamadas "máscaras de interés" para seleccionar y mantener solo los puntos relevantes.

Para generar la nube de puntos no densa, se emplearon técnicas avanzadas de detección y segmentación. Primero, se utilizó YOLOv8 para la detección de personas en las imágenes

estereoscópicas. Para este proyecto, el nivel de confianza para la detección de personas se estableció en 0.8, asegurando así la fiabilidad de las detecciones.

A partir de las detecciones realizadas con YOLOv8, se generaron máscaras de interés las cuales mantenían únicamente los píxeles correspondientes a las personas detectadas, rellenando el resto con ceros para conservar la resolución original de las imágenes. Como se aprecia en la Figura 12, se crearon dos tipos de máscaras: ROI y *Keypoints*. Las máscaras ROI se centraron en delinear el contorno de las personas detectadas y rellenar con los píxeles correspondientes dentro de ese contorno. Por otro lado, las máscaras *Keypoints* se enfocaron en los puntos clave de las personas, como articulaciones y extremidades.

Figura 12

Máscaras de interés

Nota: En la Figura 12 el recuadro de la izquierda muestra la máscara de ROI mientras que el recuadro de la derecha muestra la máscara de *Keypoints*.

Estas máscaras se aplicaron al mapa de disparidad para filtrar los píxeles correspondientes a las personas detectadas. Gracias a este filtrado, se generó la nube de puntos no densa, que contenía únicamente los puntos de interés, es decir, las personas presentes en la escena.

La generación de nubes de puntos no densas facilita un análisis más eficiente y específico, optimizando almacenamiento y recursos computacionales al momento de ser usadas. Este proceso es fundamental para aplicaciones que requieren la identificación y análisis detallado de objetos específicos dentro de una escena tridimensional en donde no se requiera o no sea necesario tener un escenario completo.

2.1.6 Cálculo de la Profundidad y Altura Estimada

El análisis de la profundidad estimada se llevó a cabo para evaluar la precisión de las nubes de puntos generadas y validar las estimaciones de distancia respecto a las cámaras estereoscópicas. Este análisis se enfocó exclusivamente en las nubes de puntos no densas generadas a partir del dataset "*Ground Truth*", con el objetivo de calcular el centroide de cada persona detectada y comparar la estimación de profundidad de dicho centroide con la distancia originalmente medida. A continuación, se detalla el proceso realizado.

Cálculo de centroides

Para el cálculo de centroides se utilizó un método diseñado para calcular el centroide de una nube de puntos tridimensional, excluyendo aquellos que se consideran ruido, es importante mencionar que para que este método funcione se debe tener más de una cantidad k de puntos, ya que este parámetro representa el número de puntos vecinos a tomar en cuenta al momento del cálculo.

Para identificar y excluir el ruido, se utiliza una estructura de datos conocida como árbol kd (k-dimensional tree). El árbol kd es una herramienta eficiente para realizar búsquedas de vecinos más cercanos en espacios multidimensionales (Bentley, 1990). Esta estructura organiza los puntos de manera jerárquica, lo que permite realizar consultas rápidas sobre las distancias entre puntos en un espacio tridimensional. Específicamente, el árbol kd se utiliza para calcular las distancias euclidianas promedio entre cada punto y sus k vecinos más cercanos, ignorando la distancia al propio punto.

Los puntos se clasifican como ruido si su distancia a los vecinos se desvía significativamente del promedio, determinado por un umbral ajustable. Los puntos que no cumplen con este criterio se consideran “ruido” y se eliminan, dejando únicamente los puntos "no ruidosos".

Finalmente, el centroide de la nube de puntos se calcula utilizando únicamente los puntos no ruidosos. Este centroide se obtiene promediando las coordenadas (*x, y, z*) de estos puntos, resultando en una estimación precisa del centroide de la nube de puntos tridimensional.

Extracción del Valor de Profundidad (Z)

Con los centroides identificados, se extrajo el valor de profundidad *Z* de cada persona en la escena. Estos valores de *Z* representaron la distancia de cada persona respecto a las cámaras estereoscópicas utilizadas para capturar la nube de puntos.

Los valores estimados y reales de *Z* se almacenaron en diversas situaciones, que representaban variaciones en las escenas donde las personas se encontraban a diferentes profundidades, para posteriormente ser comparadas.

Cálculo de la altura

El cálculo de la altura se basa en el uso del centroide previamente generado. Dado que la nube de puntos puede contener puntos clasificados como "ruido", cuya presencia podría distorsionar la medición precisa de la altura debido a sus coordenadas erráticas, se ha desarrollado una técnica para mitigar su influencia durante el cálculo.

Primero, se define un rango óptimo de confianza alrededor del centroide C, con coordenadas (x_C, y_C, z_C). Este rango se denota como m, el cual delimita un intervalo en la coordenada z donde se espera encontrar los puntos más representativos, excluyendo aquellos considerados como ruido.

Como se muestra en la Fórmula 7, el rango de confianza óptimo se define como:

$$[z_{min}, z_{max}] = [z_C - m, z_C + m] \quad (7)$$

Dentro de este rango calculado usando la Fórmula 7, se seleccionan únicamente los puntos cuya coordenada z se encuentra dentro del intervalo mencionado. A continuación, con la Fórmula 8, se calcula la diferencia entre los valores máximo y mínimo de la coordenada z de los puntos seleccionados, lo cual nos proporciona una estimación de la altura (h).

$$h = \max(y) - \min(y) \quad para \;\; z_{min} \leq z \leq z_{max} \tag{8}$$

Comparación con Datos de Referencia

Utilizando el *dataset* "*Ground Truth*", se compararon los valores reales de Z con los valores estimados por la nube de puntos para calcular el error presente en los valores estimados, mediante la Fórmula 9.

$$Error = Z_{real} - Z_{estimado} \tag{9}$$

Esta ecuación permitió cuantificar la precisión de las estimaciones de profundidad y detectar posibles desviaciones.

Visualización y Análisis del Error

Para visualizar y analizar el error en las estimaciones de profundidad, se crearon gráficas comparativas entre los valores reales y estimados de Z. Estas gráficas permitieron identificar patrones y magnitudes de error, proporcionando una visión clara de la precisión y áreas de mejora en el proceso de estimación de profundidad.

El análisis de la profundidad estimada proporcionó una base para evaluar la exactitud de las nubes de puntos generadas. Con esta evaluación, se identificaron las áreas donde se requerían ajustes adicionales. A continuación, se describe el proceso de corrección de la profundidad, enfocado en mejorar la precisión de las estimaciones realizadas.

2.1.7 Corrección Dimensional y de Profundidad

Tras la evaluación de la precisión de las estimaciones de profundidad y la identificación de errores se notó que la generación de nubes de puntos tridimensionales a partir de mapas de disparidad generados por SGBM con su respectivo método de triangulación, producían nubes de puntos con dimensiones significativamente menores en comparación con los generados por RAFT-Stereo y Selective-IGEV. Esta discrepancia dimensional afectaba la consistencia y la aplicabilidad de las nubes de puntos en contextos posteriores. Para resolver este problema, se

implementaron procesos de normalización y corrección de profundidad, detallados a continuación.

Normalización Dimensional

La normalización dimensional se realizó utilizando un enfoque sistemático para ajustar las escalas de las nubes de puntos generadas por SGBM, alineándolas con aquellas obtenidas a partir de RAFT-Stereo y Selective-IGEV, los cuales presentaban una escala mayor y más coherente. Este proceso también implicó el uso de un pivote, en este caso el origen de coordenadas, para reubicar todos los puntos de la nube. Este enfoque no solo cambió la escala dimensional de la nube, sino que también reubicó los puntos en sus nuevas posiciones de manera estandarizada, permitiendo así una mejor comparabilidad entre diferentes nubes de puntos.

- **Punto de Referencia y Factor de Escala**: La metodología incluyó la selección de un pivote usado como punto de referencia central, específicamente se usó el origen de coordenadas (0, 0, 0), el cual, si bien no tiene un peso por ser el origen, para futuras modificaciones es de gran utilidad. El punto de referencia permitió reubicar a un lugar común asegurando una transformación homogénea de la nube de puntos según un factor de escala especificado.
- **Cálculo de Nuevas Posiciones**: Se calcula la nueva posición de cada punto en la nube después de aplicar el factor de escala. Este ajuste asegura que la nube de puntos se amplíe o reduzca de manera uniforme en todas las direcciones, permitiendo una comparación directa entre las diferentes nubes de punto. Finalmente, los puntos se trasladan de vuelta a su sistema de referencia original, asegurando que la estructura y la orientación de la nube de puntos se mantengan intactas.

Corrección de la Profundidad

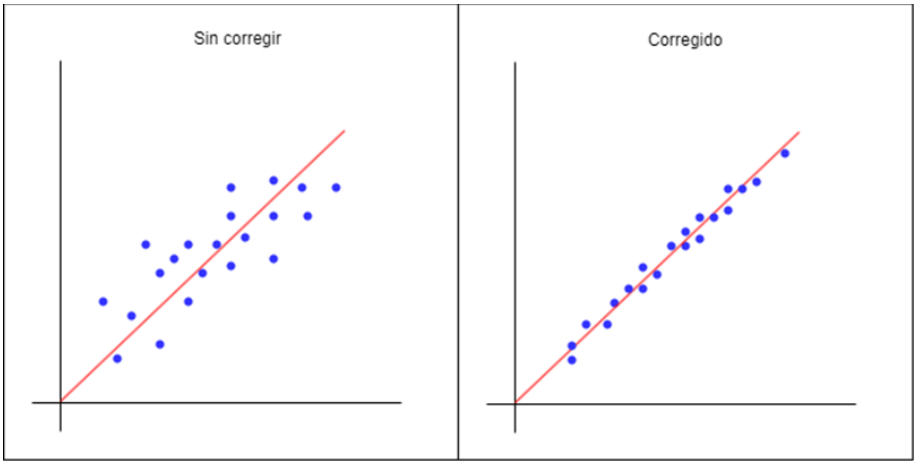
En el proceso de corrección de la profundidad, se abordaron las diferencias en las estimaciones de profundidad, particularmente aquellas surgidas después de la triangulación. Un fenómeno particular que se identificó fue el cono producto de la disparidad binocular que da paso a la diferente percepción de tamaño en los objetos (Oyama & Sato, 1967), donde los objetos

lejanos aparecían más pequeños que los cercanos, a pesar de tener tamaños reales idénticos. Este efecto, causado por la geometría de la proyección estereoscópica, hacía necesario un ajuste para contrarrestar la distorsión en la percepción de profundidad.

- **Transformación de Potencia**: Para corregir este fenómeno, se aplicó una transformación de potencia a las coordenadas de profundidad *Z*, controlada por un parámetro 'α'. Esta transformación permitió mitigar la escala de profundidad de manera no lineal, mejorando la correspondencia con los valores esperados y contrarrestando el efecto de distorsión que producía el cono cambio en el tamaño de los objetos.
- **Ajuste de Coordenadas**: Además de la corrección de la profundidad, las coordenadas *X* y *Y* se ajustaron proporcionalmente para mantener la coherencia espacial de la nube de puntos. Esto aseguró que la nube de puntos mantuviera su estructura y relaciones espaciales.
- **Aplicación de la Función de Regresión Lineal**: Para refinar aún más la precisión de las nubes de puntos, se integró una función de regresión lineal, la cual fue entrenada con 61 pares de imágenes del *dataset* "*Ground Truth*". Esta función ajustó los valores de profundidad para alinear mejor los resultados con los obtenidos por los métodos SGBM, con su respectivo filtro WLS, RAFT-Stereo y Selective-IGEV a unidades reales en escala de 100 unidades = 1 metro. La efectividad de este proceso se puede evidenciar en el ejemplo mostrado en la Figura 13.

Figura 13

Corrección mediante regresión lineal



Nota: La Figura 13 muestra el resultado de una corrección de datos usando regresión lineal.

Implementación y Resultados

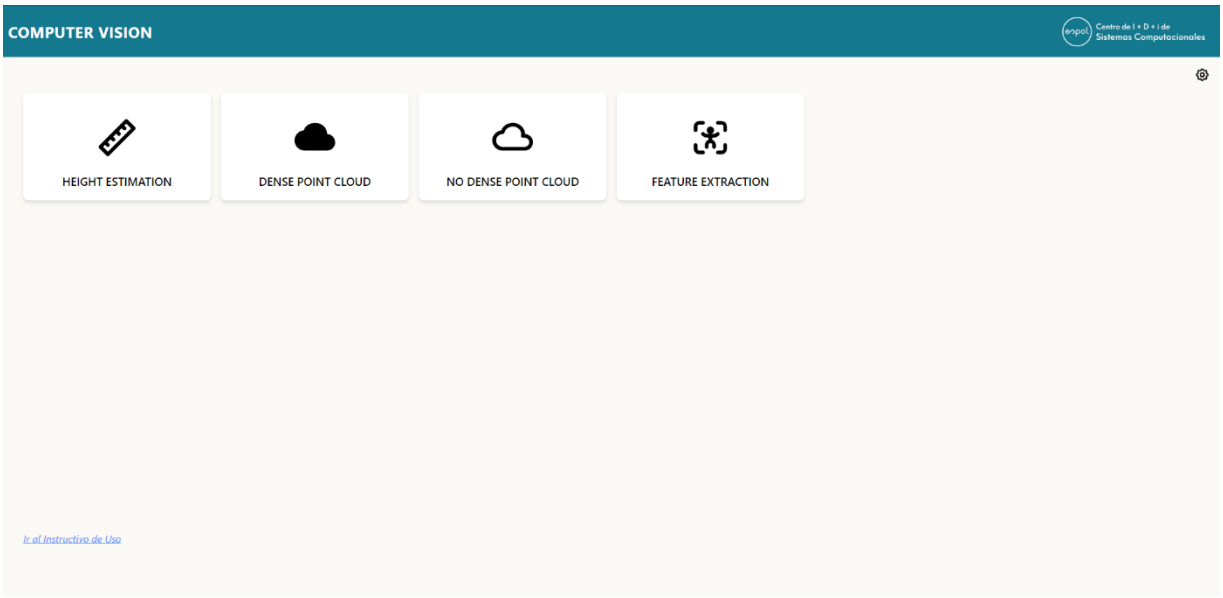
Para evaluar la eficacia de las correcciones, se compararon nuevamente los valores de profundidad corregidos con los valores del *dataset* "G*round Truth*". Las mejoras en las estimaciones de profundidad se visualizaron mediante gráficas comparativas, mostrando los valores estimados antes y después de la corrección en relación con los valores reales del *dataset* "*Ground Truth*". Estas visualizaciones ayudaron a identificar las áreas donde las correcciones fueron más efectivas y donde aún podrían requerirse ajustes adicionales.

Con la precisión de las nubes de puntos refinada y corregida, se procedió a la creación de una plataforma *demo* de visión estereoscópica la cual se puede observar en la Figura 14. Esta plataforma integra la generación de nubes de puntos corregidas, así como las diferentes configuraciones que se les puede dar, diversos módulos adicionales que se basan en estas nubes y perfiles de calibración para las diferentes cámaras y/o métodos de calibración usados.

La plataforma utiliza FastAPI para el *backend* y Astro en integración con React para el *frontend*. FastAPI es un *framework* moderno y rápido para construir APIs con Python, conocido por su eficiencia y facilidad de uso, en la Figura 15 se puede apreciar la documentación de la API. Astro React, por su parte, es una tecnología que permite construir interfaces de usuario de manera reactiva, combinando las ventajas de Astro y React para crear aplicaciones web rápidas y optimizadas.

Figura 14

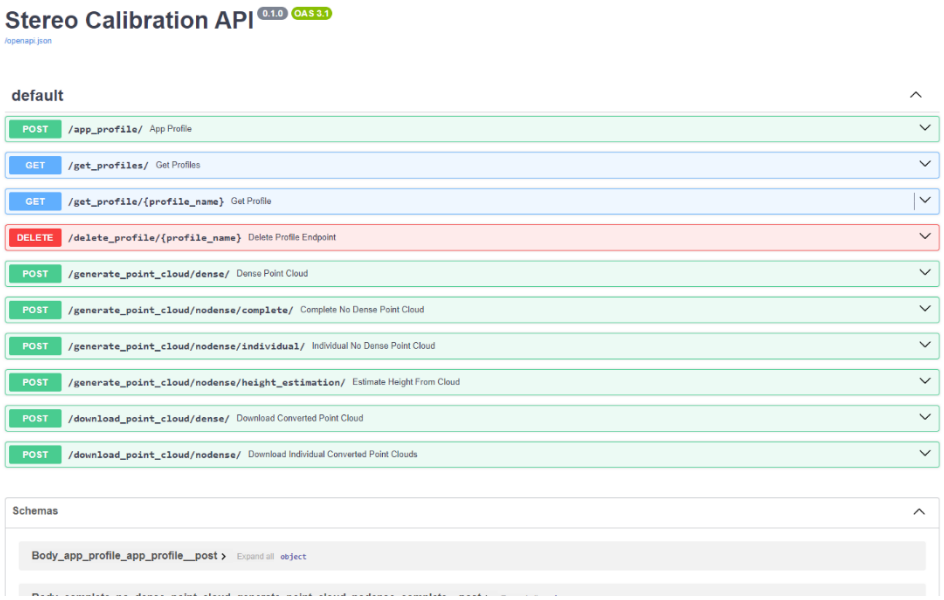
Página Principal de la plataforma desarrollada



Nota: La Figura 14 muestra la pantalla principal del *frontend* de la plataforma demo de estereovisión.

Figura 15

Página de documentación de la API Rest desarrollada



Nota: La Figura 15 muestra la documentación de los múltiples *endpoints* del *backend* de la plataforma de estereovisión.

Capítulo 3

3.1 Resultados y Análisis

En esta sección se presentan las diferentes nubes de puntos generadas y se analizan los datos resultantes obtenidos en la generación de profundidad y altura de personas generadas mediante los métodos RAFT-Stereo, Selective-IGEV y WLS-SGBM, tanto en su forma original

como tras la aplicación de corrección realizada. Nótese que, no se ha realizado un análisis de SGBM sin el filtro WLS debido a que este filtro tiene un propósito de disminución de ruido mas no una alteración de la forma en que SGBM calcula la disparidad base. Para todo este análisis se ha hecho uso de 77 pares de imágenes estereoscópicas las cuales no fueron participes del proceso para la generación de la función de regresión lineal antes mencionada.

3.1.1 Generación de Nube de Puntos

En la Figura 16 se pueden observar las nubes de puntos densas generadas mediante cada uno de los métodos, revelando diferencias notables en términos de ruido, forma y apariencia general de la nube, tanto en vista frontal como lateral.

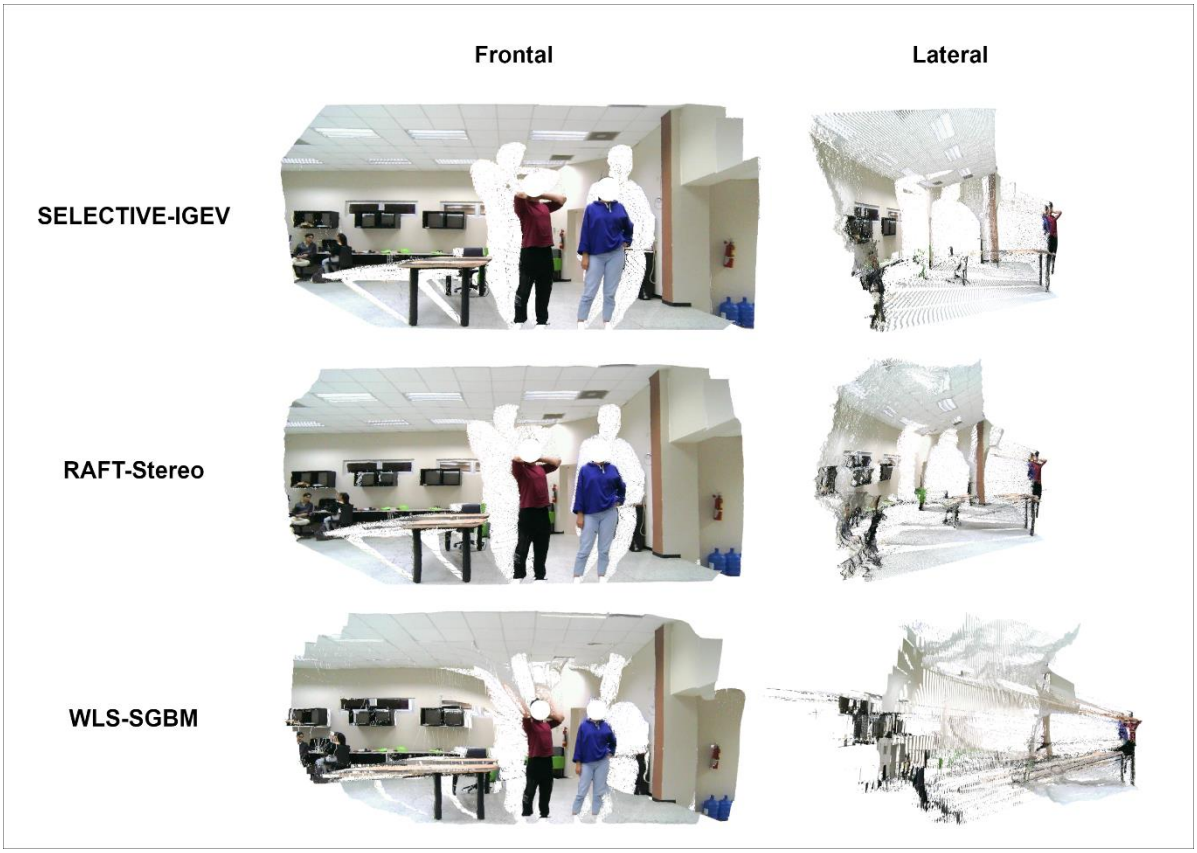
En primer lugar, el método Selective-IGEV destaca por generar la nube de puntos con la menor cantidad de ruido entre los tres métodos. En ambas vistas, los bordes de las personas, los objetos, los techos y las paredes son más definidos, con menos artefactos que distorsionen la forma de la nube. En cuanto a la forma de la nube, esta es la más coherente y cercana a la realidad, representando personas y objetos en la escena con detalles más precisos y con menos deformación en los contornos. La apariencia general de la nube de puntos generada por Selective-IGEV es mucho más nítida y organizada, con una representación espacial coherente que refleja mejor la realidad de la escena capturada.

Por otro lado, el método RAFT-Stereo muestra una cantidad de ruido ligeramente mayor en comparación con Selective-IGEV, tanto en la vista frontal como lateral. Los bordes de las personas son ligeramente menos definidos que en Selective-IGEV, aunque sigue habiendo claridad en algunas áreas. RAFT-Stereo también ofrece una estructura coherente con una ligera cantidad de ruido en las áreas planas, como las paredes, aunque persisten algunos artefactos en la representación de la profundidad. En términos de la forma de la nube, RAFT-Stereo presenta una representación bastante coherente y poco deformada de personas y objetos, aunque aún existen áreas con precisión limitada, como en los detalles de las extremidades y otros bordes más finos. En general, RAFT-Stereo ofrece una nube de puntos organizada y con una buena representación

de las estructuras espaciales, aunque con margen de mejora en términos de precisión y reducción de ruido.

Figura 16

Visualización de Nube de Puntos Densa



Nota: En la Figura 16 se puede apreciar una visualización 3D de dos vistas, una vista frontal y una vista lateral.

Finalmente, el método WLS-SGBM es el que presenta un mayor nivel de ruido en la nube de puntos generada en comparación con los métodos ML, especialmente en la vista frontal. El ruido es considerable, particularmente en los contornos de las personas y objetos en la escena, lo que resulta en bordes difusos y una nube de puntos menos precisa en cuanto a su estructura geométrica. La forma de la nube de puntos, aunque relativamente coherente con la escena, sufre de imprecisiones en los bordes de los objetos y personas, generando una representación menos definida y fiel a la realidad, con figuras más gruesas y menos detalladas. En conjunto, la nube de puntos generada por WLS-SGBM muestra una falta de cohesión en las estructuras, con

artefactos que distorsionan la geometría esperada, haciendo que la nube de puntos se vea menos organizada y precisa en comparación con los otros métodos analizados.

Por otra parte, al comparar los métodos Selective-IGEV, RAFT-Stereo y WLS-SGBM en la generación de nube de puntos no densa se observan diferencias clave que impactan la precisión y utilidad de las nubes de puntos generadas, notándose en la Figura 17.

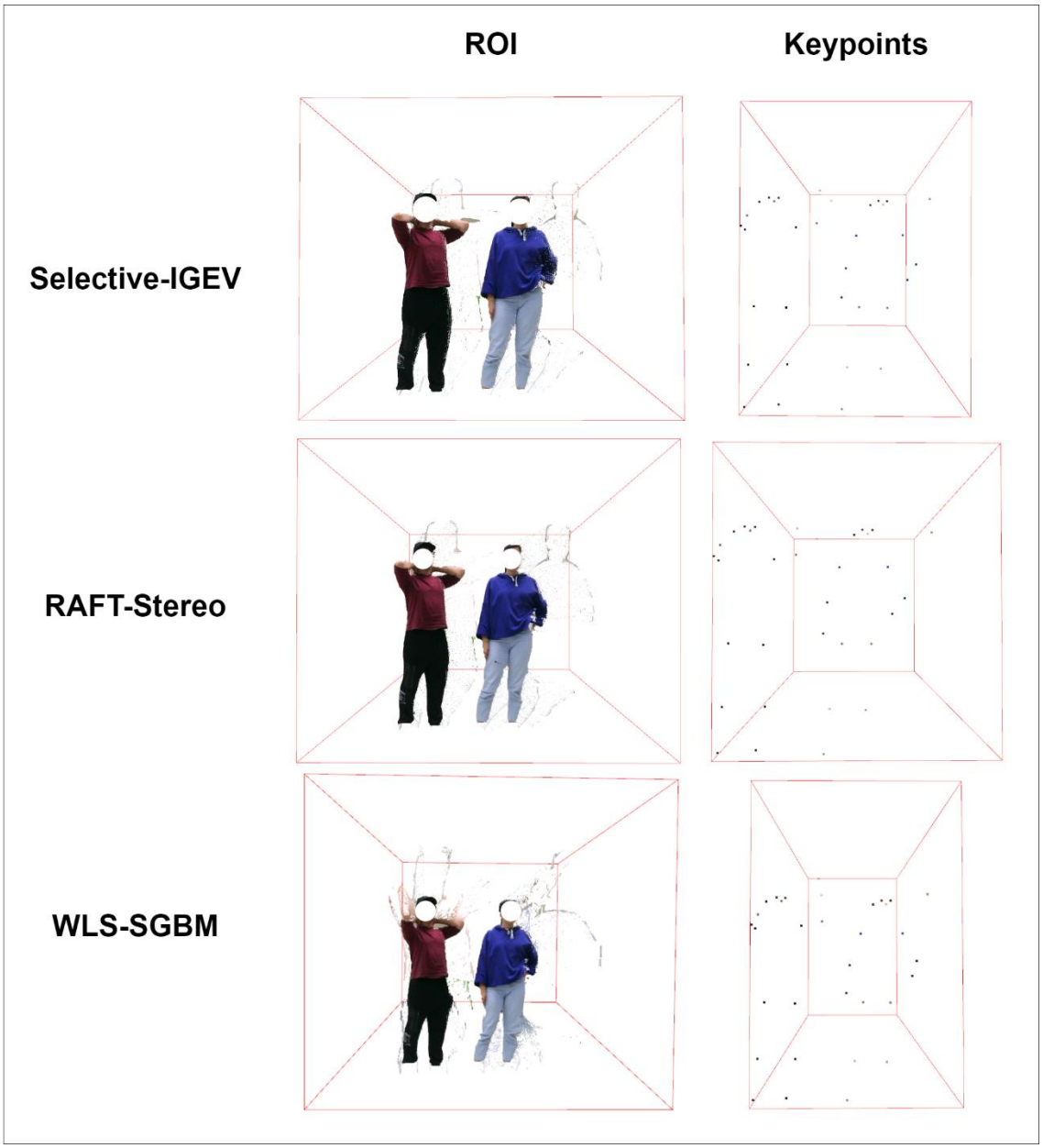
Comenzando con Selective-IGEV, el ruido en la nube de puntos es notablemente bajo en comparación con los otros métodos. Esta reducción del ruido resulta en una nube de puntos más clara y precisa, con una mejor definición de las formas y una representación espacial más coherente dentro del ROI. La claridad lograda facilita la interpretación y análisis de la nube de puntos, haciéndola especialmente adecuada para aplicaciones que requieren alta precisión en la representación de personas en una escena.

En cuanto a la distribución de *Keypoints,* Selective-IGEV destaca por su precisión. Los puntos clave se concentran principalmente en las áreas de interés y hay una menor cantidad de puntos en áreas irrelevantes. Esta precisión en la detección de *Keypoints* sugiere que Selective-IGEV es más eficaz en la identificación de características estructurales importantes, mejorando la utilidad de la nube de puntos en aplicaciones que dependen de la exactitud en la detección y representación de detalles específicos.

Por otro lado, RAFT-Stereo presenta un nivel de ruido moderado en la nube de puntos. El ruido es ligeramente más disperso que en Selective-IGEV estando presente en forma de puntos aislados que no corresponden a características estructurales significativas dentro del ROI. Esta presencia de ruido indica que RAFT-Stereo aún tiene dificultades para discriminar entre los detalles relevantes, lo que impacta la calidad general de la nube de puntos.

Figura 17

Visualización de Nube de Puntos No Densa



Nota: En la Figura 17 se muestra una visualización 3D de las dos variantes de nubes de puntos no densas, *Keypoints* y ROI.

Respecto a la distribución de *Keypoints*, RAFT-Stereo muestra una coherencia similar en comparación con Selective-IGEV. La distribución de puntos clave es más alineada con las características visibles de la escena. Esto indica que, aunque RAFT-Stereo posee falencias al momento de calcular ROI esto no se transporta al representar los *Keypoints* esto puede ser producto de que la cantidad de *Keypoints* es menor que la de ROI y estos casi siempre se encuentran en zonas con menor cantidad de ruido.

Finalmente, WLS-SGBM muestra un nivel significativo de ruido en la nube de puntos generada. Este ruido es especialmente visible en los contornos de las personas dentro de la ROI, manifestándose como puntos dispersos alrededor de los bordes y en áreas donde no hay características estructurales claras. La presencia de estos puntos ruidosos sugiere que el método puedo ser interferido por factores externos como la iluminación ambiental, lo que afecta negativamente la claridad y precisión de la nube de puntos. Dentro de la ROI, el ruido es más pronunciado, comprometiendo la definición de las formas y la coherencia general de la escena representada, lo que dificulta la interpretación de la nube de puntos en aplicaciones donde la precisión es crucial.

En cuanto a la distribución de *Keypoints*, WLS-SGBM presenta una distribución algo inconsistente. Aunque los keypoints se concentran en áreas donde se esperaría encontrar características importantes, la distribución no es uniforme como en los modelos de ML. Esta imprecisión en la detección reduce la utilidad de la nube de puntos en aplicaciones que dependen de una visualización precisa y utilidad de características geométricas.

3.1.2 Estimación de Profundidad

En esta subsección de resultados se analizará el comportamiento de la profundidad estimada utilizando las nubes de puntos no densas de *Keypoints* en conjunto con la metodología de cálculo de profundidad y altura mencionadas en la sección final de metodología a la par de los diferentes métodos, RAFT-Stereo, Selective-IGEV y WLS-SGBM, tal y como se puede observar en la Figura 18, la primera columna de gráficos corresponde a resultados sin corrección mientras que los gráficos de la segunda columna presentan la corrección dimensional y de profundidad realizada.

El análisis de la estimación de profundidad antes y después de aplicar correcciones revela diferencias importantes en el desempeño de los distintos modelos. En primer lugar, el método RAFT-Stereo sin corrección de escala muestra una sobreestimación significativa de la profundidad. Las profundidades estimadas oscilan entre 1500 y 2500 unidades, muy por encima

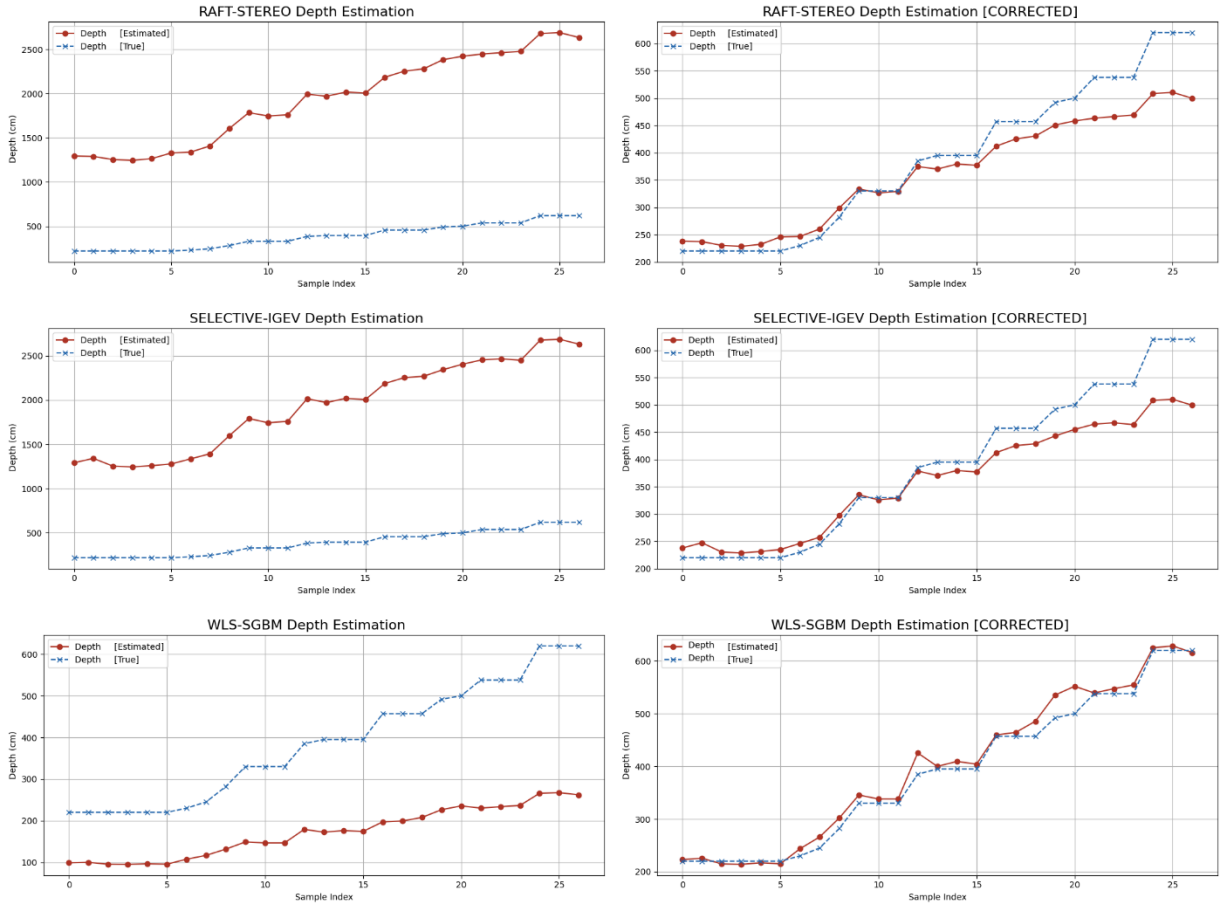
de los valores reales del "*Ground Truth*", que varían entre 200 y 600 unidades. Este sesgo alcanza hasta más de 2000 unidades en las profundidades más altas, lo que sugiere que el modelo tiene dificultades para manejar correctamente las correspondencias a grandes distancias. En consecuencia, se observa una sobreestimación sistemática en todas las muestras, con una mayor desviación en las distancias más grandes. Esto indica un sesgo inherente en el modelo que compromete su precisión, probablemente debido a la naturaleza de predicción generativa de los modelos de ML.

Por otro lado, el modelo Selective-IGEV sin corrección presenta un comportamiento similar al de RAFT-Stereo, con estimaciones que también varían entre 1500 y 2500 unidades. Y es que, al igual RAFT-Stereo, aún existe una desviación significativa, con diferencias de hasta 2000 unidades en las profundidades más altas. Esto sugiere que, aunque Selective-IGEV posee una arquitectura modificada comparándola con RAFT-Stereo, el *dataset* común usado en los modelos preentrenados sesga a ambos modelos de una forma similar.

En contraste, el método WLS-SGBM sin corrección subestima sistemáticamente la profundidad, como se ilustra en la Figura 18. En este caso, las estimaciones varían entre 100 y 300 unidades en la mayoría de los casos, lo cual, aunque más cercano a la realidad, aún presenta diferencias significativas de más de 300 unidades respecto a los valores reales. La tendencia hacia la subestimación refleja una posible falla en la escala dimensional utilizada por WLS-SGBM, lo que limita su utilidad en aplicaciones que requieren alta precisión en la estimación de profundidad en dimensiones reales.

Figura 18

Estimación de Profundidad



Nota: La Figura 18 muestra los resultados obtenidos en el cálculo de la profundidad estimada usando los diferentes métodos de cálculo de disparidad, tanto con corrección como sin ella.

Al aplicar las correcciones, RAFT-Stereo corregido muestra una mejora notable. Como se observa en la columna derecha de la Figura 18, las profundidades corregidas se alinean con el rango real de 200 a 600 unidades, reduciendo las diferencias con el "*Ground Truth*" a 120 unidades en las profundidades más altas y menos de 50 unidades en las más bajas. Esto indica que la corrección ha sido muy efectiva, aunque aún cabe notar que el error tiende a aumentar conforme aumenta la profundidad. De esta manera, las fluctuaciones se han minimizado y la tendencia hacia la sobreestimación se ha corregido sustancialmente, mejorando la confiabilidad del modelo.

De manera similar, el modelo Selective-IGEV corregido también muestra una mejora significativa tras la corrección de escala. Las estimaciones ahora oscilan entre 200 y 600 unidades, alineándose con los valores del "*Ground Truth*". Así, las correcciones han resultado en

diferencias mínimas, con menos de 50 unidades de desviación en la mayoría de las muestras al igual que RAFT-Stereo, aunque todavía persisten algunas diferencias en las mayores profundidades. Esto sugiere que, al igual que con RAFT-Stereo el proceso de corrección ha sido altamente efectivo, especialmente en las zonas de profundidad media o baja, similar a lo observado en RAFT-Stereo. Sin embargo, para ambos modelos la falta de precisión en las mayores profundidades sugiere que ambos modelos podrían beneficiarse de un entrenamiento adicional dentro de estas condiciones de mayores profundidades.

Finalmente, el modelo WLS-SGBM corregido destaca por su precisión tras la corrección de escala. Como se puede observar en la columna derecha de la Figura 18, las estimaciones corregidas se encuentran en el rango de 200 a 600 unidades, lo que representa una mejora significativa. A diferencia de los métodos basados en ML, las nubes de puntos generadas con WLS-SGBM no parecen verse afectadas por el aumento de la profundidad, lo que sugiere que su naturaleza más vinculada a un cálculo matemático directo y exacto, específicamente optimizaciones, lo que le otorga una ventaja sobre los modelos generativos.

3.1.3 Análisis de la Altura en Función de la Profundidad

En esta sección, se analiza la estimación de altura utilizando los métodos RAFT-Stereo, Selective-IGEV y WLS-SGBM, donde la altura de una persona se mantiene constante en 169 cm, mientras que la profundidad varía a lo largo de diferentes distancias. Este análisis permite evaluar cómo cada método maneja la variación de la distancia del objeto respecto a la cámara, y cómo esto afecta la precisión en la estimación de la altura. Se obtuvieron el conjunto de gráficos para análisis representado en la Figura 19, en donde los gráficos de la izquierda representan la estimación de la altura antes de aplicar la corrección, mientras que en los gráficos de la derecha la corrección ha sido aplicada.

El análisis de los métodos de estimación de altura en función de la profundidad revela diversas tendencias significativas antes y después de aplicar la corrección de escala. En primer lugar, el método RAFT-Stereo sin corrección de escala muestra una clara tendencia hacia la

sobreestimación. Por ejemplo, en distancias cortas, como de 200 a 300 cm, la altura estimada oscila entre 590 y 650 cm, lo cual es significativamente superior a la altura real de 169 cm. A medida que la distancia aumenta, la altura estimada disminuye, pero aun así permanece muy por encima de la altura real, alcanzando un mínimo de 470 cm. Esto sugiere que RAFT-Stereo, sin una corrección adecuada, sufre de errores de escala que llevan a estimaciones exageradas en distancias cortas.

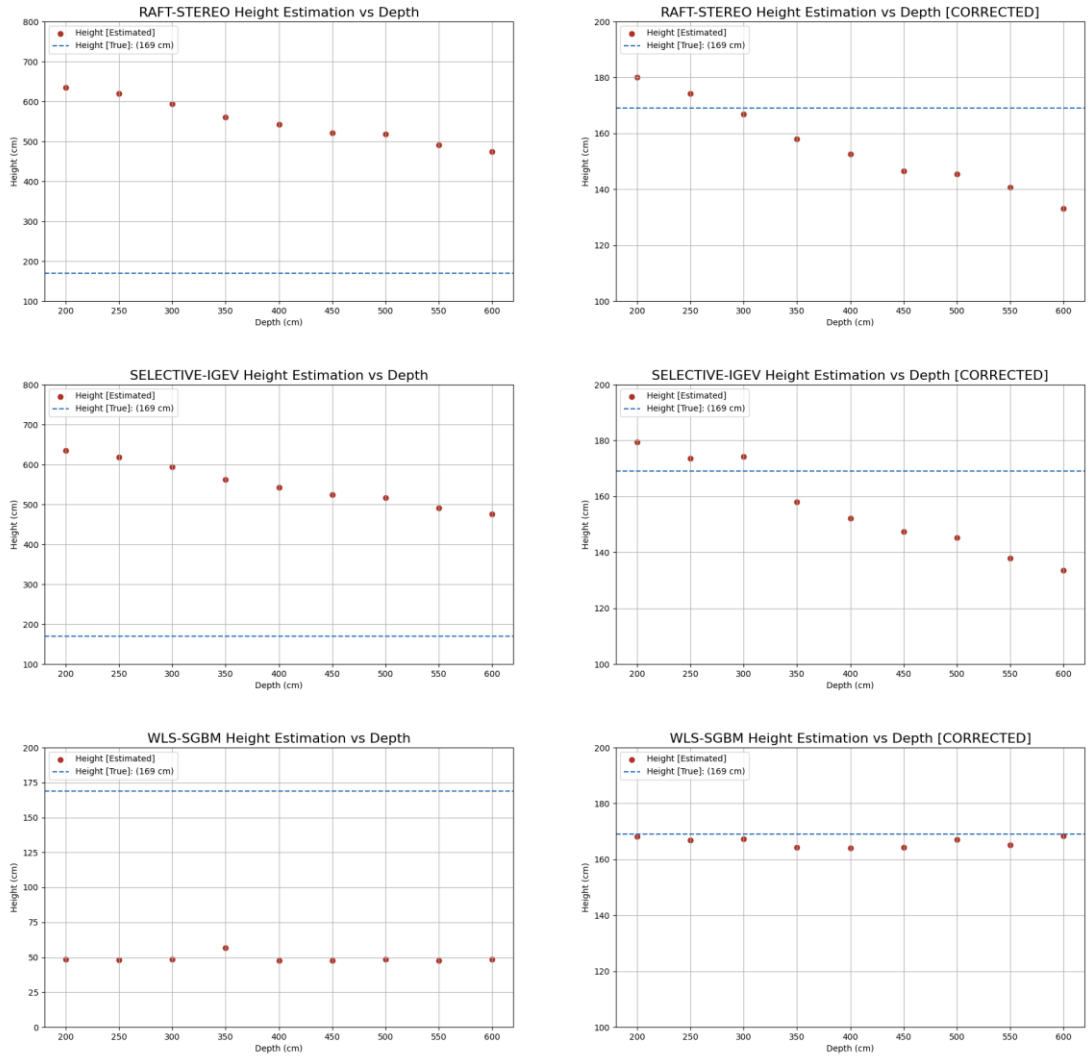
De manera similar, el método Selective-IGEV sin corrección de escala presenta un patrón comparable al de RAFT-Stereo, con alturas estimadas que también están significativamente sobreestimadas. Es decir, desde una altura máxima de 634 cm en profundidades de 200 cm, las estimaciones disminuyen gradualmente hasta aproximadamente 475 cm. Aunque Selective-IGEV tiende a tener estimaciones ligeramente menores que RAFT-Stereo, la sobreestimación sigue siendo considerable, lo que indica una fuerte sensibilidad a la falta de corrección de escala.

Por otro lado, el método WLS-SGBM sin corrección de escala se diferencia de los otros métodos al mostrar un comportamiento inverso. En este caso, las alturas estimadas están consistentemente subestimadas, oscilando entre 50 y 60 cm en todas las distancias, muy por debajo de la altura real de 169 cm. Este patrón constante sugiere que WLS-SGBM podría estar colapsando la información de profundidad, resultando en una subestimación que no refleja adecuadamente las variaciones de distancia.

Sin embargo, al aplicar la corrección de escala, RAFT-Stereo corregido muestra mejoras significativas. En efecto, la precisión en la estimación de la altura mejora considerablemente, especialmente en distancias entre 250 y 300 cm, donde las alturas estimadas se acercan mucho más a la altura real, oscilando entre 170 y 175 cm. No obstante, a profundidades mayores de 350 cm, las estimaciones comienzan a disminuir nuevamente, alcanzando un mínimo de 133 cm a 600 cm de profundidad. Aunque la corrección mitiga en gran medida la sobreestimación inicial, persisten ciertos errores, lo que sugiere que la corrección es efectiva pero no completamente suficiente.

Figura 19

Análisis de la Altura en función de la Profundidad



Nota: La Figura 19 muestra los resultados obtenidos en el cálculo de una altura de 169 cm, estimada usando los diferentes métodos de cálculo de disparidad, tanto con corrección como sin ella.

Al aplicar la corrección de escala, como se puede observar en la columna derecha de la Figura 19, Selective-IGEV corregido también experimenta una reducción significativa en los errores de estimación. Particularmente, en distancias cortas de 250 a 300 cm, las alturas estimadas se acercan a los 174 cm. Sin embargo, a mayores profundidades, la altura estimada disminuye por debajo de 160 cm, alcanzando un mínimo de 133 cm a 600 cm. Esto indica que,

aunque la corrección mejora la precisión, Selective-IGEV aún enfrenta desafíos para mantener la exactitud en distancias mayores.

Finalmente, con la corrección de escala aplicada, WLS-SGBM corregido destaca por su estabilidad y precisión. En todas las profundidades, las estimaciones se mantienen cerca del valor real de 169 cm, con variaciones mínimas entre 165 y 170 cm. Por lo tanto, este resultado subraya la robustez del método WLS-SGBM cuando se aplica la corrección adecuada, lo que lo convierte en una opción ideal para aplicaciones que requieren alta precisión en la estimación de altura y profundidad.

3.1.4 Análisis de la Altura en Función de una Profundidad Constante

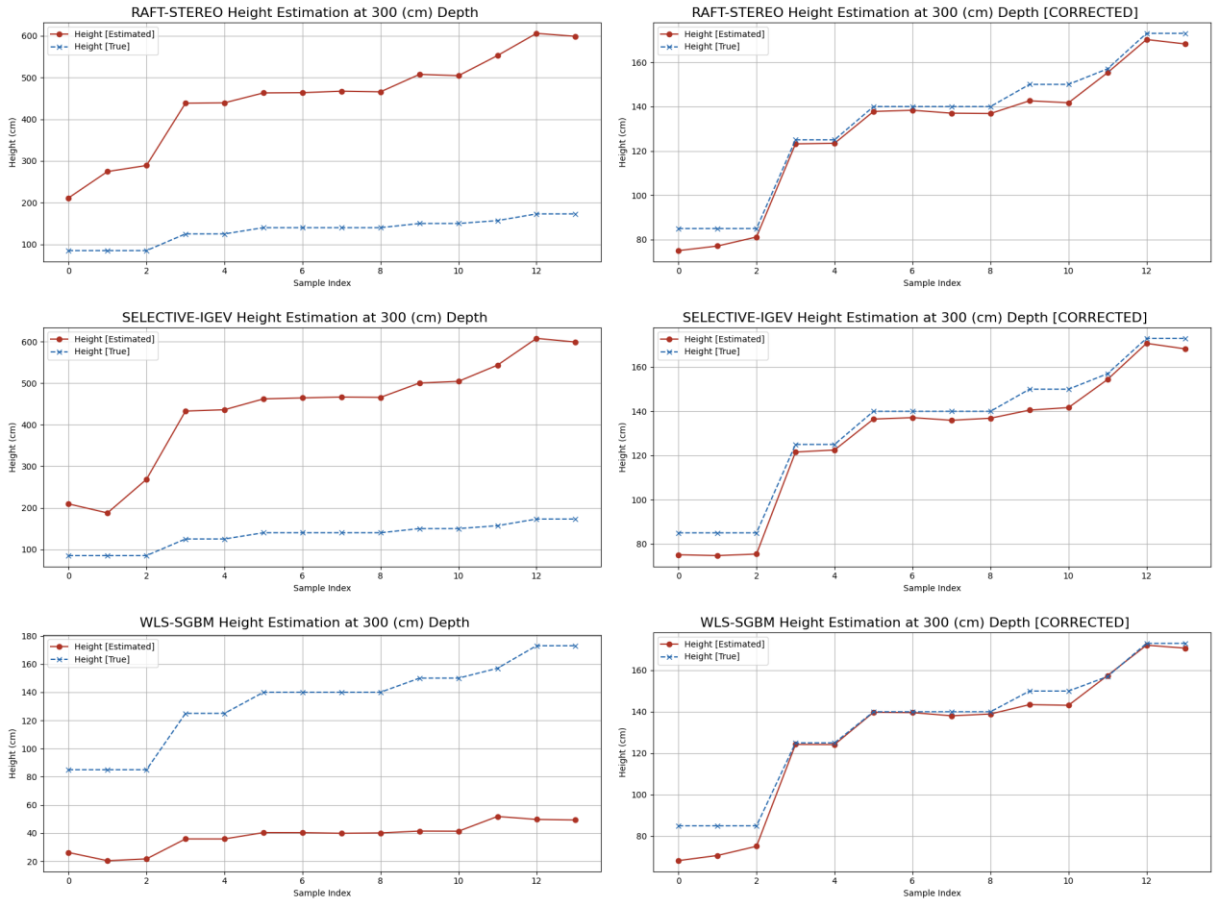
En esta sección, se analiza la estimación de altura utilizando los tres métodos antes mencionados. A diferencia del análisis anterior, en este caso la profundidad se mantiene constante en 300 cm, mientras que la altura de una persona varía. Este enfoque permite evaluar cómo cada método maneja las variaciones en altura cuando la profundidad es fija, proporcionando una visión de su precisión y consistencia. A continuación, en la Figura 20 se muestran los gráficos para el análisis persistente, en donde la primera columna corresponde a los gráficos con datos sin corregir mientras que en la segunda columna corresponde a gráficos con datos corregidos.

El análisis de la estimación de altura en función de una profundidad constante, tanto antes como después de aplicar correcciones de escala, evidencia diferencias importantes en la precisión y consistencia de los distintos métodos. En primer lugar, el método RAFT-STEREO sin corrección de escala muestra una sobreestimación significativa de la altura, como se observa en la Figura 20. En este caso, las alturas estimadas varían entre 200 cm y 600 cm, mientras que las alturas reales oscilan entre 80 cm y 170 cm. Esta discrepancia, especialmente pronunciada en las muestras con mayores alturas, revela errores que pueden alcanzar hasta 430 cm por encima de la altura verdadera. Este comportamiento sugiere que RAFT-Stereo amplifica de manera

inconsistente las diferencias de altura, particularmente como se muestra en este escenario donde la variación de altura es más notable y la profundidad se mantiene constante.

Figura 20

Análisis de la Altura en función de una Profundidad Constante



Nota: La Figura 20 muestra los resultados obtenidos en el cálculo de alturas variables a 300 cm de profundidad, usando los diferentes métodos de cálculo de disparidad, tanto con corrección como sin ella.

El método Selective-IGEV sin corrección de escala sigue un patrón similar al de RAFT-Stereo, con sobreestimaciones de altura que también oscilan entre 200 cm y 600 cm, como se puede ver en la Figura 20. Aunque este método posee una arquitectura más actual basado en módulos de atención, al igual que RAFT-Stereo las sobreestimaciones siguen siendo considerables. En particular, las discrepancias son grandes en las muestras con mayores alturas,

lo que indica que Selective-IGEV, sin corrección, también tiende a amplificar las alturas más allá de lo real aun cuando la profundidad es fija.

En contraste, el método WLS-SGBM sin corrección de escala muestra un comportamiento opuesto a RAFT-STEREO y Selective-IGEV. En lugar de sobreestimar, WLS-SGBM subestima sistemáticamente las alturas, con valores estimados que varían entre 20 cm y 60 cm. Esta subestimación generalizada sugiere que este método tiende a comprimir el rango de alturas, lo que puede ser indicativo de una limitación en su capacidad para capturar diferencias de altura cuando la profundidad es constante. Aunque es más estable que los otros métodos, la falta de ajuste a la escala real de altura es evidente.

Sin embargo, al aplicar la corrección de escala, RAFT-Stereo corregido muestra una mejora significativa en la precisión de las estimaciones de altura, como se refleja en la columna derecha de la Figura 20. Las alturas estimadas ahora oscilan entre 80 cm y 170 cm, lo cual está mucho más cerca de las alturas verdaderas. Aunque persisten pequeñas discrepancias, especialmente en las alturas mayores, la corrección ha reducido considerablemente los errores.

De manera similar, el modelo Selective-IGEV corregido también muestra una mejora notable tras la aplicación de la corrección de escala, como se evidencia en la columna derecha de la Figura 20. Las alturas estimadas ahora están más alineadas con las alturas verdaderas, variando entre 80 cm y 170 cm. La corrección ha sido efectiva en reducir las sobreestimaciones, aunque aún persisten pequeñas discrepancias, especialmente en alturas mayores. Este resultado sugiere que los modelos ML como RAFT-Stereo y Selective-IGEV pueden beneficiarse significativamente de las correcciones post-procesamiento para mejorar la precisión en la estimación de altura al menos en distancias iniciales, tal y como se muestra en este caso en concreto.

Finalmente, como de igual manera se muestra en la columna derecha de la Figura 20, el método WLS-SGBM corregido, que ya mostraba mayor estabilidad incluso sin corrección, se beneficia aún más con la aplicación de la corrección de escala. Las alturas estimadas se alinean

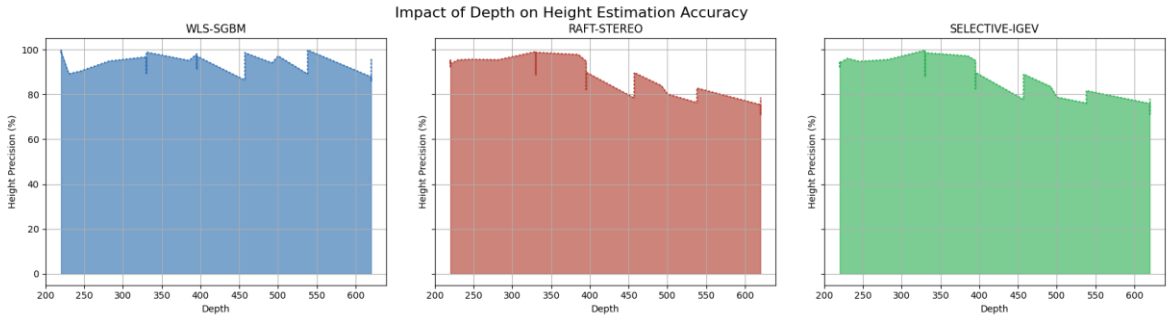
casi perfectamente con las alturas verdaderas, con variaciones mínimas entre las muestras. Este comportamiento reafirma la robustez de WLS-SGBM cuando se aplica la corrección adecuada, demostrando que este método puede ofrecer un alto grado de precisión y consistencia en la estimación de altura cuando la profundidad es constante.

3.1.5 Análisis de la Precisión en Estimación de Alturas a Profundidades Variables

Se definieron dos rangos de profundidad para evaluar la precisión de los modelos: el rango óptimo de 200 cm a 400 cm y el rango extendido de 401 cm a 600 cm. Estos rangos fueron seleccionados para observar el desempeño en condiciones óptimas y en condiciones de mayor profundidad, evidenciándose así en la Figura 21.

Figura 21

Impacto de la Profundidad en la Precisión del Cálculo de la Altura.



Nota: La Figura 21 muestra la precisión de los metodos corregidos en el cálculo de alturas variables a diferentes profundidades, utilizando los distintos métodos de cálculo de disparidad.

RAFT-Stereo por ejemplo dentro del rango óptimo de 200 cm a 400 cm posee exactamente un 93.68% de precisión promedio y aunque este modelo ML muestra una alta precisión en la estimación de la altura. Este nivel de precisión es adecuado, pero las fluctuaciones menores observadas anteriormente podrían estar afectando la consistencia del modelo en ciertas profundidades. Al extender el análisis al rango extendido de 401 cm a 600 cm de profundidad, la precisión promedio de RAFT-Stereo disminuye significativamente a 81.15% de precisión promedio. Esta caída en la precisión refuerza que el modelo tiene dificultades para mantener su desempeño en condiciones de mayor profundidad, lo cual podría estar relacionado

con la complejidad adicional que introduce este rango en la estimación de la altura la cual se vincula a la estimación de profundidad en donde los modelos ML de este caso en particular y a mostraban precedentes de baja precisión.

Siguiendo con Selective-IGEV, posee una precisión promedio de 93.62% dentro del rango optimo observable y es que, Selective-IGEV tiene una precisión promedio muy cercana a la de RAFT-Stereo, pero ligeramente inferior. Este modelo también muestra una tendencia similar al extender el análisis al rango extendido de 401 cm a 600 cm, donde la precisión promedio disminuye a 80.35%. Este comportamiento sugiere que Selective-IGEV, al igual que RAFT-Stereo, experimentaría de antecedentes de baja precisión lo cual puede estar relacionado a su *dataset* de entrenamiento en común.

Finalmente, el método WLS-SGBM posee una precisión promedio de exactamente 95.99% dentro del rango optimo, superando a ambos modelos basados en ML en términos de precisión en la estimación de la altura en el rango de 200 a 400 cm. Además, al extender el análisis al rango extendido de 401 cm a 600 cm, si bien WLS-SGBM también muestra una reducción en la precisión al igual que los modelos ML, esta es muchos menos pronunciada, alcanzando un 93.23% de precisión promedio dentro de este rango. Estos resultados siguen siendo superiores en comparación con RAFT-Stereo y Selective-IGEV, lo que nuevamente demuestra que WLS-SGBM maneja mejor las variaciones en profundidad, manteniendo un desempeño más estable en condiciones más desafiantes.

Este resultado era previsible, ya que WLS-SGBM mostraba un comportamiento cercano a los valores reales en los análisis sin corregir, aunque tendía a subestimar tanto la profundidad como la altura. Sin embargo, la corrección aplicada ha mejorado significativamente su precisión en este rango específico. A diferencia de los dos modelos basados en ML, esta mejora se mantiene de manera bastante uniforme, independientemente del aumento en la profundidad. Este rendimiento superior en este contexto particular sugiere que WLS-SGBM maneja mejor ciertas

configuraciones geométricas o que la corrección aplicada resultó ser especialmente efectiva en este modelo.

Capítulo 4

4.1 Conclusiones y Recomendaciones

4.1.1 Conclusiones

Tras aplicar meticulosamente las fases necesarias de la propuesta y analizar detenidamente los resultados obtenidos, se derivan las siguientes conclusiones primordiales que fundamentan los objetivos específicos del proyecto.

1. Se logró identificar con precisión los puntos comunes en las imágenes estéreo, permitiendo calcular la disparidad y, por ende, la profundidad de los objetos en la escena. Sin embargo, se observó que, aunque RAFT-Stereo y Selective-IGEV son capaces de capturar la variación en la disparidad, su consistencia disminuye a medida que aumenta la profundidad. Esto contrasta con SGBM, que mantiene un ritmo más coherente y predecible en el cálculo de disparidades, independientemente de la profundidad.
2. La investigación reveló diferencias significativas en la consistencia de los métodos de generación de mapas de disparidad. RAFT-Stereo y Selective-IGEV, aunque muestran un comportamiento prometedor, tienden a perder consistencia en la estimación de disparidades a mayores profundidades, lo que se traduce en variaciones erráticas que afectan la precisión general. En cambio, SGBM, basado en un enfoque más tradicional, mantiene una estabilidad superior en la estimación de disparidades a lo largo de todo el rango de profundidades.
3. Se logró generar una representación tridimensional estandarizada mediante la aplicación de procesos de normalización y corrección de profundidad. Aunque las correcciones mejoraron la precisión de RAFT-Stereo y Selective-IGEV, la mayor consistencia y estabilidad observada en SGBM sugiere que este método es más adecuado para aplicaciones que requieren una precisión constante en la visualización de geometrías espaciales.
4. La precisión de las nubes de puntos generadas fue validada a través de la medición de la altura de personas en la escena. Los métodos basados en ML, aunque mejoraron tras las correcciones, mostraron una tendencia a variaciones significativas en la profundidad y altura a mayores distancias, sugiriendo que estos errores podrían deberse a una falta de datos de

entrenamiento dentro de rangos con mayor profundidad. Por otro lado, SGBM, al mantener un ritmo constante en la estimación de disparidades, demostró ser más confiable para la medición precisa de la altura en diversas condiciones.

5. Modelos generativos como RAFT-Stereo y Selective-IGEV pueden ser capaces de competir en precisión con métodos avanzados basados en coincidencia de bloques, como es el caso de SGBM, y para ciertas aplicaciones podrían llegar a ser mejores. RAFT-Stereo y Selective-IGEV tienen un mejor desempeño al capturar una disparidad homogénea de la escena, creando una mejor representación visual de escenarios tridimensionales que SGBM aun cuando se utiliza WLS para mejorarlo.

6. Se definió una plataforma escalable que integra los módulos de generación de nubes de puntos, adaptable para futuros proyectos en visión por computador y robótica móvil. La plataforma permite una implementación flexible de diferentes métodos, asegurando que se puedan seleccionar y optimizar las técnicas más adecuadas para cada aplicación, especialmente en contextos donde la consistencia y precisión son críticas.

4.1.2 Recomendaciones

Tras culminar el plan de trabajo propuesto, se han identificado las siguientes recomendaciones principales para estudios futuros y mejoras en este campo de investigación:

1. **Evaluación de técnicas de triangulación**: Aunque este proyecto se centró en la generación de nubes de puntos utilizando la triangulación recomendada por los diferentes métodos de disparidad, se sugiere realizar un estudio más detallado sobre cómo las distintas técnicas de triangulación afectan la calidad y precisión de las nubes de puntos generadas. Se podría desarrollar un proyecto dedicado a comparar los métodos de triangulación, incluyendo la influencia de usar diferentes valores de *baseline* (ya sea calculados mediante MATLAB o documentados en la cámara de estereovisión utilizada) y analizar las diferencias que resultan de estas variaciones.

2. **Implementación de sistema en módulos móviles**: Considerando que una de las aplicaciones clave de la generación de nubes de puntos se encuentra en el campo de la robótica, se recomienda la implementación de este sistema en módulos móviles como la NVIDIA Jetson Orin NX. Este tipo de hardware, diseñado específicamente para la inteligencia artificial y el procesamiento intensivo en el borde, permite que el sistema sea embebido directamente en robots, facilitando una integración más estrecha entre la generación de nubes de puntos y las capacidades de decisión del robot.

3. **Entrenamiento adicional para los modelos de ML:** Dado que los modelos de ML mostraron inconsistencias significativas en las estimaciones de profundidad, especialmente en distancias superiores a los 400 cm, se recomienda realizar un reentrenamiento sobre el modelo pre-entrenado enfocado en estos rangos de distancia específicos, con el fin de mejorar la precisión en estas áreas. Es crucial utilizar un conjunto de datos ampliado y diversificado que refuerce la capacidad del modelo para manejar escenas a mayores profundidades.

4. **Análisis de la variabilidad de los parámetros de los lentes**: En la práctica, las diferencias entre los lentes izquierdo y derecho de las cámaras de estereovisión pueden influir significativamente en los resultados. Se recomienda investigar cómo estas variaciones impactan la generación de mapas de disparidad y explorar posibles soluciones para minimizar estos efectos. Este estudio podría incluir la evaluación de métodos para ajustar automáticamente los parámetros de los lentes en función de sus características individuales.

5. **Comparación de técnicas de calibración de cámaras**: La calibración de las cámaras es un paso crucial para la obtención de mapas de disparidad precisos. Se sugiere realizar un estudio comparativo entre diferentes técnicas de calibración, utilizando herramientas como la librería de OpenCV, MATLAB, y otras plataformas de calibración. Este análisis podría proporcionar una guía práctica sobre qué métodos son más efectivos en distintos escenarios y cómo pueden influir en la calidad final de las nubes de puntos generadas.

6. **Implementación de una plataforma web para la visión estereoscópica**: Actualmente, la plataforma desarrollada en este proyecto está diseñada para ejecutarse localmente. Un avance futuro sería llevar esta plataforma a un entorno web, permitiendo su uso remoto sin necesidad de instalar el software en cada máquina. Esto implicaría abordar desafíos relacionados con protocolos de red, ancho de banda, y optimización del rendimiento para garantizar que la plataforma funcione eficientemente en un entorno distribuido.

7. **Generación en tiempo real de nubes de puntos y altura**: Se recomienda desarrollar un sistema que permita la generación en tiempo real de nubes de puntos y estimaciones de altura. Aunque la plataforma actual permite la generación de nubes de puntos de manera discontinua a partir de imágenes capturadas en vivo, un futuro desarrollo podría incluir la actualización continua de las nubes de puntos a intervalos de tiempo establecidos. Además, se sugiere integrar un sistema de acción basado en YOLOv8 para la generación de altura en tiempo real, activando la estimación de altura solo cuando se detecten los 17 *Keypoints* de la persona en la escena. Este enfoque podría optimizar el proceso y garantizar resultados más precisos en aplicaciones donde la rapidez y la precisión son esenciales.

Referencias

Bache, R. M. (1892). Civil and Military Photogrammetry. *Proceedings of the American Philosophical Society*, *30*(138), 229-240.

Barnard, S. T., & Thompson, W. B. (1980). Disparity Analysis of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-2*(4), 333-340. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.1980.4767032

Bentley, J. L. (1990). K-d trees for semidynamic point sets. *Proceedings of the sixth annual symposium on Computational geometry*, 187-197. https://doi.org/10.1145/98524.98564

Bradski, G. R., & Kaehler, A. (2011). *Learning OpenCV: Computer vision with the OpenCV library* (1. ed., [Nachdr.]). O'Reilly.

Chen, C., & Kak, A. (1987). Modeling and calibration of a structured light scanner for 3-D robot vision. *1987 IEEE International Conference on Robotics and Automation Proceedings*, *4*, 807-815. https://doi.org/10.1109/ROBOT.1987.1087958

Deville, É., & Survey, C. T. (1895). *Photographic Surveying: Including the Elements of Descriptive Geometry and Perspective*. Government Print. Bureau.

Farbman, Z., Fattal, R., Lischinski, D., & Szeliski, R. (s. f.). *Edge-Preserving Decompositions for Multi-Scale Tone and Detail Manipulation*.

Fusiello, A., Trucco, E., & Verri, A. (2000). A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, *12*(1), 16-22. https://doi.org/10.1007/s001380050120

Geng, J. (2011). Structured-light 3D surface imaging: A tutorial. *Advances in Optics and Photonics*, *3*(2), 128. https://doi.org/10.1364/AOP.3.000128

Gonzalez Blanco, A. (1998). *Adquisición y modelado tridimensional en visión artificial mediante técnicas de luz estructurada* [Http://purl.org/dc/dcmitype/Text, Universidad Politécnica de Madrid]. https://dialnet.unirioja.es/servlet/tesis?codigo=238583

Gregory, R. L., MacKay, D. M., Whitteridge, D., & Clarke, P. G. H. (1997). Stereo vision and isoluminance. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *204*(1157), 467-476. https://doi.org/10.1098/rspb.1979.0040

Hirschmuller, H. (2005). Accurate and efficient stereo processing by semi-global matching and mutual information. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *2*, 807-814 vol. 2. https://doi.org/10.1109/CVPR.2005.56

Kim, J.-K., Park, B.-S., Kim, W., Park, J.-T., Lee, S., & Seo, Y.-H. (2022). Robust Estimation and Optimized Transmission of 3D Feature Points for Computer Vision on Mobile Communication Network. *Sensors*, *22*(21), Article 21. https://doi.org/10.3390/s22218563

Koschan, A., & Rodehorst, V. (1995). Towards real-time stereo employing parallel algorithms for edge-based and dense stereo matching. *Proceedings of Conference on Computer Architectures for Machine Perception*, 234-241. https://doi.org/10.1109/CAMP.1995.521045

Koschan, A., Rodehorst, V., & Spiller, K. (1996). Color stereo vision using hierarchical block matching and active color illumination. *Proceedings of 13th International Conference on Pattern Recognition*, *1*, 835-839 vol.1. https://doi.org/10.1109/ICPR.1996.546141

Lehmann, G. (1975). *Fotogrametría*. Reverte.

Lipson, L., Teed, Z., & Deng, J. (2021). *RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching* (arXiv:2109.07547). arXiv. http://arxiv.org/abs/2109.07547

Marr, D., Poggio, T., & Brenner, S. (1997). A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *204*(1156), 301-328. https://doi.org/10.1098/rspb.1979.0029

Menze, M., & Geiger, A. (2015). Object scene flow for autonomous vehicles. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3061-3070. https://doi.org/10.1109/CVPR.2015.7298925

Montenegro, R. D., Oliveira, A. L. I., Cabral, G. G., Katz, C. R. T., & Rosenblatt, A. (2008). A Comparative Study of Machine Learning Techniques for Caries Prediction. *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, *2*, 477-481. https://doi.org/10.1109/ICTAI.2008.138

Ogle, K. N. (1952). On the limits of stereoscopic vision. *Journal of Experimental Psychology*, *44*(4), 253-259. https://doi.org/10.1037/h0057643

Oyama, T., & Sato, F. (1967). PERCEIVED SIZE-RATIO IN STEREOSCOPIC VISION AS A FUNCTION OF CONVERGENCE, BINOCULAR DISPARITY AND LUMINANCE. *Japanese Psychological Research*, *9*(1), 1-13. https://doi.org/10.4992/psycholres1954.9.1

Remondino, F., & El-Hakim, S. (2006). Image-based 3D Modelling: A Review. *The Photogrammetric Record*, *21*, 269-291. https://doi.org/10.1111/j.1477-9730.2006.00383.x

Rusu, R. B., & Cousins, S. (2011). 3D is here: Point Cloud Library (PCL). *2011 IEEE International Conference on Robotics and Automation*, 1-4. https://doi.org/10.1109/ICRA.2011.5980567

Sankowski, W., Włodarczyk, M., Kacperski, D., & Grabowski, K. (2017). Estimation of measurement uncertainty in stereo vision system. *Image and Vision Computing*, *61*, 70-81. https://doi.org/10.1016/j.imavis.2017.02.005

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., & Westling, P. (2014). High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. En X. Jiang, J. Hornegger, & R. Koch (Eds.), *Pattern Recognition* (Vol. 8753, pp. 31-42). Springer International Publishing. https://doi.org/10.1007/978-3-319-11752-2_3

Scharstein, D., Szeliski, R., & Zabih, R. (2001). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, 131-140. https://doi.org/10.1109/SMBV.2001.988771

Schops, T., Schonberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., & Geiger, A. (2017). A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2538-2547. https://doi.org/10.1109/CVPR.2017.272

Sombekke, N. (s. f.). *Triangulation for Depth Estimation*.

Su, H., & He, B. (2011). Stereo rectification of calibrated image pairs based on geometric transformation. *International Journal of Modern Education and Computer Science*, *3*(4), 17-24. https://doi.org/10.5815/ijmecs.2011.04.03

Venkatesh, Y. V., Venkatesh, B. S., & Kumar, A. J. (2004). Stereo-disparity estimation using a supervised neural network. *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, 785-793. https://doi.org/10.1109/MLSP.2004.1423046

Wandinger, U. (2005). Introduction to Lidar. En C. Weitkamp (Ed.), *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere* (pp. 1-18). Springer. https://doi.org/10.1007/0-387-25101-4_1

Wang, X., Xu, G., Jia, H., & Yang, X. (2024). *Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching* (arXiv:2403.00486). arXiv. https://doi.org/10.48550/arXiv.2403.00486

Wolf, H. G. E. (1996). Structured lighting for upgrading 2D vision systems to 3D. *Rapid Prototyping*, *2787*, 20-25. https://doi.org/10.1117/12.248593

Žbontar, J., & LeCun, Y. (2016). Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*, *17*(65), 1-32.

Zhang, W. (2010). LIDAR-based road and road-edge detection. *2010 IEEE Intelligent Vehicles Symposium*, 845-848. https://doi.org/10.1109/IVS.2010.5548134

Zhang, Y.-J. (2021). Stereo Vision. En Y.-J. Zhang (Ed.), *Handbook of Image Engineering* (pp. 1321-1353). Springer. https://doi.org/10.1007/978-981-15-5873-3_37

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(11), 1330-1334. https://doi.org/10.1109/34.888718

Zhang, Z., Wang, Y., Jiang, T., & Gao, W. (2011). Stereoscopic learning for disparity estimation. *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, 365-368. https://doi.org/10.1109/ISCAS.2011.5937578