Universidad Zaragoza
1542

Trabajo Fin de Grado

Aprendizaje profundo y geometría para estimación densa de profundidad en pares de imágenes de ojo de pez

---

Deep learning and geometry for dense depth estimation in fisheye image pairs

Autor

Eduardo Pérez Rivasés

Directores

Jesús Bermúdez Cameo

Samuel Bruno Berenguel Baeta

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2023

RESUMEN

Este trabajo fin de grado se centra en la estimación de un mapa de profundidad denso a partir de un par de imágenes estéreo con distorsión de ‘Ojo de Pez’. La estimación de un punto a partir de su proyección en un par de imágenes estéreo es un problema geométricamente sencillo, al conocer la geometría de la cámara utilizada. En cambio, obtener un resultado denso de la imagen al completo es un problema difícil de resolver. Ya que se debe buscar a lo largo de línea epipolar para cada píxel que en este caso es curva debido a la distorsión de la imagen, la solución no es trivial debido a que el rango de búsqueda es infinito. En este trabajo se busca acotar este rango de búsqueda para así simplificar el problema. Para ello se hace uso de una red neuronal o inteligencia artificial (AI), en específico, la red MiDaS la cual estima un mapa denso de profundidad de imágenes monoculares, en el que cada píxel contiene una ‘Semilla de profundidad’, es decir, un valor de distancia por cada píxel. Este mapa denso aporta regiones de incertidumbre para cada píxel, lo que permite acotar considerablemente la búsqueda en la curva epipolar. Sin embargo, hay varias consideraciones a tener en cuenta para poder utilizar este mapa como semilla. En primer lugar, al ser una estimación monocular, el mapa denso proporcionado por la red no tiene la escala correcta. Esta escala se corrige a partir de las medidas de profundidad de emparejamientos aislados que son más fáciles de estimar. En segundo lugar, el sistema estéreo que define la cámara utilizada consta de dos lentes con gran distorsión de ‘Ojo de pez’, y MiDaS funciona correctamente en imágenes en perspectiva. Por lo tanto, en este trabajo se han definido y aplicado los modelos proyectivos que permiten rectificar y des rectificar la distorsión, y se han incluido algoritmos para obtener una reconstrucción idónea y suavizada de las imágenes.

This final degree work focuses on the estimation of a dense depth map from a pair of stereo images with 'Fisheye' distortion. The estimation of a point from its projection on a pair of stereo images is a geometrically simple problem, knowing the geometry of the camera used. On the other hand, obtaining a dense result from the whole image is a difficult problem to solve. Since it is necessary to search along the epipolar line for each pixel, which in this case is curved due to the image distortion, the solution is not trivial because the search range is infinite. In this work we seek to narrow this search range in order to simplify the problem. For this purpose, we make use of a neural network or artificial intelligence (AI), specifically, the MiDaS network which estimates a dense depth map of monocular images, in which each pixel contains a 'depth seed', i.e., a distance value for each pixel. This dense map provides uncertainty regions for each pixel, which allows to considerably narrow down the search on the epipolar curve. However, there are several considerations to take into account in order to use this map as a seed. First, being a monocular estimation, the dense map provided by the network does not have the correct scale. This scale is corrected from the depth measurements of isolated pairings that are easier to estimate. Second, the stereo system defining the camera used consists of two lenses with large 'Fisheye' distortion, and MiDaS works correctly on perspective images. Therefore, in this work, projective models have been defined and applied to rectify and de-rectify the distortion, and algorithms have been included to obtain an ideal and smoothed reconstruction of the images.

ÍNDICE

Capítulo 1

Introducción y objetivos

1.1. Motivación.

En pleno 2023, la inteligencia artificial (AI), la robótica y la visión por computador están a la orden del día, están teniendo un fuerte impacto en la sociedad y ofrece nuevas posibilidades en el entorno industrial.

La visión por computador es clave en el avance de la robótica. En el caso de este trabajo, la idea de que un robot fuese capaz de saber en todo momento a qué distancia está todo lo que le rodea dotaría a los robots de mayor independencia y sería un gran avance en la posible utilización de robots para la realización de trabajos muy tediosos o perjudiciales para la salud.

La obtención de un mapa de profundidad denso de un par de imágenes estéreo, o incluso de un video estéreo, con lente de 'Ojo de Pez' ofrece posibilidades de todo tipo. Por ejemplo, puede utilizarse para escanear objetos 3D moviendo la cámara, se pueden detectar obstáculos que se acercan, se pueden medir distancias entre puntos y obtener distancias dentro de la imagen y poder recrear mapas 3D de cualquier entorno.

Además de otras utilidades, como un control de calidad para medir distancias y ajustarse a las tolerancias deseadas en procesos de fabricación y producción

También sería interesante su aplicación en cámaras de control de calidad en procesos de fabricación; al tener un mayor campo de visión, con menos dispositivos sería capaz de analizar más piezas.

1.2. Objeto.

En este trabajo se va a abordar la estimación de profundidad de un par de imágenes estéreo de forma densa.

Estimar la profundidad de un punto a partir de su proyección en un par de imágenes estéreo es un problema que se puede resolver geométricamente de forma sencilla si conocemos la traslación y posición relativa entre las cámaras. Resolver este problema para todos los pixeles de la imagen, es decir de forma densa, es en cambio un problema más complicado de resolver porque es muy difícil emparejar cada uno de los píxeles de las dos imágenes, especialmente cuando hay poca textura. Sin embargo, sí que se pueden emparejar algunos puntos que son fáciles de identificar en ambas imágenes y que denominamos puntos característicos. Además, la propia geometría del sistema estéreo (geometría epipolar) impone la restricción epipolar que determina que, dado un punto en una imagen su correspondiente en la otra imagen se encuentra en su curva epipolar.

El problema para estimar la profundidad de forma densa se reduce entonces en una búsqueda a lo largo de la curva epipolar para cada píxel. Esa curva representa la proyección de un rayo en el que el rango de búsqueda es a priori infinito lo que hace esa búsqueda no trivial.

Por otro lado, recientemente han aparecido métodos basados en inteligencia artificial y redes neuronales que son capaces de estimar profundidad a partir de una única imagen. Esta profundidad, al contrario de la triangulación en un sistema estéreo, no se puede considerar una medida ya que es inferida a partir de los datos de ejemplo utilizados para entrenar la red. Al identificar elementos en la escena la red ha aprendido tamaños típicos de esos elementos y los puede localizar a una profundidad o a otra. Estos sistemas nos dan directamente los mapas de profundidad densos en los que los tamaños relativos son correctos, pero tienen un problema de escala.

En este proyecto se quiere utilizar una red neuronal para estimación de profundidad monocular (MiDaS) para simplificar el problema de estimación de profundidad densa en un sistema estéreo.

La idea principal es estimar una semilla de profundidad densa para cada pixel utilizando MiDaS. Como es fácil estimar emparejamientos de puntos característicos en la imagen se va a medir la profundidad de estos puntos con el par estéreo y se van a comparar estas medidas con los proporcionados por la red. Esto nos va a permitir establecer una escala con la que escalar el mapa denso de profundidad a la geometría de la escena.

Este mapa denso preliminar va a definir una región de incertidumbre para cada píxel que va a permitir reducir considerablemente el rango de búsqueda en la curva epipolar.

Como el sistema estéreo que se va a utilizar tiene lentes de ojo de pez de elevada distorsión y la red MiDaS está pensada para trabajar con imágenes perspectivas, además, se van a tener que implementar una serie de algoritmos que rectifiquen y des rectifiquen imágenes para pasar del dominio del ojo de pez al de imagen perspectiva y viceversa.

Esta memoria se estructura en seis capítulos. En el primero, la introducción, se explica cuál es la motivación para realizar este trabajo, los objetivos de este trabajo y la estructura general del software. En el segundo capítulo se definen todas las herramientas utilizadas en la realización de este trabajo. Se explicará qué cámara se utiliza y las características de las imágenes obtenidas. Finalmente, se define la fase de definición, donde se explica paso a paso las implementaciones que se van a introducir en el código para obtener lo indicado en cada fase. En el tercer capítulo se expone la fase de desarrollo. Para cada paso se implementarán explicaciones del código escrito y su seguimiento paso a paso. En el cuarto capítulo se muestran evidencias de los resultados obtenidos con el uso de imágenes y un breve análisis de los resultados obtenidos en cada parte. Finalmente, se hace un análisis de los resultados finales donde se prueba el resultado final de trabajo y se comprueba que funciona correctamente y que se tienen los resultados esperados. En el quinto capítulo se comentan posibles líneas futuras de este trabajo, las conclusiones de este y una explicación de cómo ha sido posible su realización. Finalmente, en la memoria se presenta también la bibliografía utilizada en el trabajo, una lista de figuras y los anexos.

Capítulo 2

Fase de definición

En este capítulo se especifican las herramientas y softwares utilizados. Además del dispositivo utilizado para obtener las imágenes. También, se incluye una descripción general del plan de acción a seguir en el proyecto para montar el código correctamente.

2.1. Herramientas, softwares y entornos utilizados.

En primer lugar, se ha realizado una previa formación individual en Python. Se han utilizado softwares para el desarrollo e implementación de librerías y entornos, como 'PyCharm' y 'Anaconda Navigator', respectivamente. También se cuenta con un código inicial de partida donde ya se tiene un procesamiento inicial de los píxeles de la imagen de entrada, el cual habrá que modificar para nuestro caso.

Se trabaja en el entorno virtual con los paquetes necesarios para poder utilizar la red neuronal 'MiDaS'. En caso de utilizarla en un entorno sin estos paquetes, la red neuronal no funcionaría.

En cuanto a las librerías utilizadas en Python, tenemos las siguientes: 'numpy', 'Matplotlib, 'glob', 'Pytorch', 'os', 'OpenCV', 'Image' (from PIL), 'math' y 'scipy.io'. Se escogen principalmente librerías utilizadas para visión por computador, análisis matemático y aprendizaje profundo.

2.2. Cámara Intel RealSense T265.

La cámara Intel RealSense T265 es un dispositivo electrónico de localización y mapeo simultáneo utilizado principalmente en robótica y drones. Es un dispositivo de baja potencia, pequeño y ligero con un peso de 55 gramos. Utiliza la tecnología SLAM, que es capaz de construir o actualizar un mapa de un entorno desconocido, mientras al mismo tiempo realiza un seguimiento de su propia ubicación dentro de ese entorno [1].

Cuenta con dos sensores de lente de ojo de pez con un campo de visión combinado de 163±5º, unidades de medición inercial (IMU), y una VPU, que es donde todos los algoritmos del SLAM se realizan. Es extremadamente eficiente y tiene una velocidad de procesamiento muy alta idónea para aplicaciones de realidad virtual y aumentada [1].

Figura 1. Cámara Intel RealSense T265 [1].

A partir de imágenes estéreo obtenidas de esta cámara se realiza el trabajo, se conocen los parámetros intrínsecos y extrínsecos para nuestro modelo de cámara.

2.3. Plan de acción.

Rectificación imagen ojo de pez.

En primer lugar, se transforma una imagen en perspectiva de ojo de pez en una imagen sin distorsión y de vista plana. Para ello se aplicará el modelo directo de proyección de Kannala-Brandt [5] (ver Anexo A para más detalle).

Obtención 'Semilla de profundidad'.

En segundo lugar, una vez obtenida la imagen en perspectiva plana, esta se introduce en la red neuronal 'MiDaS' [2]. El resultado es un mapa de profundidad de la imagen rectificada. Los valores obtenidos en esta imagen de profundidad no están escalados con la realidad y no proporcionan medidas reales.

Recomposición de la perspectiva de ojo de pez.

En tercer lugar, se transforma el mapa de profundidad devuelto por la red neuronal al modelo de proyección original de la cámara de ojo de pez. Para ello se utiliza el modelo de inverso de Kannala-Brandt [5]; es decir, se realiza la metodología inversa que se ha aplicado en la 1ª fase.

Obtención geométrica de profundidad y obtención de escala.

A continuación, se escogen puntos característicos de las imágenes estéreo (La imagen estéreo que se obtiene está compuesta por la imagen que proporciona la lente izquierda y por la imagen correspondiente a la lente derecha) que proporciona la cámara, los cuales son proyecciones de puntos 3D en la realidad. Haciendo uso del par de imágenes y aplicando geometría epipolar; es decir, triangulando estos puntos 3D con ambas imágenes y la propia geometría de la cámara (Distancia entre ambas lentes.) [3], se obtienen medidas con incertidumbre de las profundidades de estos puntos.

Se obtienen también los valores de profundidad proporcionados por la red neuronal en estos mismos puntos. Las imágenes utilizadas para obtener estos valores se obtienen al recomponer la perspectiva inicial, ya que la triangulación de puntos se realiza con imágenes con perspectiva ojo de pez.

Se obtiene una escala con ambas medidas de profundidad con seis puntos aleatorios de entre el conjunto de puntos característicos emparejados. Esta semilla se utiliza para ajustar el mapa de profundidad al tamaño de la escena. Este mapa de profundidad escalado por su parte se utiliza para acotar la búsqueda a lo largo de la curva epipolar para un refinamiento de la medida de profundidad. El resto de puntos característicos se utilizan para evaluar la viabilidad del mapa de profundidad escalado como semilla.

Obtención rango de búsqueda.

Finalmente, se obtiene el rango de búsqueda en la curva epipolar donde se encuentra el punto real. [3] Tomando como entrada un punto cualquiera de la imagen izquierda se define una región en la imagen derecha donde se acotará la búsqueda de la curva epipolar.

Para definir un rango razonable para el tamaño de este área se proyectan 24 puntos característicos de la imagen izquierda en la imagen derecha de los que previamente se

conoce su localización en la imagen derecha y por lo tanto se puede estimar su profundidad. Los errores de posicionamiento entre los puntos proyectados y los mismos puntos calculados con la información proporcionada por la red neuronal nos van a dar una estimación del rango a utilizar en la búsqueda acotada.

En base a un estudio de los errores, se obtiene un rango de búsqueda que agiliza el proceso de emparejamiento de puntos 3D. Conociéndose los limites donde se encuentra el punto real, se acota el problema y por lo tanto su coste computacional.

Capítulo 3

Fase de desarrollo

En esta capitulo se detalla el funcionamiento del código implementado para la realización de cada fase del procedimiento a seguir.

Rectificación imagen ojo de pez.

Para aplicar el modelo de proyección de Kannala-Brandt [Anexo A] en la imagen con perspectiva de ojo de pez se sigue el siguiente procedimiento.

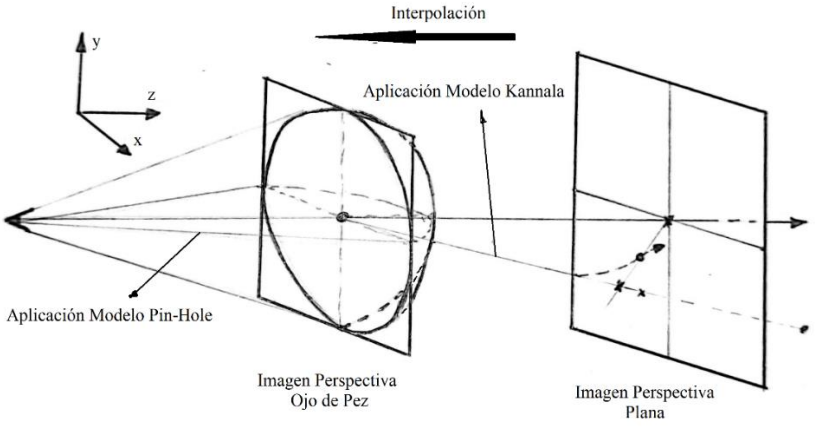


Figura 2. Esquema 3D para rectificar la imagen de entrada.

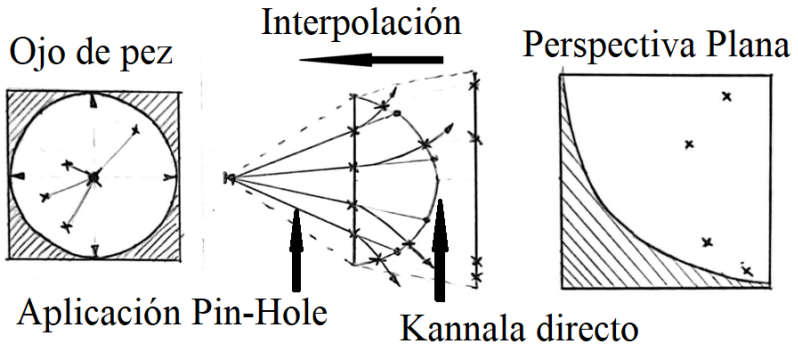


Figura 3. Esquema 2D para rectificar la imagen de entrada.

Se define primero la imagen en perspectiva. Esta se trata como una matriz con su respectivo número de columnas y de filas (1024 x 1024), la cual tiene 3 canales, RGB (Red-Green-Blue). Se aplica el modelo de proyección Pin-Hole [4] a esta matriz de entrada y se obtienen los rayos 3D de la imagen desde la referencia definida en el modelo [Figura 2 y 3].

Para reconstruir la imagen obtenida mediante el modelo Pin-Hole [Anexo A] al tamaño original de entrada en perspectiva plana, es necesario aplicar una interpolación bilineal [Figura 4]. A continuación, a los rayos 3D de la imagen perspectiva se le aplica el modelo de Kannala-Brandt, y se obtienen los píxeles de estos rayos con su distorsión en ojo de pez. Con estos píxeles se obtienen los cuatro índices de la interpolación bilineal a aplicar en la imagen original de entrada de 848 x 800 píxeles.

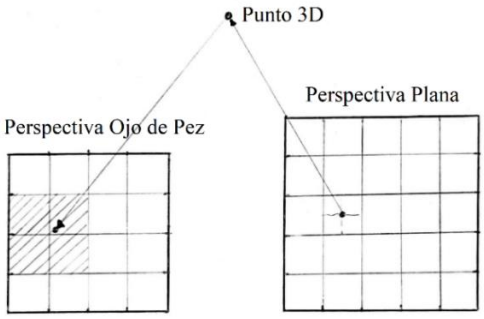


Figura 4. Interpolación bilineal para reconstruir la perspectiva plana.

Este método se basa en la interpolación lineal en dos direcciones. En primer lugar, se realiza una interpolación lineal en la dirección horizontal y luego en la dirección vertical. La interpolación bilineal utiliza cuatro píxeles adyacentes para calcular el valor del píxel interpolado [Anexo B]. Se reconstruye la imagen original al tamaño de la nueva imagen en perspectiva plana (1024 x 1024) con estos nuevos pixeles interpolados ya calculados, a los cuales ya se les ha aplicado la distorsión.

Obtención 'Semilla de profundidad'

En este paso, se hace uso de la librería numpy, OpenCV y Pytorch de Python. También se utiliza la red neuronal de MiDaS [2] [Anexo C], que hace uso de pyTorch, una librería de deep learning que facilita el entrenamiento y ejecución de redes neuronales.

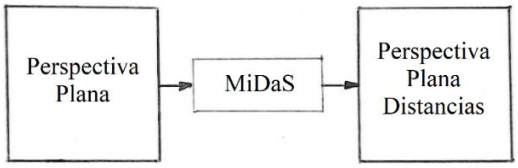


Figura 5. Procedimiento que seguir con la red neuronal.

En primer lugar, se toma una red ya entrenada para la estimación monocular de profundidad, definiendo su tamaño. En este caso se escoge el tamaño grande (precisión alta y velocidad de deducción baja).

En segundo lugar, se utiliza la red neuronal con las imágenes iniciales en perspectiva plana y se obtienen las nuevas imágenes con el mismo formato de tamaño, pero con la diferencia de que la imagen ahora únicamente tiene un canal. Los valores que da este canal son distancias, pero estas no están escaladas a valores referenciados a la realidad. Esto se debe a que a partir de una única imagen no se puede medir profundidad por lo que la red neuronal tiene que inferir el mapa de relieve de la imagen de entrada a partir de los datos de entrenamiento, es decir asignando valores típicos relacionados con los elementos que puede identificar.

Recomposición de la perspectiva de ojo de pez.

Se quiere obtener una estimación densa de la profundidad en la imagen de ojo de pez. Para ello se va a des-rectificar el mapa de profundidad obtenido de la red aplicando el modelo inverso de proyección de Kannala-Brandt [Anexo A] en la imagen de profundidad obtenida en la segunda fase se sigue el siguiente procedimiento, y se hace uso de las librerías de numpy, OpenCV de Python.

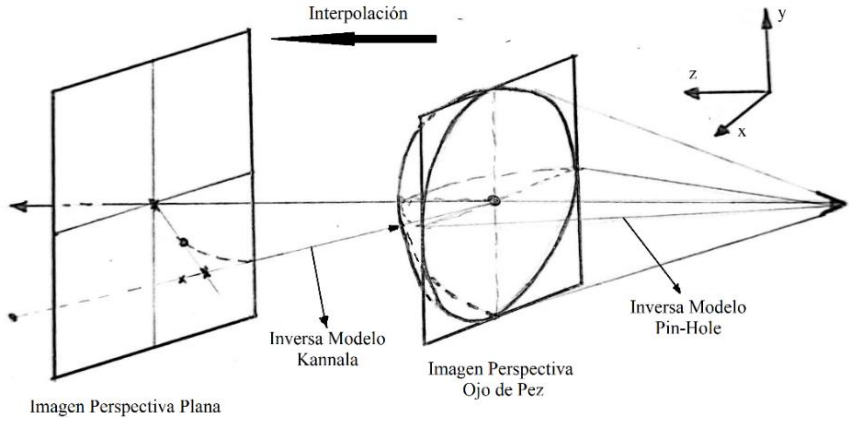


Figura 6. Esquema 3D para recomponer la perspectiva ojo de pez.

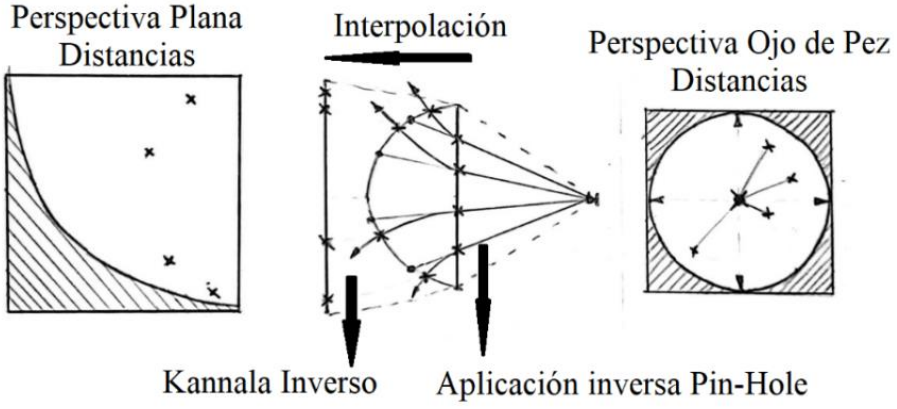


Figura 7. Esquema 2D para recomponer la perspectiva ojo de pez.

En primer lugar, se aplica el modelo inverso de proyección con los píxeles de la imagen con el tamaño original (848 x 800), los parámetros intrínsecos y parámetros de distorsión para obtener los rayos 3D que origina la imagen que da la cámara [Figura 6 y 7].

A continuación, se obtienen los píxeles 2D a partir de los rayos 3D y su relación con la perspectiva plana [Figura 6 y 7]. (Esta relación depende directamente del tamaño de la imagen de entrada), con los que se obtendrán los índices de la interpolación bilineal.

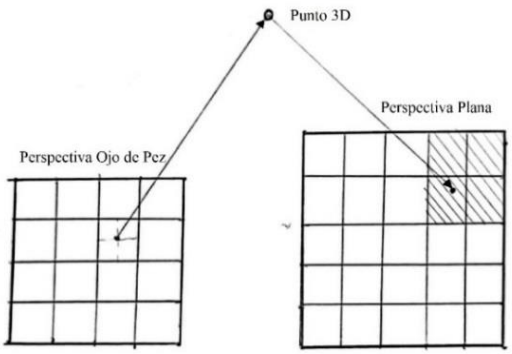


Figura 8. Interpolación bilineal para reconstruir la perspectiva ojo de pez.

De igual manera que para la primera fase, pero ahora haciendo el proceso inverso, se aplica la interpolación [Anexo B] a la imagen plana de profundidad y se reconstruye la misma imagen de profundidad. Pero ahora, con el tamaño original y con la distorsión original de ojo de pez.

Obtención geométrica de profundidad y obtención de escala.

Además de la estimación monocular de profundidad densa obtenida de la red neuronal Midas, se pueden utilizar el par de imágenes estéreo para medir profundidad.

Esta medición requiere identificar cada punto de la imagen izquierda en la imagen derecha lo que es fácil para determinados puntos característicos pero difícil en regiones en las que no hay textura o en regiones similares entre sí.

Sin embargo, una vez emparejados algunos puntos característicos vamos a tener mediciones que se pueden utilizar para escalar el mapa de profundidad denso obtenido como salida la red Midas.

En la fase actual del proceso se utilizan puntos correspondientes en ambas imágenes. Estos puntos se emparejarían de forma automática utilizando métodos clásicos de detección y emparejamiento de características o métodos basados en Deep Learning como SuperGlue [6]. En este trabajo se van a emparejar de forma manual y se van a utilizar para obtener una estimación de la escala a aplicar al mapa denso de profundidad.

En primer lugar, se obtienen los rayos 3D de ambas imágenes para estos puntos mediante el modelo inverso de proyección de Kannala-Brandt [Anexo A]. A continuación, se definen las matrices de translación y rotación entre la referencia de la lente izquierda y de la derecha.

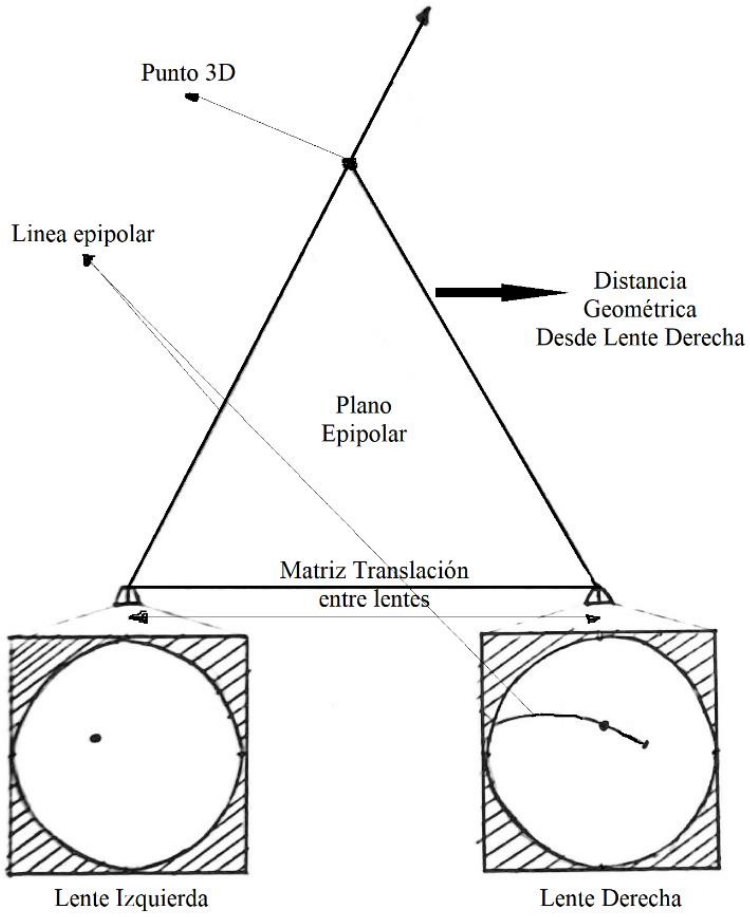


Figura 9. Geometría epipolar para obtener distancias geométricas.

Posteriormente, mediante la triangulación de los mismos puntos entre ambas imágenes se obtienen los puntos 3D en el sistema de referencia de la cámara izquierda. La tercera coordenada de estos puntos 3D corresponde con la medida de profundidad en ese píxel. A continuación, se extraen los valores de profundidad que da la red neuronal a los mismos puntos que estamos evaluando geométricamente.

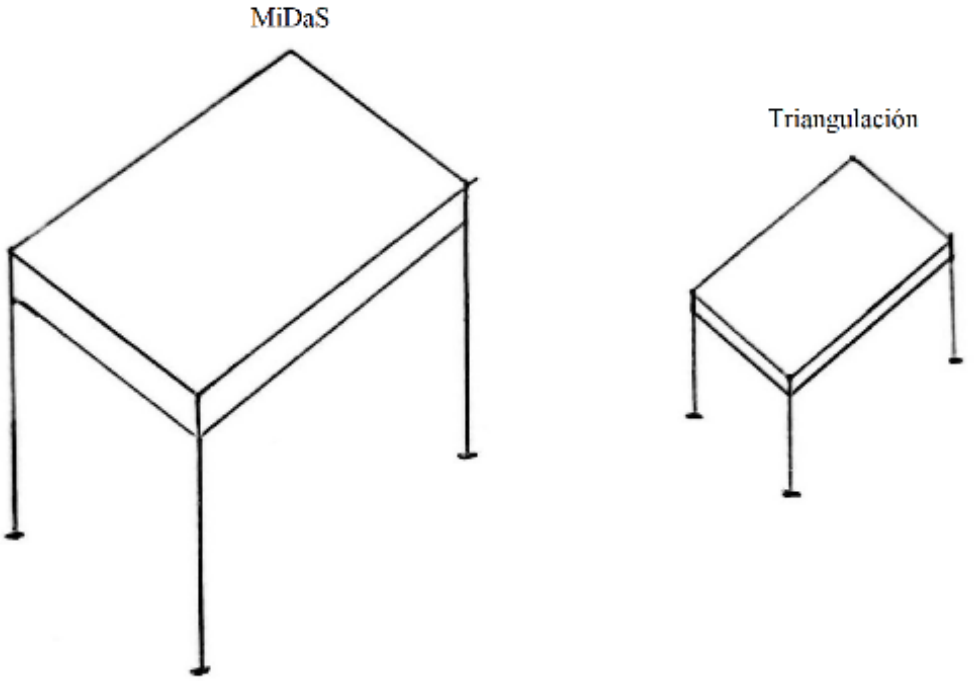


Figura 10. Comparación de tamaños a escala entre los valores que aporta MiDaS y la geometría epipolar.

Finalmente, se debe corregir la distancia proporcionada por la red ya que como se ve en la figura 10 la red representa las distancias de manera gigantesca. Para ello se aplica la escala a la lente izquierda, la cual se obtiene realizando una media de las escalas obtenidas en seis puntos, de los cuales se tiene una alta confianza de su posición y por eso se usan. La escala de cada punto se obtiene dividiendo la medida calculada mediante triangulación entre la medida extraída de la red para ese mismo punto.

Obtención rango de búsqueda.

En este último paso se busca obtener los intervalos de un rango de búsqueda de la posición real de los puntos en la línea epipolar (en la imagen derecha) dada la posición de un punto de la imagen izquierda. El objetivo es afinar el emparejamiento de puntos 3D a partir de las imágenes estereo. Para ello, se toman 24 puntos, de los que se conoce el emparejamiento y posición 3D, para estudiar cómo varían las distancias residuales entre la proyección del punto y su posición real, y definir un rango de búsqueda acorde para todos los posibles puntos de la imagen. El rango de búsqueda se entiende gráficamente como el área donde se encuentra el punto real y, por lo tanto, donde el emparejamiento es perfecto.

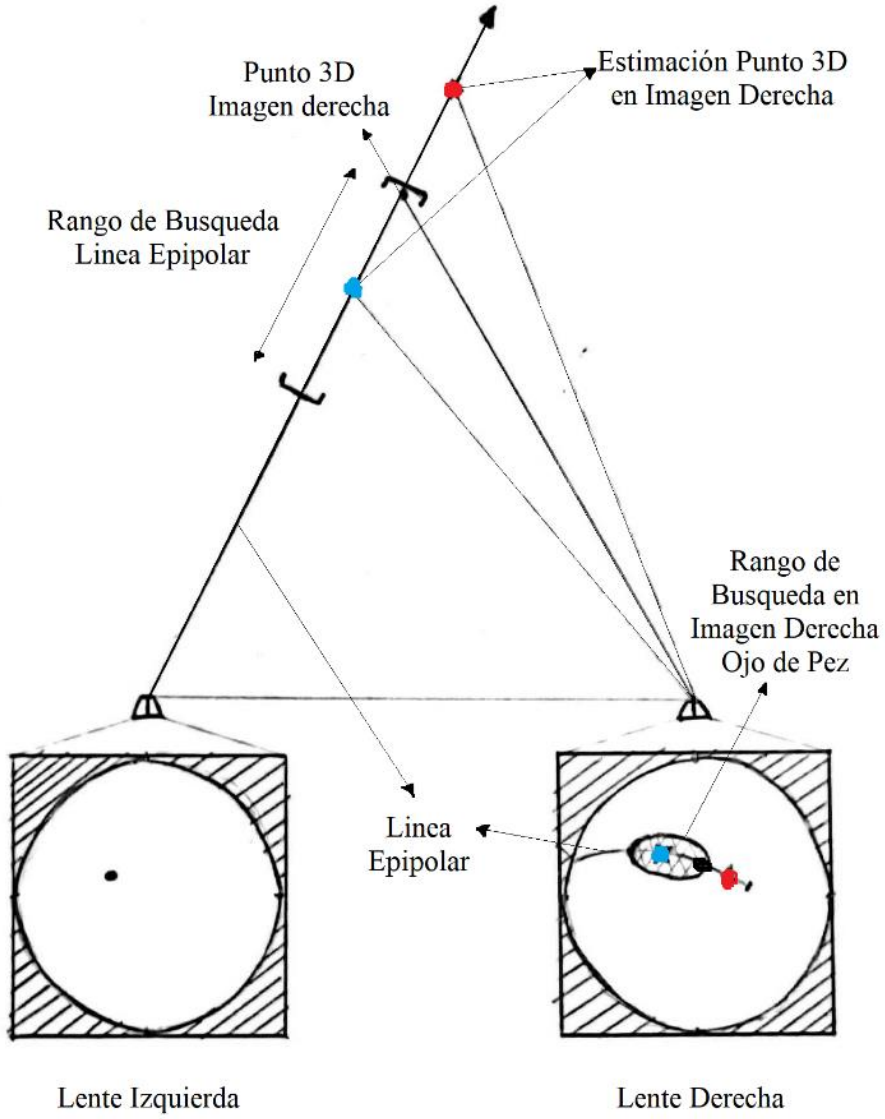


Figura 11. Representación del rango de búsqueda en la imagen y en la geometría epipolar.

En primer lugar, una vez que ya se ha normalizado la imagen de profundidad con el uso de la escala obtenida anteriormente se obtienen los puntos 3D que definen el total de la imagen izquierda de profundidad con la escala aplicada. Posteriormente, se proyectan estos puntos 3D en la referencia de la cámara derecha y a partir de ellos se obtienen los píxeles correspondientes a esos rayos en la imagen derecha.

Finalmente, una vez proyectados los píxeles de la cámara izquierda en la derecha, se procede con una comparación de posición de los 24 puntos. Se comparan las posiciones ya conocidas de esos puntos en la imagen derecha con las posiciones de esos mismos puntos provenientes de la imagen izquierda que se han proyectado en la imagen derecha. Se estudian las diferencias de posicionamiento de píxeles de los puntos mencionados anteriormente y se define un cierto rango de búsqueda que se ajuste bien.

De esta manera, la proyección de cualquier punto de la imagen izquierda en la imagen derecha se tomará como el centro de la elipse que recogerá el área donde se encuentra el punto real; es decir, el área donde el emparejamiento de puntos es correcto.

Capítulo 4

Fase de evaluación.

En este capítulo se van a analizar los resultados obtenidos con ejemplos de cada una de las fases. También se hace una evaluación global del proceso.

4.1. Análisis de resultados por fases.

Rectificación imagen ojo de pez.

En esta primera fase, lo primero que se presenta es la imagen original, la cual cuenta con una distorsión de ojo de pez, tanto de la lente izquierda como de la lente derecha.

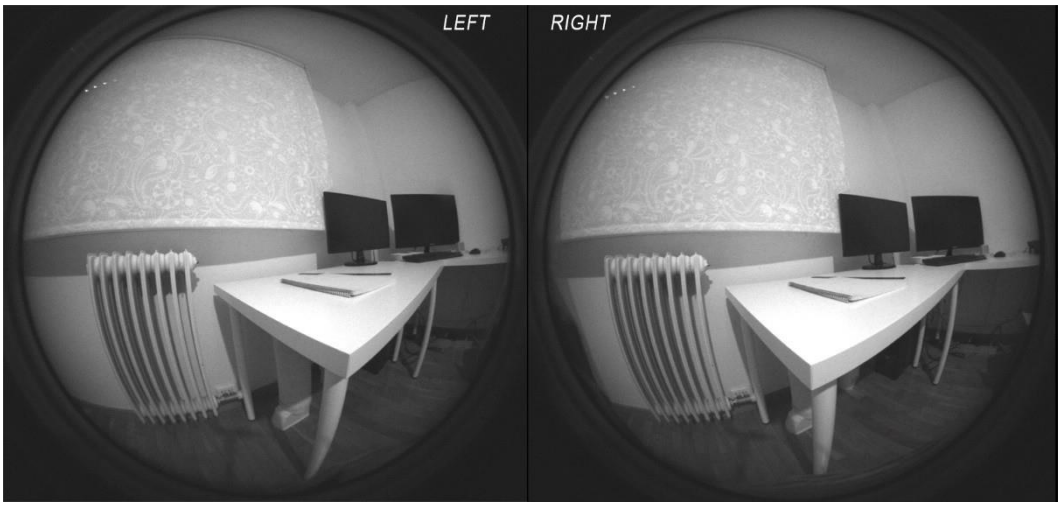


Figura 12. Imagen obtenida con la cámara Intel RealSense T265, lente izquierda y derecha.

A continuación, se muestra la sección de la imagen original que se selecciona para aplicar el modelo de proyección.

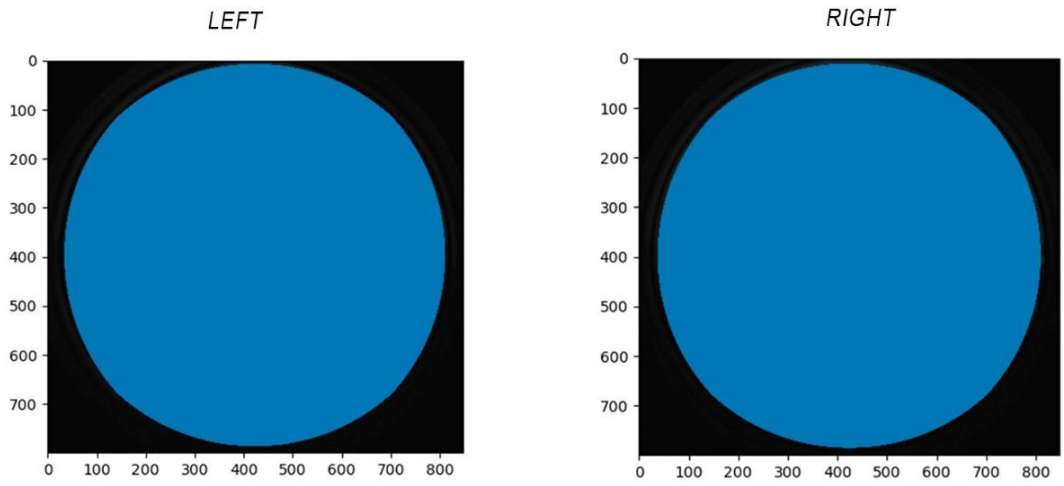


Figura 13. Selección del área de la a rectificar mediante modelo Pin-Hole, lente izquierda y derecha.

Se puede observar cómo se selecciona únicamente la zona de la imagen con información de la imagen con perspectiva de ojo de pez; es decir, la zona que no es negra. Finalmente, se presentan las imágenes originales de entrada en perspectiva plana.

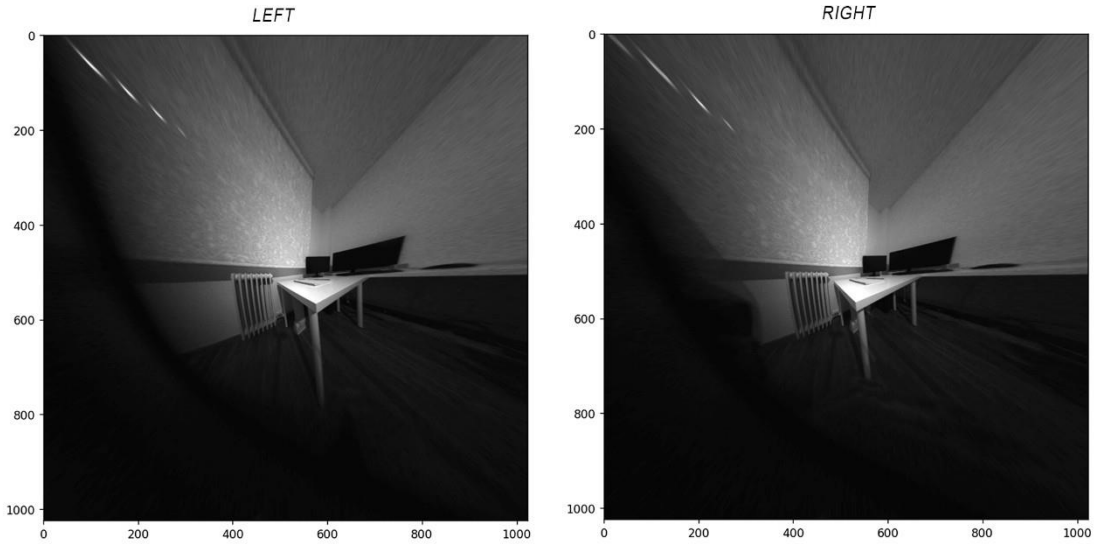


Figura 14. Imágenes en perspectiva plana, lente izquierda y derecha.

Se observa que aparece una parte de la imagen en perspectiva plana en negro, esto se debe a que se está cogiendo algo de negro de la imagen original y al aplicar el modelo de Kannala-Brandt aparece esa zona oscura que representa más o menos un cuarto de la imagen. Pero esto no es algo que vaya a afectar en gran medida al resultado final.

Obtención 'Semilla de profundidad'.

Para esta segunda fase donde se obtiene la 'semilla de profundidad' se continua con la última imagen obtenida en la primera fase, la imagen perspectiva [Figura 4].

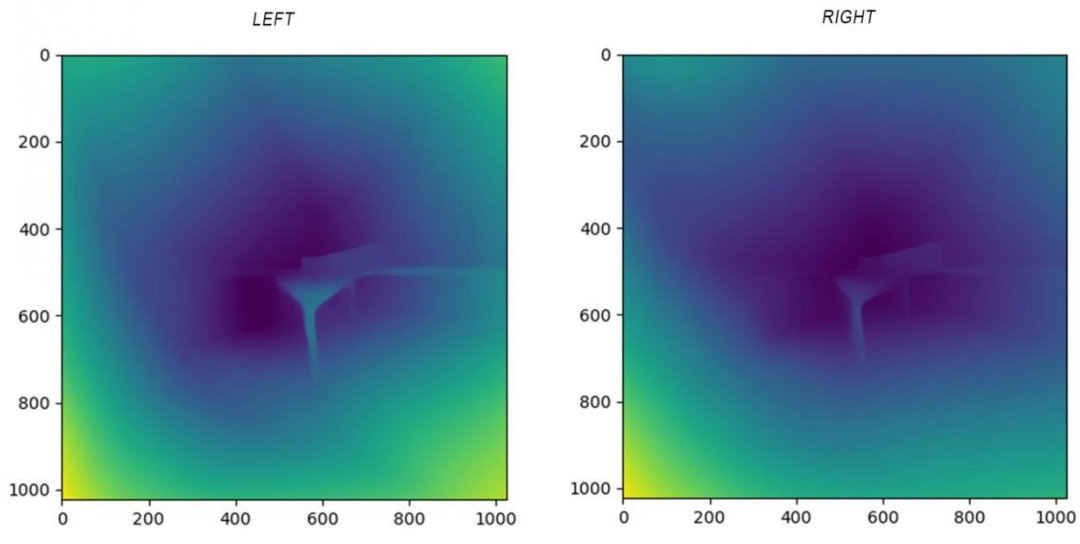


Figura 15. Imágenes de profundidad en perspectiva plana, lente izquierda y derecha.

Se ha utilizado un tamaño de red grande. Se ajusta bien este tamaño de red para la obtención del mapa de relieve de la imagen de entrada.

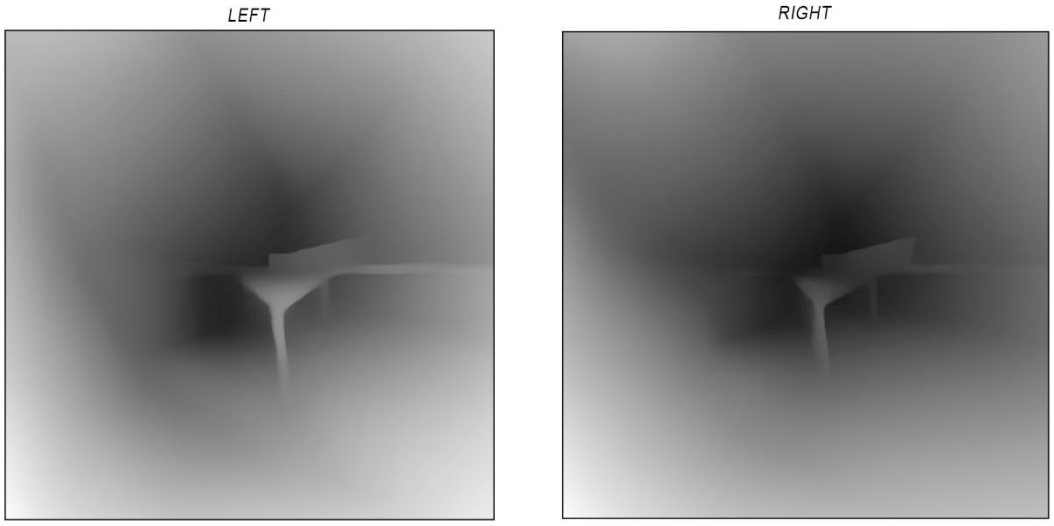


Figura 16. Imágenes de profundidad en perspectiva plana en escala de grises, lente izquierda y derecha.

Se observa ahora la dupla de imágenes en perspectiva plana y de profundidad en escala de grises.

Recomposición de la perspectiva de ojo de pez.

En esta tercera fase se recompone la perspectiva plana de la imagen obtenida anteriormente y se obtiene la imagen en perspectiva de ojo de pez con la aplicación de la red neuronal.

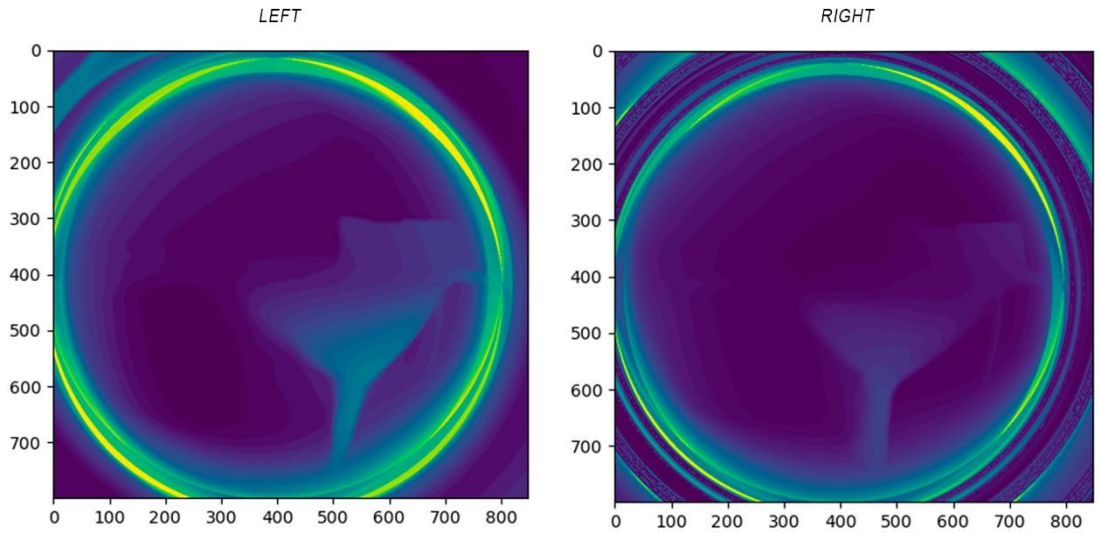


Figura 17. Imágenes de profundidad en perspectiva ojo de pez, lente izquierda y derecha.

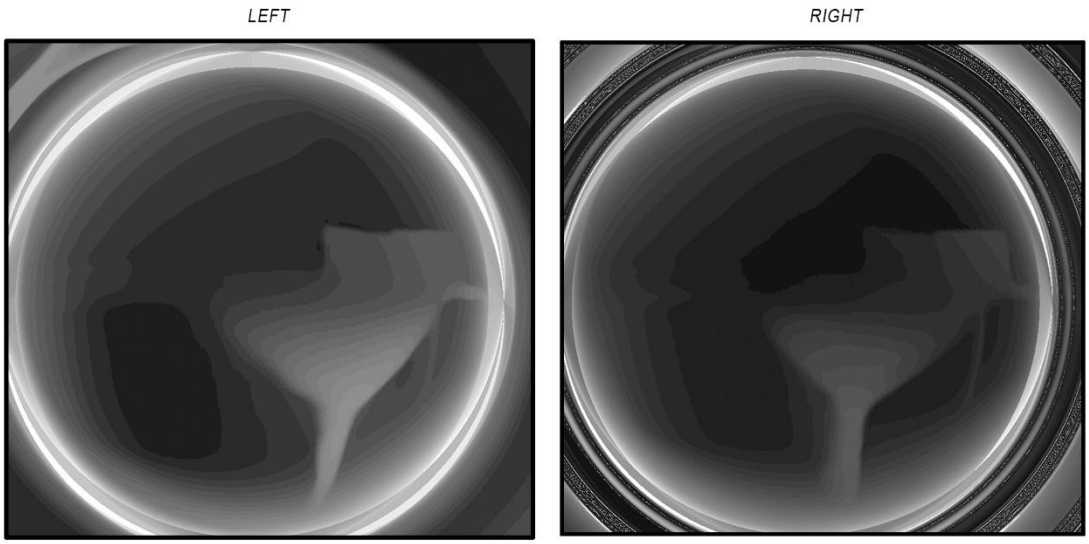


Figura 18. Imágenes de profundidad en perspectiva ojo de pez en escala de grises, lente izquierda y derecha.

Obtención geométrica de profundidad y obtención de escala.

Para esta fase, el primer paso es exponer donde están los puntos de los cuales se han obtenido sus profundidades mediante triangulación tanto en la imagen original de entrada [Figura 2] y la última imagen obtenida en la tercera fase [Figura 7].

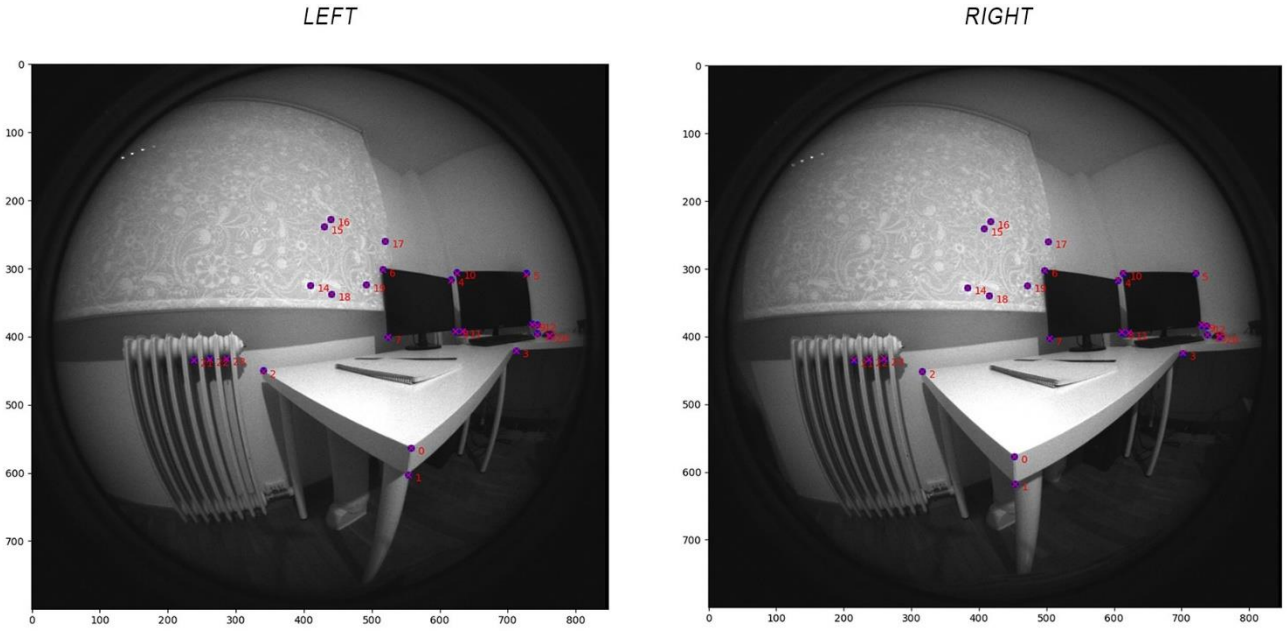


Figura 19. Puntos definidos en imágenes originales, lente izquierda y derecha.

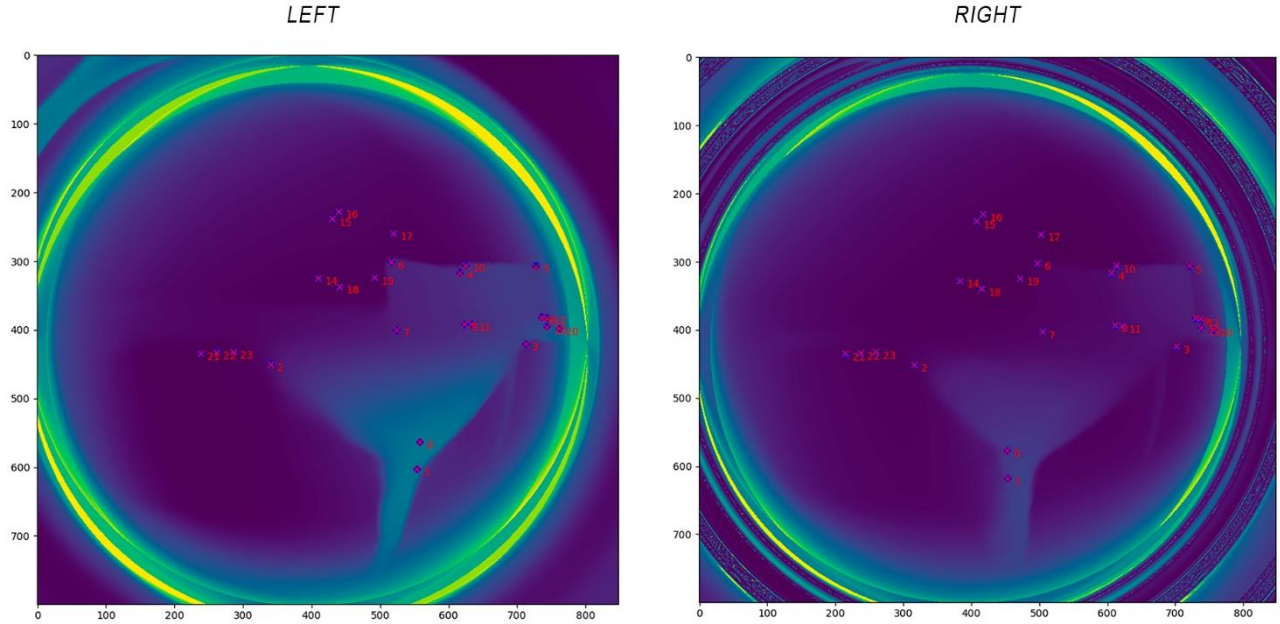


Figura 20. Puntos definidos en imágenes de profundidad en perspectiva de ojo de pez, lente izquierda y derecha.

Puntos	Distancias Triangulación (m)	Distancias Red Neuronal	Distancia Red neuronal con escala (m)	Error distancia Triangulación vs Red con escala (m)
1	0,13156	15,18522	1,19417	1,06261
2	0,12844	15,2301	1,19770	1,06926
3	0,56949	4,65435	0,36602	0,20347
4	0,38916	7,98596	0,62802	0,23886
5	0,76163	7,30427	0,57441	0,18722
6	0,42558	8,73546	0,68696	0,26138
7	0,73562	5,94573	0,46757	0,26805
8	0,70388	6,0021	0,47201	0,23187
9	0,70726	7,01156	0,55139	0,15587
10	0,4161	8,46605	0,66577	0,24967
11	0,74418	6,14183	0,48300	0,26118
12	0,70853	7,22554	0,56822	0,14031
13	0,34678	8,58461	0,67510	0,32832
14	0,36098	11,31426	0,88976	0,52878
15	0,57776	3,22761	0,25382	0,32394
16	0,63648	3,69344	0,29045	0,34603
17	0,64353	3,75052	0,29494	0,34859
18	0,75447	3,28982	0,25871	0,49576
19	0,63132	3,15864	0,24840	0,38292
20	0,6982	3,09232	0,24318	0,45502
21	0,26748	12,69697	0,99849	0,73101
22	0,40282	2,82769	0,22237	0,18045
23	0,4303	2,9164	0,22935	0,20095
24	0,45554	3,16826	0,24915	0,20639

Tabla 1. Profundidades y error en cálculos de profundidades para los 24 puntos.

Se observan 24 puntos y sus distancias o profundidades, unas se han obtenido mediante triangulación y otras se han obtenido directamente del mapa denso de profundidad que proporciona la red neuronal [Tabla 1]. Se tienen en cuenta únicamente seis puntos, los cuales se encuentran a distintas profundidades y en distintas zonas de la imagen, para el posterior cálculo de la escala. *Puntos = [1, 2, 5, 6, 20, 22]*.

Entonces, se formula la escala a aplicar a toda la imagen de profundidad como la proporción que hay entre la medida de profundidad obtenida a partir de la red neuronal 'MiDaS' y la obtenida mediante el método de triangulación, únicamente teniendo en cuenta los seis puntos anteriormente mencionados, con los que se tiene en cuenta de manera general toda la información que aporta la imagen de entrada y a su vez la imagen obtenida con la red. El valor final de la escala a aplicar para la imagen izquierda será una media de los valores obtenidos con los seis puntos.

Este valor de escala [Tabla 2] es un valor orientativo para acotar la solución de la última fase del trabajo.

Escala Izquierda	0,0786403

Tabla 2. Escala para imagen profundidad lente izquierda.

Obtención rango de búsqueda.

En la siguiente imagen se representa la distancia que hay entre los puntos originales de la imagen y la proyección de esos mismos puntos proveniente de la imagen izquierda.

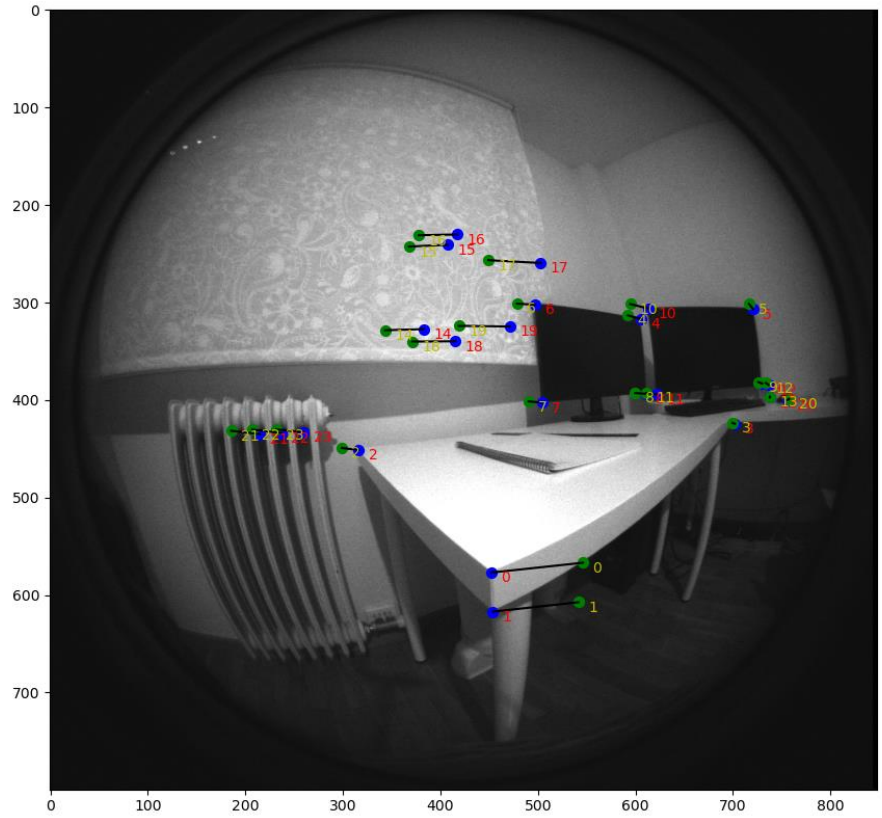


Figura 21. Distancias residuales entre los puntos para la imagen derecha.

Los puntos azules son los ya conocidos y propios de la imagen derecha, obtenidos de forma manual. Los puntos verdes son los puntos provenientes de la imagen izquierda proyectados en la imagen derecha. La línea negra que une los puntos es la distancia residual; es decir, el error de posicionamiento cuando se hace la proyección. En la linea epipolar se encuentran las proyecciones (Puntos verdes) y los puntos reales (Puntos azules), pero esta linea no se representa en la figura 21.

La calidad de la proyección del punto en comparación con su posición real es totalmente dependiente del valor de la escala que se aplica en la imagen izquierda.

La obtención de una escala que sea buena tanto para puntos a corta distancia, media distancia y larga distancia es complicado, ya que el mapa de profundidad obtenido tampoco es tan detallado y este no es perfecto. Dependiendo del valor de la escala las proyecciones de ciertos puntos serán mejores.

Las proyecciones de los puntos se encuentran a la izquierda del punto real cuando en la línea epipolar la estimación de distancia del punto es menor que la distancia real del punto. En cambio, cuando esta estimación es mayor que la del punto real, la proyección del punto se encuentra a la derecha del punto real. Esto sucede porque para obtener el rango de búsqueda, se conoce la distancia y la posición exacta de los puntos en la imagen derecha, ya que se han obtenido de forma manual.

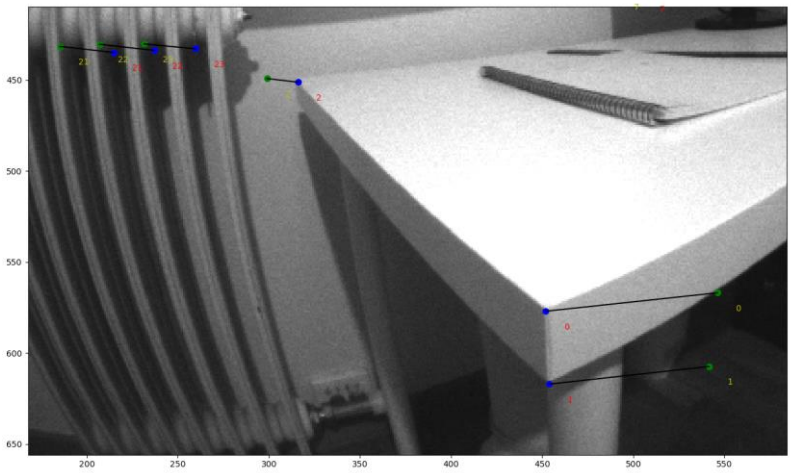


Figura 22. Recorte 1 en distancias residuales entre los puntos para la imagen derecha.

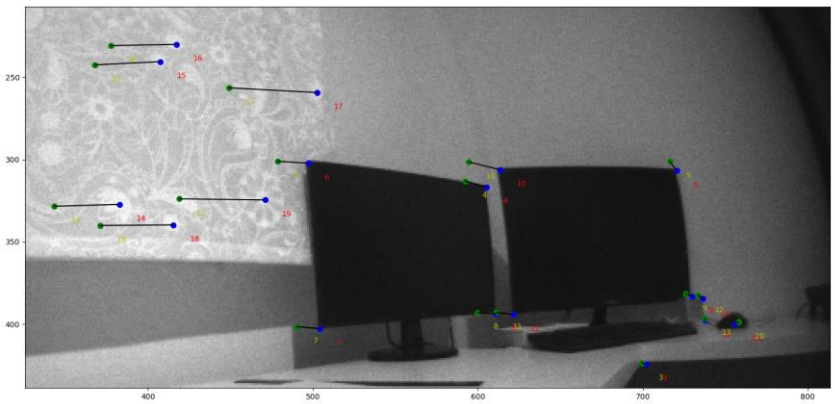


Figura 23. Recorte 2 en distancias residuales entre los puntos para la imagen derecha.

El rango de búsqueda vendrá definido por el área de una elipse con centro en la proyección del punto, la anchura y la altura será la máxima distancia entre las proyecciones obtenidas y los puntos reales de estas proyecciones [Tabla 3]. Se tiene en cuenta un nivel de confianza para asegurar que se recoge toda el área necesaria y que ningún emparejamiento se quede fuera de esta área.

Máxima Distancia Horizontal (Pixeles)	94,18
Máxima Distancia Vertical (Pixeles)	10,11

Tabla 3. Máximas distancias entre puntos y sus proyecciones en pixeles.

El nivel de confianza se aplica a la distancia vertical y horizontal máxima entre los puntos originales y sus proyecciones estimadas en la imagen derecha, con lo que se define el rango máximo de anchura y altura del área de la elipse. Se implementa para asegurar que el área definida contiene el punto real y su estimación en la curva epipolar, la cual atraviesa el área de búsqueda. Su valor se ha definido mediante prueba y error, hasta que se ha encontrado un valor adecuado; en nuestro caso, se ha escogido finalmente un nivel de confianza de 4. Es una manera sencilla de asegurar que el rango de búsqueda siempre contenga el punto real que se busca.

Anchura (Pixeles)	Altura (Pixeles)	Nivel Confianza	Área ($Pixeles^2$)	% Reducción
94,18	10,11	1	747,82	99,89
188,36	20,22	2	2991,30	99,56
282,54	30,33	3	6730,42	99,01
376,72	40,44	4	11965,19	98,24
470,9	50,55	5	18695,61	97,24

Tabla 4. Rangos de búsqueda según el nivel de confianza.

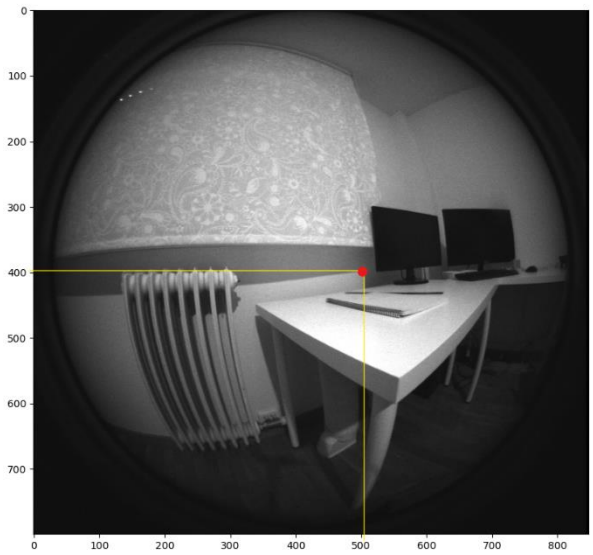
En esta tabla 4, se observa como en función del valor que se le de al nivel de confianza. Incremente el área de búsqueda. Vemos que hasta un nivel de confianza de 5 el porcentaje de reducción de área de búsqueda se reduce en mínimo un 97 % respecto del total de la imagen que tiene un área de 678400 pixeles cuadros.

4.2. Análisis de resultado final.

Como resultado final, obtenemos un proceso automático donde se escogerá cualquier punto de la imagen izquierda y se obtendrá el rango de búsqueda para la iteración de geometría epipolar en la imagen derecha. Es decir, se obtiene el área de la imagen donde se sabe que se encuentra el punto correspondiente al elegido en la imagen izquierda.

Para ello se define en primer lugar, el punto que se va a escoger de la imagen izquierda.

Figura 24. Punto en la imagen izquierda y posición [x,y]=[503,396].



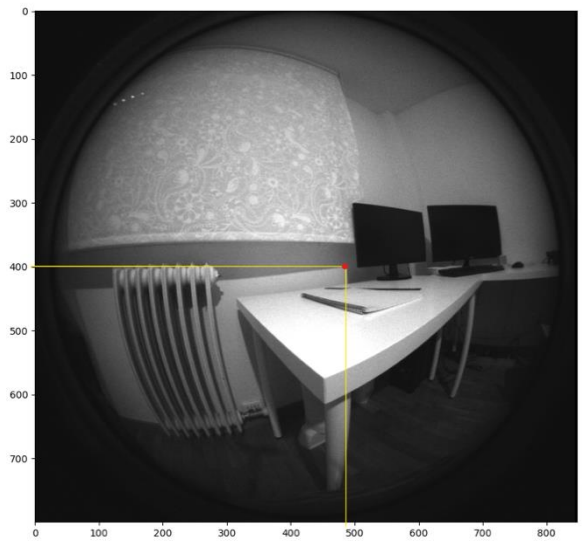


Figura 25. Punto en la imagen derecha y posición [x,y]=[486,399].

Entonces, se obtiene la proyección de ese punto proveniente de la imagen izquierda en la imagen derecha (Punto verde) y se representa el área de búsqueda donde se encuentra el punto (Punto azul) al que hay que emparejar el punto inicial de partida de la imagen izquierda [Figura 16].

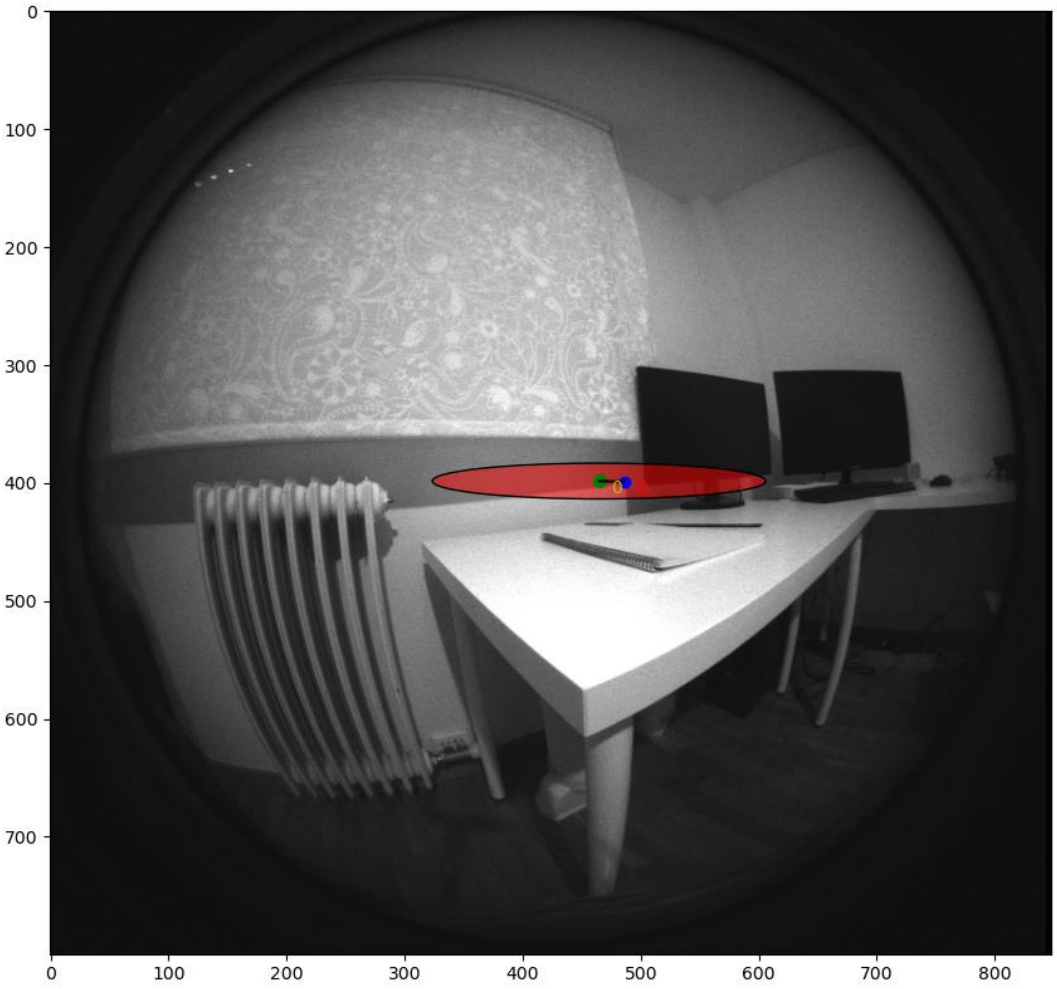


Figura 26. Área de búsqueda para emparejamiento en la imagen derecha punto inicial. NC = 3.

El área de búsqueda en la aplicación de iteración de geometría epipolar para el emparejamiento de puntos es de 6731,10 píxeles al cuadrado, lo que supone una reducción del área de búsqueda total del 99% respecto de la imagen completa. Se aplica un nivel de confianza de 3 para el rango de anchura y de altura de la elipse. Que se ve que es bueno para recoger el punto real en la zona de emparejamiento.

Ahora, vamos a probar con los puntos donde más difiere la proyección de esos mismos puntos con la posición real (Por ejemplo, el punto cero). Se mantienen el mismo nivel de confianza y se ve que el rango de búsqueda es un poco justo para este tipo de puntos que están a una distancia corta [Figura 17].

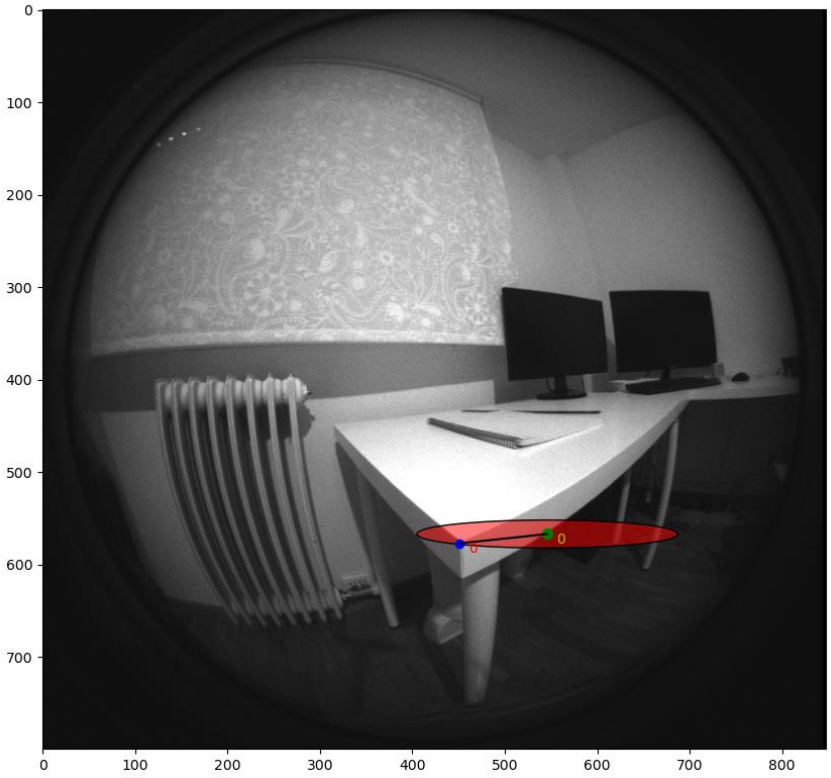


Figura 27. Área de búsqueda para emparejamiento en la imagen derecha punto 0. NC = 3.

Se reajusta el nivel de confianza a 4 para asegurar que en estos puntos más cercanos se recoja el punto real en el área de búsqueda [Figura 18]. El área de búsqueda es de 11966,40 pixeles al cuadrado y una reducción del área de búsqueda total del 98,24 % respecto de la imagen completa.

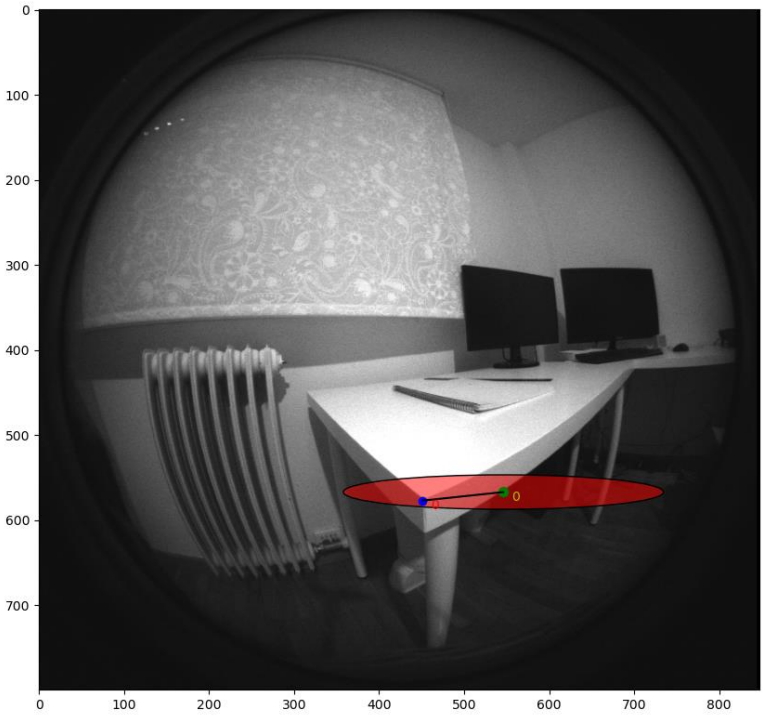


Figura 28. Área de búsqueda para emparejamiento en la imagen derecha punto 0. NC = 4.

Capítulo 5

Conclusiones y líneas futuras.

5.1. Conclusión.

Como conclusión de este trabajo, se puede decir que ha tenido un resultado positivo y que tiene varias líneas de futuro, algunas más ambiciosas que otras. A lo largo de este trabajo se ha hecho un tratamiento de imágenes en términos de transformación de modelos proyectivos mediante rectificación de imágenes. Además, se ha integrado una red neuronal llamada 'MiDaS' que es capaz de obtener una primera estimación de la profundidad de una imagen plana, lo que agiliza mucho el proceso de estimación de la profundidad de toda la imagen obtenida. También se aplica geometría epipolar con el par de imágenes para triangular puntos y obtener medidas reales de profundidades en unos puntos determinados, con lo que se consigue un primer valor estimado de la escala a aplicar a los valores de profundidad que se obtienen de la red 'MiDaS'. Finalmente, se obtiene un intervalo en el que sabremos dónde se encuentra un punto de la imagen izquierda en la imagen derecha. Este intervalo se define como una elipse en la figura y en esa zona se encuentra el punto buscado; es decir, el emparejamiento.

5.2. Líneas futuras.

Como líneas futuras de este trabajo destaca una por encima del resto, la cual es automatizar este proceso de obtención de la profundidad total de la imagen. El primer paso sería automatizar el emparejamiento de los puntos caracteristicos que ahora se hace de forma manual. El segundo paso que seguir sería realizar una búsqueda epipolar de todos los puntos 3D de la imagen. Se definirían las líneas epipolares de todos estos puntos y haciendo uso del rango de búsqueda obtenido en este trabajo se podría acotar zonas de la línea epipolar donde se encuentra el punto en la realidad. Este resultado permitiría automatizar una estimación densa de profundidad.

Otras proyecciones más ambiciosas en este proyecto sería extender la estimación de profundidades a videos que permitiese fusionar un mapa denso utilizando el sistema de odometría visual del sistema estéreo.

Como posibles aplicaciones y futuros desarrollos de este proyecto sería la incorporación de este sistema como sensor de un robot lo que permitiría que la propia máquina fuese capaz de saber a qué distancia está todo lo que le rodea. También hay que decir que existen otros dispositivos para obtener profundidades más rápidamente y que son más precisos (Por ejemplo, el sensor LIDAR). La ventaja de este dispositivo es que con único sensor se realizan diferentes tareas, lo que reduce costes.

5.3. Realización del trabajo.

Para la realización de este trabajo, tengo que agradecer mucho la implicación, participación y seguimiento del director y codirector del TFG. También quiero mencionar que este TFG se ha trabajado en paralelo al que realiza mi compañero César Rodríguez,

sobre cálculo de profundidades en imágenes de ojo de pez únicamente utilizando geometría epipolar.

Para la realización de este trabajo se ha contado con códigos ya hechos y material de teoría sobre modelos de proyección y cálculos de profundidades utilizando geometría epipolar. Estas fuentes han sido proporcionadas por los directores del trabajo.

Con el uso de estos códigos y estas fuentes, se ha desarrollado el código que sustenta este trabajo. En él se ha obtenido como resultado un rango de búsqueda de emparejamiento de puntos, a partir de una primera estimación de la profundidad utilizando la escala obtenida y la red neuronal. Se llega a obtener una medida aproximada de profundidad en cualquier punto.

Capítulo 6

Bibliografía.

[1] Intel RealSense. (s.f.). Tracking Camera T265. Recuperado de https://www.intelrealsense.com/tracking-camera-t265/

[2] Liu, Y., Wang, Y., and Zhang, X. (2021). QRNN-MIDAS: A novel quantile regression neural network for mixed sampling frequency data. Neurocomputing, vol. 457, p. 84-105. Recuperado de https://www.sciencedirect.com/science/article/abs/pii/S0925231221009012

[3] Gallego, A. J. (2009). Detección de objetos y estimación de su profundidad mediante un algoritmo de estéreo basado en segmentación. Recuperado de https://www.researchgate.net/profile/Antonio-Javier-Gallego-2/publication/39435989_Deteccion_de_objetos_y_estimacion_de_su_profundidad_mediante_un_algoritmo_de_estereo_basado_en_segmentacion/links/02e7e51aa035accbae000000/Deteccion-de-objetos-y-estimacion-de-su-profundidad-mediante-un-algoritmo-de-estereo-basado-en-segmentacion.pdf

[4] R. Szeliski, Computer Vision: Algorithms and Applications, 2nd ed., 2021.

[5] KANNALA, J. and BRANDT, S. A generic camera model and calibration method for conventional, wideangle, and fish-eye lenses. IEEE transactions on pattern analysis and machine intelligence, 2006, vol. 28, no 8, p. 1335-1340.

[6] Sarlin, P.E., DeTone, D., Malisiewicz, T. and Rabinovich, A.," SuperGlue: función de aprendizaje que coincide con redes neuronales gráficas", arXiv:1911.11763 [cs], Nov. 2019. [Online]. Recuperado de https://arxiv.org/abs/1911.11763

Anexo A. Modelos proyectivos de cámara.

Se hace uso de dos modelos proyectivos para la realización de este trabajo; el modelo Pin-Hole y el modelo de Kannala-Brandt. Se define en este anexo la definición de cada modelo, su representación y su modelaje matemático.

Modelo Pin-Hole

El modelo directo Pin-Hole es una función que transforma un punto 3D en el sistema de referencia de la cámara a un punto en la imagen de coordenadas (En este caso, pixeles). Los rayos son rectos al chocar en la imagen, y la imagen se rota 180 º.

$$F : R^3 \rightarrow R^2$$

Punto 3D:

$$X_p = \begin{pmatrix} x_p & y_p & z_p \end{pmatrix}^T$$

Punto en imagen de coordenadas:

$$X_c = (u \quad v)^T$$

Desarrollo de la función:

$$X_c = \begin{bmatrix} u \\ v \end{bmatrix} = F(X_p) = \begin{bmatrix} \frac{f}{d_u} * \frac{x_p}{z_p} + u_c \\ \frac{f}{d_v} * \frac{y_p}{z_p} + v_c \end{bmatrix} = \begin{bmatrix} f_u * \frac{x_p}{z_p} + u_c \\ f_v * \frac{y_p}{z_p} + v_c \end{bmatrix} \quad (1.1)$$

Donde f es la distancia focal de la cámara, u_c y v_c son las coordenadas en la imagen del centro óptico, d_u y d_v es el tamaño de un píxel en la imagen.

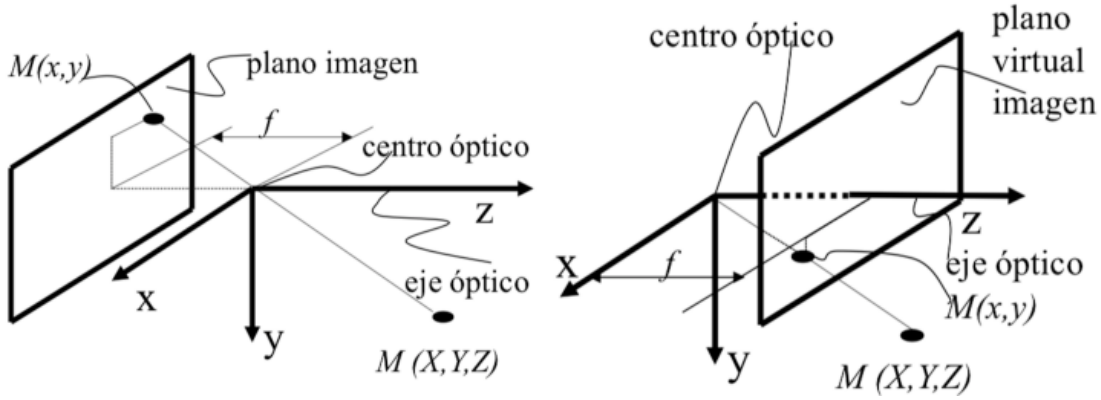


Figura 29. Funcionamiento de la cámara Pin-Hole. Se representa en la izquierda la proyección en plano de la imagen real, y en la derecha la proyección en el plano de la imagen virtual.

El modelo inverso de Pin-Hole es la función inversa, la cual se desarrolla de la siguiente manera:

$$Q : R^2 \rightarrow R^3 = F^{-1}$$

Desarrollo de la función:

$$X_p = \begin{bmatrix} x_p \\ y_p \\ z_p \end{bmatrix} = F^{-1}(X_c) = \begin{bmatrix} \frac{u - u_c}{f_u} * s \\ \frac{v - v_c}{f_v} * s \\ s \end{bmatrix} = \begin{bmatrix} \frac{u - u_c}{f_u} \\ \frac{v - v_c}{f_v} \\ 1 \end{bmatrix} \quad (1.2)$$

Se toma el valor de s como uno, ya que en nuestro caso se considera que la profundidad en coordenadas es para todos los pixeles de la imagen la misma; es decir, la unidad.

Modelo Kannala-Brandt

El modelo directo de Kannala-Brandt se ajusta bien a camara tipo 'Ojo de pez', pero en si mismo propone un modelo de cámara genérico. El modelo de proyección se considera que es radialmente simétrico. En este caso, los rayos se curvan para chocar en la imagen. El modelo describe las diferentes proyecciones de la siguiente manera:

$$r(\theta) = f(\theta + k_1\theta^3 + k_2\theta^5 + k_3\theta^7 + k_4\theta^9) \quad (2.1)$$

Donde θ es el ángulo que forma el rayo proyectante con el eje Z, $r(\theta)$es la distancia desde el punto proyectado hasta el centro de la imagen, $f$es la distancia focal de la cámara y $k_i$son los coeficientes de distorsión de la cámara.

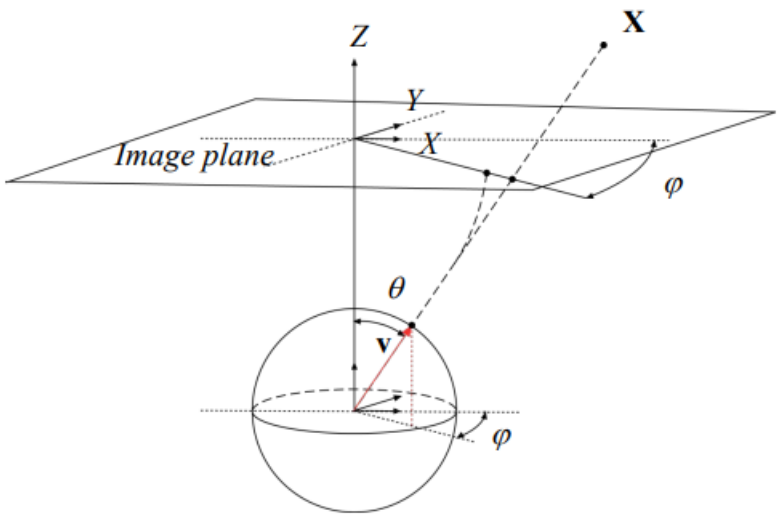


Figura 30. Representación del modelo de proyección Kannala-Brandt.

Se obtiene una buena aproximación de las diferentes curvas de proyección al tomar las cinco primeras potencias de θ, ya que se consiguen suficientes grados de libertad.

Se obtiene entonces una función que transforma un punto 3D definido de forma esférica en el sistema de referencia de la cámara a un punto en la imagen de coordenadas (En este caso, pixeles).

$$F : R^3 \rightarrow R^2$$

Punto 3D:

$$X_p = (r(\theta) * \cos\varphi \quad r(\theta) * sen\,\varphi \quad 1)^T$$

Punto en imagen de coordenadas:

$$X_c = (u \quad v)^T$$

Desarrollo de la función:

$$X_c = \begin{bmatrix} u \\ v \end{bmatrix} = F(X_p) = K_c * X_p = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} r(\theta) * \cos\varphi \\ r(\theta) * sen\,\varphi \\ 1 \end{bmatrix} \quad (2.2)$$

Donde K_c es una matriz que esta compuesta por los parámetros intrínsecos de la cámara.

El modelo inverso de Kannala-Brandt se basa en obtener los puntos 3D a partir de los puntos en la imagen de coordenadas. Se define la función inversa:

$$Q : R^2 \rightarrow R^3 = F^{-1}$$

Desarrollo función:

$$X_p = \begin{bmatrix} x_p \\ y_p \\ z_p \end{bmatrix} = F^{-1}(X_c) = K_c^{-1} * X_c \quad (2.3)$$

Anexo B. Interpolación Bilineal.

Se utiliza la interpolación bilineal para reconstruir cierta imagen de un tamaño a partir de otra imagen de distinto tamaño. En este anexo se explica como se obtiene la interpolación y las razones por las que se utiliza este método para recomponer la imagen.

El método de interpolación permite el cálculo de nuevos datos a partir de un conjunto de valores conocidos. En el caso de la interpolación bilineal se tiene en cuenta los valores en los pixeles conocidos que rodean a uno dado en una vecindad de los 2x2 píxeles más cercanos.

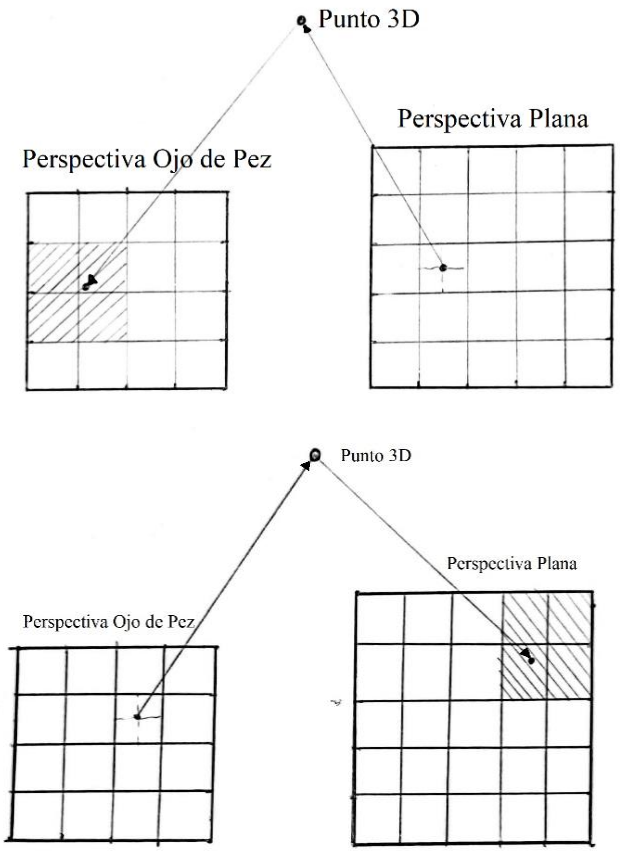


Figura 31. Representación de la aplicación de la interpolación bilineal tanto para recomponer la perspectiva plana como la de ojo de pez.

Se toma el promedio ponderado de estos 4 pixeles y se calcula el valor interpolado. Se utiliza este tipo de interpolación ya que se obtiene un resultado más suavizado que utilizando otros métodos más simples, como el método del píxel más cercano. El tiempo de procesado es mayor que para métodos más simples.

En este trabajo se utiliza este tipo de interpolación para reconstruir las imágenes porque la perspectiva de ojo de pez y la perspectiva plana tienen distinto tamaño, entonces los rayos que van hacia la imagen vacía a reconstruir desde el punto 3D no coinciden con el centro del píxel de la imagen que tiene información y se toma en cuenta una media ponderada de los 4 pixeles mas cercanos a donde cae el rayo. Se mantiene considerablemente la resolución de la imagen original.

Anexo C. Red neuronal MiDaS.

Se utiliza la red neuronal MiDaS en este trabajo y en este anexo se va a explicar que es exactamente una red neuronal y que tiene de especial el uso de la red MiDaS frente a otras redes que tienen la misma finalidad.

Las siglas de MiDaS provienen del siguiente acrónimo 'Monocular Depth Estimation of Dynamic Scenes'. Es una red neuronal profunda que estima la profundidad de escenas dinámicas a partir de imagen monoculares; es decir, imágenes que se capturan con una sola cámara o lente. Es capaz de estimar la distancia entre los objetos en una imagen.

La estructura de esta red neuronal es bastante compleja y consta de varias capas. MiDaS utiliza una arquitectura de red neuronal profunda basada en codificación de características y la decodificación de profundidad. La red se entrena utilizando un conjunto de datos de imagen monoculares y profundidades correspondientes. Durante el entrenamiento, la red aprende a mapear las características de las imágenes a las profundidades correspondientes.

La arquitectura de MiDaS consta de tres componentes principales:

- La extracción de características: Se realiza mediante una red neuronal convolucional (CNN) ya entrenada, la cual extrae características útiles de las imágenes.
- La codificación de características: Se realiza mediante una serie de capas convolucionales y no lineales que transforman las características extraídas en un espacio latente. El espacio latente es un espacio de alta dimensión que se utiliza para representar las características de las imágenes en una forma más compacta y significativa.
- La decodificación de profundidad: Se realiza mediante una serie de capas convolucionales y no lineales que transforman el espacio latente en un mapa de profundidad.

La razón de usar esta red en vez de otras que tienen la misma utilidad se basa en que MiDaS es mejor en términos de precisión y velocidad. La red usada en este trabajo ya está entrenada previamente.

Lista de figuras.

Lista de tablas.