Estimación de la Posición de Objetos en el Plano con una Sola Imagen

Jairo Enrique Ramírez Sánchez, Adrián Augusto Ferrer Orgaz

Tecnológico de Monterrey, Escuela de Ingeniería y Ciencias.
{A01750443,A01749394}@tec.mx

Resumen Las fotografías son representaciones bidimensionales del mundo tridimensional cuya proyección pierde uno de los aspectos más importantes: la profundidad. Se torna relevante contar con algoritmos eficientes con capacidad de obtener la mayor cantidad de información del mundo tridimensional a partir de dicha representación. Con esto, en el presente trabajo presentamos un primer acercamiento a un modelo completo de estimación de la posición de objetos en el plano, con el cual no solo se aproxima la profundidad, sino también la distancia lateral con respecto a la cámara. Nuestro modelo utiliza el algoritmo de detección de objetos YOLOv3 para asignar un cuadro delimitador que permita conocer el ancho y la posición aparente del objeto en la imagen. En seguida, un modelo de ajuste racional para aproximar la profundidad del objeto; dicha profundidad es alimentada a una composición de modelos lineales que toman en cuenta la perspectiva para estimar la posición lateral. Finalmente, tomando como base el error promedio de la estimación en las pruebas, se asigna un área de alta probabilidad de encontrar el objeto. De los experimentos realizados, se obtuvo un error absoluto medio de 0,16 m con $\sigma = 0{,}07$ para la estimación de la distancia lateral y 21,3 m con $\sigma = 0{,}1$ para la estimación de la profundidad. Así, el uso de un modelo matemático ajustado a cada objeto establece a nuestro trabajo como una opción de bajo costo computacional y de fácil escalamiento.

Keywords: visión por computadora, estimación de la posición en el plano, YOLOv3, estimación de profundidad

1. Introducción

El problema de la estimación de profundidad y reconstrucción tridimensional a partir de imagen y video es un problema fundamental de la visión por computadora. Consiste en extraer la mayor cantidad de información en un modelo tridimensional a partir de un medio bidimensional donde se pierde la noción de la profundidad. Los campos de aplicación se pueden extender desde el desarrollo virtual de espacios, navegación, sorteo de obstáculos, arqueología, reconstrucción digital, entre otros. Existen dos principales acercamientos para la solución del problema dependiendo de la cantidad de imagen por muestra disponible, cada una con sus propias marcas y señales fundamentales a considerar (tabla 1) [1].

Tabla 1: Aproximaciones a la reconstrucción tridimensional a partir de imágenes.

Cantidad de imágenes	**Marcas y señales fundamentales**
Dos o más	Disparidad binocular, parallax en moción, silueta, estructura de moción, desenfoque
Una	Perspectiva lineal, forma y sobreado, dispersión atmosférica

Para cualquiera de los casos anteriores existe una lista de parámetros de estandarización que se deben de considerar al tomar las capturas de imágenes. Por un lado existen los parámetros intrínsecos, que se refiere a las especificaciones técnicas de la cámara. Por otro lado existen los parámetros extrínsecos, que incluyen la información relevante a la orientación, inclinación y posición de esta en el momento de la captura [1].

El problema de la estimación de la profundidad en imágenes suele ser abordado con la generación de un mapa de densidad, ya sea con o sin la necesidad de datos etiquetados [2]. De aquí se pueden realizar tareas para extraer información relevante como la determinación de coordenadas tridimensionales [1, 2]. En el caso de la estimación de profundidad a partir de una sola imagen se consideran las siguientes marcas y señales de mayor importancia [2]:

1. Oclusión
2. Perspectiva
3. Tamaño aparente vs real
4. Textura
5. Dispersión atmosférica
6. Patrones de iluminación
7. Altura

Para la percepción humana, estudios muestran que la comparación entre el tamaño aparente vs el real es muy importante para la estimación visual de la profundidad [2]. Al tener conocimiento previo del tamaño real promedio del objeto en cuestión solemos poder estimar la distancia a la que se encuentran. Debido al potencial computacional y de generalización en datos no estructurados, el uso de las Redes Neuronales Convolucionales (CNN) ha mostrado grandes avances y constan como la herramienta común para la solución del problema en su versión de visión stereo. Además los trabajos suelen ser enfocados en bases de datos con ambientes cerrados o abiertos [2–4]. Esta información es valiosa puesto que al modelar inteligencia artificial se suelen abordar los problemas de forma análoga a la cual el humano percibe la realidad y aprende de su entorno. En este caso, atacando ciertas marcas y señales específicas estudiadas en el proceso humano de estimación de profundidad de objetos.

En el presente trabajo se enfoca en la estimación de la posición de diversos objetos en el plano a partir de una sola imagen mediante el análisis de el tamaño aparente en contraste con el real. Son propuestos dos modelos matemáticos para modelar tanto la profundidad como el desplazamiento lateral en tomando a la posición de la cámara el como punto de coordenadas (0,0).

La investigación se encuentra dividida de la siguiente manera: la sección 2 aborda la revisión del trabajo previo sobre la estimación de la distancia de objetos a partir de imágenes. La sección 3 contiene la descripción del modelo propuesto, así como las fases del proceso. La sección 4 explica las condiciones en las que fueron llevados a cabo los experimentos que derivaron en la determinación de los modelos de profundidad y de desplazamiento lateral. Por su parte, la sección 5 muestra los resultados obtenidos para las pruebas realizadas. Finalmente, en la sección 6 se abordan las conclusiones y el trabajo futuro.

2. Trabajos Relacionados

En [4] se realizó una exploración del comportamiento a alto nivel con el cual una CNN de nombre *MonoDepth* estima la profundidad en imágenes únicas. La experimentación se realizó a base de modificaciones y alteraciones de imágenes con las que fue entrenada (provenientes del dataset KITTI) con base en las marcas y señales influyentes en el proceso de estimación humano. Dentro de los resultados relevantes, se encuentra que el proceso de estimación de la red usa principalmente la posición vertical subordinada del tamaño aparente. Además, la estimación puede generalizar a objetos que no se encuentran en el set de entrenamiento, ya que la estimación no depende en sí del objeto, si no de las circunstancias que lo rodean. Además, en [5] se concluye que en el proceso de estimación de la profundidad no solo influye la posición vertical, pero también el fondo de la imagen. Es decir, incluso en caso de tratarse de un ambiente cerrado o abierto las condiciones de el fondo es factor de gran importancia en la estimación de la profundidad.

En [6] se resolvió el problema en el caso de visión stereo, siendo el objetivo la reconstrucción tridimensional para la navegación robótica y aumentar los grados de interacción robot-ambiente. El set de datos utilizados consta de una serie de fotografías de un tipo de tablero de damas a diferentes distancias en un ambiente parecido a un garage. Se utilizó un algoritmo de programación dinámica para el cálculo del campo de disparidad entre cada imagen de cada muestra para realizar la estimación de la profundidad.

En [3] se realiza una mezcla de ambas aproximaciones de cantidad de imágenes. Se utiliza tanto cámaras stereo como mono. El problema a tratar es la detección de obstáculos a larga y corta distancia, desde centímetros hasta un kilómetro, en una infraestructura de trenes. La implementación consta de una combinación de visión térmica, nocturna y visión convencional stereo; la fusión de las entradas permite una mejor captación de información. La detección de obstáculos en si se realiza con el sistema stereo con ayuda de la arquitectura de detección de objetos de Google YOLO, de los cuales se obtuvo la información referente a su posición y caja delimitadora. Posteriormente se calcularon las características relevantes para finalmente ser ingresadas en la arquitectura DisNet y realizar la estimación de la profundidad. El entrenamiento supervisado de la arquitectura se realizó con una base de datos de imágenes tomadas en lineas de tren similares a las cuales se planea realizar la implementación final.

Existen otras integraciones de diferentes elementos para realizar la detección de objetos y marcas y señales. Un ejemplo es [7] donde se realiza detección de obstáculos con visión por computadora y un láser. Otro ejemplo es [8] donde se utiliza el algoritmo de aprendizaje Campo Aleatorio de Markov para capturar marcas y señales monoculares y posteriormente integrarse a un sistema de estimación stereo. Sin embargo, también existen otras configuraciones en las bases de datos utilizadas, como ya fue mencionado anteriormente el estudio suele basarse en un tipo de ambiente, ya sea exteriores o interiores. En [9] se utiliza aprendizaje supervisado estimando mapas de profundidad con una sola imagen. Una particularidad de este trabajo es su uso de imágenes con ambientes interiores y exteriores, contrario a la aproximación usual.

3. Propuesta

El modelo propuesto está dividido en 5 etapas que se muestran en la figura 1.

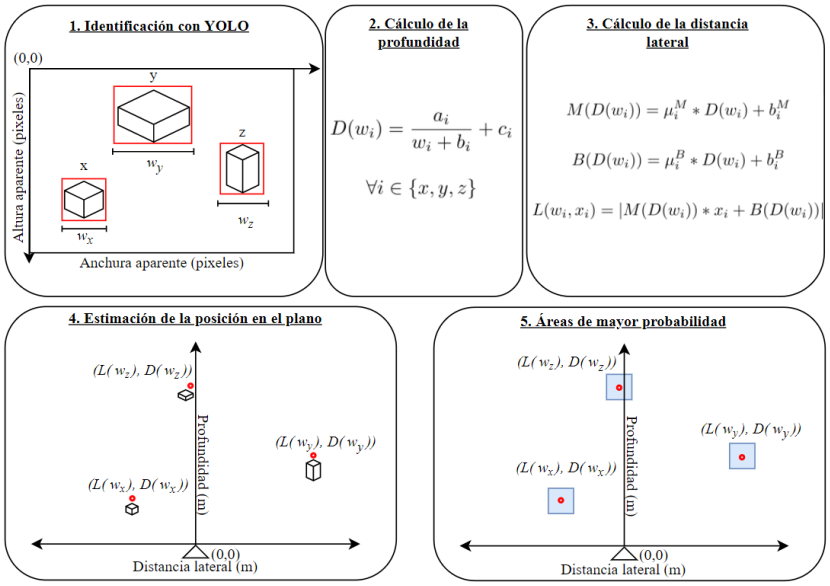


Figura 1: Diagrama del proceso propuesto para la estimación de la posición en el plano.

En primer lugar, una imagen tomada por un celular es ingresada al algoritmo de detección de objetos de YOLOv3 [10], con la finalidad de extraer los recuadros delimitadores de los objetos de interés. Las coordenadas extraídas del reconocimiento son escaladas por las dimensiones de la imagen para obtener valores

entre 0 y 1 que sean invariables a las dimensiones de la fotografía original. Para el caso del eje X de la fotografía, los valores adicional al escalamiento, se recorren 0.5, para obtener una representación de $-0{,}5$ a 0,5 cuyo 0 coincida con la cámara.

En seguida, se estima la profundidad $D(w_i)$ (coordenada en el eje y) a la que se encuentra cada objeto $i \in T$ siendo T el set de todos los objetos que han sido reconocidos por YOLOv3 en la imagen:

$$D(w_i) = \frac{a_i}{w_i + b_i} + c_i \tag{1}$$

Donde:
w_i es la anchura relativa del objeto medida en píxeles escalados entre 0 y 1.
a_i, b_i y c_i son los parámetros del modelo racional, lo cuales se deben de ajustar para cada objeto i.

Posteriormente, $D(w_i)$ es utilizada para realizar el cálculo de la distancia lateral $L(w_i)$ (coordenada en el eje x), la cual se obtiene con la siguiente ecuación:

$$\begin{aligned} L(w_i, x_i) &= |M(D(w_i)) * x_i + B(D(w_i))| \\ M(w_i) &= \mu_i^M * D(w_i) + b_i^M \\ B(w_i) &= \mu_i^B * D(w_i) + b_i^B \end{aligned} \tag{2}$$

Donde:

x_i: es la distancia aparente en píxeles del centro de la imagen al centro del objeto i.
w_i: es la anchura aparente en píxeles del objeto i.
$M(w_i)$: es el modelo lineal que obtiene el valor de la pendiente para el cálculo de la distancia lateral.
$B(w_i)$: es el modelo lineal que obtiene el valor del corte con el eje para el cálculo de la distancia lateral.

Tanto la función $M(w_i)$ como $B(w_i)$ deben de ser ajustadas para cada objeto i. Más detalles sobre cómo fue determinado dicho modelo y su uso se incluyen en la sección n.

Finalmente, para presentar la estimación de la posición del objeto, es graficada una región de probabilidad considerar el error medio de estimación en en el eje X (E_x) y en el eje Y (E_y). Dicha región es el rectángulo cuyos vértices son:

$$\begin{bmatrix} (x - E_x, y + E_y) & (x + E_x, y + E_y) \\ (x - E_x, y - E_y) & (x + E_x, y - E_y) \end{bmatrix} \tag{3}$$

Así no se brinda la estimación como un punto aislado, sino como una región de alta probabilidad tomando en cuenta dichos errores de estimación.

4. Experimentos

La principal marca atacada en el presente es el tamaño aparente de los objetos en las imágenes. Para realizar los experimentos se generó una base de datos con imágenes, posteriormente se presentan las características de la configuración de captura.

Creación del set de entrenamiento Fue seleccionado una lista de objetos identificables con el algoritmo de detección YOLOv3 en dos categorías: objetos de tamaño cercano a la dimensión humana (referido como grupo A): personas, mesas, sillas, bicicletas; y objetos de fácil manipulación manual (referido como grupo B): manzanas, plátanos, vasos y tazas. Para ambos grupos de objetos, fueron tomadas 10 fotografías de variaciones del mismo objeto, en diferentes lugares y condiciones de iluminación, ejecutando variaciones en la profundidad y distancia lateral tomando a la cámara como el punto de referencia $(0, 0)$. Dichas distancias se muestran en la tabla 2:

Tabla 2: Profundidad y distancia lateral a la que fueron tomadas las fotografías.

Grupo	**Profundidad (m)**			**Distancia lateral (m)**		
	Min	**Max**	**Intervalo (m)**	**Min**	**Max**	**Intervalo (m)**
A	1	6	1	0	2	0.5
B	0.2	1	0.2	0	0.3	0.1

Las cámara con la que se tomaron las capturas pertenece a un celular Samsung modelo A30, las especificaciones técnicas de esta se encuentran en la tabla 3.

Tabla 3: Especificaciones de la cámara utilizada para realizar las capturas.

Megapixeles	**Apertura**	**Distancia Focal**
16	F1.7	27 mm

Obtención de características Cada una de las imágenes del set de entrenamiento anterior fueron procesadas con el algoritmo de detección YOLOv3 y se extrajo la anchura en píxeles de la caja delimitadora calculada por este (figura 2). Se procede a almacenar la anchura promedio del objeto para cada uno de estos a cada distancia.

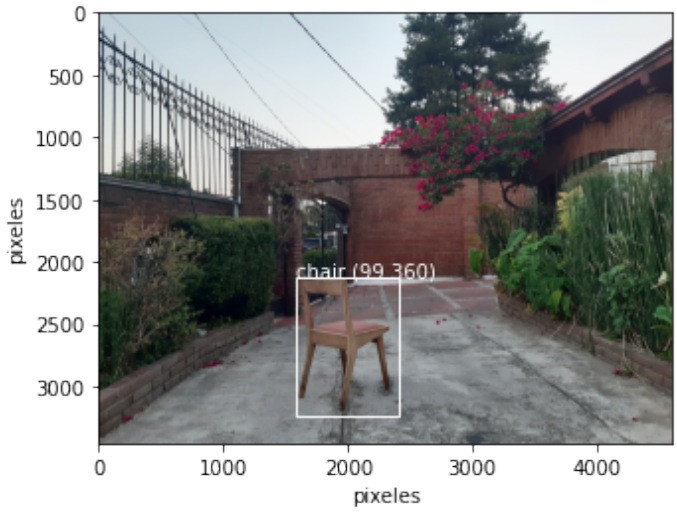
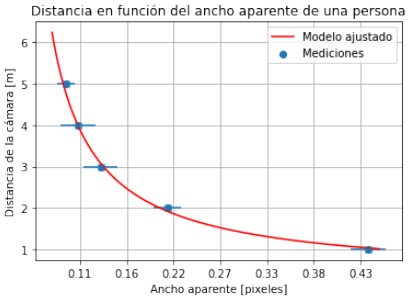


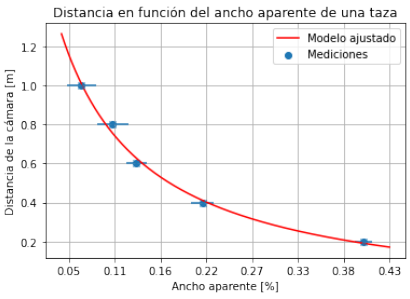
Figura 2: Ejemplo de detección del algoritmo YOLO en el set de datos A. Cámara a 0.5m de altura; silla a 3m de profundidad y 0.5m de distancia lateral.

4.1. Modelo racional de estimación de profundidad

Con las anchuras en píxeles promedio obtenidas anteriormente se procede a optimizar los parámetros para cada objeto según el modelo de la ecuación 1, con el algoritmo de optimización no restringida Broyden-Fletcher-Goldfarb-Shanno con la biblioteca de Python Scipy [11]. La función objetivo elegida para el proceso de optimización fue el error cuadrado. Un ejemplo se muestra en la figura 3).



(a) Modelo ajustado para un objeto del grupo A.



(b) Modelo ajustado para un objeto del grupo B.

Figura 3: Ejemplos del ajuste de los modelos racionales de profundidad obtenidos.

4.2. Modelo de estimación de la distancia lateral

Para el caso de la distancia lateral, se analizó la distancia relativa en píxeles del centro del objeto al centro en su componente x para cada una de las profundidades. Un ejemplo para un objeto del grupo A se muestra en la figura 4:

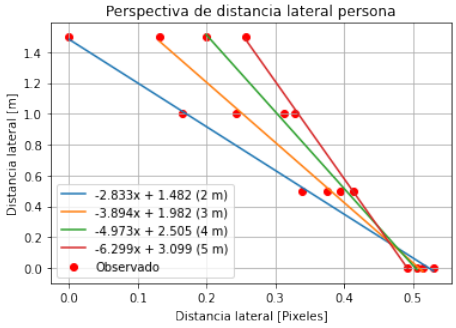
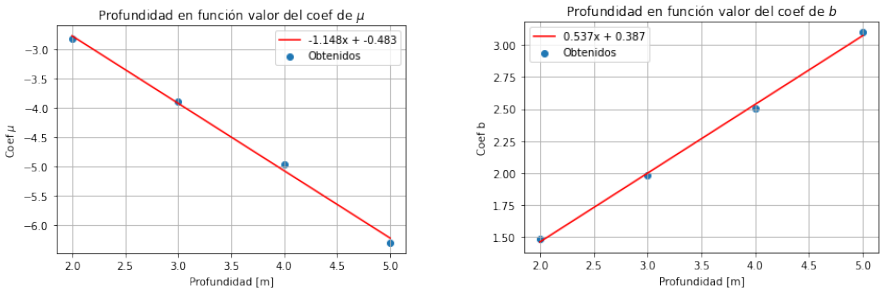


Figura 4: Perspectiva de distancia lateral en función de la distancia.

Es destacable la naturaleza del comportamiento lineal de la distancia lateral, el cual tiene ligeras modificaciones en la pendiente en función de la distancia a la que se encuentre. Más aún, los parámetros de las ecuaciones, es decir, la pendiente μ y el corte con el eje b de dichas ecuaciones se comportan de manera lineal, como se muestra en la figura 5.



(a) (b)

Figura 5: Ajuste de modelo lineal en función de la profundidad para los coeficientes del modelo de distancia lateral.

Con lo cual, es factible crear un modelo lineal que aproxime los coeficientes de la ecuación para el desplazamiento lateral, en función de la profundidad.

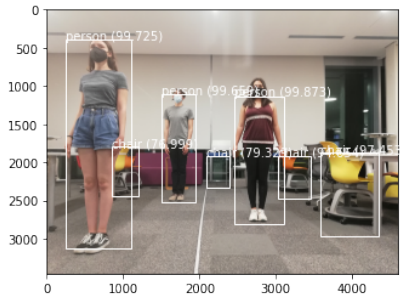
5. Resultados

Como resultado de los experimentos realizados se encontró la relación del error de estimación en función de la distancia a la que se encontraba el objeto. Se muestran en la tabla 4.

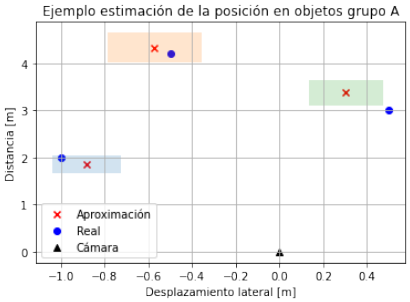
Tabla 4: Errores promedio según el intervalo de de distancia

Distancia (m)	**Error promedio en el eje Y (m)**	**Error promedio en el eje X (m)**	**Grupo del objeto**
(0 - 0.5]	0.046	0.045	B
(0.5 - 1]	0.062	0.052	B
(1 - 2]	0.158	0.18	B
(2 - 3]	0.175	0.24	A
(3 - 4]	0.191	0.26	A
(4 - 5]	0.234	0.31	A
(5 - 6]	0.26	0.35	A

Conocer este error de estimación permite realizar predicciones dentro de un intervalo de confianza y así, poder denotar regiones de probabilidad donde encontrar el objeto en lugar de un solo punto, como se muestra en la figura 8.



(a) Fotografía de ingreso al modelo. Solo las personas son consideradas para estimar la posición.



(b) Plano de estimación de la posición real vs aproximada con las áreas de mayor probabilidad.

Figura 6: Demostración para objetos del grupo A

Finalmente, para probar la precisión del modelo presentado, fueron capturas 50 fotografías de diferentes distribuciones de objetos en el plano, variando la cantidad de objetos presente en la imagen, la distribución (profundidad y distancia lateral), condiciones de luz y tipos de objetos. Algunos ejemplos se muestran en la figura 7.

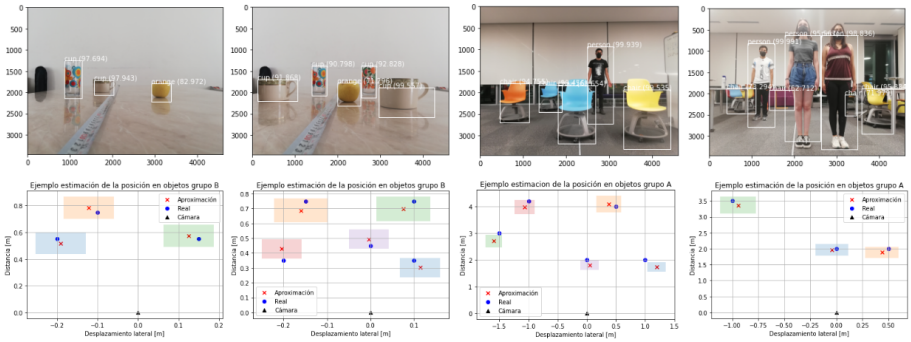


Figura 7: Ejemplos de pruebas.

Se obtuvo un error absoluto medio promedio de 0,16 m con $\sigma = 0{,}07$ para la estimación de la distancia lateral y 21,3 m con $\sigma = 0{,}1$ para la profundidad.

5.1. Contraste con un trabajo relacionado

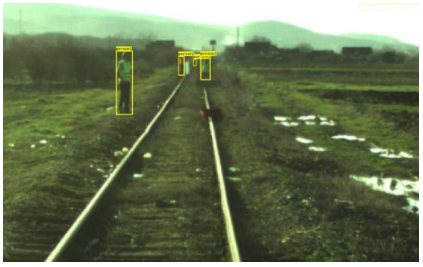
En [3] fue presentado un sistema de estimación de la profundidad (DisNet) utilizando redes neuronales cuyo funcionamiento fue demostrado sobre dos imagenes del H2020 Shift2Rail project SMART [12]. Este trabajo y sus resultados son tomados como referencia para realizar una comparación del presente modelo. Los resultados se muestran en la tabla 5.

Tabla 5: Comparación del presente modelo con DisNet para la estimación de la profundidad

Figura	**Objeto**	**Rail scene**			**Distancia lateral**
		Real	**DisNet**	**Propio**	
9(a)	Persona 1	100 m	101.89 m	114.02 m	1.46 m
	Persona 2		99.44 m	117.43 m	4.12 m
9 (b)	Persona 1	50 m	54.26 m	51.44 m	4.26 m
	Persona 2	150 m	167.59 m	172.81 m	3.83 m
	Persona 3	100 m	133.26 m	128.57 m	1.49 m
	Persona 4	300 m	338.51 m	492.67 m	3.51 m

Como se puede observar, el modelo de ajuste racional proporciona aproximaciones adecuadas, generalizando incluso para distancias mucho mayores con las que fue entrenado. Por la parte de la distancia lateral, el dataset no presenta un punto de comparación para este dato.

(a) Persona 1 y 2 a 100m de distancia



(b) Persona 1 a 50m, 2 a 100m, 3 a 150m y 4 a 300m

Figura 8: Estimación de la profundidad en Railway Scene - Distance estimation from the single RGB identificación

6. Conclusiones

Como se ha demostrado a lo largo del presente trabajo, la estimación de la posición a partir de una sola imagen es un rubro en donde se han propuesto diversas aproximaciones con diferentes costos computacionales y complejidades. El modelo presentado es un primer acercamiento a la construcción de un sistema completo de estimación en el plano de la posición con una sola imagen. Tomando como base los resultados, podemos asegurar que cuenta con las bondades de ser escalable - al requerir de un set considerablemente pequeño de datos para ajustar los modelos de regresión, de bajo costo computacional y preciso al obtener un error absoluto promedio de 0,16 m con $\sigma = 0{,}07$ para la estimación de la distancia lateral y 21,3 m con $\sigma = 0{,}1$ para la profundidad; dichas características lo posicionan como una opción totalmente viable para ser aplicado en sistemas de planificación de rutas o robots de bajo costo.

Como trabajo a futuro se planea extender el estudio a más objetos y en mayor diversidad de posiciones que permitan robustecer el sistema cada vez más. Adicionalmente, se proyecta incluir un algoritmos de optimización para encontrar la mejor ruta que permita a la cámara desplazarse entre los objetos que tiene en frente.

Referencias

1. Aharchi, M., Ait Kbir, M.: A Review on 3D Reconstruction Techniques from 2D Images. (2020) 510–522
2. Mertan, A., Duff, D.J., Unal, G.: Single image depth estimation: An overview (2022)
3. Haseeb, M.A., Guan, J., Ristić, D., Gräser, A.: DisNet : A novel method for distance estimation from monocular camera. 10th Planning, Perception and Navigation for Intelligent Vehicles (2018)
4. Van Dijk, T., De Croon, G., Nl, G.C.H.E.D.: (How do neural networks see depth in single images?)

5. Epstein, W.: Perceived depth as a function of relative height under three background conditions (1966)
6. Ann, N.Q., Achmad, M.S., Bayuaji, L., Daud, M.R., Pebrianti, D.: Study on 3D scene reconstruction in robot navigation using stereo vision. In: Proceedings - 2016 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS 2016, Institute of Electrical and Electronics Engineers Inc. (2017) 72–77
7. F. Jiménez, J. E. Naranjo, J.J.A.F.G.A.P.J.M.A.: Advanced driver assistance system for road environments to improve safety and efficiency (2016)
8. N. Bernini, M. Bertozzi, L.C.M.P.M.S.: Real-time obstacle detection using stereo vision for autonomous ground vehicles: A survey (2014)
9. A. Saxena, H.·Sung, A.Y.N.: 3-d depth reconstruction from a single still image (2007)
10. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. (2018)
11. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17** (2020) 261–272
12. of Rail Transport, S.S.A.: Shift2rail project smart (2014) `http://www.smartrail-automation-project.net`.