EXPLORANDO LOS FUNDAMENTOS DE LA RECONSTRUCCIÓN 3D CON STRUCTURE FROM MOTION: IMPLEMENTACIÓN Y PRUEBAS CON MONUMENTOS DE LA UIS

Paula Uzcátegui León
Ingeniería de Sistemas, Universidad Industrial de Santander, Colombia,
paula2211475@correo.uis.edu.co

Andrea Juliana Parra Ariza
Ingeniería de Sistemas, Universidad Industrial de Santander, Colombia,
andrea2210062@correo.uis.edu.co

Jorge Andrey Garcia Vanegas
Ingeniería de Sistemas, Universidad Industrial de Santander, Colombia,
jorge2180115@correo.uis.edu.co

Ramiro Santiago Avila Chacón
Ingeniería de Sistemas, Universidad Industrial de Santander, Colombia,
ramiro2191920@correo.uis.edu.co

Hoover Fabián Rueda Chacón, PhD.
Ingeniería de Sistemas, Universidad Industrial de Santander, Colombia,
hfarueda@uis.edu.co

Resumen.

Introducción: *Structure from Motion (SfM)* es una técnica fotogramétrica utilizada para la reconstrucción 3D, que estima la estructura tridimensional de una escena a partir de múltiples imágenes bidimensionales. **Objetivo**: Este artículo implementa los principios de SfM para reconstruir digitalmente estatuas y monumentos en la Universidad Industrial de Santander (UIS). **Metodología:** Las imágenes RGB se capturan en diversas locaciones de la UIS y luego se procesan siguiendo un pipeline de SfM que incluye la calibración de la cámara, procesamiento y emparejamiento de puntos clave, estimación de las poses de la cámara, triangulación en 3D, y ajuste de paquete. **Resultados:** Los resultados muestran nubes de puntos que representan las reconstrucciones 3D de las estatuas, así como su postprocesamiento en blender. **Conclusiones:** Cada etapa del pipeline de SfM presenta oportunidades de optimización, como lo muestran las últimas implementaciones del estado del arte.

Palabras clave: Reconstrucción 3D; Structure from Motion (SfM); Geometría epipolar; Calibración; Monumentos.

1. Introducción

La reconstrucción 3D es el proceso de capturar escenas del mundo real y crear representaciones digitales tridimensionales. Existen diversas técnicas para llevar a cabo este proceso, incluyendo, la fotogrametría, LiDAR, luz estructurada, y Structure from Motion (SFM) (Ham, Wesley, & Hendra, 2019). SFM es una técnica que utiliza una serie de imágenes bidimensionales de una escena para reconstruir su estructura tridimensional. A diferencia de otras técnicas, una sola imagen carece de la información necesaria para una reconstrucción 3D, ya que al capturar una escena se proyecta desde tres dimensiones a dos, perdiendo la información de profundidad (Aharchi & Ait Kbir, 2020).

La ventaja de SFM, que se basa en los principios de la fotogrametría, es que no requiere conocer de antemano las diferentes posiciones de la cámara, lo que la convierte en una opción económica y eficiente, sin necesidad de equipos especializados. Esta técnica ha sido aplicada en diversos campos, incluyendo la reconstrucción de superficies corporales para el monitoreo de pacientes (Abreu de Souza et al., 2023), el análisis de estructuras forestales en monitoreo remoto (Frey et al., 2018), y la reconstrucción de estructuras históricas, como en el proyecto “Building Rome in a Day” (Agarwal et al., 2011).

Aunque los principios básicos de SFM están bien establecidos, cada etapa del proceso ofrece oportunidades para la optimización, especialmente mediante el uso de nuevas tecnologías como el aprendizaje profundo (deep learning). En este trabajo, presentamos una implementación de los principios fundamentales de SFM, con el objetivo de comprender en profundidad cada fase del proceso y explorar posibles mejoras.

2. Metodología

El proceso de reconstrucción 3D mediante SFM se puede dividir en dos etapas principales: primero, la adquisición de imágenes de la escena desde múltiples ángulos y la calibración de la cámara, para asegurar la precisión geométrica; segundo, la estimación de la matriz fundamental, la reconstrucción inicial de la escena utilizando un par de imágenes y la integración progresiva de imágenes adicionales para expandir y densificar la nube de puntos 3D, mejorando la representación completa de la estructura. A esta etapa la llamaremos SfM.

2.1 Adquisición de imágenes y calibración de la cámara

La adquisición de las imágenes se realiza en el campus central de la UIS en un entorno abierto para el monumento “Lección de Geometría” y entornos cerrados para las estatuas. El sensor utilizado para la captura de las imágenes fue el de la cámara del celular Iphone XR con características de una cámara gran angular de 12MP, apertura de f/1.8, lente de seis elementos y fotos panorámicas (hasta 63MP).

Posteriormente, se procede a la calibración de la cámara, para determinar la relación exacta entre los puntos 3D del mundo real y sus proyecciones 2D correspondientes en la imagen capturada, es necesario tomar en cuenta las posibles distorsiones o efectos visuales que aplique la lente y demás partes de la cámara a las fotos.

Matemáticamente, esta relación se representa por las Ecuaciones 1 y 2 (Zhang, 2000).

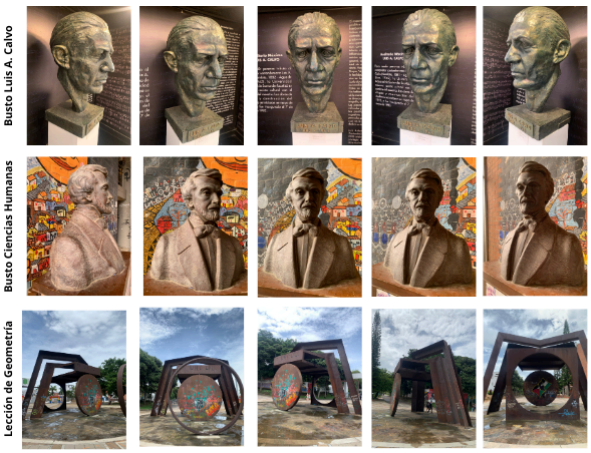
$$s\tilde{m} = A[Rt]\tilde{M} \quad (1)$$

$$\begin{aligned} \hat{x} &= x + x[k_1(x^2+y^2) + k_2(x^2+y^2)^2] \\ \hat{y} &= y + y[k_1(x^2+y^2) + k_2(x^2+y^2)^2] \end{aligned} \quad (2)$$

Para (1) $\tilde{m}$ es el vector $m = [u, v]$ aumentado que representa un punto 2D, $\tilde{M}$ es el vector $M = [x, y, z]$ aumentado que representa un punto 3D, s un factor de escala arbitrario, (R, t) son los parámetros extrínsecos de rotación y traslación de la cámara y A sus parámetros intrínsecos. En cuanto a (2) (x, y) son las coordenadas ideales de la imagen, $(\hat{x}, \hat{y})$ las distorsionadas, k_1 y k_2 los coeficientes de distorsión radial

Figura 1

Adquisición de las imágenes para las estatuas y monumento



Fuente: Elaborada por los autores

La idea de calibrar la cámara con la que se tomaron las fotos es obtener los parámetros A (que llamaremos K) y k para corregir la proyección de los puntos.

Se empleó el algoritmo de calibración con un patrón de tablero de ajedrez, propuesto por Zhang (1998) y que continúa siendo el estándar en su campo hoy en día. Su proceso consiste en lo siguiente: Imprimir y pegar el patrón en una superficie plana y tomar múltiples imágenes desde diferentes ángulos y posiciones, detectar los puntos en que interseccionan las casillas en cada imagen, estimar los parámetros intrínsecos y extrínsecos, estimar los coeficientes de distorsión radial, y optimizar los parámetros minimizando el error entre los puntos reales y

sus proyecciones. Se utilizaron las implementaciones de este método hechas por MATLAB y con OpenCV.

2.2. SFM

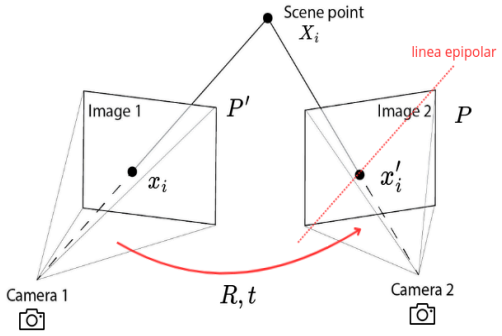
Structure From Motion se basa en los mismos principios de la fotogrametría estéreo, pero se diferencia en que las posiciones de las cámaras no son conocidas de antemano. Para determinar estas posiciones, se aplican las reglas de la geometría epipolar (ver Figura 2). El objetivo es calcular las matrices de proyección de las cámaras, denotadas como P y P ′, las cuales están definidas por la rotación y traslación de una cámara con respecto a la otra. La geometría epipolar establece la relación entre dos imágenes de la misma escena capturadas desde diferentes puntos de vista, conforme a las ecuaciones en (3) (Hartley y Zisserman, 2003).

$$x_i = PX_i \qquad (3)$$
$$x_i' = P'X_i$$

Donde x_i y x_i' son píxeles que corresponden al mismo punto Xi en la escena real.

Figura 2

Representación de geometría epipolar



Fuente: Imagen adaptada de Epipolar Geometry (Qureshi, 2020) con algunas modificaciones

Para calcular las matrices de proyección de las cámaras, la geometría epipolar introduce el concepto de la Matriz Fundamental, que contiene toda la información que describe la geometría de la escena (Ecuación 4).

$$x^T F x' = 0 \qquad (4)$$

Una vez obtenida la matriz fundamental, se puede calcular la matriz esencial, la cual está relacionada con la matriz de los parámetros intrínsecos de la cámara, K (como se muestra en la Ecuación 5). Al realizar la descomposición en valores singulares (SVD) de la matriz esencial (Abdi, 2007), es posible determinar las matrices de rotación y traslación de una cámara con respecto a la otra.

$$K^T F K = 0 \qquad (5)$$

Para encontrar la matriz fundamental F y, en última instancia, los parámetros de las cámaras, es necesario identificar puntos x_i y x_i' en las imágenes que correspondan al mismo punto en la escena, lo que se conoce como el **problema de correspondencia**. SfM cuenta con una serie de pasos bien definidos para resolver este problema, los cuales se detallan a continuación. Los algoritmos mencionados han sido implementados utilizando la biblioteca de Python OpenCV.

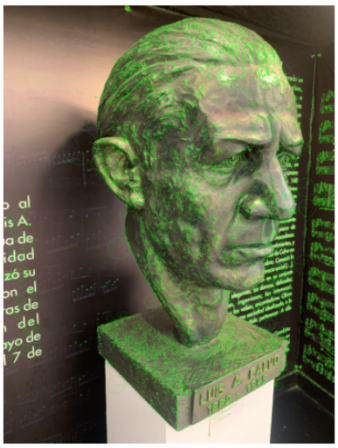
Extracción de características: Este paso consiste en encontrar puntos clave en la imagen. Un punto clave es un punto de interés que posee propiedades únicas o distintivas, como una esquina, un borde, un cambio de textura, u otro tipo de punto que sea estable y repetible bajo diferentes condiciones. Para ello, existen algoritmos como SIFT (*Scale-Invariant Feature Transform*), ORB (*Oriented FAST and Rotated BRIEF*), y SURF (*Speeded Up Robust Features*) (Bansal, Kumar, y Kumar, 2021), los cuales buscan características distintivas en la imagen que sean estables frente a cambios de escala y rotación. En nuestra implementación, utilizamos SIFT (ver Figura 3), que está implementado en la librería y es capaz de detectar miles de puntos clave, especialmente en áreas con texturas y colores variados, o en las esquinas.

Emparejamiento de puntos: Una vez identificados los puntos clave en ambas imágenes, es necesario emparejar aquellos que probablemente correspondan al mismo punto de la escena. Los algoritmos de detección de puntos clave mencionados anteriormente también generan descriptores para cada punto, los cuales son vectores que ofrecen una descripción única basada en su vecindario local. El emparejamiento de puntos se realiza buscando los descriptores más cercanos entre sí. En OpenCV, esto se puede hacer mediante un método de fuerza bruta o utilizando el algoritmo **FLANN** (*Fast Library for Approximate Nearest Neighbors*) (Muja y Lowe, 2014).

En nuestra implementación, utilizamos FLANN (ver Figura 4), lo que permite reducir la cantidad de puntos clave a aquellos más útiles para la reconstrucción.

Figura 3.

Puntos clave (verde) obtenidos con SIFT para la estatua de Luis A. Calvo.



Fuente: Elaborada por autores.

Figura 4.

Emparejamiento de puntos clave de dos imágenes consecutivas de la estatua de Luis A. Calvo.

Fuente: Elaborada por autores.

Cálculo de la matriz fundamental: El siguiente paso consiste en computar la matriz

fundamental resolviendo un sistema de ecuaciones basado en los puntos emparejados. Para ello, se utiliza el algoritmo 8Points (Hartley, 1997), el cual, como su nombre indica, requiere un mínimo de ocho puntos para calcular la matriz. Cuanto más precisos sean los emparejamientos de puntos, más exacta será la matriz fundamental resultante. En OpenCV, existen versiones optimizadas del método de 8Points que emplean técnicas como **RANSAC** (*Random Sample Consensus*) (Fischler y Bolles, 1981) o **LMEDS** (*Least Median of Squares*) (Rousseeuw, 1984) para eliminar los valores atípicos en los emparejamientos.

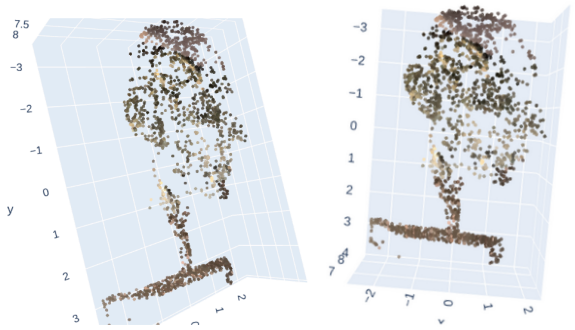
Una vez obtenida la matriz fundamental, se calcula la matriz esencial, junto con las matrices de rotación y traslación de una cámara con respecto a la otra, siguiendo las ecuaciones mencionadas anteriormente.

Después de resolver el problema de correspondencia, se aborda el problema de reconstrucción para determinar la posición tridimensional de cada píxel. SfM también define pasos específicos para llevar a cabo esta tarea:

Triangulación: Una vez conocidas las posiciones de las cámaras y las correspondencias de puntos, se utiliza la triangulación para calcular las posiciones tridimensionales de los puntos en la escena. Esto se logra proyectando rayos desde cada cámara a través de los puntos correspondientes en las imágenes. La intersección de estos rayos determina la posición 3D de los puntos en la escena. Este proceso se implementó mediante una función de OpenCV y, como se muestra en la Figura 5, permite obtener una reconstrucción del perfil de la estatua con 2.281 puntos.

Figura 5.

Nube de puntos inicial obtenida a partir de dos imágenes. A la izquierda se muestra una vista frontal de la nube de puntos, mientras que a la derecha se presenta una vista rotada hacia la izquierda.



Fuente: Elaborada por autores.

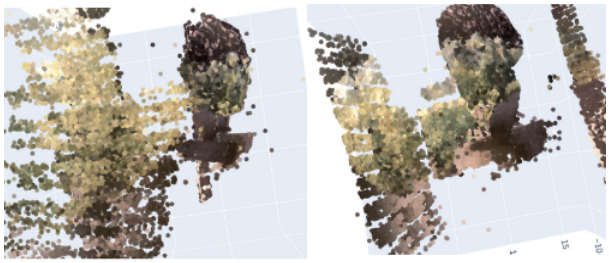
Agregando múltiples vistas: Para ampliar la nube de puntos obtenida mediante la triangulación de dos imágenes, es necesario determinar la posición de la siguiente imagen en relación con la escena 3D reconstruida hasta el momento. Este es un problema ampliamente estudiado en el campo de la visión por computadora. En OpenCV, se utiliza el algoritmo *Perspective-n-Point* (PnP) (Lepetit y Fua, 2006) para calcular la posición de la siguiente cámara (a partir de puntos 2D)

con respecto a la triangulación de las vistas previas (puntos 3D). Una vez determinada la posición de la nueva cámara, se realiza nuevamente la triangulación con la imagen anterior, concatenando los nuevos puntos a la nube de puntos existente. Como se muestra en la Figura 6, este enfoque permite obtener una reconstrucción detallada de la estatua utilizando SfM con 27 imágenes y 72.563 puntos, como se puede observar en la Tabla 1.

Además, cada vez que se realiza una triangulación, se aplica un proceso de optimización conocido como *Bundle Adjustment*. Este proceso busca minimizar el error de reproyección entre la nube de puntos tridimensional obtenida y las imágenes bidimensionales originales de las que se derivó

Figura 6.

Reconstrucción de la estatua de Luis A. Calvo con múltiples vistas.



Fuente: Elaborada por autores.

3. Resultados

Se lograron implementar de manera efectiva los principios fundamentales de Structure from Motion (SfM) utilizando Python, replicando exitosamente cada etapa del pipeline, desde la calibración de la cámara hasta la generación de nubes de puntos tridimensionales. Esta implementación no requirió el uso de equipos especializados ni posiciones de cámara previamente conocidas, consiguiendo un resultado económico, flexible y fácil de recrear. En los bustos de ciencias humanas y de Luis A. Calvo, los resultados fueron especialmente satisfactorios, obteniendo reconstrucciones lo suficientemente buenas para distinguirlas a simple vista; en cambio, en la lección a la geometría, se obtuvieron muy pocos puntos, y no hubo un buen resultado. En cuanto la obtención de los modelos 3D se hizo uso de blender y su sistema para modificar la geometría de un objeto: *Geometry Nodes*, más específicamente con los nodos Puntos a Volumen y Volumen a Malla. Luego de definir un radio para el volumen que ocupa cada punto en la nube y establecer la cantidad de voxeles (equivalente tridimensional de un píxel) que rellenarán determinado volumen, este se transforma en una superficie o malla según un umbral de densidad; entre más bajo el umbral se cubrirán las áreas menos densas. Las estructuras resultantes se muestran en la Figura 7; la primera hecha con un total de 410800 voxeles, un radio de 0.2 metros y un umbral del 0.1 y su siguiente con 157000 voxeles, un radio de 0.1 metros y un umbral de 1.

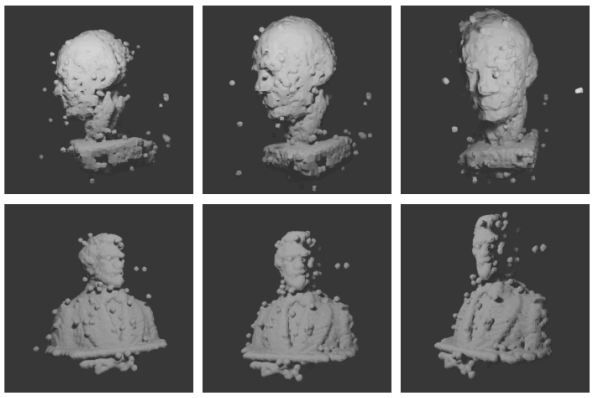
Tabla 1: *Número de puntos obtenidos por el número de imágenes para los monumentos probados*

Estatuas	Imágenes	Puntos obtenidos
Luis. A Calvo	27	72.563
C. Humanas	16	162.306
Lección de geometría	19	3.244

Fuente: Elaborada por los autores

Figura 7

Renderizado de la nube de puntos en Blender



Fuente: Elaborada por autores.

3. Discusión

Nuestro método logró reconstrucciones precisas en imágenes con buena visibilidad y múltiples texturas distintivas, pero enfrentó desafíos debido a sombras y texturas repetitivas, lo que causó emparejamientos incorrectos en algunos casos. Aunque utilizamos algoritmos tradicionales y SfM en Python, se sugiere que métodos modernos, como el de Sarlin, DeTone, Malisiewicz, y Rabinovich (2019), mejorarían el emparejamiento en condiciones difíciles. La calibración previa de la cámara permitió geometrías más precisas, acercándonos al modelo euclidiano, en comparación con trabajos anteriores como el de Beardsley, Torr, y Zisserman (1996). Para optimizar tiempos, nos enfocamos en reconstrucciones parciales, mejorando la eficiencia a costa de algunos detalles.

Es importante destacar que los parámetros como el número de vóxeles y el radio por punto utilizados durante el renderizado en Blender variaron considerablemente entre las diferentes estructuras. En el caso de los bustos, la nube de puntos generada fue lo suficientemente densa y precisa como para permitir un buen proceso de renderizado. Sin embargo, en el monumento "Lección de Geometría", la calidad de los puntos obtenidos fue más baja, lo que afectó negativamente la reconstrucción final.

4. Conclusiones

En este artículo, se utiliza la técnica de *Structure from Motion* (SfM) para realizar reconstrucciones 3D de diferentes monumentos en la Universidad Industrial de Santander (UIS) a partir de imágenes bidimensionales RGB. Los principios básicos de SfM, utilizando algoritmos tradicionales, han demostrado ser eficientes, exhaustivos y con tiempos de procesamiento reducidos. Sin embargo, cada etapa del proceso presenta oportunidades de optimización, desde la identificación y emparejamiento de puntos

clave hasta el postprocesamiento de la nube de puntos, como se evidencia en mejoras implementadas en el estado del arte. Finalmente, UISFM se pone a disposición del público como una herramienta de código abierto para facilitar la realización de reconstrucciones 3D por otros usuarios.

Agradecimientos

Agradecemos al Semillero de Investigación Hands-On Computer Vision por su valiosa dirección y apoyo durante el desarrollo de este proyecto.

Referencias

Abdi, H. (2007). Singular value decomposition (svd) and generalized singular value decomposition. Encyclopedia of measurement and statistics, 907(912), 44. Descargado de https://www.cimat.mx/~alram/met_num/clases/Abdi-SVD2007-pretty.pdf

Abreu de Souza, M., Alka Cordeiro, D. C., Oliveira, J. d., Oliveira, M. F. A. d., y Bonafini, B. L. (2023). 3d multi-modality medical imaging: Combining anatomical andinfrared thermal images for 3d reconstruction. Sensors, 23(3). Descargado de https://www.mdpi.com/1424-8220/23/3/1610 doi: 10.3390/s23031610

Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., y Szeliski, R. (2011, oct). Building rome in a day. Commun. ACM, 54(10), 105–112. Descargado de https://doi.org/10.1145/2001269.2001293 doi:10.1145/2001269.2001293

Aharchi, M., y Ait Kbir, M. (2020). A review on 3d reconstruction techniques from 2d images. En M. Ben Ahmed, A. A. Boudhir, D. Santos, M. El Aroussi, y İ. R. Karas (Eds.), Innovations in smart cities applications edition 3 (pp. 510–522). Cham: Springer International Publishing. doi: 10.1007/978-3-030-37629-1_37

Bansal, M., Kumar, M., y Kumar, M. (2021). 2d object recognition: A comparative analysis of sift, surf and orb feature descriptors. Multimedia Tools and Applications, 80, 18839–18857. doi: 10.1007/s11042-021-10646-0

Beardsley, P., Torr, P., y Zisserman, A. (1996, 01). 3d model acquisition from extended image sequences. En Computer vision — eccv ’96 (Vol. 1065, pp. 683–695). Springer Berlin Heidelberg. doi: 10.1007/3-540-61123-1_181

Fischler, M. A., y Bolles, R. C. (1981, jun). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM, 24(6), 381–395. doi: 10.1145/358669.358692

Frey, J., Kovach, K., Stemmler, S., y Koch, B. (2018). Uav photogrammetry of forests as a vulnerable process. a sensitivity analysis for a structure from motion rgb-image pipeline. Remote Sensing, 10(6). Descargado de https://www.mdpi.com/2072-4292/10/6/912 doi: 10.3390/rs10060912

Ham, H., Wesley, J., y Hendra, H. (2019). Computer vision based 3d reconstruction : A review. International journal of electrical and computer engineering (Malacca, Malacca), 9(4), 2394-. doi: 10.11591/ijece.v9i4.pp2394-2402

Hartley, R. (1997, junio). In defense of the eight-point algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(6), 580–593. doi: 10.1109/34.601246

Hartley, R., y Zisserman, A. (2003). Multiple view geometry in computer vision (2nd ed. ed.). Cambridge, UK ;: Cambridge University Press. doi: 10.1017/cbo9780511811685

Lepetit, V., y Fua, P. (2006, septiembre). Keypoint recognition using randomized trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(9), 1465–1479. doi: 10.1109/tpami.2006.188

Muja, M., y Lowe, D. G. (2014, noviembre). Scalable nearest neighbor algorithms for high dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(11), 2227–2240. doi: 10.1109/tpami.2014.2321376

Qureshi, F. (2020). Epipolar geometry. Ontario Tech University. Descargado 2024-09-09, de https://csundergrad.science.uoit.ca/courses/cv-notes/notebooks/21-epipolar-geometry.html

Rousseeuw, P. J. (1984, diciembre). Least median of squares regression. Journal of the American Statistical Association, 79(388), 871–880. doi: 10.1080/01621459.1984.10477105

Sarlin, P.-E., DeTone, D., Malisiewicz, T., y Rabinovich, A. (2019). Superglue: Learning feature matching with graph neural networks. arXiv. doi: 10.48550/ARXIV.1911.11763

Zhang, Z. (1998). Determining the epipolar geometry and its uncertainty: A review. International Journal of Computer Vision, 27(2), 161–195. doi: 10.1023/a:1007941100561

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