ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE SISTEMAS

UNIDAD DE TITULACIÓN

ESTABILIZACIÓN DE VIDEO EN BASE A LA DISCRIMINACIÓN DE PUNTOS DE INTERÉS UTILIZANDO MAPAS DE PROFUNDIDAD

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGISTER DE INVESTIGACIÓN EN COMPUTACIÓN

HERNANDO HERIBERTO MANOTOA SANDOVAL
Hernando.manotoa@epn.edu.ec

Director: Ph.D. Wilbert G. Aguilar
wilbert.aguilar@epn.edu.ec

Co-director: Ph.D. Sang Guun Yoo
sang.yoo@epn.edu.ec

2022

APROBACIÓN DEL DIRECTOR

Como director del trabajo de titulación ESTABILIZACIÓN DE VIDEO EN BASE A LA DISCRIMINACIÓN DE PUNTOS DE INTERÉS UTILIZANDO MAPAS DE PROFUNDIDAD desarrollado por HERNANDO HERIBERTO MANOTOA SANDOVAL, estudiante de la MAESTRÍA EN COMPUTACIÓN, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral.

Wilbert G. Aguilar

DIRECTOR

APROBACIÓN DEL CO-DIRECTOR

Como director del trabajo de titulación ESTABILIZACIÓN DE VIDEO EN BASE A LA DISCRIMINACIÓN DE PUNTOS DE INTERÉS UTILIZANDO MAPAS DE PROFUNDIDAD desarrollado por HERNANDO HERIBERTO MANOTOA SANDOVAL, estudiante de la MAESTRÍA EN COMPUTACIÓN, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral.

Sang Guun Yoo
CO-DIRECTOR

DECLARACIÓN DE AUTORÍA

Yo, HERNANDO HERIBERTO MANOTOA SANDOVAL, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Firmado electrónicamente por:
HERNANDO HERIBERTO MANOTOA SANDOVAL

Hernando Heriberto Manotoa Sandoval

DEDICATORIA

A mis padres y esposa por su apoyo incondicional durante todo este proceso académico, agradezco todo su empuje y ánimos para dar todo mi esfuerzo ante las adversidades que se han presentado en este camino.

Finalmente dedico esta tesis a las personas que fueron compañeros de clase, como también a los profesores que compartieron sus conocimientos y por su paciencia, eternamente agradecido por haber compartido este proceso.

AGRADECIMIENTO

A mi esposa por ser mi compañera en todas las metas que me propongo, en especial es este proceso que es tan importante para mí. Agradezco infinitamente por su paciencia en este proceso.

ÍNDICE DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

El objetivo de la estabilización de video es optimizar la calidad del video, siendo importante eliminar los movimientos involuntarios de la cámara, hasta la actualidad los métodos existentes no discriminan los puntos de interés como paso inicial para la estabilización de video.

El objetivo principal de este estudio es definir un método de estabilización de video que discrimine puntos de interés utilizando mapas de profundidad. Se inicia por analizar varios métodos basados en CNN que permitan la obtención de un mapa de profundad robusto. Donde, el método que utiliza el modelo entrenado MiDaS v2.1que es utilizado para primera parte del algoritmo que consiste en tomar cada fotograma y obtener un mapa de profundidad.

En la segunda parte de algoritmo, se define un algoritmo base de estabilización de video que recibe como entrada una secuencia de fotogramas, toma como entrada las los fotogramas aplicados una máscara definida por el mapa de profundidad. Para lograr la estabilización de video utilizamos un método validado de obtención de mapas de profundidad en combinación con un método de estabilización de video.

El algoritmo propuesto en este trabajo fue evaluado en 24 videos en 4 diferentes escenarios. Nuestro algoritmo logra una mejora en el ITF del 5% comparado entre métrica de los videos estabilizados con el método propuesto y los videos estabilizados que no utilizan una máscara de profundidad como parte de su método.

Palabras clave: Estabilización de video, mapas de profundidad, visión por computadora, aprendizaje automático

ABSTRACT

The objective of video stabilization is to optimize the quality of the video, being important to eliminate involuntary movements of the camera, until now the existing methods do not discriminate the points of interest as an initial step for video stabilization.

The main objective of this study is to define a video stabilization method that discriminates points of interest using depth maps. It begins by analyzing several methods based on CNN that allow obtaining a robust depth map. Where, the method used by the MiDaS v2.1 trained model is used for the first part of the algorithm that consists of taking each frame and obtaining a depth map.

In the second part of the algorithm, a base video stabilization algorithm is defined that receives a sequence of frames as input, takes as input the applied frames a mask defined by the depth map. To achieve video stabilization, we use a validated depth mapping method in combination with a video stabilization method.

The algorithm proposed in this work was evaluated in 24 videos in 4 different scenarios. Our algorithm achieves an improvement in the ITF of 5% compared between the metrics of the videos stabilized with the proposed method and the stabilized videos that do not use a depth mask as part of their method.

Keywords: Video stabilization, depth maps, computer vision, machine learning

Capítulo 1

1. INTRODUCCIÓN

Durante la adquisición de secuencias de imágenes mediante cámaras digitales, teléfonos inteligentes o sistemas de dinámica compleja como robots, pueden ocasionarse movimientos involuntarios provocando la perdida en la calidad de fotogramas continuos a causa de la vibración [1][2]. La estabilización de video es una etapa de preprocesamiento que mitiga esta situación al eliminar o reducir los errores producidos por movimientos involuntarios del dispositivo de grabación [3].

La estimación de la profundidad a partir de imágenes en 2D es un paso decisivo en la reconstrucción de escenas, el reconocimiento, la segmentación y la detección de objetos en 3D [4]. La profundidad es un requisito previo clave para realizar múltiples tareas como la percepción, la navegación y la planificación [5][6]. La estabilización de video basada en la estimación de profundidad tiene como objetivo obtener una representación de la estructura espacial de una escena, recuperando la forma tridimensional y la apariencia de los objetos en las imágenes [3][7][8].

Generalmente, la mayoría de las escenas tienen grandes variaciones estructurales y de textura, oclusiones de objetos y detalles geométricos significativos [9]. Todos estos factores contribuyen a dificultar la estimación precisa de la profundidad. El problema se enmarca en como predecir un mapa de profundidad denso para cada píxel, dada una imagen RGB como entrada.

El presente trabajo de tesis tiene como objetivo el desarrollo de un algoritmo de estabilización de video utilizando mapas de profundidad para discriminar los de puntos de interés cercanos al campo de visión, que contengan objetos con y sin movimiento. Para ello, primero se obtiene el mapa de profundidad de la secuencia de fotogramas empleando Redes Neuronales Convolucionales (CNN) en un modelo validado, posteriormente se realiza la discriminación de los puntos de interés en base al mapa de profundidad. Los puntos de interés obtenidos se utilizarán para realizar la estabilización de video mediante métodos de visión por computadora como Motion Estimation [10][11].

1.1. Antecedentes

Uno de los campos de inteligencia artificial es la visión por computadora que se asocia al análisis de imágenes y videos. Donde, la estabilización de video es un método importante para mejorar secuencias de fotogramas [12]. La estabilización de video tiene como objetivo optimizar la calidad del video, eliminando los movimientos involuntarios de la cámara [13][14].

La tarea de estimación de profundidad requiere de una imagen RGB de entrada y genera un fotograma de profundidad [15] y contiene información sobre la distancia de los objetos desde el punto de vista que generalmente es la cámara que toma la imagen.

En otras palabras, el mapa de profundidad contiene datos de las coordenadas vectoriales de la orientación normal de la superficie desde el plano de la imagen [16]. Las aplicaciones de estimación de profundidad incluyen afinar partes borrosas de una imagen, una excelente representación de escenas 2D y 3D, autos autónomos, comprensión de la robótica, cirugía asistida por robot, mapeo de sombras en gráficos 3D por computadora, entre otros [5][17][18].

Los métodos de estabilización de video basados en características utilizan para la detección y descripción de características: Scale-Invariant Feature Transform (SIFT) [19], Speeded-Up Robust Features (SURF) [20], Oriented FAST and Rotated BRIEF (ORB) [21], Features from Accelerated Segment Test (FAST) [22], Binary Robust Independent Elementary Features (BRIEF) [23], Binary Robust Invariant Scalable Keypoints (BRISK) [24], Kanade–Lucas–Tomasi Feature Tracker (KLT) [25]. Son métodos que pueden ser combinados para el proceso de detección y descripción de características [26][27].

En la estabilización de video, existe el proceso de coincidencia de características que está definido por la comparación de descriptores de dos conjuntos de características, este proceso puede ser refinado por el algoritmo RANSAC [28] que ha sido usado en los últimos años [27][29][30] para eliminar valores atípicos, que son datos que no encajan en el modelo de estabilización. Finalmente, se realiza la transformación de movimiento del fotograma a ser estabilizado, en esta parte final al igual que la anterior se ayuda de métodos de filtrado de datos como Filtro de Kalman o Filtro Gaussiano para realizar un proceso de suavizado de movimiento [31][32][33].

1.2. Planteamiento del problema

Los métodos presentados hasta la actualidad utilizan como parámetro de entrada del método de estabilización de video un fotograma completo [30][34][35] como inicio del proceso de detección y descripción de características.

El problema consiste en que los puntos cercanos al campo de visión de la imagen pueden ser distorsionados por objetos en movimiento y por movimientos involuntarios de la cámara, por lo mismo la propuesta consiste en la obtención de mapas de profundidad robustos, dado una secuencia de fotogramas como entrada. El mapa de profundidad es usado para generar una máscara de imagen que discrimina puntos de interés en regiones cercanas al campo de visión que contengan objetos en movimiento. Por tanto, presentamos la aplicación de técnicas de visión por computador en el análisis, diseño e implementación de un algoritmo de estabilización de video utilizando la discriminación de los puntos de interés utilizando mapas de profundidad generados desde fotogramas.

Para ello, se utilizará los conjuntos de datos NYU-Depth V2 y Make3D vinculadas a imágenes de una variedad de escenas grabadas por cámaras RGB para probar la obtención de mapas de profundidad, luego el método de mapas de profundidad será integrado a la estabilización de video definida.

1.3. Pregunta(S) de investigación

¿Qué métodos de estabilización de video discriminan puntos de interés utilizando mapas de profundidad?

¿Qué métodos de extracción de características son utilizados en estabilización de video?

¿Qué métodos se utilizan para la generación de mapas de profundidad o estimación de profundidad?

1.4. Objetivo general

Realizar la estabilización de video en base a la discriminación de puntos de interés utilizando mapas de profundidad.

1.5. Objetivos específicos

Realizar revisión sistemática de literatura (RSL) sobre estabilización de video en base a la discriminación de puntos de interés utilizando mapas de profundidad. La RSL nos permitirá referenciar literatura relacionada y resultados de trabajos relevantes entender la problemática general y el estado del arte.

Utilizar las imágenes de los datasets NYU-Depth V2 [12] y Make3D [13] vinculadas a una variedad de escenas grabadas por cámaras RGB que proporcionan imágenes con alta resolución.

Definir el algoritmo de estabilización de video en ambientes controlados discriminando puntos de interés utilizando mapas de profundidad.

Comparar los resultados obtenidos con otros métodos de estabilización de video.

1.6. Hipótesis

Es posible mejorar la estabilización de un video mediante la discriminación de puntos de interés en una secuencia de fotogramas basado en la distancia de objetos cercanos utilizando mapas de profundidad.

1.7. Marco Teórico

En esta sección se exponen conceptos relacionados a estabilización de video, mapas de profundidad, visión por computadora, que ayudan a mejorar el entendimiento en el trabajo a realizar.

Estabilización de video

La estabilización de video se utiliza para evitar la pérdida de calidad visual al reducir los movimientos no deseados de un dispositivo de captura de video sin influir en los objetos

en movimiento. Esto es particularmente esencial en los dispositivos de imágenes portátiles, estos se ven más afectados por las sacudidas debido a su tamaño o peso. La inestabilidad en los fotogramas generalmente es causada por el movimiento no deseado de la mano durante la acción de la cámara, mientras que las fluctuaciones de posición no deseadas de la cámara dan como resultado secuencias de imágenes inestables. El uso de técnicas de estabilización de video garantiza una alta calidad visual y secuencias de video estables incluso en condiciones no óptimas.

Los métodos de estabilización de video basado en características están definidos por los siguientes subprocesos:

- Estimación de movimiento
- Eliminación de valores atípicos
- Modelado de movimiento de cámara
- Corrección de movimiento de cámara
- Síntesis de vídeo

A. Estimación de movimiento (Motion estimation)

La estimación de movimiento tiene como objetivo recuperar los movimientos presentes en la secuencia de video. Luego, los parámetros de movimiento de la cámara se estiman analizando la correspondencia espacio-temporal entre fotogramas consecutivos. Los píxeles (o bloques de píxeles) del primer fotograma se comparan con los píxeles (o bloques de píxeles) del siguiente fotograma mediante una medida de similitud o una métrica de distancia.

Los métodos utilizados en la estimación de movimientos son: coincidencia de bloques [36], basado en pixeles [37], basado en características SIFT,SURF,ORB) [38], KLT [25]

B. Eliminación de valores atípicos (Outlier removal)

Todos los movimientos observados en la secuencia de video se ven afectados por el movimiento de la cámara, solo una fracción podría ser adecuada para determinar el movimiento efectivo de la cámara. De hecho, la presencia de objetos en movimiento en la escena captada puede ser una fuente de errores, haciendo que la discriminación entre los diversos movimientos (cámara/objetos) sea una tarea importante.

También pueden ocurrir errores al determinar los movimientos en la secuencia de video porque algunos movimientos pueden ser demasiado complejos para un modelo de movimiento dado. Existen dos tipos de análisis para eliminación de valores atípicos: cuadro a cuadro [39][40] y video en directo [41][42][43].

C. Modelado de movimiento de cámara

A parte de los valores atípicos, los movimientos restantes son el resultado del movimiento de la cámara. Por lo tanto, se pueden utilizar para modelar o aproximar el movimiento de la cámara. Para ello se pueden utilizar dos estrategias. En la mayoría de los trabajos, el modelado de la cámara en movimiento se basa en modelos geométricos que describen el proceso físico de capturar una escena con una cámara fotográfica sin lente. Los modelos aplicados hasta la actualidad son modelos 2D [44][45][46], modelos 3D [47][48][49] y modelos perceptivos [41][50][46].

D. Corrección de movimiento de cámara

Una vez que se ha modelado el movimiento de la cámara, se deben determinar nuevos movimientos de cámara para mejorar la calidad de la síntesis de video. Los aspectos más problemáticos del movimiento de la cámara son las sacudidas de alta frecuencia, que causan una incomodidad visual considerable. Los métodos utilizados para la corrección de movimiento son ajuste de ruta [43][42][41], filtración [47][51][52], aproximación de rango bajo [52] y CNN [50][44][45][53].

E. Síntesis de vídeo

Una vez que se ha calculado el movimiento de la cámara rectificada, se genera un nuevo video correspondiente a esta ruta. Este paso depende de la elección del modelo de cámara. La mayoría de los modelos geométricos describen el movimiento original y corregido de cada píxel. Los procesos de reconstrucción utilizados hasta la actualidad son: denso [53][45][44], disperso [42][39] ,y content-preserving warps (CPW -) [41][54].

Mapas de profundidad

Un mapa de profundidad para una sola imagen RGB es la tarea de estimar la profundidad a partir de una sola imagen dada. Específicamente, para cada píxel en la imagen RGB, se necesita estimar un valor métrico de profundidad.

Los métodos utilizados para la generación de mapas de profundidad son:

- Red neuronal convolucional
- Multitarea
- Uso de datos sintéticos
- Aprendizaje no supervisado
- Conjuntos de datos
- Aprendizaje desde anotaciones de profundidad relativa

A. Red neuronal convolucional

Es una red neuronal convolucional (CNN), las CNN se basan en neuronas que están organizadas en capas y, por lo tanto, pueden aprender representaciones jerárquicas. Las neuronas entre capas están conectadas a través de pesos y sesgos. La capa inicial es la capa de entrada, por ejemplo, datos de teledetección, y la última capa es el resultado, como una clasificación predicha de un mapa de profundidad.

En el medio hay capas ocultas que transforman el espacio de características de la entrada de manera que coincida con la salida. Las CNN incluyen al menos una capa convolucional como capa oculta para explotar patrones [55]. El objetivo de una CNN es realizar tareas de regresión y clasificación. En la Figura 1 se muestra la estructura de una CNN.

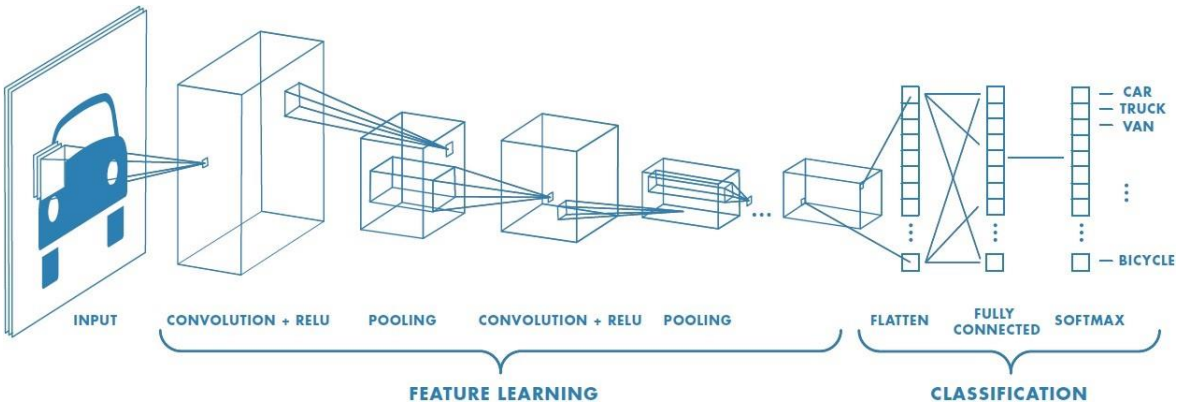


Figura 1: Esquema de una CNN [56]

B. Multitarea

Se enfoca en mejorar los resultados y aumentar la robustez de las soluciones adquiridas, estas se definen como tareas auxiliares. Las tareas auxiliares que pueden realizarse sobre

una imagen son: segmentación semántica [57], estimación normal de superficie [58], estimación de contorno [59], y ego-motion [60].

C. Uso de datos sintéticos

El entrenamiento se realiza de forma supervisada con datos etiquetados, los datos sintéticos eliminan la costosa recolección de datos reales ya que se pueden crear datos sintéticos de manera automática en gran cantidad y en gran diversidad, siendo una de las formas de superar la necesidad de datos etiquetados. Si quieren aprender a generar datos sintéticos consulte aquí [61].

D. Aprendizaje no supervisado

El aprendizaje no supervisado utiliza imágenes estéreo o grabaciones de video con pequeños cambios en las posiciones de la cámara entre fotogramas, ya que dos fotogramas consecutivos pueden considerarse imágenes estéreo. Si se necesita más información revisar [62][63]

Visión por computadora

Uno de los campos de la inteligencia artificial (IA) es la visión por computadora que permite que los equipos computacionales y los sistemas adquieran información significativa de imágenes, videos y otras entradas visuales, y procesos basados en esa información [64].

Aprendizaje automático

Aprendizaje automático es una disciplina científica en el campo de la IA que crea sistemas que aprenden automáticamente. Aprender en este contexto significa identificar patrones complejos entre muchos datos [65].

Mascara de imagen

Para definir la máscara de un fotograma se debe manipular el umbral. El umbral de un fotograma es definido para mostrar una imagen en blanco y negro, se definen tres tipos: umbral simple, umbral adaptativo y binarización de Otsu. El fotograma utilizado para probar los tipos de definir un umbral se tomó de [66].

A. Umbral simple:

El umbral simple utiliza en cada píxel, se aplica el mismo valor de umbral. Si el valor del píxel es menor que el umbral, se establece en 0, de lo contrario, se establece en un valor máximo [67]. En la Figura 1, se muestran los resultados de aplicar los métodos de umbral simple.

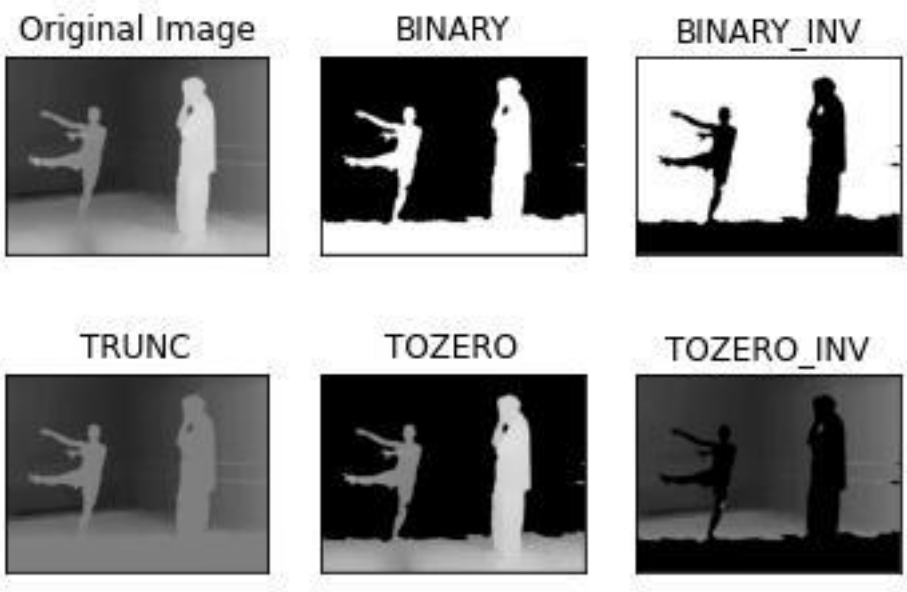


Figura 1: Resultados de aplicar los métodos de umbral simple

B. Umbral adaptativo

El umbral adaptativo define en su algoritmo determinar el umbral para un píxel en función de una pequeña región a su alrededor. Entonces se obtiene diferentes umbrales para diferentes regiones del mismo fotograma, lo que da mejores resultados para fotogramas

con iluminación variable [67]. En la Figura 2, se muestra los resultados de aplicar los métodos de umbral adaptativo a un fotograma.

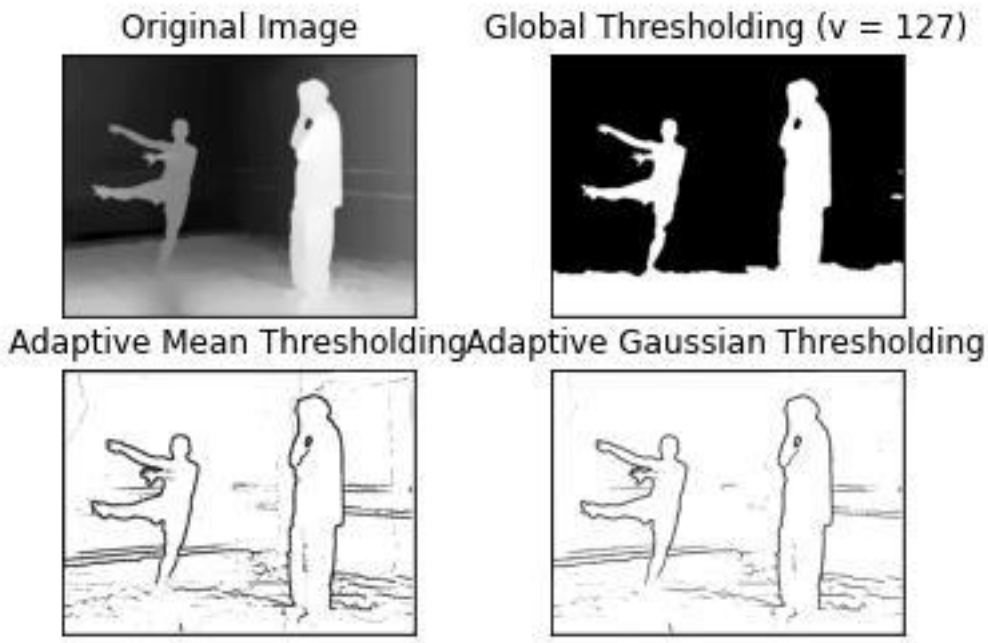


Figura 2: Resultados de aplicar los métodos de umbral adaptativo a un fotograma

C. Binarización de Otsu

La Binarización de Otsu en su algoritmo determina el umbral para un píxel en función de una pequeña región a su alrededor. Entonces obtenemos diferentes umbrales para diferentes regiones del mismo fotograma en base a una función media y una función de suma ponderada gaussiana, lo que da mejores resultados para fotogramas con iluminación variable [67]. En la Figura 3, se muestra el resultado de aplicar los el método de binarización de Otsu.

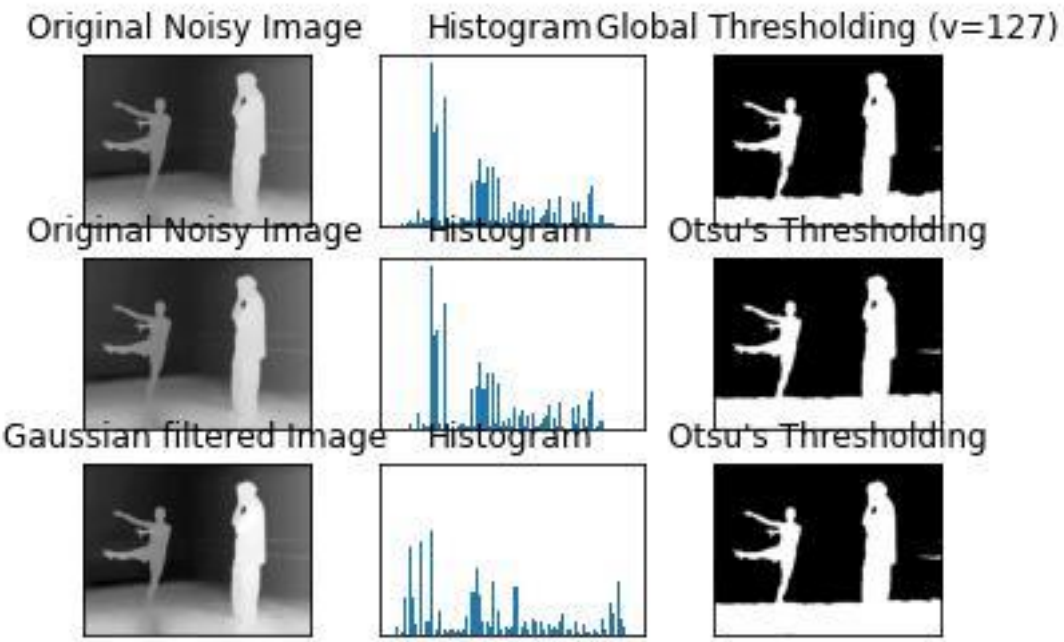


Figura 3: Resultado de aplicar el método de binarización de Otsu

Capítulo 2

2. REVISION SISTEMATICA DE LA LITERATURA

La Revisión Sistemática de Literatura (RSL) acerca de la estabilización de video en base a la discriminación de puntos de interés utilizando mapas de profundidad tiene como objetivo obtener evidencia a partir de artículos científicos almacenados en repositorios digitales, que debe ser sistemática, reproducible y auditable para formular preguntas de investigación sobre un área temática o fenómeno de interés y para buscar, seleccionar, analizar y comunicar toda la investigación relevante, básica o aplicada, necesaria para responder a estas preguntas.

La RLS se basa en el proceso definido en la metodología de Bárbara Kitchenham [68], [69], se definen los siguientes pasos fundamentales:

- Identificar la necesidad de la revisión,
- Definir preguntas de investigación,
- Definir palabras claves para el problema de estudio,
- Estrategias de búsqueda,
- Definir fuentes bibliográficas de búsqueda,
- Criterios de inclusión,
- Criterios de exclusión, y
- Resultados primarios de búsqueda.

2.1. Identificar la necesidad de la revisión

Con la presente RSL, se busca determinar técnicas y métodos utilizados para la estabilización de video en base a la discriminación de puntos de interés utilizando mapas de profundidad, o temas relacionados. Además, que utilicen CNN para generación de mapas de profundidad y/o visión por computadora para estabilización de video. La RSL es la base para definir un algoritmo que realice la estabilización de video que realice la discriminación de puntos de interés a base de mapas de profundidad y determinar si podrá ser eficiente como los métodos de estabilización de video basados en características, tomando en cuenta los recursos disponibles que sirvan de guía para la investigación.

2.2. Definir preguntas de investigación

Como un mecanismo de guiar el desarrollo de la RSL, se ha planteado preguntas de investigación, de tal manera que en el actual trabajo se enfoque en entregar respuesta a estas preguntas, conservando siempre el enfoque y perspectiva determinados. En la Tabla 1, se muestra las preguntas de investigación definidas.

Tabla 1: Preguntas de investigación

CODIGO	=
P1	¿Qué métodos de estabilización de video utilizan la discriminación de puntos de interés mediante mapas de profundidad?
P2	¿Qué métodos permiten la generación de una imagen de profundidad teniendo como entrada imagen simple?
P3	¿Qué bases de datos son utilizadas en los métodos de estimación de profundidad o mapas de profundidad?
P4	¿Qué métricas se utilizan para evaluar los resultados de estabilización de video?
P5	¿Qué técnicas de extracción de características de una imagen son utilizados en loa métodos de estabilización de video?

2.3. Definir palabras claves para el problema de estudio

Basándonos en las preguntas de investigación, se ha definido las siguientes palabras claves, las cuales permiten construir las cadenas de búsqueda. Las palabras clave en cuestión son:

- video stabilization
- depth estimation
- depth prediction
- depth maps
- computer vision
- machine learning

Se han añadido “depth estimation” y “depth prediction” porque son utilizadas en publicaciones de procesamiento de imagen y video para determinar distancias de objetos o profundidad de objetos dentro de una imagen [62][16][63].

2.4. Estrategias de búsqueda

Al combinar las palabras claves y utilizando los operadores "AND" y/o "OR”, se definen las cadenas de búsqueda en formato genérico, estas son utilizadas según el formato utilizado en las diferentes fuentes bibliográficas de búsqueda, las cadenas de búsqueda se muestran en la Tabla 2.

Tabla 2: Cadenas de búsqueda

CODIGO	CADENA DE BUSQUEDA
CB01	("video stabilization" AND "depth estimation")
CB02	("video stabilization" AND "depth maps")
CB03	("video stabilization" AND "depth prediction")
CB04	("depth estimation" AND ("Convolutional neural network" OR CNN))
CB05	("depth maps" AND ("Convolutional neural network" OR CNN))
CB06	("depth prediction" AND ("Convolutional neural network" OR CNN))
CB07	("video stabilization" AND ("SIFT" OR "SURF" OR "ORB" OR "Features Based"))

2.5. Fuentes bibliográficas de búsqueda

Las fuentes bibliográficas seleccionadas son las siguientes bibliotecas virtuales:

- ACM Digital Library (https://dl.acm.org/)
- IEEExplore (https://ieeexplore.ieee.org/)
- Springer Link (https://link.springer.com/)
- ScienceDirect (https://www.sciencedirect.com/)

2.6. Criterios de inclusión

Luego de haber detallado las cadenas de búsqueda a utilizar el siguiente paso fue establecer los criterios de inclusión.

- Criterios de inclusión
  - Publicaciones científicas realizadas en revistas en la disciplina de Ciencias de la computación,
  - Artículos que utilicen para la estabilización de video características SIFT, SURF u ORB dentro de su algoritmo,
  - Artículos que utilicen mapas de profundidad, estimación de profundidad o predicción de profundidad en imágenes simples dentro del algoritmo de estabilización de video,
  - Artículos que utilicen CNN para obtener mapas de profundidad, estimación de profundidad o predicción de profundidad en imágenes simples 2D, y
  - Artículos en inglés

2.7. Criterios de exclusión

Luego de haber detallado las cadenas de búsqueda a utilizar el siguiente paso fue establecer los criterios de exclusión.

- Criterios de exclusión
  - Artículos que no aportan con métodos de extracción de características SIFT, SURF u ORB para estabilización de video,
  - Artículos que no tengan incluido la frase “video stabilization” en el título, palabras clave o resumen de investigación,
  - Artículos científicos con metodología que no sea clara o específica,
  - Reseñas de literatura, capítulos de libros, tesis, informes técnicos, propuestas de investigación, conferencias o manuales, revisiones,
  - Artículos donde utilicen cámaras estéreo para obtención de imágenes, y
  - Artículos publicados entre los años 2016 – 2021 y artículos duplicados.

2.8. Resultados primarios de búsqueda

Los resultados de las cadenas de búsqueda después de haber aplicado algunos criterios de exclusión como publicaciones entre los años 2016 – 2021 y que pertenezcan al área de ciencias de la computación y de ser posible que pertenezcan a las subáreas inteligencia artificial y visión por computadora, se muestra a continuación la información correspondiente a cada una de las bases de datos definidas.

En la Tabla 3, se muestran los resultados que se obtuvieron en la búsqueda dentro de la biblioteca IEEE, dando un total de106 publicaciones.

Tabla 3: Resultados primarios de búsqueda en IEEE

FUENTE	CODIGO	CADENA DE BUSQUEDA	TOTAL
IEEE	CB01	("All Metadata":"video stabilization" AND "All Metadata":"depth estimation")	**1**
	CB02	("All Metadata":"video stabilization" AND "All Metadata":"depth maps")	**0**
	CB03	("All Metadata":"video stabilization" AND "All Metadata":"depth prediction")	**0**
	CB04	("All Metadata":"depth estimation" AND ("All Metadata":"Convolutional neural network" OR "All Metadata":CNN))	**65**
	CB05	("All Metadata":"depth maps" AND ("All Metadata":"Convolutional neural network" OR "All Metadata":CNN))	**26**
	CB06	("All Metadata":"depth prediction" AND ("All Metadata":"Convolutional neural network" OR "All Metadata":CNN))	**10**
	CB07	("All Metadata":"video stabilization" AND ("All Metadata":"SIFT" OR "All Metadata":"SURF" OR "All Metadata":"ORB" OR "All Metadata":"Features Based"))	**4**
		TOTAL	**106**

En la **¡Error! No se encuentra el origen de la referencia.**, se muestran los resultados que se obtuvieron en la búsqueda dentro de la biblioteca ACM, dando un total de162 publicaciones.

Tabla 4: Resultados primarios de búsqueda en ACM

FUENTE	CODIGO	CADENA DE BUSQUEDA	TOTAL
ACM	CB01	(AllField:("video stabilization") AND AllField:("depth estimation"))	**1**
	CB02	(AllField:("video stabilization") AND AllField:("depth maps"))	**1**
	CB03	(AllField:("video stabilization") AND AllField:("depth prediction"))	**0**
	CB04	(AllField:("depth estimation") AND (AllField:("Convolutional neural network") OR AllField:(CNN)))	**56**
	CB05	(AllField:("depth maps") AND (AllField:("Convolutional neural network") OR AllField:(CNN)))	**80**
	CB06	(AllField:("depth prediction") AND (AllField:("Convolutional neural network") OR AllField:(CNN)))	**16**
	CB07	(AllField:("video stabilization") AND (AllField:("SIFT") OR AllField:("SURF") OR AllField:("ORB") OR AllField:("Features Based")))	**8**
		TOTAL	**162**

En la Tabla 5, se muestran los resultados que se obtuvieron en la búsqueda dentro de la biblioteca Springer, dando un total de 222 publicaciones.

Tabla 5: Resultados primarios de búsqueda en Springer

FUENTE	CODIGO	CADENA DE BUSQUEDA	TOTAL
Springer	CB01	(All:"video stabilization" AND All:"depth estimation")	**3**
	CB02	(All:"video stabilization" AND All:"depth maps")	**4**
	CB03	(All:"video stabilization" AND All:"depth prediction")	**0**
	CB04	(All:"depth estimation" AND (All:"Convolutional neural network" OR All:CNN))	**59**
	CB05	(All:"depth maps" AND (All:"Convolutional neural network" OR All:CNN))	**122**
	CB06	(All:"depth prediction" AND (All:"Convolutional neural network" OR All:CNN))	**20**
	CB07	(All:"video stabilization" AND (All:"SIFT" OR All:"SURF" OR All:"ORB" OR All:"Features Based"))	**14**
		TOTAL	**222**

En la Tabla 6,se muestran los resultados que se obtuvieron en la búsqueda dentro de la biblioteca Elsevier, dando un total de 1069 publicaciones.

Tabla 6: Resultados primarios de búsqueda en Elsevier

FUENTE	CODIGO	CADENA DE BUSQUEDA	TOTAL
Elsevier	CB01	(All:"video stabilization" AND All:"depth estimation")	**152**
	CB02	(All:"video stabilization" AND All:"depth maps")	**152**
	CB03	(All:"video stabilization" AND All:"depth prediction")	**152**
	CB04	(All:"depth estimation" AND (All:"Convolutional neural network" OR All:CNN))	**153**
	CB05	(All:"depth maps" AND (All:"Convolutional neural network" OR All:CNN))	**153**
	CB06	(All:"depth prediction" AND (All:"Convolutional neural network" OR All:CNN))	**153**
	CB07	(All:"video stabilization" AND (All:"SIFT" OR All:"SURF" OR All:"ORB" OR All:"Features Based"))	**154**
		TOTAL	**1069**

Finalmente, en la Tabla 7, se muestran los resultados que se obtuvieron en la búsqueda dentro de la biblioteca ScienceDirect, dando un total de 907 publicaciones, al igual que en la tabla anterior el resultado en esta se da porque no permite la búsqueda con filtro de rango en años, el filtro se realiza de forma manual

Tabla 7 Resultados primarios de búsqueda en ScienceDirect

FUENTE	CODIGO	CADENA DE BUSQUEDA	TOTAL
ScienceDirect	CB01	(All:"video stabilization" AND All:"depth estimation")	**6**
	CB02	(All:"video stabilization" AND All:"depth maps")	**7**
	CB03	(All:"video stabilization" AND All:"depth prediction")	**0**
	CB04	(All:"depth estimation" AND (All:"Convolutional neural network" OR All:CNN))	**305**
	CB05	(All:"depth maps" AND (All:"Convolutional neural network" OR All:CNN))	**584**
	CB06	(All:"depth prediction" AND (All:"Convolutional neural network" OR All:CNN))	**5**
	CB07	(All:"video stabilization" AND (All:"SIFT" OR All:"SURF" OR All:"ORB" OR All:"Features Based"))	
		TOTAL	**907**

2.9. Selección de estudios relevantes

En la RSL no se han publicado trabajos de estabilización de video que realicen discriminación de puntos de interés utilizando mapas de profundidad o estimación de profundidad, en vista de esto se eligió los trabajos de investigación que aborden los temas de estabilización de video basados en extracción de características mediante SIFT, SURF y ORB. Además, trabajos de investigación sobre mapas de profundidad y estimación de profundidad que utilicen CNNs para la generación de los mismos.

P1: ¿Qué métodos de estabilización de video utilizan la discriminación de puntos de interés mediante mapas de profundidad?

La estabilización de video ha sido un tópico muy activo de investigación en los últimos años. Se han propuesto muchos enfoques que se describen en los artículos de investigación, cabe mencionar que en los métodos utilizados no utilizan discriminación de puntos de interés en base a mapas de profundidad, por tanto, se han seleccionado los trabajos que utilizan extracción de características y métodos que utilizan CNNs para obtener mapas de profundidad de un fotograma.

Es necesario combinar los métodos de estabilización de video y métodos de mapas de profundidad para proponer un método a ser evaluado en este trabajo de investigación.

P2: ¿Qué métodos permiten la generación de un mapa de profundidad teniendo como entrada imagen simple?

El mapa de profundidad en una imagen o video es el proceso de medir la distancia entre objetos y el observador. Se puede realizar utilizando diferentes métodos, en el enfoque de este trabajo de ha obtenido los siguientes métodos basados en CNN: Convolutional Neural Network (CNN), Residual Network (ResNet) based CNN y Residual Network (ResNet) based U-Net (CNN).

P3: ¿Qué conjunto de datos son utilizados en los métodos de obtención de mapas de profundidad?

Un conjunto de datos, es una colección de imágenes que se usan para el entrenamiento y la evaluación de algoritmos de aprendizaje automático, en este caso de los métodos de

obtención de mapas de profundidad, las utilizadas según la RSL son: NYU Depth V2, NYU Depth, KITTI, Make3D, SUN RGB-D, NYU Depth, Eigen Split, KITTI Split, Cityscape, KITTI, DIW, ETH3D, Sintel y TUM.

Los métodos de estimación de profundidad o mapas de profundidad han utilizado un conjunto de datos o una combinación de los mismos para validar el rendimiento del método propuesto en cada artículo de la RSL.

P4: ¿Qué métricas se utilizan para evaluar los resultados de estabilización de video?

Las métricas para evaluación de estabilización de video son un conjunto de indicadores que se pueden usar para medir la eficiencia de un algoritmo. Estas métricas se pueden utilizar para comparar diferentes algoritmos y también para ajustar los parámetros del algoritmo.

Las métricas que son utilizadas en la RSL son: Mean Squared Error (MSE) Root Mean Square Error (RMSE), Peak Signal to Noise Ratio (PSNR), Structural SIMilarity Index (SSIM), Inter-frame Transformation Fidelity (ITF) y Percentage of Pixels Held PH (%) . Estas métricas se usar para evaluar el efecto de cambios en un algoritmo. Además, existen otros factores a tener en cuenta son el tamaño, la resolución del video, el número de Frames per Second (FPS).

P5: ¿Qué técnicas de extracción de características de una imagen son utilizados en loa métodos de estabilización de video?

La extracción de características es un método utilizado para obtener características relevantes de una imagen. Estas características pueden ser utilizadas para representar la imagen de una manera más eficiente para ser analizada, para detectar ciertos objetos o patrones.

Los métodos de detección y descripción de características utilizados en la RSL son: Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF), Oriented FAST and Rotated BRIEF (ORB), Features from Accelerated Segment Test (FAST), Binary Robust Independent Elementary Features (BRIEF), Binary Robust Invariant Scalable Keypoints (BRISK)

La Figura 4, presenta la fase de RLS. Se encontraron 2488 artículos después de realizar la actividad de estrategia de búsqueda aplicando las cadenas de búsqueda, el rango de anos entre 2016 y 2021 y que sean artículos de investigación. Una vez aplicados los criterios de inclusión y exclusión faltantes, se excluyeron 2370 artículos, dando un total de118 artículos como resultado del proceso de filtrado.

Posteriormente, en la etapa de validación se realiza la revisión del contenido de los artículos y se eligió los artículos más relevantes de acuerdo al contenido del resumen y a la explicación de la metodología utilizada, quedando un total de 20 artículos, y luego de aplicar la técnica Forward Snowballing se seleccionaron 12 artículos para tener un total de 32 artículos que serán la guía para el proyecto de investigación. En la RSL no se obtuvieron métodos que discriminen puntos de interés en base a mapas de profundad para realizar el proceso de estabilización de video, por lo cual, en la

Tabla 8 se presenta los estudios primarios de estabilización de video que fueron seleccionados como base para el planteamiento de la metodología.

Tabla 8: Estudios primarios seleccionados de estabilización de video

ID	Autores	Titulo	Año
PEV01	Aguilar et al.	Real-time model-based video stabilization for micro aerial vehicles	2016
PEV02	Aguilar et al.	Onboard video stabilization for rotorcrafts	2017
PEV03	Arsalan et al.	Panoramic video stabilization based on rotational damping filter	2016
PEV04	Fang et al.	A video stabilization algorithm based on affine SIFT	2018
PEV05	Hu et al.	Digital video stabilization based on multilayer gray projection	2018
PEV06	Kaur et al.	Video stabilization for an aerial surveillance system using sift and surf	2016
PEV07	Ke et al.	Efficient Real-Time Video Stabilization with a Novel Least Squares Formulation	2021
PEV08	Li et al.	Real-time feature-based video stabilization on FPGA	2016
PEV09	Liu et al.	Video Stabilization Algorithm Based on Kalman Filter and Homography Transformation	2018
PEV10	Mm et al.	Fast tunnel monitoring video stabilization method based on improved ORB feature	2018
PEV11	Pedrini et al.	Combination of local feature detection methods for digital video stabilization	2018
PEV12	Qiu et al.	Stabilization Algorithm Based on Improved Motion Model for Jittery Video in Minimally Invasive Surgery	2019
PEV13	Sharif et al.	Improved Video Stabilization using SIFT-Log Polar Technique for Unmanned Aerial Vehicles	2019
PEV14	Xiong et al.	Antiblurry dejitter image stabilization method of fuzzy video for driving recorders	2017

En la Tabla 9,se presenta los estudios primarios de mapas de profundidad que fueron seleccionados como base para el planteamiento de la metodología.

Tabla 9: Estudios primarios seleccionados de mapas de profundidad

ID	AUTOR	TITULO	AÑO
PEV01	Laina et al.	Deeper depth prediction with fully convolutional residual networks	2016
PEV02	Cao et al.	Estimating depth from monocular images as classification using deep fully convolutional residual networks	2017
PEV03	Ma et al.	Depth estimation from single image using CNN-residual network	2017
PEV04	Fu et al.	Deep ordinal regression network for monocular depth estimation	2018
PEV05	Zhang et al.	Joint Task-Recursive Learning for Semantic Segmentation and Depth Estimation	2018
PEV06	Kim et al.	Deep monocular depth estimation via integration of global and local predictions	2018
PEV07	Chen et al.	Towards Scene Understanding: Unsupervised Monocular Depth Estimation with Semantic-Aware Representation	2019
PEV08	Song at al.	Depth estimation from a single image using guided deep network	2019
PEV09	Fu et al.	Monocular depth estimation based on multi-scale graph convolution networks	2019
PEV10	Zhou et al.	Joint object detection and depth estimation in multiplexed image	2019
PEV11	Bhat et al.	Adabins: Depth estimation using adaptive bins	2020
PEV12	Khan et al.	Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer	2020
PEV13	Lu et al.	Taskology: Utilizing Task Relations at Scale	2020
PEV14	Hambarde et al.	S2DNet: Depth estimation from single image and sparse samples	2020
PEV15	Liu et al.	A contextual conditional random field network for monocular depth estimation	2020
PEV16	Mathew et al.	Monocular depth estimation with SPN loss	2020
PEV17	Zhang et al.	Joint Depth and Defocus Estimation From a Single Image Using Physical Consistency	2021
PEV18	Choi et al.	Self-Supervised Monocular Depth Estimation with Extensive Pretraining	2021

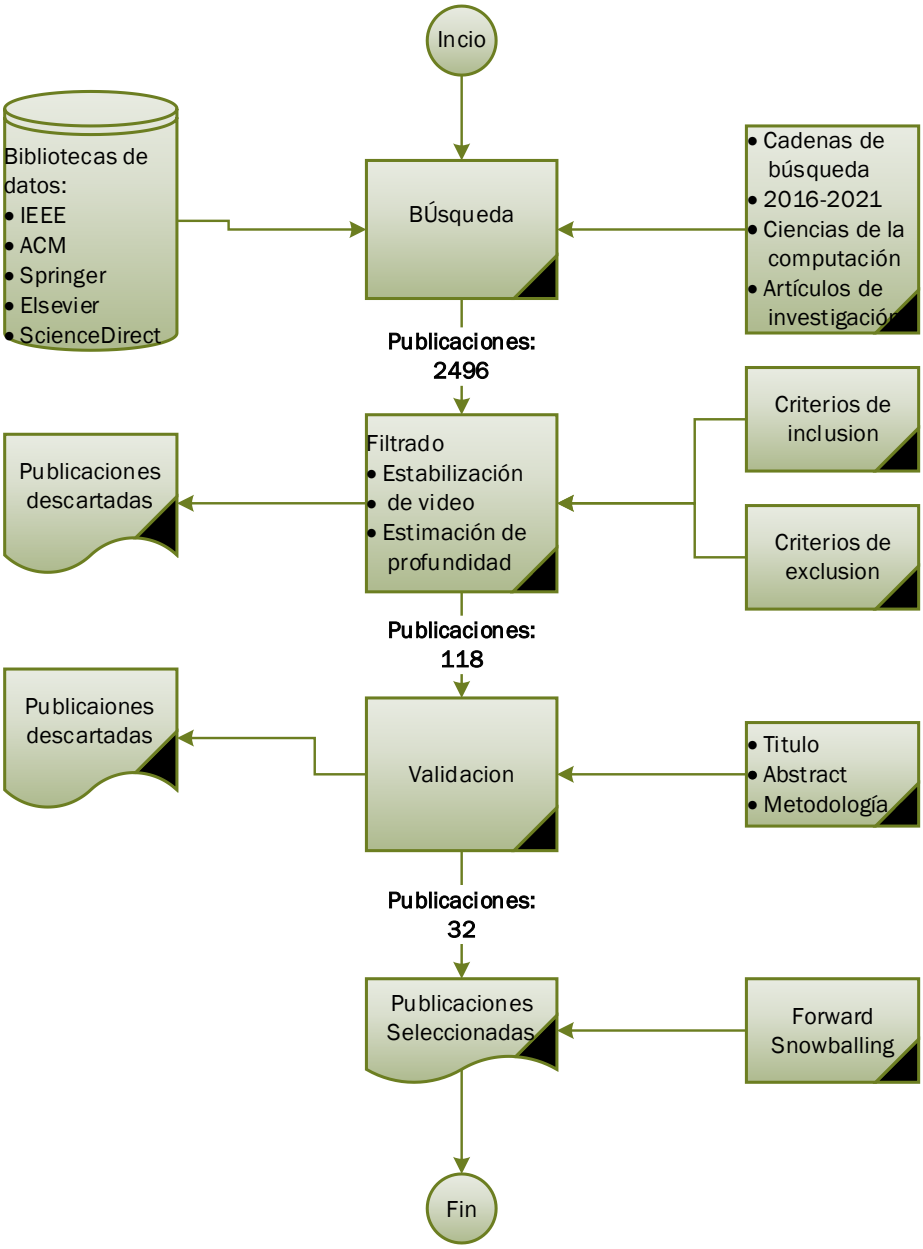


Figura 4: Proceso de revisión sistemática de literatura (RSL)

2.10. Síntesis de datos

En la RSL se concluyó que no se han propuesto métodos que utilicen mapas de profundidad para discriminar puntos de interés obtenidos de una imagen para la estabilización de video.

La Tabla 10, presenta los métodos, bases de datos utilizadas para los métodos de mapas de profundidad y en la en la Tabla 11, se muestran las características, descriptores, coincidencias, compensación de movimiento y métodos asociados para mejorar coincidencia de características y compensación de movimiento. Además, se, como también las métricas que utilizan para validar el rendimiento en los métodos de estabilización de video.

El diseño de las tablas se basa en los hallazgos realizados a partir de las preguntas de investigación (P1, P2, P3, P4 y P5). Los datos generados se presentan, analizan e interpretan en la sección de “Resultados de la RSL”.

Tabla 10: Métodos y conjuntos de datos utilizados en los métodos de mapas de profundidad

ID	MÉTODO	CONJUNTO DE DATOS
PEV01	CNN	NYU Depth V2
PEV02	CNN	NYU Depth and KITTI
PEV03	ResNet based CNN	NYU Depth and KITTI
PEV04	CNN	KITTI, Make3D and NYU Depth v2
PEV05	ResNet based CNN	NYU-Depth v2 and SUN RGB-D
PEV06	CNN	NYU Depth and KITTI, Make3D
PEV07	DispNet variant CNN	Eigen Split, KITTI Split and Cityscapes Dataset
PEV08	ResNet based CNN	NYU depth v2, Make3D, and Cityscape and KITTI
PEV09	ResNet based CNN	KITTI
PEV10	ResNet based CNN	KITTI
PEV11	CNN	SUN RGB-D,NYU Depth and KITTI
PEV12	Resnet CNN	DIW, ETH3D, Sintel, KITTI, NYU and TUM
PEV13	U-Net(CNN) with ResNet18	Cityscapes and KITTI
PEV14	CNN	NYU Depth and KITTI
PEV15	Resnet CNN	NYU Depth and KITTI and Make3D

PEV16	Resnet CNN	KITTI
PEV17	Resnet CNN	KITTI
PEV18	Resnet CNN	KITTI

Tabla 11: Métodos y métricas de evaluación de los métodos de estabilización de video.

ID	FOTOGRAMA	CARACERÍSTICAS	KEYPOINTS-DESCRIPTORS	COINCIDENCIA	COMPENSACION-FILTRO	METRICAS
PEV01	COMPLETO	SURF	SURF	MATCH-RANSAC	WARP- KALMAN	ITF, RMSE
PEV02	COMPLETO	SURF	BRIEF	MATCH-RANSAC	WARP- KALMAN	ITF, RMSE, PSNR
PEV03	COMPLETO	SURF	SURF	MATCH-RANSAC	WARP- DAMPING	ITF
PEV04	COMPLETO	ASIFT	SIFT	MATCH	WARP- Gaussian low pass filter	-
PEV05	COMPLETO	-	-	-	Traslacion, rotacion y escala KALMAN	ITF, speed
PEV06	COMPLETO	SIFT- DoG-LoG	SIFT	MATH- BUCLE	WARP	PNSR, MSE
PEV07	COMPLETO	SURF	SURF-FLANN	MATCH-RANSAC- Least Squares	WARP	ITF, ms
PEV08	COMPLETO	ORB	ORB	MATCH-remove matches	WARP	PSNR, MSE
PEV09	COMPLETO	SURF	SURF	MATCH-KNN	WARP	ITF
PEV10	COMPLETO	ORB-KSW	ORB	MATCH	WARP	ITF
PEV11	COMPLETO	MSER-FAST	FAST	MATCH-RANSAC	WARP- Gaussian	PSNR-SIMM-PH (%)
PEV12	COMPLETO	SURF	SURF	MATCH	WARP- Gaussian	PSNR
PEV13	COMPLETO	SIFT-LoG	SIFT-DoG	MATCH	WARP- KALMAN	-
PEV14	COMPLETO	FAST	BRISK	MATCH-RANSAC	WARP	ITF, FPS

2.11. Resultados de la RSL

En la RSL se utilizan varios métodos de estabilización de video basados en extracción de características. EL porcentaje de artículos que usan estos métodos de extracción de características se muestran en la Figura 5. Donde, el 36% de los artículos revisados define el método SURF como extractor de características. El 14% de los artículos revisados define el método ORB como extractor de características. En porcentajes iguales del 7% de los artículos revisados definen a los métodos ASIFT, SURF-Log, SIFT, ASIFT como extractores de características y finalmente el 22% de los artículos revisados define otros métodos como extractores de características.

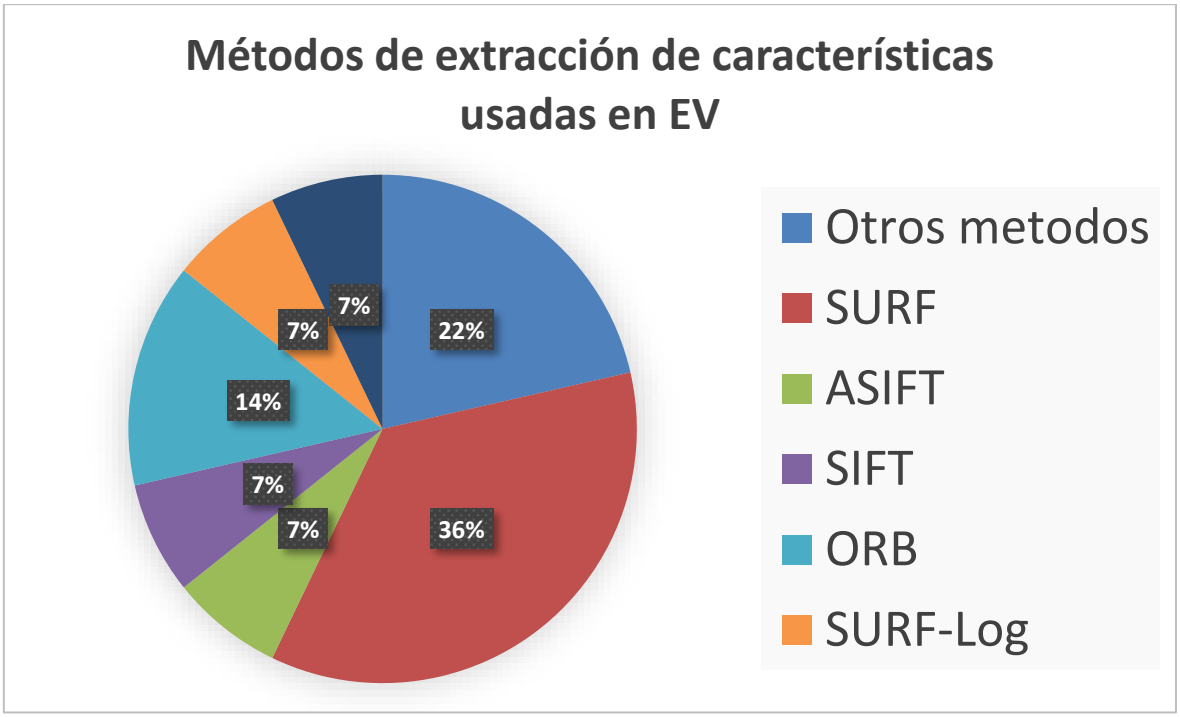


Figura 5: Métodos de extracción de características usadas en EV

En la Figura 6, en base a los artículos de EV que utilizan una o varias métricas de evaluación para un video estabilizado, se presenta que la métrica ITF tiene el 36% de uso como métrica de evaluación. El 30% define a PSNR como métrica de evaluación. En porcentajes iguales del 3% define a FPS, SSIM y PH (%) como métrica de avaluación. Finalmente, el 17% define a MSE como métrica de evaluación.

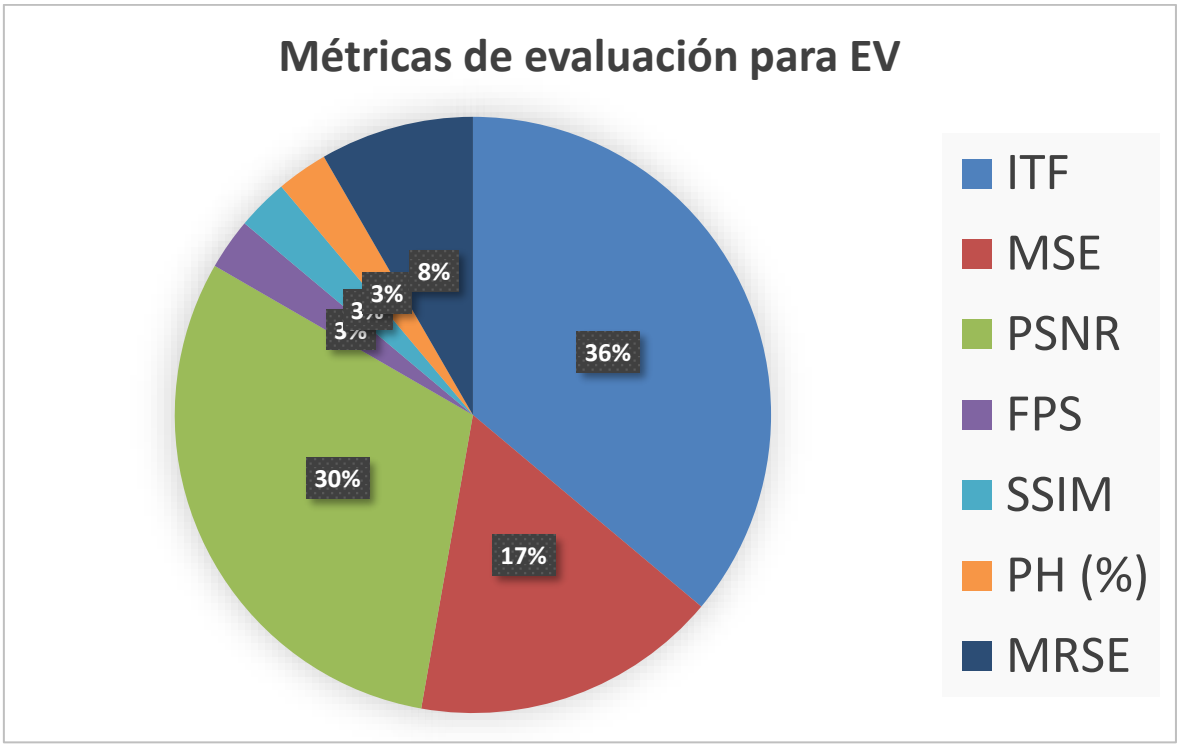


Figura 6: Métricas de evaluación para EV

En la Figura 7, se presenta los conjuntos de datos utilizados para los procesos de entrenamiento, testeo y validación en lo métodos de detección de mapas de profundidad. El 37% de artículos utiliza el conjunto de datos KITTI. El 37% de artículos utiliza el conjunto de datos KITTI. El 15% de artículos utiliza el conjunto de datos NYU Depth. En porcentajes iguales al 10% de artículos utiliza el conjunto de datos NYU Depth V2 y Make3D. También, en porcentajes iguales al 3% de artículos utiliza el conjunto de datos DIW, NYU Depth, THH3D y Sintel. El 2% de artículos utiliza el conjunto de datos KITTI Split y Eigen Split. El 7% de artículos utiliza el conjunto de datos Cityscape. Finalmente, El 5% de artículos utiliza el conjunto de datos SUN RGB-D.

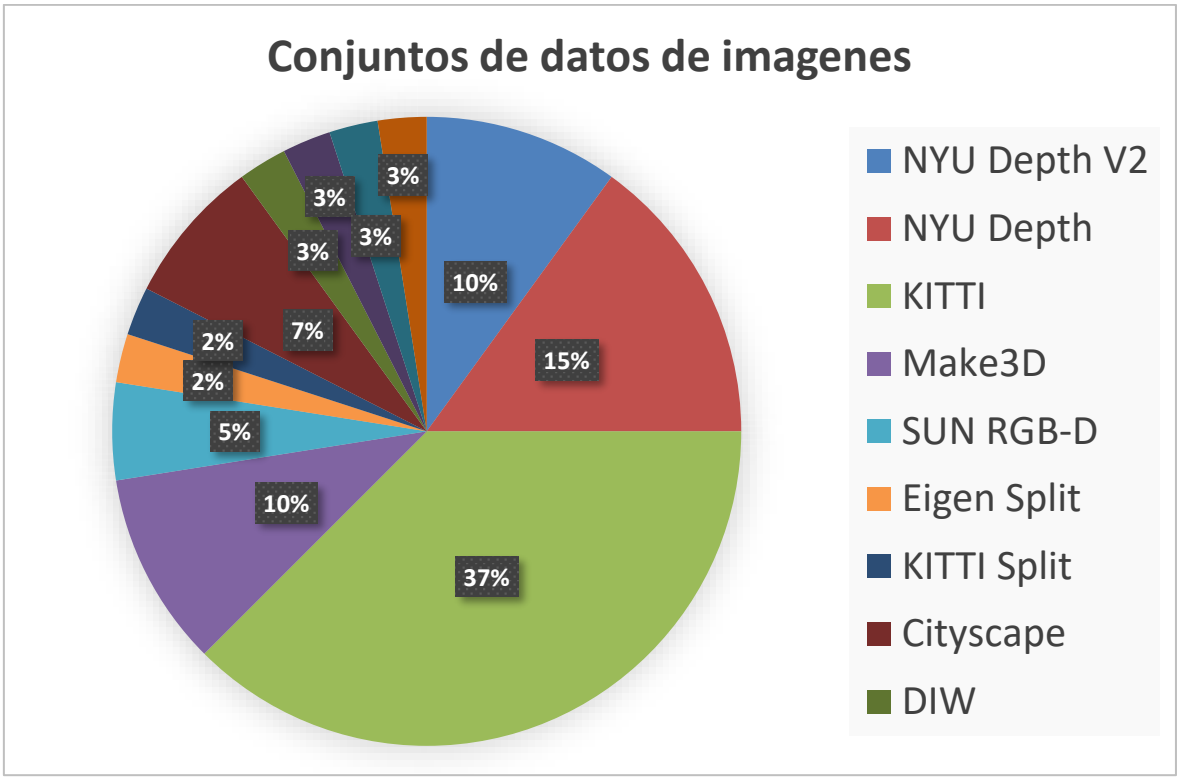


Figura 7: Conjuntos de datos de imágenes

En la Figura 8, se muestra el método que reflejan en los artículos de la RS, donde el 33% de artículos define a CNN como método para obtener el mapa de profundidad. El 5% de artículos define a DispNet variant CNN como método para obtener el mapa de profundidad. El 6% de artículos define a ResNet based U-NET(CNN) como método para obtener el mapa de profundidad, seguido del 56% de artículos define a ResNet based CNN como método para obtener el mapa de profundidad.

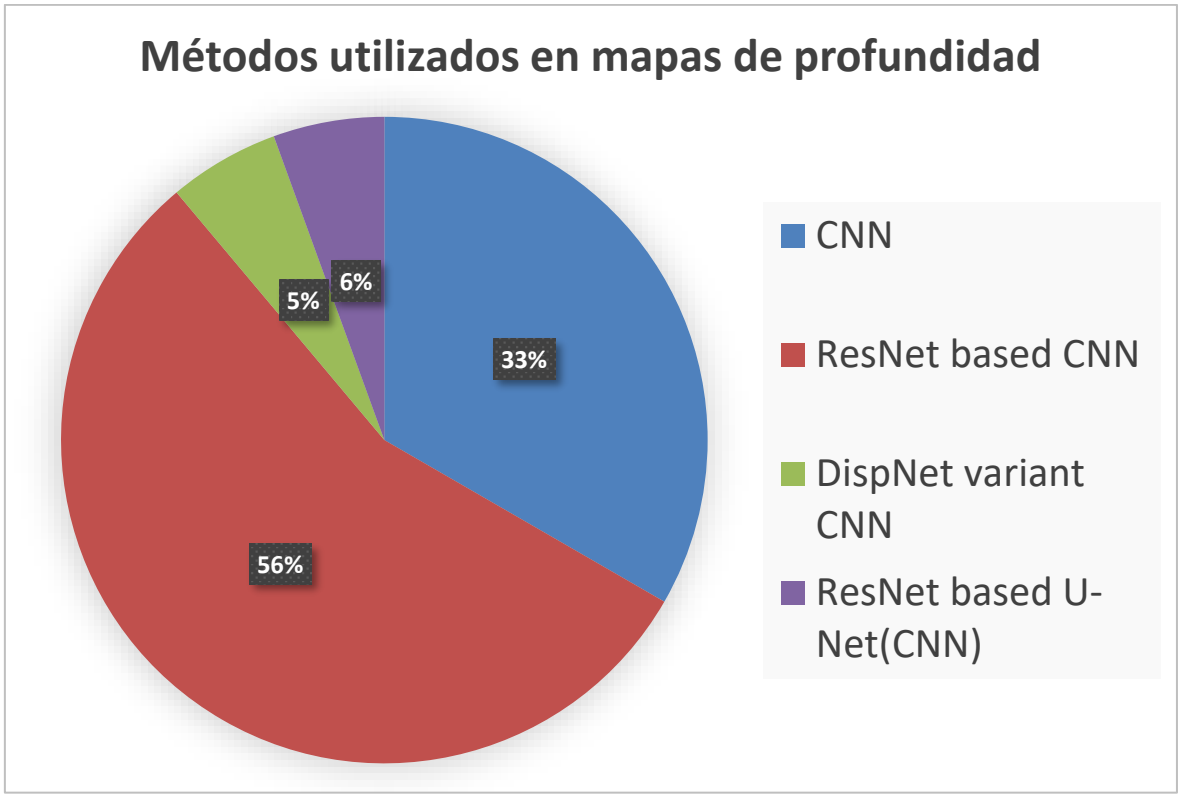


Figura 8: Métodos utilizados en mapas de profundidad

Capítulo 3

3. METODOLOGÍA

En este capítulo se describen los materiales utilizados para definir el método propuesto de estabilización de video, los procesos de estabilización de video basados en características, los procesos son: estimación de movimiento, filtrado de movimiento y corrección de movimiento. También se describe un paso que se agregara a la imagen de entrada al estabilizador de movimiento. Finalmente se describe los métodos para obtener mapas de profundad basados en CNN.

3.1. Materiales

Equipo

Para procesar las tareas de este trabajo de investigación se ha utilizado una laptop con las siguientes características: Procesador Inter(R) Core(TM) i5-10500H CPU @ 2.50GHz, Tarjeta gráfica NVIDIA GeForce RTX 3060 Laptop GPU, Memoria RAM de 8 GB y Disco de estado sólido de 500 GB.

Lenguaje de programación

El lenguaje de programación utilizado es Python en la versión 3.7.13, es fundamental utilizar esta versión debido a que el modelo de detección de mapas de profundidad trabaja con librerías que no son compatibles con versiones superiores.

Librerías

Las librerías principales de Python utilizadas para este trabajo de investigación son:

LIBRERÍA	VERSIÓN	CARACTERISTICA
Torch	1.1.0	Aceleración de GPU y redes neuronales profundas
cv2	4.4.0	Mejor rendimiento con GPU e integridad
vidstab	1.7.4	Clase utilizada para la estabilización de video
numpy	1.21.5	Objeto de matriz N-dimensional

Conjunto de datos

La base de datos utilizada para evaluar el algoritmo de estabilización es de Nghia et al. [70]. Se utilizo el siguiente algoritmo para la clasificación del conjunto de datos, se definió dos conjuntos de datos: con iluminación alta e iluminación baja.

Algoritmo:

- Recibe como entrada una imagen, la dimisión y el umbral,
- Cambiar de tamaño a la imagen según la dimensión recibida,
- Convertir el espacio de color a formato LAB y se extrae el canal L,
- Normalizar el canal L dividiendo todos los pixeles para el pixel de mayor valor,
- Retorna un valor “True” si la media es mas grande que el umbral, caso contrario “False”
- “True” significa que tiene iluminación baja y “False” que tiene iluminación baja.

Estos dos grupos de datos se dividió en subgrupos de escenas que contengan objetos en movimiento y objetos estáticos. Así, se definieron los siguientes grupos:

- Conjunto de escenas con iluminación alta y objetos en movimiento,
- Conjunto de escenas con iluminación alta y objetos estáticos,
- Conjunto de escenas con iluminación baja y objetos en movimiento, y
- Conjunto de escenas con iluminación baja y objetos en estáticos.

3.2. Estabilización de video general

La estabilización de video tiene el propósito de eliminar los movimientos involuntarios de la cámara, para este trabajo se considera la estimación de puntos de interés, búsqueda de correspondencia y la compensación de movimiento. En la Figura 9, se muestra flujo general de un método de estabilización de video [70].

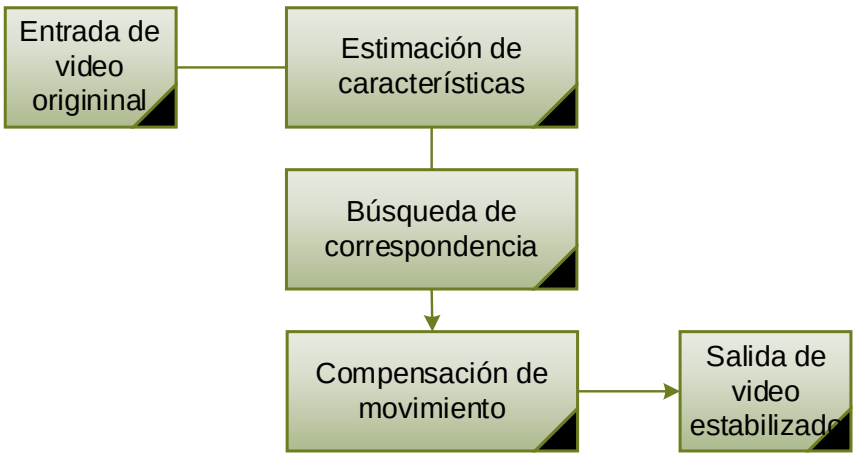


Figura 9: Flujo general de estabilización de video

El algoritmo funciona definido por Nghia et al. [70] de la siguiente manera:

- Estimar las características donde se define SIFT como extractor de características, luego obtiene los puntos característicos y los descriptores de las características
- Se realiza la búsqueda de correspondencias en base a dos fotogramas consecutivos
- Encuentre la transformación del fotograma anterior al actual utilizando el flujo óptico para todos los fotogramas. La transformación solo consta de tres parámetros: dx, dy, da $(ángulo)$. Básicamente, una transformación euclidiana rígida, sin escalar y sin compartir.
- Acumule las transformaciones para obtener la "trayectoria" para el ángulo formado por x, y, en cada cuadro.
- Suavice la trayectoria usando una ventana promedio deslizante. El usuario define el radio de la ventana, donde el radio es el número de fotogramas utilizados para suavizar.
- Cree una nueva transformación tal que:
  $nueva\_transformacion = transformacion + (trayectoria\_suavizada \ - trayectoria)$
- Aplica la nueva transformación al video.

Estimación de características

En el proceso de detección y extracción de características en fotogramas, se utilizan diferentes métodos para detectar las características relevantes en una imagen. Los principales métodos utilizados para la detección de características son: Scale Invariant Feature Transform (SIFT) [35][71] [72], Speed Up Robust Feature (SURF) [26][73][30][74][34], Robust Independent Elementary Features (Brief), Oriented Fast, Rotated Brief (ORB) [75][76].

Este proceso obtiene como resultado los puntos característicos y los descriptores de las características del fotograma. En la Figura 10, se muestra el flujo de los métodos de estabilización de video basados en características.

Figura 10: Flujo de estimación de características

En la Figura 11, se muestra las características tomadas de 5 fotogramas consecutivos de dos escenas diferentes

Figura 11: Detección de características

Búsqueda de coincidencia de características

La búsqueda de coincidencias es el proceso que toma el descriptor de una característica en el primer fotograma f_k y se compara con todas las demás características en el segundo fotograma f_{k+1} utilizando algún cálculo de distancia, el resultado es el más cercano.

Existe el proceso de coincidencia de fuerza bruta que se puede realizar con descriptores SIFT, SURF y ORB. Existe una diferencia que para ORB no se puede realizar la prueba de relación que discrimina las coincidencias de características encontradas en base a un parámetro de distancia [77], así se elige las mejores coincidencias y elimine las ruidosas. En la Figura 12 se muestra el flujo de coincidencias de características

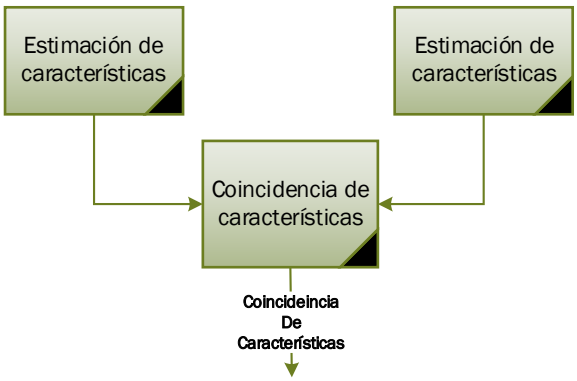


Figura 12: Flujo de búsqueda de coincidencia de características

En la Figura 13, se muestra la coincidencia de características tomadas de 5 pares de fotogramas consecutivos de dos escenas diferentes.

Figura 13: Coincidencia de características

El ajuste de la transformación geométrica se realiza en base a las coincidencias de características. La deformación entre fotogramas se puede expresar matemáticamente mediante la transformación geométrica que relaciona los puntos de interés:

$$I_t = {H_t}^{-1} * I_{sp} \quad (1)$$

Donde I_{sp} es un conjunto de puntos de interés de la imagen de referencia y la imagen no compensada, respectivamente. Y H_t es la matriz de transformación geométrica.

Se utiliza el modelo de traslación, que se refiere al movimiento de la imagen cuando el movimiento del dispositivo de captura es de traslación, paralelo al plano de la imagen. El modelo es el siguiente:

$$H_t = \begin{bmatrix} 1 & 0 & tx \\ 0 & 1 & ty \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

$$H_t = \begin{bmatrix} s * \cos(\Phi) & -s * sen(\Phi) & t_x \\ s * sen(\Phi) & s * \cos(\Phi) & t_y \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

Donde t_x y t_y son traslaciones, y Φ es el ángulo de rotación de “Roll”.

Se utiliza este modelo, ya que de la matriz se extraen los valores de las variables t_x, t_y, y Φ, para que puedan ser procesados utilizando el Filtro de Kalman implementado anteriormente, donde se comparan los valores reales con los anteriores, estimando una intención de movimiento segura. El modelo afín estima cuatro parámetros, dos desplazamientos en el plano paralelo de la imagen, como el modelo de traslación, rotación “Roll” y escala s, que son proporcionales al movimiento en la orientación del eje “Roll” [78].

Compensación de movimiento

La compensación de movimiento es el proceso de calcular el modelo de movimiento definido por la traslación, rotación y ángulo de rotación en el plano. En la Figura 14 se muestra el flujo de compensación de movimiento.

Figura 14: Flujo de compensación de movimiento

En la Figura 15, se muestra la compensación de movimiento tomadas de 5 fotogramas consecutivos de dos escenas diferentes, el lado izquierdo de la escena corresponde al video original y lado derecho corresponde al video estabilizado

Figura 15: Compensación de movimiento

Un punto muy importante es la acumulación de movimiento que se define por la multiplicación de las matrices de movimiento entre los fotogramas. En la ecuación (4) se representa la ecuación de acumulación de movimiento en $t-1$ y se denota como MA_{t-1}, representada en función de su traslación $t(x,y)$, rotación Φ y escala s.

$$MA_{t-1} = \begin{bmatrix} s * \cos(\Phi) & -s * sen(\Phi) & t_x \\ s * sen(\Phi) & s * \cos(\Phi) & t_y \\ 0 & 0 & 1 \end{bmatrix} \quad (4)$$

Donde, la acumulación de movimiento de un siguiente fotograma se representa como en la ecuación (5).

$$MA_t = \begin{bmatrix} s * \cos(\Phi) & -s * sen(\Phi) & t_x \\ s * sen(\Phi) & s * \cos(\Phi) & t_y \\ 0 & 0 & 1 \end{bmatrix} * MA_{t-1} \quad (5)$$

3.3. Estabilización de video basado en discriminación de características

Los métodos de estabilización de video existentes realizan el proceso de estimación de características, búsqueda de coincidencias y compensación de movimiento, se propone en este trabajo realizar un proceso adicional en el fotograma antes que ingresa al método de estabilización de video.

Detección del mapa de profundidad

La estimación de profundidad es un tema muy tratado en el área de visión por computadora, y es especial importancia para la robótica. Esto se debe a que permite a los robots observar e interactuar con el mundo que les rodea. Para una estimación de profundidad se utiliza CNN [79][80][4][81][82][83], ResNet [84][85][86][87][88][89][90][91][92][93], U-Net [94] y DispNet[95], todos estos métodos están basados en CNN de los cuales ResNet es la método utilizado.

A. Red Residual

Residual Networks (ResNets) está basada en CNN que aprenden funciones residuales con referencia a las entradas de la capa, en lugar de aprender funciones no referenciadas. En lugar de esperar que cada una de las capas apiladas se ajuste directamente a un mapeo subyacente deseado, las redes residuales permiten que estas capas se ajusten a un mapeo residual. Apilan bloques residuales [96] uno encima del otro para formar una red [96].

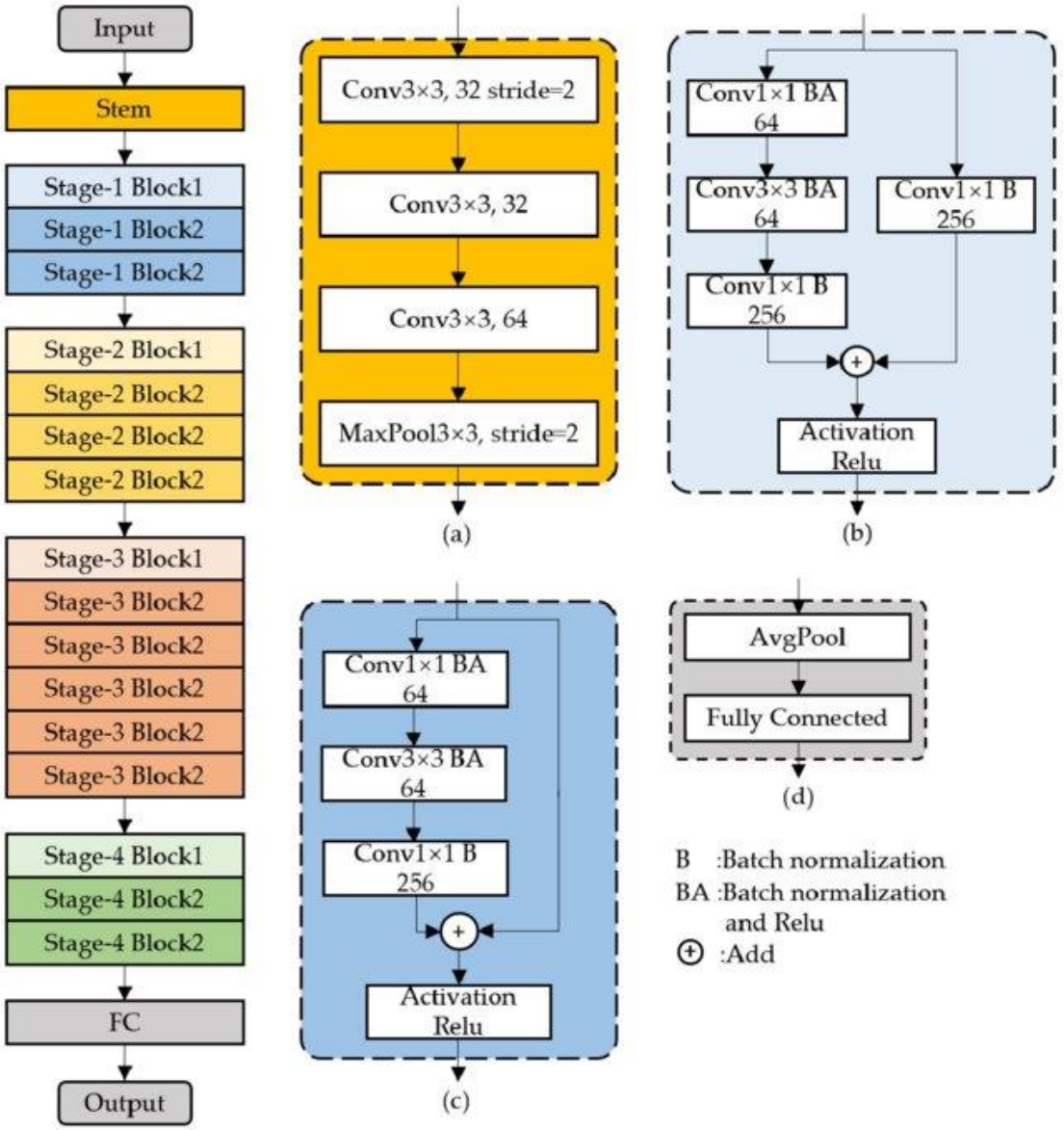


Figura 16: La arquitectura de ResNet-50-vd. (a) bloque de tallo; (b) Etapa 1-Bloque 1; (c) Etapa 1-Bloque 2; (d) Bloque FC [97].

Utilizando los mapas de profundidad obtenidos utilizando un método basado en CNN, combinado con algún método de enmascaramiento se obtendrá un nuevo fotograma que no contenga las características discriminadas para iniciar el proceso de estabilización de video general mostrado en la Figura 9. Además, en la Figura 17, se muestra el flujo de método propuesto para discriminación de puntos de interés en base a mapas de profundidad.

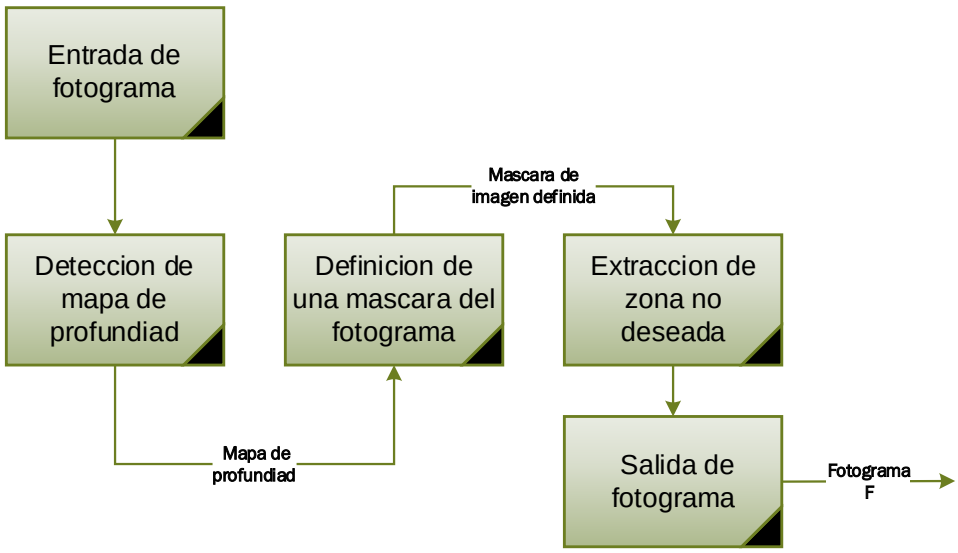


Figura 17: Flujo de método propuesto para discriminación de puntos de interés en base a mapas de profundidad.

Detección de mapa de profundidad: Esta definido por Ranftl et al. [89], que utiliza combinaciones de bases de datos para probar el método basado en ResNet [98]. En la Figura 18, se muestra la los mapas de profundidad obtenidos de 10 fotogramas elegidos al azar de la base de datos definida [99] definida por Shuaicheng et al. [39], el lado izquierdo de la escena corresponde al fotograma original y lado derecho corresponde al mapa de profundidad detectado.

Figura 18: Detección de mapas de profundidad [95]. (a) Fotograma original (b) Mapa de profundidad

Definición de mascara del fotograma: Luego de obtener el mapa de profundidad se procese a realizar el procedo de binarización de Otsu [100] a partir del mapa de profundidad, es uno de los métodos más efectivos para trabajar en base del umbral de un fotograma.

En la Figura 19, se muestra la los la máscara de imagen definida a partir del mapa de profundidad de 10 fotogramas elegidos al azar de la base de datos [99] definida por Shuaicheng et al. [39], el lado izquierdo de la escena corresponde al mapa de profundidad y lado derecho corresponde a la máscara definida en base al mapa de profundidad

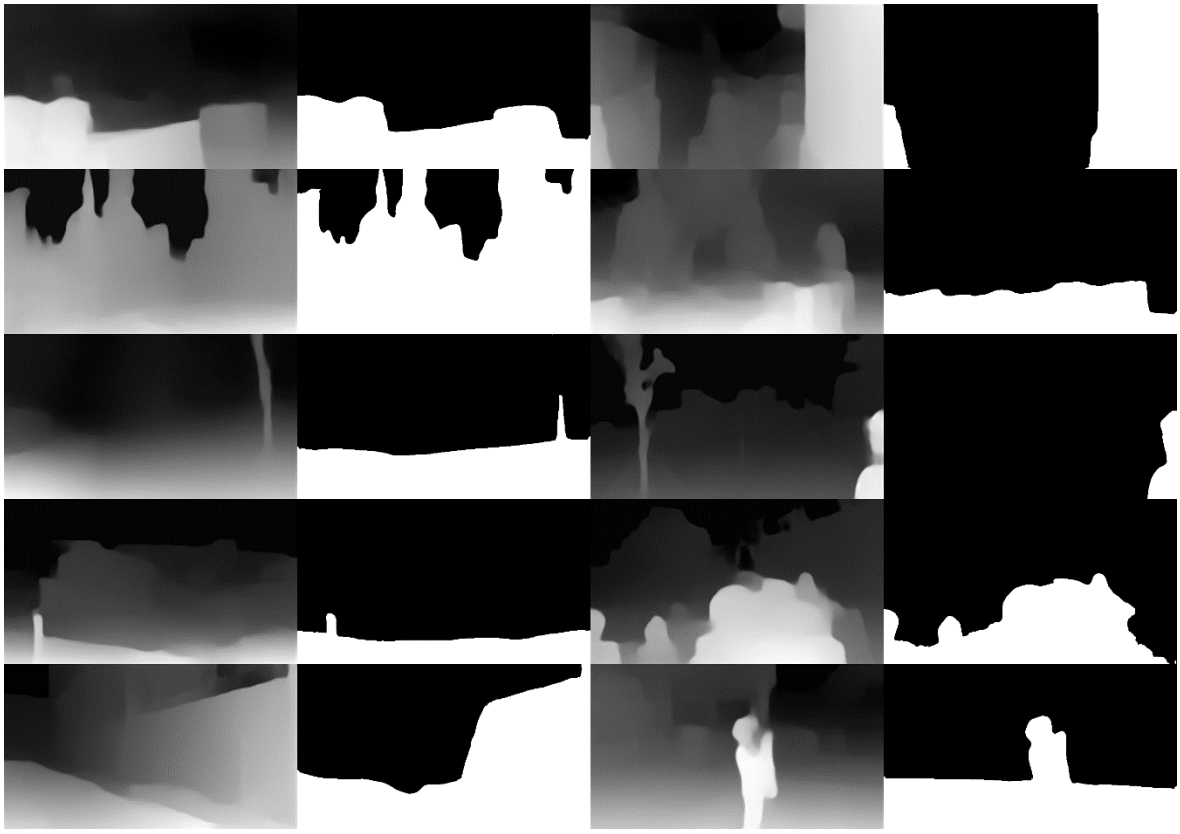


Figura 19: Mapa de profundidad aplicado la binarización de Otsu.

Extracción de la zona no deseada: Para la discriminación de los puntos de interés o eliminación de la zona no desea se realiza aplicando en el fotograma original la máscara definida en base al mapa de profundidad.

En la Figura 20, se muestra los resultados de 10 fotogramas aplicado una máscara definida a base de un mapa de profundidad, dando un fotograma resultante, este fotograma es el dato de entrada para el método de estabilización de video básico definido en un punto anterior.

Salida del fotograma: El método propuesto para la discriminación de puntos de interés en base a mapas de profundidad definido en la Figura 17, entrega un fotograma resultado de este proceso para iniciar con el método de estabilización de video básico. En la Figura 21, se muestra algunas salidas en secuencia de 5 fotogramas, que son entrada para el método de estabilización de video básico.

Figura 20: Extracción de la zona no deseada

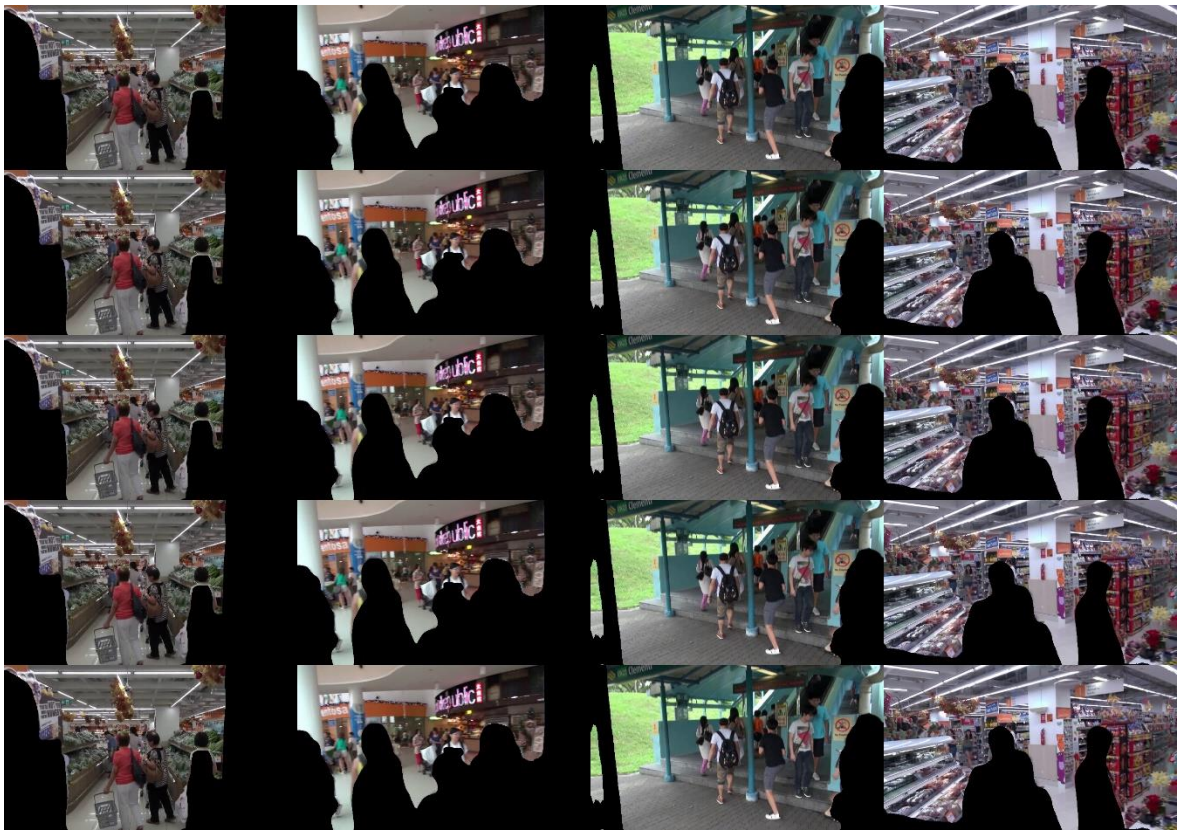
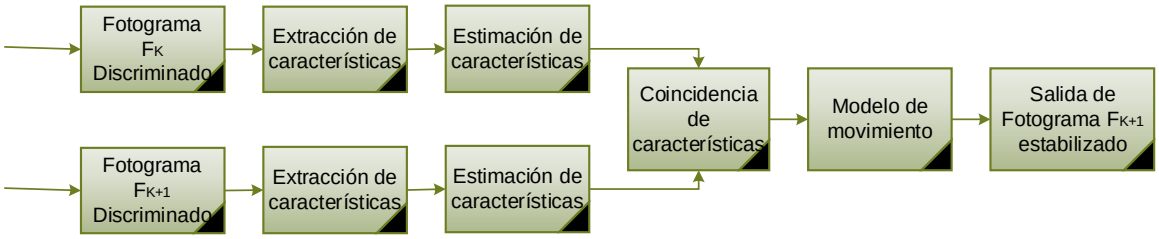


Figura 21: Salida de fotogramas

En la RSL de estabilización de video se han propuesto métodos que no realizan discriminación de puntos de interés en base a mapas de profundidad. En la Figura 22, se muestra el método propuesto de estabilización de video que realiza la discriminación de puntos de interés en base a mapas de profundidad.

(a) Flujo de método de discriminación de puntos de interés en base a mapas de profundidad



(b) Flujo de método de estabilización de video general

Figura 22: Flujo de método de estabilización de video que discrimina puntos de interés en base a mapas de profundidad. (a) Flujo de método de discriminación de puntos de interés en base a mapas de profundidad. (b) Flujo de método de estabilización de video general

El método de estabilización de video básico recibe como entrada la secuencia de fotogramas discriminado los puntos de interés en base a los mapas de profundidad de cada fotograma. En la Figura 23,Figura 24 y Figura 25 se muestra los resultados de realizar el proceso de estimación de características, coincidencia de características y compensación de movimiento respectivamente.

Figura 23: Estimación de características

Figura 24: Coincidencia de características

Figura 25: Compensación de movimiento

3.4. Métricas de evaluación

Para evaluar la efectividad del enfoque propuesto, adoptamos el ITF que ha sido ampliamente utilizado para medir la suavidad temporal para el desempeño de la estabilización de video. Un valor alto de ITF [101][26] [35] a menudo indica un video de buena calidad. La ITF se puede obtener mediante:

$$ITF = \frac{1}{T_f - 1} \sum_{k=1}^{T_f - 1} PSNR(k) \tag{6}$$

Donde T_f es el total de fotogramas de la secuencia de video y $PSNR(k)$ es la relación señal/ruido máximo entre dos fotogramas consecutivos (f_k, f_{k+1}) que se define como:

$$PSNR(f_t, f_{t+1}) = 10log_{10} \frac{I_{max}}{MSE(k)} \tag{7}$$

donde I_{max} es la intensidad máxima de píxeles de un cuadro y es igual a 255 en imágenes en escala de grises de 8 bits. El $MSE(k)$ es el valor de desviación del nivel de gris de píxel correspondiente entre fotogramas adyacentes y su fórmula de cálculo es:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|f_t(x,y) - f_{t+1}(x,y)\|^2 \tag{8}$$

donde m y n son la alto y ancho de la imagen, $I_{k+1}(i,j)$ y $I_k(i,j)$ son los valores de gris del fotograma f_{k+1} y f_k.

3.5. Resultados

De acuerdo con las tres fórmulas anteriores, calculamos el valor ITF del video original, el video con método de estabilización de video básico y el video con el método de estabilización de video propuesto, se han definido cuatro grupos para evaluar el método propuesto, a continuación se muestran los resultados.

En la Tabla 12, se muestra los resultados de conjunto de escenas con alta iluminación y objetos en movimiento, donde la puntuación ITF del conjunto de escenas evaluado sobre el método propuesto se encuentra de 2,85 a 13,82. Se debe mencionar que en una de las escenas el ITF de método básico fue mayor que el ITF del método propuesto por 0.03.

Tabla 12: Resultados de las pruebas de evaluación con ITF de escenas con alta iluminación y objetos en movimiento

ESCENA \ PROCESO	ORIGINAL	EV BÁSICO	EV PROPUESTO
Video1	3,07	3,44	**3,70**
Video2	3,44	3,69	**4,03**
Video3	4,06	4,75	**5,05**
Video4	11,20	13,26	**13,82**
Video5	6,93	7,87	**8,18**
Video6	2,48	2,94	**3,00**
Video7	5,84	**7,57**	7,54
Video8	11,15	13,02	**13,71**
Video9	6,15	7,75	**7,94**
Video10	4,72	5,31	**5,60**
Video11	5,61	6,53	**6,68**
Video12	6,46	7,79	**7,87**
Video13	7,02	8,17	**8,42**
Video14	4,10	4,66	**4,78**
Video15	2,79	3,20	**3,44**
Video16	6,95	8,26	**8,68**
Video17	5,40	6,01	**6,17**
Video18	7,92	8,81	**9,12**
Video19	2,97	3,54	**3,58**
Video20	2,85	3,21	**3,38**
Video21	2,55	3,15	**3,30**
Video22	4,35	4,99	**5,51**
Video23	5,10	5,81	**5,97**
Video24	2,54	2,84	**2,95**

En la Tabla 13, se muestra los resultados de conjunto de escenas con alta iluminación y objetos estáticos, donde la puntuación ITF del conjunto de escenas se encuentra de 4,06 a 14,99.

Tabla 13: Resultados de las pruebas de evaluación con ITF de escenas con alta iluminación y objetos estáticos

ESCENA \ PROCESO	ORIGINAL	EV BÁSICO	EV PROPUESTO
Video1	5,02	5,70	**6,11**
Video2	6,89	7,47	**7,89**
Video3	7,31	8,66	**8,94**
Video4	5,82	6,81	**6,99**
Video5	5,67	6,28	**6,69**
Video6	9,33	9,96	**10,50**
Video7	4,52	4,94	**5,26**
Video8	3,16	3,86	**4,06**
Video9	7,12	8,30	**8,72**
Video10	6,37	7,69	**8,00**
Video11	2,78	3,03	**4,14**
Video12	8,55	9,40	**9,95**
Video13	5,57	6,87	**7,24**
Video14	5,62	6,48	**7,35**
Video15	10,72	14,14	**14,99**
Video16	9,13	10,33	**11,31**
Video17	8,63	10,62	**11,00**
Video18	5,87	6,69	**7,04**
Video19	9,12	11,37	**12,47**
Video20	4,86	6,12	**6,82**
Video21	5,33	5,90	**6,23**
Video22	6,24	6,88	**7,39**
Video23	5,79	6,36	**7,13**
Video24	7,43	8,44	**8,63**

En la Tabla 14, se muestra los resultados de conjunto de escenas con baja iluminación y objetos en movimiento, donde la puntuación ITF del conjunto de escenas se encuentra de 2,23 a 16,73. Se debe mencionar que en una de las escenas el ITF de método básico fue mayor que el ITF del método propuesto por 0.02.

Tabla 14: Resultados de las pruebas de evaluación con ITF de escenas con baja iluminación y objetos en movimiento

ESCENA \ PROCESO	ORIGINAL	EV BÁSICO	EV PROPUESTO
Video1	4,71	5,30	**5,41**
Video2	3,67	4,17	**4,23**
Video3	3,90	4,42	**4,68**
Video4	3,09	3,74	**3,87**
Video5	5,89	6,80	**7,26**
Video6	5,37	6,32	**6,42**
Video7	11,17	13,65	**14,11**
Video8	2,01	2,36	**2,45**
Video9	4,64	5,54	**5,69**
Video10	6,58	8,17	**8,21**
Video11	4,33	5,55	**5,60**
Video12	4,72	5,77	**5,88**
Video13	13,01	16,38	**16,73**
Video14	4,09	4,92	**4,99**
Video15	9,15	11,18	**11,38**
Video16	6,21	7,13	**7,31**
Video17	3,91	4,53	**4,72**
Video18	6,17	6,90	**7,19**
Video19	8,06	9,68	**9,83**
Video20	2,94	3,17	**3,38**
Video21	6,47	7,63	**7,78**
Video22	1,92	2,28	**2,33**
Video23	2,94	3,49	**3,67**
Video24	1,77	**2,25**	2,23

En la Tabla 15, se muestra los resultados de conjunto de escenas con baja iluminación y objetos estáticos, donde la puntuación ITF del conjunto de escenas se encuentra de 2,67 a 10,92.

Tabla 15: Resultados de las pruebas de evaluación con ITF de escenas con baja iluminación y objetos estáticos

ESCENA\ PROCESO	ORIGINAL	EV BÁSICO	EV PROPUESTO
Video1	3,62	4,43	**4,80**
Video2	2,39	2,68	**2,75**
Video3	2,29	2,49	**2,67**
Video4	5,29	6,13	**6,35**
Video5	4,34	4,74	**4,84**
Video6	6,38	6,70	**7,05**
Video7	5,49	6,41	**6,78**
Video8	5,81	6,40	**7,24**
Video9	4,42	4,66	**4,93**
Video10	3,14	3,78	**4,02**
Video11	8,40	10,33	**10,92**
Video12	2,40	3,03	**3,05**
Video13	6,52	7,65	**7,78**
Video14	7,98	9,15	**10,21**
Video15	6,20	6,70	**7,11**
Video16	6,54	7,04	**7,44**
Video17	7,40	8,24	**8,70**
Video18	5,14	**6,45**	6,41
Video19	5,89	6,41	**6,65**
Video20	5,35	5,81	**6,05**
Video21	6,57	7,05	**7,79**
Video22	5,32	5,78	**6,08**
Video23	4,37	4,73	**5,03**
Video24	7,41	8,20	**9,57**

En la Tabla 16, se muestra los resultados de grupo de escenas, donde el promedio de la puntuación ITF del conjunto de escenas se encuentra de 6,35 a 8,12. Después de analizar los resultados sobre el método propuesto de cada conjunto de datos presenta un incremento de 0,17 a 0,52 sobre el ITF del método de estabilización de video básico.

Tabla 16: Resultados de las pruebas de evaluación con ITF según los grupos de escenas analizados

GRUPO ESCENAS\ PROCESO	ORIGINAL	EV BÁSICO	EV PROPUESTO
brillante con objetos en movimiento	5,24	6,11	**6,35**
brillante con objetos estáticos	6,54	7,60	**8,12**
oscuro con objetos en movimiento	5,28	6,30	**6,47**
oscuro con objetos estáticos	5,36	6,04	**6,43**
	5,60	6,51	**6,84**

3.6. Discusiones

Después de explorar en contenido de los artículos relacionados a mapas de profundidad involucra el uso de métodos basados en CNN que es muy utilizados en los últimos trabajos publicados. Sobre los métodos de estabilización de video basado en características, no se explica porque usan siempre la imagen completa como entrada al algoritmo de estabilización.

Los métodos de estabilización de video no han pasado de moda, ya que existen investigadores fieles que siguen realizando mejoras en los métodos de estabilización de video basado en características. Los métodos SIFT, SURF y ORB son muy utilizados en muchos trabajos como parte del algoritmo de estabilización.

Los métodos de estabilización de video basados en la discriminación de puntos de interés utilizando mapas de profundidad, según la RSL no existe publicaciones en las bibliotecas digitales definidas para esta investigación, talvez una diferente definición de las cadenas de búsqueda origine mejores resultados.

La definición del algoritmo propuesto de estabilización de video se lo realiza en base a una librería de estabilización de video [70] combinada con un método de mapas de profundidad robusto [89], siendo intermediario el método de binarización de Otsu [100] utilizado para

definir una máscara de fotograma en base al mapa de profundidad y aplicado en el fotograma completo de entrada al método de estabilización de video.

Los resultados del trabajo realizado se aprecia de manera que el algoritmo propuesto para estabilización de video basado en la discriminación de puntos de interés utilizando mapas de profundidad mejora los resultados respecto a un método de estabilización de video [70] definido, se realiza esta comparación debido a que no existen métodos de estabilización de video que discriminen puntos de interés en base a mapas de profundidad.

Capítulo 4

4. CONCLUSIONES

4.1. Objetivos de investigación: Resumen de los hallazgos y conclusiones

Objetivos de investigación: Resumen de los hallazgos y conclusiones

Existen algunas preguntas de investigación que se han ilustrado a través de la investigación. Primero: ¿Qué métodos de estabilización de video utilizan la discriminación de puntos de interés mediante mapas de profundidad? En segundo lugar: ¿Qué métodos permiten la generación de una imagen de profundidad teniendo como entrada imagen simple? Tercero: ¿Qué bases de datos son utilizadas en los métodos de estimación de profundidad o mapas de profundidad? Cuarto: ¿Qué métricas se utilizan para evaluar los resultados de estabilización de video? Finalmente: ¿Qué técnicas de extracción de características de una imagen son utilizados en loa métodos de estabilización de video?

En las siguientes subsecciones se aborda estas preguntas alineadas a los objetivos de investigación planteados. Se inicia con lo relevante a la revisión de literatura realizada. Además, se evalúan los descubrimientos experimentales y se proponen trabajos futuros.

Estabilización de video en base a la discriminación de puntos de interés utilizando mapas de profundidad

Conclusiones

Los métodos de estabilización de video basado en características logran el objetivo de mejorar la calidad del video al minimizar o eliminar el movimiento no deseado entre fotogramas, el método de estabilización de puede ser representado en forma de diagrama.

Los mapas de profundidad tienen como objetivo estimar la profundidad de un objeto en una imagen, siendo muy útil para definir una máscara de imagen basada en una medida de umbral y poder discriminar las regiones y objetos cercanos al campo de visión.

La RSL realizada mostro que en todos los métodos de estabilización de video basados en características se utiliza la secuencia de fotogramas completos, confirmando que

discriminar puntos de interés en base a mapas de profundidad es una propuesta ya definida en esta investigación.

Conjuntos de datos utilizados en detección de mapas de profundidad y estabilización de video.

Conclusiones

Los conjuntos de datos son esenciales para el entrenamiento de una CNN, y se ha demostrado según la RSL ser muy utilizados los métodos de detección de mapas de profundidad basados en CNN, por lo tanto, los conjuntos de datos son muy importantes.

Los métodos de estabilización de video basados en características utilizan conjuntos de videos agrupados por tipos para evaluar su rendimiento, para este trabajo se clasifico en 4 grupos en base a la intensidad de luz existente en los videos.

Algoritmo de estabilización de video en ambientes controlados discriminando puntos de interés utilizando mapas de profundidad

Conclusiones

Se ha definido un algoritmo de estabilización de video que toma como entrada un fotograma discriminado los puntos de interés de objetos cercanos al campo de visión en movimiento o estáticos. La discriminación de puntos de interés se realiza en base a ResNet basado en CNN que ha permitido obtener los mapas de profundidad y definir una mascara de fotograma para realizar este proceso, y así definir el fotograma discriminado los puntos de interés para seguir con el proceso de estabilización de video.

Comparar los resultados obtenidos con otros métodos de estabilización de video basados en características.

Conclusiones

Los resultados obtenidos sobre el grupo de escenas han demostrado que el método propuesto mejora la estabilización de video utilizando como entrada fotogramas donde se discriminan punto de interés cercanos, el promedio de la evaluación ITF del conjunto de escenas se encuentra de 6,35 dB a 8,12 dB.

Los resultados sobre el método propuesto de cada conjunto de datos presenta un incremento de 0,17dB a 0,52 dB sobre el ITF del método de estabilización de video básico, Finalmente, se obtiene un incremento de rendimiento promedio de 5% sobre el método de estabilización de video básico [70].

Recomendaciones y trabajos futuros

Antes de iniciar un proyecto que involucre mapas de profundidad o estimación de profundidad debe analizarse los requerimientos computacionales para poder ejecutar los modelos entrenados disponibles, ya que estos modelos necesitan procesamiento con GPU para funcionar de forma eficiente con un rendimiento de Frames por Segundo (FPS) aceptables.

Existen algunos aspectos de rendimiento de FPS respecto a la detención de mapas de profundidad, podrían analizarse en el futuro para incrementar el rendimiento del algoritmo actual.

Es importante en este punto, estudiar otros métodos que puedan realizar la detección de mapas de profundidad con un menor costo computacional, así mejorar las aplicaciones de este algoritmo en tiempo real.

REFERENCIAS BIBLIOGRÁFICAS

[1] ChoiJinsoo and K. So, “Deep Iterative Frame Interpolation for Full-frame Video Stabilization,” *ACM Trans. Graph.*, vol. 39, no. 1, Jan. 2020, doi: 10.1145/3363550.

[2] W. G. Aguilar and C. Angulo, “Real-time video stabilization without phantom movements for micro aerial vehicles,” *Eurasip J. Image Video Process.*, vol. 2014, no. 1, pp. 1–13, Sep. 2014, doi: 10.1186/1687-5281-2014-46.

[3] R. Szeliski, *Computer Vision: Algorithms and Applications*, Ilustrada. 2010. [Online]. Available: https://books.google.com.ec/books?id=bXzAlkODwa8C

[4] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep Ordinal Regression Network for Monocular Depth Estimation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011. doi: 10.1109/CVPR.2018.00214.

[5] K. Bipin, V. Duggal, and K. Madhava Krishna, “Autonomous navigation of generic monocular quadcopter in natural environment,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1063–1070. doi: 10.1109/ICRA.2015.7139308.

[6] G. Nur, S. Dogan, H. K. Arachchi, and A. M. Kondoz, “Impact of depth map spatial resolution on 3D video quality and depth perception,” in *2010 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, 2010, pp. 1–4. doi: 10.1109/3DTV.2010.5506315.

[7] A. Saxena, M. Sun, and A. Ng, “Make3D: Depth Perception from a Single Still Image.,” in *Proceedings of the National Conference on Artificial Intelligence*, 2008, vol. 3, pp. 1571–1576.

[8] H. Kumar, S. Gupta, and K. S. Venkatesh, “Hole correction in estimated depth map from single image using color uniformity principle,” in *International Conference on Digital Signal Processing, DSP*, Nov. 2017, vol. 2017-Augus. doi: 10.1109/ICDSP.2017.8096058.

[9] A. Bhoi, “Monocular Depth Estimation: A Survey,” Jan. 2019, Accessed: Sep. 27, 2021. [Online]. Available: https://arxiv.org/abs/1901.09402v1

[10] W. Guilluy, L. Oudre, and A. Beghdadi, “Video stabilization: Overview, challenges and perspectives,” *Signal Process. Image Commun.*, vol. 90, p. 116015, Jan. 2021, doi: 10.1016/J.IMAGE.2020.116015.

[11] L. Hlmg, K. Reddy, and S.Akhila, “A survey on video stabilization algorithms,” *Int. J. Adv. Inf. Sci. Technol.*, vol. Vol 31, 2014.

[12] A. Saxena, M. Sun, and A. Y. Ng, “Make3D: Learning 3D scene structure from a single still image,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, 2009, doi: 10.1109/TPAMI.2008.132.

[13] N. Kumar *et al.*, “VIDEO STABILIZATION: AN IN-DEPTH SURVEY,” *Int. J. Technol. Res. Eng.*, vol. 5, no. 10, 2018, Accessed: Oct. 01, 2021. [Online]. Available: www.ijtre.com

[14] H. B. Ghorpade and S. K. Jagtap, “A Survey: Video Stabilization Techniques for Handheld Videos,” *Curr. Trends Signal Process.*, vol. 7, no. 2, pp. 8–12, Sep. 2017, Accessed: Oct. 01, 2021. [Online]. Available: http://stmjournals.com/index.php?journal=CTSP&page=article&op=view&path%5B%5D=8617

[15] S. Gur and L. Wolf, “Single image depth estimation trained via depth from

defocus cues," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, vol. 2019-June, pp. 7675–7684. doi: 10.1109/CVPR.2019.00787.

[16] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, vol. 2015 Inter, pp. 2650–2658. doi: 10.1109/ICCV.2015.304.

[17] K. Dawson-Howe, *A Practical Introduction to Computer Vision with OpenCV*. 2014. Accessed: Sep. 27, 2021. [Online]. Available: http://www.amazon.com/Practical-Introduction-Computer-Imaging-Technology/dp/1118848454/ref=sr_1_6?s=books&ie=UTF8&qid=1415059357&sr=1-6&keywords=opencv

[18] A. Gallego Sanchez, P. Compañ Rosique, C. Villagrá, R. Molina, and P. Arques, "Detección de objetos y estimación de su profundidad mediante un algoritmo de estéreo basado en segmentación," *II Congr. Español Informática*, no. November 2015, 2007, Accessed: Sep. 27, 2021. [Online]. Available: https://www.researchgate.net/publication/39435989_Deteccion_de_objetos_y_estimacion_de_su_profundidad_mediante_un_algoritmo_de_estereo_basado_en_segmentacion

[19] "OpenCV: Introduction to SIFT (Scale-Invariant Feature Transform)." https://docs.opencv.org/3.4/da/df5/tutorial_py_sift_intro.html (accessed Jun. 19, 2022).

[20] "OpenCV: Introduction to SURF (Speeded-Up Robust Features)." https://docs.opencv.org/3.4/df/dd2/tutorial_py_surf_intro.html (accessed Jun. 19, 2022).

[21] "OpenCV: ORB (Oriented FAST and Rotated BRIEF)." https://docs.opencv.org/3.4/d1/d89/tutorial_py_orb.html (accessed Jun. 19, 2022).

[22] "OpenCV: FAST Algorithm for Corner Detection." https://docs.opencv.org/3.4/df/d0c/tutorial_py_fast.html (accessed Jun. 19, 2022).

[23] "OpenCV: BRIEF (Binary Robust Independent Elementary Features)." https://docs.opencv.org/3.4/dc/d7d/tutorial_py_brief.html (accessed Jun. 19, 2022).

[24] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2548–2555, 2011, doi: 10.1109/ICCV.2011.6126542.

[25] Y. C. Lee, K. W. Tseng, Y. T. Chen, C. C. Chen, C. S. Chen, and Y. P. Hung, "3D Video Stabilization with Depth Estimation by CNN-based Optimization," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10616–10625. doi: 10.1109/CVPR46437.2021.01048.

[26] W. G. Aguilar, D. Loza, L. Segura, A. Ibarra, and ..., "Onboard video stabilization for rotorcrafts," *… Intell. Robot. …*, 2017, doi: 10.1007/978-3-319-65292-4_60.

[27] J. Y. Xiong, M. Dai, C. L. Zhao, and ..., "Antiblurry dejitter image stabilization

method of fuzzy video for driving recorders," *KSII Trans. …*, 2017, [Online]. Available: https://www.koreascience.or.kr/article/JAKO201721948859829.page

[28] K. G. Derpanis, "Overview of the RANSAC Algorithm," 2010.

[29] M. R. e Souza and H. Pedrini, "Combination of local feature detection methods for digital video stabilization," *Signal, Image Video Process.*, vol. 12, no. 8, pp. 1513–1521, Nov. 2018, doi: 10.1007/S11760-018-1307-8.

[30] J. Ke, A. J. Watras, J. J. Kim, H. Liu, and ..., "Efficient Real-Time Video Stabilization with a Novel Least Squares Formulation," *ICASSP 2021-2021 …*, 2021, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9414545/

[31] M. K. Yuan, L. Q. Dai, D. M. Yan, L. Q. Zhang, J. Xiao, and X. P. Zhang, "Fast and Error-Bounded Space-Variant Bilateral Filtering," *J. Comput. Sci. Technol.*, vol. 34, no. 3, pp. 550–568, May 2019, doi: 10.1007/S11390-019-1926-8.

[32] H. Kawade and A. M. Deshpande, "Implementation of Video Stabilization Algorithm for Surveillance System," in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 2017, pp. 1–4. doi: 10.1109/ICCUBEA.2017.8463700.

[33] J. Dong and H. Liu, "Video Stabilization for Strict Real-Time Applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 716–724, 2017, doi: 10.1109/TCSVT.2016.2589860.

[34] Y. Qiu, X. Li, S. Ding, S. Yang, L. Li, and H. Zhang, "Stabilization Algorithm Based on Improved Motion Model for Jittery Video in Minimally Invasive Surgery," 2019. Accessed: Jun. 09, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/8868352/

[35] M. Fang, H. Li, and S. Si, "A video stabilization algorithm based on affine SIFT," 2018. Accessed: Jun. 09, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/8346332/

[36] K. S. Rao, A. V. Paramkusam, N. K. Darimireddy, and A. Chehri, "Block Matching Algorithms for the Estimation of Motion in Image Sequences: Analysis," *Procedia Comput. Sci.*, vol. 192, pp. 2980–2989, Jan. 2021, doi: 10.1016/J.PROCS.2021.09.070.

[37] M. D. Ansari, S. P. Ghrera, and V. Tyagi, "Pixel-Based Image Forgery Detection: A Review," *http://dx.doi.org/10.1080/09747338.2014.921415*, vol. 55, no. 1, pp. 40–46, Jan. 2014, doi: 10.1080/09747338.2014.921415.

[38] S. Poddar, R. Kottath, and V. Karar, "Motion Estimation Made Easy: Evolution and Trends in Visual Odometry," *Stud. Comput. Intell.*, vol. 804, pp. 305–331, 2019, doi: 10.1007/978-3-030-03000-1_13.

[39] S. Liu, L. Yuan, P. Tan, and J. Sun, "Bundled camera paths for video stabilization," *ACM Trans. Graph.*, vol. 32, no. 4, Jul. 2013, doi: 10.1145/2461912.2461995.

[40] J. Sánchez and J. M. Morel, "Motion smoothing strategies for 2D video stabilization," *SIAM J. Imaging Sci.*, vol. 11, no. 1, pp. 219–251, Jan. 2018, doi: 10.1137/17M1127156.

[41] Y. J. Koh, C. Lee, and C. S. Kim, "Video Stabilization Based on Feature Trajectory Augmentation and Selection and Robust Mesh Grid Warping,"

IEEE Trans. Image Process., vol. 24, no. 12, pp. 5260–5273, Dec. 2015, doi: 10.1109/TIP.2015.2479918.

[42] Y. S. Wang, F. Liu, P. S. Hsu, and T. Y. Lee, “Spatially and temporally optimized video stabilization,” *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 8, pp. 1354–1361, 2013, doi: 10.1109/TVCG.2013.11.

[43] S. Liu, L. Yuan, P. Tan, and J. Sun, “SteadyFlow: Spatially smooth optical flow for video stabilization,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4209–4216, Sep. 2014, doi: 10.1109/CVPR.2014.536.

[44] J. Yu and R. Ramamoorthi, “Robust video stabilization by optimization in CNN weight space,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, vol. 2019-June, pp. 3795–3803. doi: 10.1109/CVPR.2019.00392.

[45] J. Choi and I. S. Kweon, “Deep Iterative Frame Interpolation for Full-frame Video Stabilization,” *ACM Trans. Graph.*, vol. 39, no. 1, 2020, doi: 10.1145/3363550.

[46] M. Zhao and Q. Ling, “PWStableNet: Learning Pixel-Wise Warping Maps for Video Stabilization,” *IEEE Trans. Image Process.*, vol. 29, pp. 3582–3595, 2020, doi: 10.1109/TIP.2019.2963380.

[47] E. Ringaby and P. E. Forssén, “Efficient Video Rectification and Stabilisation for Cell-Phones,” *Int. J. Comput. Vis. 2011 963*, vol. 96, no. 3, pp. 335–352, Jun. 2011, doi: 10.1007/S11263-011-0465-8.

[48] F. Liu, M. Gleicher, H. Jin, and A. Agarwala, “Content-preserving warps for 3D video stabilization,” *ACM Trans. Graph.*, vol. 28, no. 3, Jul. 2009, doi: 10.1145/1531326.1531350.

[49] G. Zhang, W. Hua, X. Qin, Y. Shao, and H. Bao, “Video stabilization based on a 3D perspective camera model,” *Vis. Comput.*, vol. 25, no. 11, pp. 997–1008, Nov. 2009, doi: 10.1007/S00371-009-0310-Z.

[50] M. Wang *et al.*, “Deep Online Video Stabilization with Multi-Grid Warping Transformation Learning,” *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2283–2292, May 2019, doi: 10.1109/TIP.2018.2884280.

[51] A. Goldstein and R. Fattal, “Video stabilization using epipolar geometry,” *ACM Trans. Graph.*, vol. 31, no. 5, Sep. 2012, doi: 10.1145/2231816.2231824.

[52] P. Rawat and J. Singhai, “Adaptive Motion Smoothening for Video Stabilization,” *Int. J. Comput. Appl.*, vol. 72, pp. 14–20, 2013.

[53] M. Zhao and Q. Ling, “PWStableNet: Learning Pixel-Wise Warping Maps for Video Stabilization,” *IEEE Trans. Image Process.*, vol. 29, pp. 3582–3595, 2020, doi: 10.1109/TIP.2019.2963380.

[54] J. Hu, D. Q. Zhang, H. Yu, and C. W. Chen, “Multi-objective content preserving warping for image stitching,” *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2015-August, Aug. 2015, doi: 10.1109/ICME.2015.7177505.

[55] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, “Review on Convolutional Neural Networks (CNN) in vegetation remote sensing,” *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 24–49, Mar. 2021, doi: 10.1016/J.ISPRSJPRS.2020.12.010.

[56] R. Prabhu, “Understanding of Convolutional Neural Network (CNN) Deep Learning,” *Medium.Com*, pp. 1–11, 2018, Accessed: Jun. 04, 2022. [Online]. Available: https://medium.com/@RaghavPrabhu/understanding-of-

convolutional-neural-network-cnn-deep-learning-99760835f148
[57] F. Lateef and Y. Ruichek, “Survey on semantic segmentation using deep learning techniques,” *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019, doi: 10.1016/j.neucom.2019.02.003.
[58] K. Klasing, D. Althoff, D. Wollherr, and M. Buss, “Comparison of surface normal estimation methodsfor range sensing applications,” in *Proceedings - IEEE International Conference on Robotics and Automation*, 2009, pp. 3206–3211. doi: 10.1109/ROBOT.2009.5152493.
[59] J. S. Park and J. H. Han, “Contour motion estimation from image sequences using curvature information,” *Pattern Recognit.*, vol. 31, no. 1, pp. 31–39, Jan. 1998, doi: 10.1016/s0031-3203(97)00031-9.
[60] N. H. Khan and A. Adnan, “Ego-motion estimation concepts, algorithms and challenges: an overview,” *Multimed. Tools Appl.*, vol. 76, no. 15, pp. 16581–16603, Aug. 2017, doi: 10.1007/S11042-016-3939-4.
[61] Viraf, “Create A Synthetic Image Dataset — The ‘What’, The ‘Why’ and The ‘How’ | by Viraf | Towards Data Science,” *Towards Data Science*, 2020. https://towardsdatascience.com/create-a-synthetic-image-dataset-the-what-the-why-and-the-how-f820e6b6f718 (accessed Jun. 04, 2022).
[62] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-Janua, pp. 6602–6611. doi: 10.1109/CVPR.2017.699.
[63] R. Garg, B. G. Vijay Kumar, G. Carneiro, and I. Reid, “Unsupervised CNN for single view depth estimation: Geometry to the rescue,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9912 LNCS, pp. 740–756. doi: 10.1007/978-3-319-46484-8_45.
[64] IBM, “What is Computer Vision_ _ IBM.” 2019. Accessed: Jun. 04, 2022. [Online]. Available: https://www.ibm.com/topics/computer-vision
[65] CleverData, “¿Qué es Machine Learning? – Cleverdata,” *¿Que es Machine Learning?*, 2019.
[66] P. Lai, A. Ortega, C. C. Dorea, P. Yin, and C. Gomila, “Improving view rendering quality and coding efficiency by suppressing compression artifacts in depth-image coding,” in *Visual Communications and Image Processing 2009*, Jan. 2009, vol. 7257, p. 72570O. doi: 10.1117/12.810546.
[67] “OpenCV: Image Thresholding.” https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.html (accessed Jun. 16, 2022).
[68] J. R. Tillaguango Jiménez, “Revisión Sistemática de Literatura: Análisis de viabilidad para la detección y diagnóstico de Covid-19, aplicando modelos de Inteligencia Artificial (IA),” *CEDAMAZ*, vol. 11, no. 2, pp. 142–151, Dec. 2021, doi: 10.54753/cedamaz.v11i2.1183.
[69] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering - A systematic literature review,” *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.
[70] Nghia Ho, “Simple video stabilization using OpenCV | Nghia Ho.”

http://nghiaho.com/?p=2093 (accessed Jun. 22, 2022).
[71] J. Kaur and A. K. Bathla, “Video stabilization for an aerial surveillance system using sift and surf,” in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Oct. 2016, pp. 742–747. doi: 10.1109/NGCT.2016.7877509.
[72] M. Sharif, S. Khan, T. Saba, M. Raza, and A. Rehman, “Improved Video Stabilization using SIFT-Log Polar Technique for Unmanned Aerial Vehicles,” 2019. Accessed: Jun. 09, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/8716427/
[73] A. Arsalan, M. Majid, and S. M. Anwar, “Panoramic video stabilization based on rotational damping filter,” 2016. Accessed: Jun. 09, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/7429891/
[74] C. Liu, X. Li, and M. Wu, “Video Stabilization Algorithm Based on Kalman Filter and Homography Transformation,” *Lect. Notes Data Eng. Commun. Technol.*, vol. 6, pp. 308–313, 2018, doi: 10.1007/978-3-319-59463-7_31.
[75] J. Li, T. Xu, and K. Zhang, “Real-Time Feature-Based Video Stabilization on FPGA,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 907–919, 2017, doi: 10.1109/TCSVT.2016.2515238.
[76] Z. Mm, S. Jian, S. Dihua, and T. Yi, “Fast tunnel monitoring video stabilization method based on improved ORB feature,” in *2018 Chinese Control And Decision Conference (CCDC)*, 2018, pp. 5856–5861. doi: 10.1109/CCDC.2018.8408155.
[77] “OpenCV: Feature Matching.” https://docs.opencv.org/4.x/dc/dc3/tutorial_py_matcher.html (accessed Jun. 17, 2022).
[78] “3 Desarrollo 3.1 Vector de posición”.
[79] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” *Proc. - 2016 4th Int. Conf. 3D Vision, 3DV 2016*, pp. 239–248, Dec. 2016, doi: 10.1109/3DV.2016.32.
[80] Y. Cao, Z. Wu, and C. Shen, “Estimating Depth From Monocular Images as Classification Using Deep Fully Convolutional Residual Networks,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3174–3182, Nov. 2018, doi: 10.1109/TCSVT.2017.2740321.
[81] Y. Kim, H. Jung, D. Min, and K. Sohn, “Deep Monocular Depth Estimation via Integration of Global and Local Predictions,” *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4131–4144, 2018, doi: 10.1109/TIP.2018.2836318.
[82] S. F. Bhat, I. Alhashim, and P. Wonka, “AdaBins: Depth Estimation Using Adaptive Bins,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4008–4017, 2021, doi: 10.1109/CVPR46437.2021.00400.
[83] P. Hambarde and S. Murala, “S2DNet: Depth Estimation From Single Image and Sparse Samples,” *IEEE Trans. Comput. Imaging*, vol. 6, pp. 806–817, 2020, doi: 10.1109/TCI.2020.2981761.
[84] H. Amini Amirkolaee and H. Arefi, “Monocular depth estimation with geometrical guidance using a multi-level convolutional neural network,” *Appl. Soft Comput. J.*, vol. 84, Nov. 2019, doi: 10.1016/J.ASOC.2019.105714.
[85] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, “Joint task-recursive

learning for semantic segmentation and depth estimation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11214 LNCS, pp. 238–255, 2018, doi: 10.1007/978-3-030-01249-6_15/COVER/.

[86] S. M. Shankaranarayana, K. Ram, K. Mitra, and M. Sivaprakasam, "Fully Convolutional Networks for Monocular Retinal Depth Estimation and Optic Disc-Cup Segmentation," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 4, pp. 1417–1426, 2019, doi: 10.1109/JBHI.2019.2899403.

[87] J. Fu, J. Liang, and Z. Wang, "Monocular Depth Estimation Based on Multi-Scale Graph Convolution Networks," *IEEE Access*, vol. 8, pp. 997–1009, 2020, doi: 10.1109/ACCESS.2019.2961606.

[88] C. Zhou, Y. Liu, Q. Sun, and P. Lasang, "Joint Object Detection and Depth Estimation in Multiplexed Image," *IEEE Access*, vol. 7, pp. 123107–123115, 2019, doi: 10.1109/ACCESS.2019.2936126.

[89] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022, doi: 10.1109/TPAMI.2020.3019967.

[90] J. Liu, Q. Li, R. Cao, W. Tang, and G. Qiu, "A contextual conditional random field network for monocular depth estimation," *Image Vis. Comput.*, vol. 98, Jun. 2020, doi: 10.1016/J.IMAVIS.2020.103922.

[91] A. Mathew and J. Mathew, "Monocular depth estimation with SPN loss," *Image Vis. Comput.*, vol. 100, Aug. 2020, doi: 10.1016/J.IMAVIS.2020.103934.

[92] A. Zhang and J. Sun, "Joint Depth and Defocus Estimation From a Single Image Using Physical Consistency," *IEEE Trans. Image Process.*, vol. 30, pp. 3419–3433, 2021, doi: 10.1109/TIP.2021.3061901.

[93] H. Choi, "Self-Supervised Monocular Depth Estimation With Extensive Pretraining," *IEEE Access*, vol. 9, pp. 157236–157246, 2021, doi: 10.1109/ACCESS.2021.3129628.

[94] Y. Lu *et al.*, "Taskology: Utilizing Task Relations at Scale".

[95] P. Y. Chen, A. H. Liu, Y. C. Liu, and Y. C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 2619–2627, Jun. 2019, doi: 10.1109/CVPR.2019.00273.

[96] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Jun. 2016, doi: 10.1109/CVPR.2016.90.

[97] S. Wang, X. Xia, L. Ye, and B. Yang, "Automatic detection and classification of steel surface defect using deep convolutional neural networks," *Metals (Basel).*, vol. 11, no. 3, pp. 1–23, Mar. 2021, doi: 10.3390/met11030388.

[98] Paperwithcode, "ResNet Explained | Papers With Code." https://paperswithcode.com/method/resnet (accessed Jun. 23, 2022).

[99] "Video DataBase." http://liushuaicheng.org/SIGGRAPH2013/database.html (accessed Jun. 23, 2022).

[100] F. A. Khan *et al.*, "Computer-aided diagnosis for burnt skin images using deep

convolutional neural network," *Multimed. Tools Appl.*, vol. 79, no. 45–46, pp. 34545–34568, Dec. 2020, doi: 10.1007/S11042-020-08768-Y.
[101] W. G. Aguilar and C. Angulo, "Real-Time Model-Based Video Stabilization for Microaerial Vehicles," *Neural Process. Lett.*, vol. 43, no. 2, pp. 459–477, Apr. 2016, doi: 10.1007/s11063-015-9439-0.