INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN Y
DESARROLLO DE TECNOLOGÍA DIGITAL


**DOCTORADO EN CIENCIAS EN SISTEMAS DIGITALES**

# RECONOCIMIENTO DE ROSTRO Y EXPRESIONES UTILIZANDO VISIÓN MULTI-OCULAR

**TESIS**

QUE PARA OBTENER EL GRADO DE
**DOCTORADO EN CIENCIAS EN SISTEMAS DIGITALES**

PRESENTA

**M. C. MARTIN GONZALEZ RUIZ**

BAJO LA DIRECCIÓN DE

**DR. VÍCTOR HUGO DÍAZ RAMÍREZ**
**DR. RIGOBERTO JUÁREZ SALAZAR**

**DICIEMBRE 2024** **TIJUANA, B.C., MÉXICO**

SIP-13
REP 2017

# INSTITUTO POLITÉCNICO NACIONAL
## SECRETARIA DE INVESTIGACIÓN Y POSGRADO

*ACTA DE REGISTRO DE TEMA DE TESIS*
*Y DESIGNACIÓN DE DIRECTOR DE TESIS*

Ciudad de México 08 de julio del 2021

El Colegio de Profesores de Posgrado de Centro de Investigación y Desarrollo de Tecnología Digital en su Sesión
(Unidad Académica)

Ordinaria No. 7 celebrada el día 08 del mes julio de 2021 conoció la solicitud presentada por el (la) alumno (a):

| Apellido Paterno: | Gonzalez | Apellido Materno: | Ruiz | Nombre (s): | Martin |
|---|---|---|---|---|---|

Número de registro: A 2 1 0 1 9 9

del Programa Académico de Posgrado: DCSD - Doctorado en Ciencias en Sistemas Digitales

Referente al registro de su tema de tesis; acordando lo siguiente:

1.- Se designa al aspirante el tema de tesis titulado:

Reconocimiento de rostro y expresiones utilizando visión multi-ocular

Objetivo general del trabajo de tesis:

Desarrollar un sistema opto-digital multi-ocular para el reconocimiento confiable del rostro y expresiones con tolerancia a perturbaciones.

2.- Se designa como Directores de Tesis a los profesores:

Director: Dr. Víctor Hugo Díaz Ramírez 2° Director: Dr. Rigoberto Juárez Salazar

No aplica: ☐

3.- El Trabajo de investigación base para el desarrollo de la tesis será elaborado por el alumno en:

Centro de Investigación y Desarrollo de Tecnología Digital

que cuenta con los recursos e infraestructura necesarios.

4.- El interesado deberá asistir a los seminarios desarrollados en el área de adscripción del trabajo desde la fecha en que se suscribe la presente, hasta la aprobación de la versión completa de la tesis por parte de la Comisión Revisora correspondiente.

Director(a) de Tesis

Dr. Víctor Hugo Díaz Ramírez

Aspirante

M. en C. Martin Gonzalez Ruiz

2° Director de Tesis (en su caso)

Dr. Rigoberto Juárez Salazar

Presidente del Colegio

Dr. Julio César Rolón Garrido

S.E.P.
INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN Y
DESARROLLO DE TECNOLOGÍA
DIGITAL
DIRECCIÓN

SIP-14
REP 2017

# INSTITUTO POLITÉCNICO NACIONAL

## SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

### Dirección de Posgrado

*ACTA DE REVISIÓN DE TESIS*

En la Ciudad de Tijuana siendo las 14:00 horas del día 11 del mes de diciembre del 2024 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Posgrado de: Centro de Investigación y Desarrollo de Tecnología Digital para examinar la tesis titulada:

Reconocimiento de rostro y expresiones utilizando visión multi-ocular

del (la) alumno (a):

| Apellido Paterno: | Gonzalez | Apellido Materno: | Ruiz | Nombre (s): | Martin |
|---|---|---|---|---|---|

Número de boleta: A 2 1 0 1 9 9

Alumno del Programa Académico de Posgrado: DCSD - Doctorado en Ciencias en Sistemas Digitales

Una vez que se realizó un análisis de similitud de texto, utilizando el software antiplagio, se encontró que el trabajo de tesis tiene __25__ % de similitud. **Se adjunta reporte de software utilizado.**

Después que esta Comisión revisó exhaustivamente el contenido, estructura, intención y ubicación de los textos de la tesis identificados como coincidentes con otros documentos, concluyó que en el presente trabajo **SI** [ ] **NO** [x] **SE CONSTITUYE UN POSIBLE PLAGIO.**

**JUSTIFICACIÓN DE LA CONCLUSIÓN:** *(Por ejemplo, el % de similitud se localiza en metodologías adecuadamente referidas a fuente original)*

El reporte de similitud elaborado por el software Turnitin arrojó un porcentaje de similitud total del 25%. De este porcentaje, el 20% corresponde a fuentes de internet, el 14% corresponde a publicaciones y el 17% corresponde a trabajos entregados del estudiante. El reporte enumera 300 fuentes con el mayor número de coincidencias, sin embargo, ninguna de ellas supera el 1%, lo que indica que las coincidencias encontradas corresponden a palabras comunes utilizadas frecuentemente en trabajos relacionados. Después de analizar detalladamente el reporte de similitud, se concluye que el trabajo de tesis elaborado por el estudiante es original y corresponde al trabajo de investigación desarrollado.

Finalmente y posterior a la lectura, revisión individual, así como el análisis e intercambio de opiniones, los miembros de la Comisión manifestaron **APROBAR** [x] **SUSPENDER** [ ] **NO APROBAR** [ ] la tesis por **UNANIMIDAD** [x] o **MAYORÍA** [ ] en virtud de los motivos siguientes:

El trabajo de tesis cumplió al 100% con el objetivo de investigación. Además, se cumplieron con la totalidad de los requisitos del programa de posgrado. Finalmente, el estudiante mostró un desempeño académico sobresaliente y una productividad científica destacada.

**COMISIÓN REVISORA DE TESIS**

Director de Tesis
Dr. Víctor Hugo Díaz Ramírez
Nombre completo y firma

Director de Tesis
Dr. Rigoberto Juárez Salazar
Nombre completo y firma

Dr. Eduardo Javier Moreno Valenzuela
Nombre completo y firma

Dr. Ricardo Ramón Pérez Alcocer
Nombre completo y firma

Dr. Luis Tupak Aguilar Bustos
Nombre completo y firma

Dr. Julio César Rolón Garrido
Nombre completo y firma
**PRESIDENTE DEL COLEGIO DE PROFESORES**

S E P
INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN Y DESARROLLO DE TECNOLOGÍA DIGITAL
DIRECCIÓN

# INSTITUTO POLITÉCNICO NACIONAL
## SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

*CARTA DE AUTORIZACIÓN DE USO DE OBRA PARA DIFUSIÓN*

En la Ciudad de México el día _11_ del mes de diciembre del año 2024, el que suscribe Martin Gonzalez Ruiz alumno del programa DCSD-Doctorado en Ciencias en Sistemas Digitales con número de registro A210199, adscrito al Centro de Investigación y Desarrollo de Tecnología Digital manifiesta que es autor intelectual del presente trabajo de tesis bajo la dirección del Dr. Víctor Hugo Diaz Ramírez y del Dr. Rigoberto Juárez Salazar y cede los derechos del trabajo titulado Reconocimiento de rostro y expresiones utilizando visión multi-ocular, al Instituto Politécnico Nacional, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expresado del autor y/o directores. Este puede ser obtenido escribiendo a las siguientes direcciones de correo. mgonzalezr1800@alumno.ipn.mx, vdiazr@ipn.mx, y rjuarezsa@conahcyt.mx. Si el permiso se otorga, al usuario deberá dar agradecimiento correspondiente y citar la fuente de este.

Martin Gonzalez Ruiz

Nombre completo y firma autografá del (de la) estudiante

*Ut est rerum omnium magister usus*
***Gaius Iulius Caesar***

# Dedicatoria

*A mi madre y hermanos*
*A Anayatzint my sunshine*
Sin ustedes esto no habría sido posible

# Agradecimientos

Agradezco de manera sincera al Dr. Víctor Hugo Díaz Ramírez por brindarme la oportunidad de formar parte de su equipo de investigación, compartir sus conocimientos, por la confianza brindada y ser un gran mentor a lo largo de este proceso de formación.

Mi profunda gratitud al Dr. Rigoberto Juarez Salazar por la atención brindada y su constante asesoría durante el desarrollo de este trabajo de investigación.

Agradezco a los miembros del comité tutorial Dr. Eduardo Javier Moreno Valenzuela, Dr. Ricardo Ramón Peréz Alcocer y Dr. Luis Tupak Aguilar Bustos por sus observaciones y aportaciones realizadas en este trabajo.

Un agradecimiento especial a los doctores Miguel Cazorla, Francisco Gómez, Félix Escalona, Germán González y miembros del grupo de investigación *RoViT* de la Universidad de Alicante, por recibirme, acompañarme y aconsejarme durante mi proceso de movilidad de investigación.

Adicionalmente, me gustaría agradecer a las personas que de manera indirecta contribuyeron al desarrollo de esta investigación, a mis compañeros de laboratorio, personal administrativo, colaboradores y amigos.

Finalmente, extiendo un agradecimiento a las dependencias que financiaron a lo largo de estos años esta investigación, al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT), la Secretaría de Investigación y Posgrado (SIP), Secretaría de Innovación e Integración Social del Instituto Politénico Nacional (IPN), la Sociedad Internacional de Óptica y Fotónica (*SPIE*). Al Instituto Politécnico Nacional-Centro de Investigación y Desarrollo de Tecnología Digital (CITEDI) por haber proporcionado un espacio de trabajo y equipo necesario para el desarrollo de esta investigación.

# Reconocimiento de rostro y expresiones utilizando visión multi-ocular

## Resumen

En este trabajo se presenta un método para reconocimiento de expresiones del rostro, utilizando sistemas de visión multi-ocular. Primero, imágenes estéreo multi-oculares son rectificadas por medio de un método basado en el algoritmo de enjambre de partículas, el cual minimiza errores de distorsión y cumple con las restricciones epipolares. Las imágenes rectificadas son utilizadas para estimar la profundidad de la escena observada. La profundidad se determina a través de un método de asociacón estéreo multilínea base ajustable, utilizando la disparidad entre imágenes multi-oculares. Inicialmente, un mapa de disparidad es estimado entre la imagen de referencia y la imagen más cercana. El mapa obtenido es utilizado para predecir el mapa de disparidad ente la imagen de referencia y la siguiente imagen más cercana, este proceso se realiza de manera iterativa hasta llegar a la imagen más alejada. Posteriormente, el mapa de disparidad es postprocesado para remover información estimada errónemanete utilizando un enfoque de interpolación basado en la teoría de Bayes. Despúes, la disparidad procesada y los parámetros físicos del sistema son utilizados para encontrar la distribución geométrica de profundidad en la escena. Finalmente, esta información de profundidad en conjunto con la imagen de referencia son utilizadas como entradas de una arquitectura de redes neuronales convolucionales para extraer y clasificar características de expresiones del rostro. El desempeño del método propuesto para el reconocimiento de expresiones del rostro es evaluado en imágenes de base de datos utilizando medidas de desempeño objetivas.

**Palabras claves**: Visión multi-ocular, rectificación multi-ocular, visión estéreo, multiliínea base, reconocimiento de expresiones faciales, redes neuronales convolucionales.

# Face and expression recognition using multi-ocular vision

## Abstract

In this work, we present a method for facial expression recognition using multi-ocular vision systems. First, multi-ocular stereo images are rectified using the particle swarm algorithm, which minimizes distortion errors and meets the epipolar constraints. The rectified images are used to estimate the depth of the observed scene. The depth is estimated using the disparity between multi-ocular images through a multi-ocular stereo multi-ocular association method. Initially, a disparity map is estimated between the reference image and the closest image. The disparity map obtained is used to predict the disparity map between the reference image and the next closest image. The method is performed iteratively until the farthest image is reached. Subsequently, the disparity map is post-processed to remove erroneously estimated information using an interpolation approach based on a Bayesian function. Afterwards, the processed disparity and the physical system parameters are used to find the geometric depth distribution of the scene. Finally, the depth information and reference image are used as inputs to a convolutional neural network architecture to extract and classify features of face expressions. The performance of the proposed method for face expression recognition is evaluated on database images using database and objective performance measures.

**Keywords**: Multi-ocular vision, image rectificacion, stereo vision, multibaseline, facial expression recognition, convolutional neural network.

# Índice general

# Índice de figuras

# Índice de cuadros

# Lista de Acrónimos

**FER** *Facial Expression Recognition*

**FACS** *Facial Actions Coding System*

**AU** *Actions Unit*

**LBP** *Local Binary Patterns*

**LPQ** *Local Phase Quantization*

**HOG** *Histogram of Oriented Gradiemt*

**SIFT** *Scale Invariant Feature Transform*

**SVM** *Support Vector Machines*

**3D** Tridimensional

**2D** Bidimensional

**1D** Unidimensional

**SVD** *Singular Value Decomposition*

**CNN** *Convolutional Neural Networks*

**SAD** *Sum of Absolute Difference*

**SSD** *Sum of Squared Difference*

**NCC** *Normalized Cross Correlation*

**RT** *Rank Transform*

**CT** *Census Transform*

**ASW** *Adaptive Support Weight*

**LBF** *Local Binary Features*

**AAM** *Active Appearance Model*

**SDM** *Supervised Decent Method*

**Tanh** Tangente hiperbólica

**ReLu** Unidad Lineal Rectificada

**STROMI** *Stereo-Rectification for Optimized multi-ocular Images*

**PSO** *Particle Swarm Optimization*

**AMC** *Adaptive Morphological Correlation*

**BDMR** *Binary Dissimilarity-to-Matching Ratio*

**DHB** *Disparity Hole-filling unsing Bayesian function*

**BAMS** *Baseline Adjustable Multi-ocular Stereo*

**MCL-RS** *Media CommLab Real Stereo*

**RANSAC** *Randon Sample Consensus*

**Asp.Rat.** *Aspect Ratio*

**IWCT** *Inproved Weighted Census Transform*

**ADCT** *Absolute Difference Census Transform*

**SPR** *Stereo Phase rectification*

**MAE** *Mean Absolute Difference*

**MSE** *Mean Squared Error*

**PSNR** *Peak Signal-to-Noise Ratio*

**RMS** *Root Mean Square*

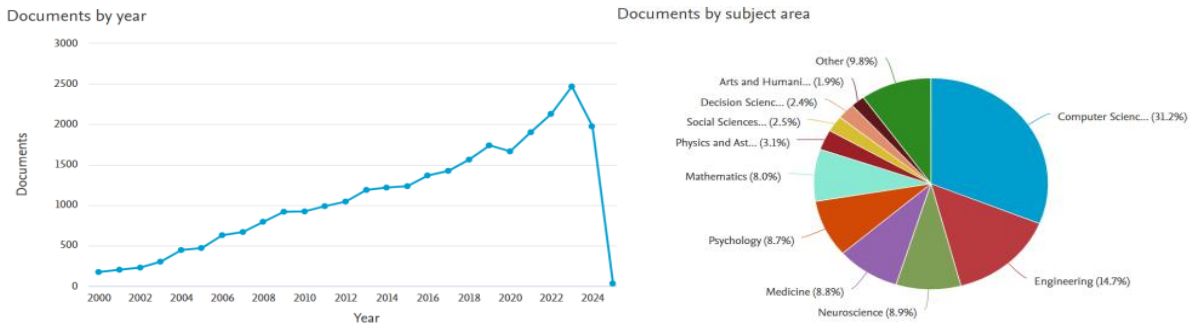
Capítulo 1

# Introducción

## 1.1. Introducción

La comunicación interpersonal es una de las actividades más esenciales entre humanos. Habitualmente, esta actividad se puede realizar de forma oral, escrita o visual, a través del uso de algún lenguaje. Las expresiones del rostro humano causadas por gesticulaciones durante un proceso de comunicación, permiten inferir el estado de ánimo de una persona [2]. El reconocimiento de expresiones del rostro humano, tiene gran utilidad en diversas aplicaciones tecnológicas en beneficio de la sociedad.

La investigación y desarrollo de técnicas para el reconocimiento de expresiones faciales tienen un gran impacto comercial, científico y social en diferentes disciplinas del conocimiento como las ciencias del comportamiento, neurología e inteligencia artificial. Para darnos una idea de la relevancia científica de este tema, en la Fig. 1.1.1 se muestra la distribución de documentos indexados en la plataforma *Scopus* desde el año 2000 hasta el 2024. Adicionalmente, esta figura presenta la distribucón por área de interés, destacando las ciencias de la computación, ingenería y neurociencias.




(*a*) (*b*)

**Figura 1.1.1:** Documentos registrados en la base de datos *Scopus* relacionados con FER. (a) Distribución de documentos por año. (b) Distribución por área de interés.

El reconocimiento, interpretación y procesamiento de emociones basadas en enfoques asistidos por computadora, pueden explotarse con gran éxito en aplicaciones tecnológicas de vanguardia, como la detección del nivel de concentración de estudiantes en sistemas de educación en línea [3–5], entretenimiento interactivo [6, 7], monitoreo médico de pacientes a distancia [8, 9], y operadores de sistemas de navegación aérea, entre muchos otras.

En el estado del arte del reconocimiento de emociones basado en expresiones del rostro, se identifican emociones universales como enojo, disgusto, miedo, felicidad, tristeza y sorpresa [10]. A través de los años, se han propuesto diferentes enfoques para el reconocimiento automático de expresiones faciales (FER, por sus siglas en inglés *Facial Expression Recognition*) utilizando sistemas de visión por computadora con el fin de detectar características faciales, y asignarlas de manera confiable en alguna categoría correspondiente de expresión facial. Estos sistemas toman en cuenta las emociones universales anteriormente mencionadas para clasificar la información procesada [11–13].

En la actualidad existe una gran cantidad de enfoques para reconocer secuencias de movimientos musculares faciales. Esto permite realizar una asociación precisa de movimientos musculares y expresiones emocionales [2]. El sistema de codificación de acción facial (FACS, por sus siglas en inglés *Facial Action Coding System*), presenta una clasificación estandarizada basada en la anatomía humana, para describir los movimientos musculares faciales visualmente distinguibles. En este sistema, las expresiones faciales se descomponen en componentes individuales denominadas unidades de acción (AU, por sus siglas en inglés *Action Units*). Para cada AU, existe un movimiento de los músculos faciales que generalmente coincide con regiones del rostro como contornos del ojo, contornos de la boca, cejas y nariz. Una expresión facial se compone de una secuencia de AUs. La Fig. 1.1.2 muestra ejemplos de la composición de distintas expresiones faciales con AUs. Las FACS permiten describir expresiones faciales de forma objetiva, por medio de la identificación de cambios visibles del tejido facial. El reconocimiento de expresiones faciales por medio de las FACS requiere contar con una alta calidad de video y un proceso de entrenamiento exhaustivo para la discriminación de acciones faciales.

Los métodos tradicionales para FER pueden dividirse en cuatro etapas, como se muestra en la Fig. 1.1.3. Estas etapas son adquisición de imágenes, preprocesamiento, extracción de características faciales y codificación automática de acción facial.

Durante la etapa de preprocesamiento, las imágenes de entrada se preparan para cumplir con los requerimientos necesarios para ser procesadas. Algunas tareas que se realizan en esta etapa son: localización del rostro [14, 15], localización de características faciales [16, 17], alineación de rostro, entre otros.

La etapa de extracción de características, tiene como objetivo convertir un conjunto de píxeles de la imagen de entrada en un nivel superior de representación numérica del movimiento, la apariencia o el espacio de estructuras faciales. Esta etapa es esencial para reducir la dimensión de los datos de entrada y minimizar las variaciones no deseadas ocasionadas por cambios de iluminación, desenfoque o errores de alineación. Los métodos de extracción de características, pueden clasificarse ampliamente en las siguientes dos categorías: métodos basados en sistemas prediseñados y métodos basados en sistemas de aprendizaje [13]. Los sistemas prediseñados son elaborados manualmente para extraer información de utilidad, de acuerdo con la experiencia del diseñador. Dentro de estos sistemas, podemos encontrar sistemas de extracción de características basados en apariencia, donde se describe el color o la textura de la región de interés. Por ejemplo,

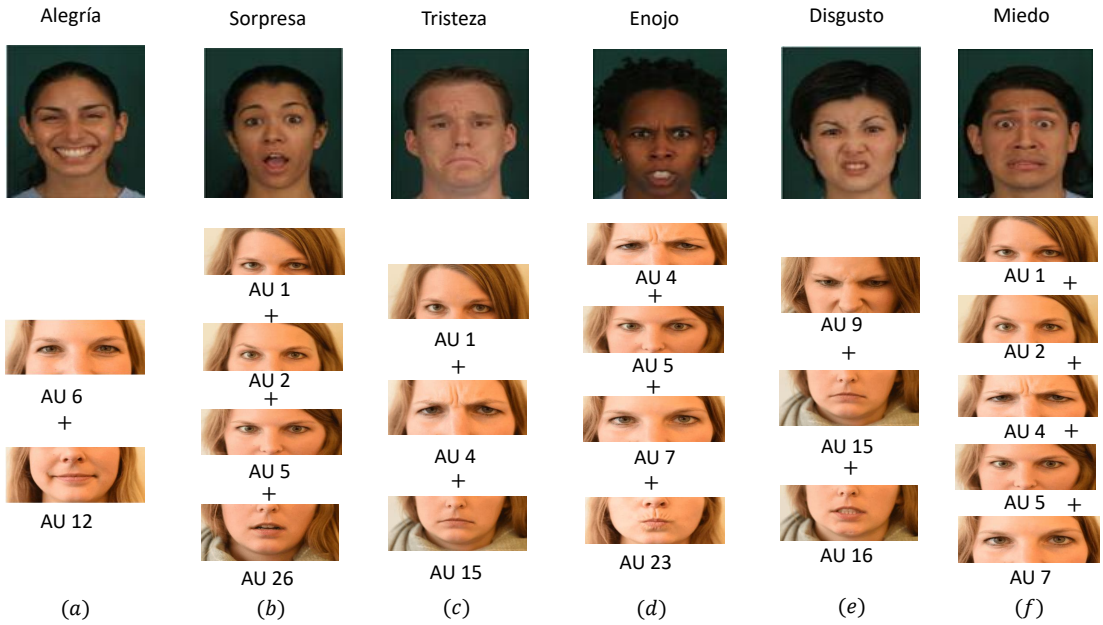



**Figura 1.1.2:** Composición de expresiones faciales con AU. (a) Expresión facial de alegría compuesta por las AU 6 y 12. (b) Expresión facial de sorpresa compuesta por las AU 1, 2, 5 y 26. (c) Expresión facial de tristeza compuesta por las AU 1, 4 y 15. (d) Expresión facial de enojo compuesto por las AU 4, 5, 7 y 23. (e) Expresión facial de disgusto compuesto por las AU 9, 15 y 16. (f) Expresión facial de miedo compuesto por las AU 1, 2, 4 ,7, 20 y 26.

en [18–20] se utilizan los niveles de intensidad de los píxeles sin procesar para describir las características del rostro. Sin embargo, este método es sensible a variaciones ocasionadas por ruido e iluminación. Una alternativa es el uso de bancos de filtros sobre las regiones de interés, para extraer las características necesarias. En la literatura científica, se reporta el uso de filtros basados en ondeletas de Gabor [21, 22], transformada de coseno [23] o características tipo Haar [24]. Una de las desventajas de estos métodos es que presentan un alto costo computacional. Una alternativa eficiente a estos problemas es el uso de técnicas de textura local, como patrones locales binarios (LBP, por sus siglas en inglés *Local Binary Patterns*) [25, 26], o cuantización local de fase (LPQ, por sus siglas en inglés *Local Phase Quantization*) [27]. Por otro lado, los descriptores basados en gradientes como histograma de gradiente orientado (HOG, por sus siglas en inglés *Histogram of Oriented Gradient*) [28], o transformación de características invariantes a escala (SIFT, por sus siglas en inglés *Scale-Invariant Feature Transform*) [29], permiten representar patrones locales con mayor invarianza a perturbaciones geométricas como rotaciones, escalamientos y deformaciones. Los sistemas de extracción de características geométricas realizan mediciones de distancias, deformaciones o curvaturas, como por ejemplo el movimiento de cejas o boca, los cuales generan una deformación facial [30, 31]. Las características geométricas son sencillas de registrar, además, son independientes de las condiciones de iluminación.

Por otro lado, el enfoque basado en técnicas de aprendizaje máquina ha sido ampliamente utilizado en años recientes para la extracción de características. Estas técnicas, requieren de una etapa de entrenamiento sin supervisión, que permite el manejo de una mayor cantidad de datos [32]. El desempeño de un sistema FER basado en aprendiza-

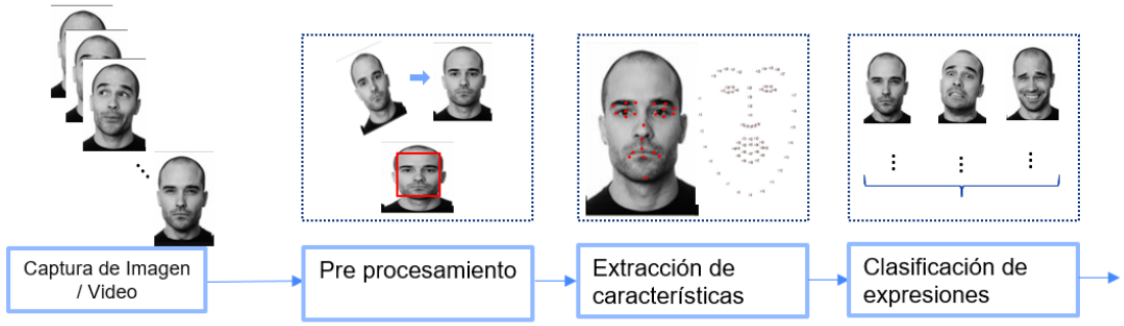



**Figura 1.1.3:** Etapas de un sistema de reconocimiento facial.

je automático, depende en gran medida de la disponibilidad de grandes bases de datos durante la etapa de entrenamiento.

La clasificación es la última etapa de los sistemas FER. En esta etapa, las emociones o estados de ánimo del sujeto de prueba son inferidas al asignar a una categoría correspondiente el conjunto de características extraídas. Las emociones se representan como puntos en un espacio de características, donde cada dimensión corresponde de un rasgo expresivo. Con ello, se pueden definir emociones sin supervisión y discriminar expresiones con diferencias sutiles. La mayoría de estos sistemas FER utilizan un modelo de múltiples clases tomando como referencia las emociones descritas en [10]. Algunas de las técnicas mayormente utilizadas para la clasificación de patrones son máquinas de soporte vectorial (SVM, por sus siglas en inglés *Support Vector Machines*) [33, 34], bosques al azar [35], vecinos cercanos [36], k-vecinos cercanos [37] o redes neuronales [38].

Durante la década pasada, los sistemas FER hacían uso de información bidimensional ($2D$) [39]. Varios métodos exitosos fueron reportados en la literatura científica; sin embargo, pueden presentar limitaciones en la presencia de cambios de escala, rotaciones fuera de plano, e iluminación no homogénea. La estimación de expresiones faciales es más precisa usando información $3D$ que $2D$. Por lo tanto, los sistemas basados en información $3D$ se han convertido en un área atractiva de investigación. La adquisición de información tridimensional del rostro requiere de un sistema de visión multi-ocular especializado que puede ser activo o pasivo [40]. La precisión, velocidad y densidad de la adquisición de información es crucial para el funcionamiento del sistema. Numerosas técnicas para la captura de información tridimensional del rostro han sido exploradas durante décadas, como escaneo 3D, captura de movimiento basada en marcadores, sistemas de luz estructurada o sistemas basados en imágenes estéreo [41]. A pesar de estos avances, adquirir representaciones faciales con alta fidelidad siguen presentando retos importantes. Un ejemplo consiste en sistemas de escaneo 3D donde se puede obtener una alta resolución facial, pero solo en poses estáticas. Los sistemas de captura basada en marcadores, permite capturar movimientos dinámicos faciales. Sin embargo, fallan en el registro de detalles faciales. En la literatura científica se reporta progreso en métodos basados en visión multi-ocular, donde es posible capturar expresiones faciales dinámicas 3D con alta fidelidad, resolución y consistencia. Sin embargo, aún es necesario desarrollar más investigación para mejorar el desempeño de estos sistemas, en términos de efectividad, robustez o velocidad [42].

Adicionalmente, a pesar de la existencia de diferentes técnicas efectivas para la construcción de sistemas FER, aún existen importantes retos a resolver, como la respuesta a variaciones de la posición de la cabeza, cambios de iluminación, oclusiones parciales,

presencia de fondo en la imagen, entre otros.

En este trabajo de tesis, se propone el desarrollo de un sistema opto-digital multi-ocular para el reconocimiento confiable del rostro y expresiones con tolerancia a perturbaciones. Se pueden identificar dos aspectos principales, la captura de imágenes tridimensionales del rostro a través de un sistema de visión multi-ocular, y el desarrollo de un algoritmo para el reconocimiento facial y de expresiones en imágenes 3D. Se contempla el planteamiento de modelos del rostro y expresiones con el fin de desarrollar una solución más precisa, en relación con los métodos existentes. Finalmente, se considera la exploración de herramientas de aprendizaje automático para tareas de alta generalización, como la extracción de rasgos y clasificación.

## 1.2. Motivación

El desarrollo de sistemas avanzados de visión computacional multi-ocular en combinación con técnicas opto-digitales para el procesamiento de imágenes y reconocimiento de patrones coadyuvan al avance científico y tecnológico en temas de vanguardia como seguridad, interacción humano-computadora, entretenimiento, educación a distancia, terapia médica, monitoreo de enfermedades neurodegenerativas y realidad aumentada.

## 1.3. Hipótesis

Es posible desarrollar un método confiable y eficiente para el reconocimiento del rostro y expresiones faciales basado en un sistema de visión multi-ocular, modelos tridimensionales del rostro y herramientas computacionales avanzadas.

## 1.4. Objetivo

El objetivo general es desarrollar un sistema opto-digital multi-ocular para el reconocimiento confiable del rostro y expresiones con tolerancia a perturbaciones.

Los objetivos específicos de esta tesis se enlistan a continuación.

- Desarrollar un método para la digitalización del rostro basado en visión multi-ocular.
- Desarrollar un método para el reconocimiento confiable del rostro y expresiones con tolerancia a perturbaciones utilizando técnicas opto-digitales.
- Evaluar el desempeño del sistema desarrollado y comparar su desempeño con métodos existentes a través de medidas objetivas.

## 1.5. Contribuciones científicas

Como resultado del presente trabajo de tesis, se han desarrollado diferentes aportaciones científicas originales para resolver problemas de clasificación de imágenes y reconstrucción tridimensional, derivando en los siguientes productos de investigación.

### Artículos sometidos a publicación

- Diaz-Ramirez, V. H., Gonzalez-Ruiz, M., Juarez-Salazar, R., & Cazorla, M.(2024, Noviembre). Reliable disparity estimation using multiocular vision with adjustable baseline. Sensors

### Artículos publicados

- Diaz-Ramirez, V. H., Gonzalez-Ruiz, M., Kober, V., & Juarez-Salazar, R. (2022). Stereo image matching using adaptive morphological correlation. Sensors, 22(23), 9050.
  DOI: https://doi.org/10.3390/s22239050

- Diaz-Ramirez, V. H., Juarez-Salazar, R., Gonzalez-Ruiz, M., & Adeyemi, V. A. (2023). Restoration of Binocular Images Degraded by Optical Scattering through Estimation of Atmospheric Coefficients. Sensors, 23(21), 8918.
  DOI: https://doi.org/10.3390/s23218918

### Memorias de congreso internacional

- Gonzalez-Ruiz, M., Diaz-Ramirez, V. H., & Juarez-Salazar, R.(2022, October). Multi-baseline stereo vision for three-dimensional object reconstruction. In Optics and Photonics for Information Processing XVI (Vol. 12225, pp. 55-62). SPIE.
  DOI: https://doi.org/10.1117/12.2633119.

- Diaz-Ramirez, V. H., Gonzalez-Ruiz, M., & Juarez-Salazar, R. (2022, October). Camera pose estimation based on local image correlation. In Optics and Photonics for Information Processing XVI (Vol. 12225, pp. 95-100). SPIE.
  DOI: https://doi.org/10.1117/12.2633732

- Brito-Muñoz, Y., Diaz-Ramirez, V. H., Gonzalez-Ruiz, M., & Juarez-Salazar, R. (2023, October). Performance evaluation of facial landmark detection methods. In Optics and Photonics for Information Processing XVII (Vol. 12673, pp. 24-30). SPIE.
  DOI: https://doi.org/10.1117/12.2677146

- Gonzalez-Ruiz, M., Diaz-Ramirez, V. H., & Juarez-Salazar, R. (2023, October). Three-dimensional object reconstruction using multi-ocular vision. In Optics and Photonics for Information Processing XVII (Vol. 12673, pp. 79-86). SPIE.
  DOI: https://doi.org/10.1117/12.2676947

- Diaz-Ramirez, V. H., Gonzalez-Ruiz, M., & Juarez-Salazar, R. (2024, September). Binocular vision-based depth estimation in scattering media. In Optics and Photonics for Information Processing XVIII (Vol. 13136, pp. 59-65). SPIE.
  DOI: https://doi.org/10.1117/12.3028316

- Gonzalez-Ruiz, M., Díaz-Ramírez, V. H., Cazorla, M., & Juarez-Salazar, R. (2024, September). A comparative study of facial feature classification methods. In Optics and Photonics for Information Processing XVIII (Vol. 13136, pp. 116-124). SPIE.
  DOI: https://doi.org/10.1117/12.3027562

**Otras actividades relevantes de formación profesional**

- SPIE Optics + Photonics Student Conference Support(2023)
- Proyecto con financiamiento "Digitalización multi-ocular del rostro", derivado de la convocatoria *Proyectos Desarrollo tecnólogico e inovación para alumnos del IPN 2022*, Secretaría de Investigación y Posgrado del Instituto Politécnico Nacional.
- Estancia de investigación en el grupo de investigación *Robotic & Tridimensional vision Research Group* de la Universidad de Alicante (UA), Alicante, España.

**Retribución Académica**

- Miembro revisor de revista JCR *Engineering Research Express*, *Measurement science and technology* de la editorial *IOP Science*, *Journal of Electronic Imaging*, *Optical Engineering*, de la editorial SPIE y *Remote sensing* de la editorial MDPI.
- Conferencia regional "*Reconocimiento de expresiones del rostro para educación de calidad*", impartida en el Seminario de Primavera sobre Sistemas Digitales (2022).
- Conferencia nacional "*Sistemas de visión multi-cámara: ventajas y aplicaciones*", impartida en el Seminario de Divulgación General de la Ciencia: Sistemas de Visión Digital 3D y Procesamiento Opto-digital (2022).
- Taller "*¿Cómo pueden ver las computadoras?"* impartido en la Semana Nacional del Conocimiento, Museo interactivo el trompo (2023).
- Promoción de los programas de posgrado de CITEDI-PN en los eventos de divulgación *Expo Profesiográfica* y conferencia *Procesamiento de imágenes y visión por computadora: oportunidades de investigación en el CITEDI* (2023).
- Conferencia nacional "*Digitalización facial basada en visión multi-ocular*", impartida en el Seminario de Divulgación de Ciencia Básica en Sistemas Opto-digitales de Visión 3D (2023).
- Taller de preparación para categoría Universitaria en la modalidad "*FinderBot*", impartido en el Museo interactivo el trompo (2024).
- Participación en *International Workshop MEEBAI*, organizado por el Proyecto PROMETEO MEEBAI y la Cátedra UNESCO de Educación, Investigación e Inclusión Digital, en la Universidad de Alicante (2024).
- Miembro revisor de *2025 IEEE Symposium on Computational Intelligence in Image, Signal Processing and Synthetic Media* organizado por IEEE.
- Conferencia internacional "*Reconstrucción tridimensional basada en visión multi-ocular*", impartidaen en la feria de investigación de Ciencia y Tecnología, Universidad Popular del César, Colombia (2024).

## 1.6. Organización de la tesis

El presente documento está organizado de la siguiente manera. En el capítulo 2, se presentan los fundamentos teóricos utilizados en esta investigación. Primero, se aborda la teoría de geometría epipolar, matriz fundamental, rectificación de imágenes estéreo, estimación de profundidad, y visión estéreo multilínea base. Estos conceptos, son fundamentales para el proceso de reconstrucción tridimensional utilizando sistemas de visión estéreo. Adicionalmente, se describen brevemente diferentes métodos reportados en la literatura para estimación de mapas de disparidad. Finalmente, se describen algunos métodos de extracción de puntos característicos del rostro y las etapas de la arquitectura de redes neuronales convolucionales para el problema de clasificación.

En el capítulo 3, se presentan dos aportaciones principales. La primer aportación consiste en un sistema de reconstrucción tridimensional del rostro basado en visión multi-ocular. Para desarrollar esta tarea, se propone un método novedoso de rectificación estéreo multi-ocular. El método propuesto está diseñado para rectificar $n$ imágenes estéreo, optimizando una función objetivo que minimiza los errores de rectificación y la distorsión inducida. También, se propone un nuevo método de estimación de mapas de disparidad basado en correlación morfológica adaptativa, minimizando un nuevo críterio formulado denominado relación binaria de disimilaridad y coincidencia. Adicionalmente, se desarrolló un método de post-procesamiento para mapas de disparidad, diseñado específicamente para recuperar información en zonas ocluidas o con estimaciones erróneas. Este método, utiliza una interpolación basada en la similitud de los valores de intensidad de la imágen de la escena y su proximidad espacial. Por último, se extiende el método de estimación de disparidad binocular a un sistema multi-ocular utilizando un rango de búsqueda optimizado basado en la relación de separación de cámaras. Los límites de este rango de búsqueda, se establecen dinámicamente considerando la separación entre cámaras y la dispersión de disparidad esperada, en cada punto de la escena. Posteriormente, la reconstrucción tridimensional se obtiene por medio del método lineal homogéneo, que requiere el mapa de disparidad estimado de las imágenes multi-oculares y las matrices de parámetros de las cámaras. La segunda aportación de este trabajo de tesis consiste en un clasificador de expresiones del rostro utilizando arquitecturas de redes neuronales convolucionales. La red neuronal convolucional recibe dos entradas de información: imágenes de intensidad y distribución de profundidad del rostro. Para procesar la información de profundidad se genera una máscara de sorporte basada en descomposición binaria de los puntos característicos del rostro. Los tensores resultantes son concatenados en un solo vector para ser clasificado.

En el capítulo 4, se presentan los resultados obtenidos de la implementación de los métodos propuestos. Adicionalmente, para validar el desempeño del sistema propuesto se realizó una comparación con métodos reportados en la literatura utilizando medidas de desempeño objetivas.

Finalmente, en el capítulo 5 se presentan las conclusiones de esta investigación y las áreas de oportunidad para generalizar y robustecer el trabajo de investigación futuro.

Capítulo 2

# Marco teórico

En este capítulo se presentan los fundamentos teóricos para planteamiento y solución del problema de reconocimiento de expresiones del rostro utilizando sistemas de visión multi-ocular. Se presenta un estudio de las técnicas relevantes de clasificación de expresiones del rostro. Finalmente, se presentan conceptos generales de visión multi-ocular y reconstrucción tridimensional utilizando visión estéreo.

## 2.1. Planteamiento del problema

Los métodos FER pueden ser divididos de acuerdo a la manera de extracción de características del rostro; estos pueden ser métodos convencionales y métodos basados en aprendizaje profundo [43]. Los métodos convencionales dependen en menor medida de grandes cantidades de datos y *hardware* especializado. Sin embargo, al requerir diseño por separado de la etapa de extracción y clasificación, estos modelos no pueden ser optimizados de manera simultánea. Son dependientes del desempeño de cada componente individual para presentar una respuesta robusta. Los métodos basados en aprendizaje profundo reducen su dependencia de la etapa de extracción de características, debido a la integración de la etapa de clasificación [44]. Estos enfoques permiten una predicción directa de cada emoción, presentando respuestas más robustas ante imágenes con diferentes elementos. Sin embargo, requeiren una cantidad basta de imágenes etiquetadas para el entrenamiento y son susceptibles a sobre ajustar el modelo de clasificación.

La etapa de extracción de características en un sistema de visión para el reconocimiento de patrones, permite identificar información visual relevante donde la localización espacial se encuentra definida. Esta información se puede asociar a un vecindario de píxeles o a la búsqueda de una estructura específica en la imagen, como esquinas, regiones o formas [12]. La etapa de extracción de características faciales juega un rol de gran relevancia en el desempeño del sistema de reconocimiento. Una extracción deficiente de las características de la imagen influye significativamente en la obtención de resultados erróneos; incluso con el uso de excelentes clasificadores. Los métodos basados en extracción de características faciales tienen como objetivo utilizar regiones de la imagen del rostro para sustraer información de cambios sutiles como arrugas o surcos de las cejas, ojos, nariz y boca [45].

La etapa de clasificación es un componente esencial en los sistemas FER. En esta

etapa se realiza el análisis de las gesticulaciones del rostro para asignarlas a una categoría particular correspondiente a una emoción. La mayoría de sistemas FER utilizan modelos de clasificación multiclase para identificar la expresión dentro de un conjunto establecido.

Las redes neuronales convolucionales (CNN, por sus siglas en inglés *Convolutional Neural Networks*) [46] son una arquitectura de aprendizaje profundo ampliamente utilizada para tareas de extracción de características y clasificación. La robustez de redes convolucionales en la tarea de clasificación es elevada siempre y cuando los datos de prueba sean similares a los datos de entrenamiento. Sin embargo, la tarea de clasificación se vuelve más compleja cuando las imágenes presentan cambios en la pose. El uso de la información tridimensional del rostro proporciona mayor robustez ante los cambios de pose e iluminación [47].

Los sistemas de reconocimiento de expresiones del rostro $2D$ pueden presentar problemas de precisión cuando existen variaciones en las expresiones faciales, posiciones de la cabeza, oclusiones, entre otros factores [48]. Por lo tanto, el uso de información $3D$ es conveniente en estos sistemas para robustecer la respuesta ante micro-expresiones o variaciones de la estructura facial [43].

En los últimos años, dispositivos basados en proyección de luz estructurada han sido utilizados para recuperar la información tridimensional de una escena. Los sistemas de luz estructurada, estan conformados por un arreglo cámara-proyector [49]. Un factor clave en los métodos de luz estructurada son el diseño del patrón a proyectar, así como el número de patrones necesarios. Esto determina la velocidad, resolución y precisión del sistema [50]. Los sistemas de proyección de patrones sinusoidales son altamente precisos, pero requieren un mínimo de tres patrones de franjas para medir superficies complejas. Para implementaciones con imágenes dinámicas, una solución es implementar el método de perfilometría por transformada de Fourier [51]; sin embargo, este enfoque es muy sensible al ruido y a las variaciones de la textura; aspectos comunes presentes en imágenes del rostro.

Dispositivos basados en múltiples sensores han sido utilizados para recuperar la información tridimensional de una escena. Un ejemplo de este tipo de sensor es el *Microsoft Kinect*. Este dispositivo es fácil de operar y presenta buenos resultados en términos generales. Sin embargo, está sujeto a problemas sistemáticos como ruido o ambigüedad, relacionados con el sensor utilizado. Estos dispositivos tienen un desempeño satisfactorio a una distancia aproximada de $5-7$ metros, pero son sensibles en ambientes externos [52].

Las técnicas de visión estéreo permiten realizar mediciones en todas las áreas de la escena observada sin realizar un proceso de escaneo; por este motivo, estos métodos son adecuados para aplicaciones de tiempo real. Los sistemas de visión estéreo pueden ser más apropiados para el problema de reconocimiento de expresiones del rostro, ya que permiten obtener reconstrucciones $3D$ con buena resolución a partir de una sola toma, y pueden ser utilizados en ambientes internos y externos [53].

## 2.2. Visión multi-ocular

El uso de múltiples cámaras proporcionan información que no puede ser observada por una sola cámara. Un sistema de visión multi-ocular permite recuperar la información tridimensional de una escena a partir de dos o más imágenes estéreo [54]. Los sistemas de visión estéreo representan una alternativa atractiva para reconstruir la información

tridimensional de una escena de manera compacta, y efectiva, con capacidad de adaptación a diferentes condiciones ambientales [55].

En el caso de un arreglo binocular de cámaras, la información de profundidad de la escena se determina a partir de la correspondencia de cada píxel en el par de imágenes estéreo. La precisión en la reconstrucción tridimensional con arreglos binoculares depende de la distancia de separación entre cámaras. Una distancia de separación amplia mejora la resolución de disparidad, produciendo mayor precisión en la reconstrucción. Sin embargo, esta configuración incrementa la probabilidad de estimaciones de disparidad erróneas debido a las diferencias de perspectiva inducidas. Por otra parte, una distancia de separación corta reduce la probabilidad de estimaciones erróneas. No obstante, se disminuye la resolución de disparidad, comprometiendo la precisión en la reconstrucción.

Un proceso previo fundamental para reducir el costo computacional de los métodos de estimación de correspondencia estéreo es la etapa de rectificación [1]. El proceso de rectificación consiste en realizar una transformación proyectiva a las imágenes estéreo de entrada con el fin de que sus líneas epipolares resulten paralelas.

En este capítulo se describen los fundamentos teóricos para realizar el proceso de reconstrucción tridimensional usando sistemas de visión estéreo. Se presenta el análisis matemático del proceso de adquisición, rectificación, correspondencia y reconstrucción tridimensional utilizando sistemas de visión estéreo. Finalmente, se analiza un sistema de visión estéreo basado en una configuración multilínea base.

### 2.2.1. Modelo de cámara *pinhole*

El modelo de cámara *pinhole* describe la relación geométrica entre los puntos observados de una escena (3D) y su proyección en el plano imagen (2D). El modelo *pinhole*, representa uno de los modelo de cámara más populares debido a su efectividad y sencillez [56]. Este modelo considera que la cámara solo poseé una pequeña apertura por donde pasa la luz. En este modelo, el centro de proyección se considera el centro óptico, mientras que la línea perpendicular que pasa a través del plano imagen y el centro óptico se denomina eje óptico. Adicionalmente, la distancia entre el centro óptico y el plano imagen es conocida como longitud focal $f$, como se muestra en la Fig. 2.2.1.

Utilizando el modelo de cámara *pinhole*, un punto en el espacio $P_d = [X, Y, Z]^T$ en las coordenadas de referencia de la cámara, se proyecta en el plano imagen como

$$s(x, y) = \mathcal{H}^{-1}[KP_d], \tag{2.2.1}$$

donde $\mathcal{H}^{-1}$ es el operador inverso de coordenadas homogéneas [57] y $K$ es la matriz de parámetros intrínsecos de la cámara. Comúnmente, la matriz $K$ se define como

$$K = \begin{bmatrix} \frac{1}{s_x} & \sigma & \frac{\tau_x}{fs_x} \\ 0 & \frac{1}{s_y} & \frac{\tau_y}{fs_y} \\ 0 & 0 & \frac{1}{f} \end{bmatrix}, \tag{2.2.2}$$

donde $s_x$ y $s_y$ son el ancho y largo del píxel, respectivamente, $\tau_x$ y $\tau_y$ son las coordenadas del punto principal (el punto de intersección entre el plano imagen y el eje óptico), y $\sigma$ es un valor de sesgo del píxel.

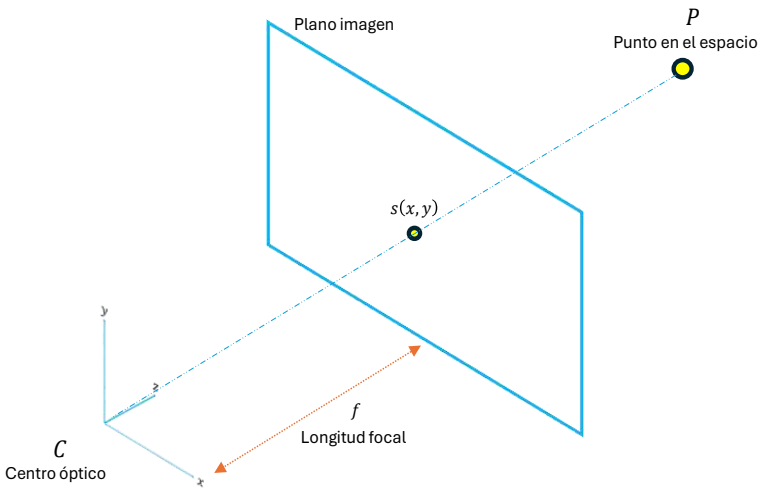



**Figura 2.2.1:** Modelo de cámara *pinhole*. Un punto del espacio es observado y registrado por la cámara en el plano imagen.

Si consideramos que la cámara está en una posición arbitraria del sistema coordenado del mundo real. La posición y orientación de la cámara estarán definidas por un vector de traslación $t$ y una matriz de rotación $R$. El punto $P_d$ es observado desde la cámara como

$$P_d = L\mathcal{H}[P], \tag{2.2.3}$$

donde $P = [X, Y, Z]^T$, $\mathcal{H}$ es el operador de coordenadas homogéneas, y $L$ es la matriz de parámetros extrínsecos de la cámara, definida como [57]

$$L = [R^T, -R^T t]. \tag{2.2.4}$$

Finalmente, la proyección de puntos espaciales al plano imagen, puede realizarse al considerar las Ecs. (2.2.1) y (2.2.3) como

$$s(x, y) = \mathcal{H}^{-1}[KL\mathcal{H}[P]] = \mathcal{H}^{-1}[\mathcal{C}\mathcal{H}[P]], \tag{2.2.5}$$

donde $\mathcal{C} = KL$ es conocida como la matriz de cámara.

### 2.2.2. Geometría epipolar

Sea $P_n = \{[X_n, Y_n, Z_n]^T | n = 1, \ldots, N_n\}$ un conjunto de puntos de una escena. Estos puntos son capturados por un arreglo de cámaras estéreo $\{C_i \ | i = 1, 2\}$ y se denotan como $\{s_{i,n}(x, y) \ | n = 1, \ldots, N_n\}$ en el plano imagen $I_i(x, y)$, como se muestra en la Fig. 2.2.2. La relación entre $P_n$ y $s_{i,n}(x, y)$ puede describirse como

$$s_{i,n}(x, y) = \mathcal{H}^{-1}[\mathcal{C}\mathcal{H}[P_n]]. \tag{2.2.6}$$

La geometría epipolar describe la intersección entre dos planos con respecto al punto de observación [58]. La correspondencia entre los puntos $s_{1,n}(x, y)$ y $s_{2,n}(x, y)$ se establece por medio del plano epipolar formado por la recta que une el centro óptico de $C_1$ y $C_2$

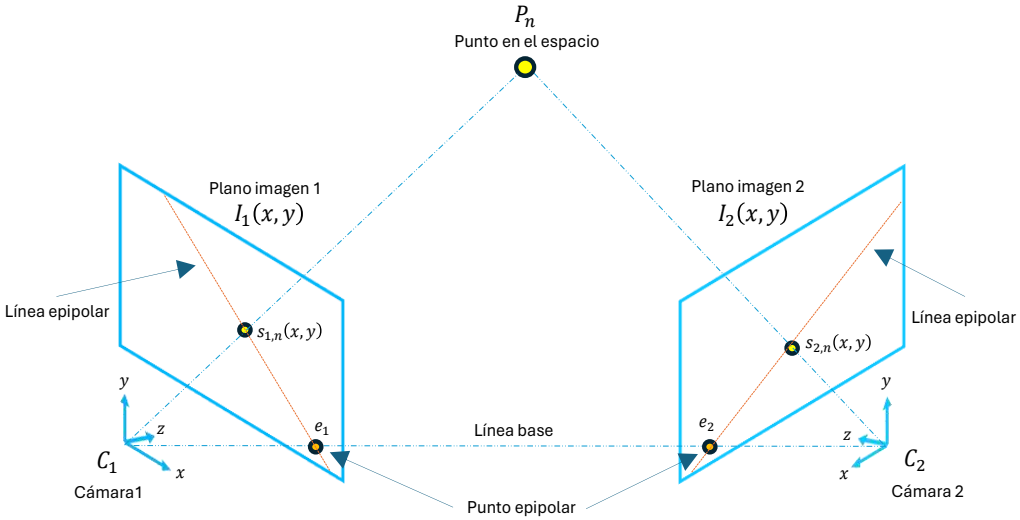



**Figura 2.2.2:** Configuración de un sistema de visión estéreo.

(denominada línea base B) y las líneas proyectadas del centro óptico de $C_1$ y $C_2$ hacia $P_n$. El punto de intersección entre la línea $B$ y el plano $I_i(x, y)$ es denominado como punto epipolar $e_i$. La línea que pasa por el punto $e_i$ y la coordenada $s_{i,n}(x, y)$ en el plano $I_i(x, y)$ es conocida como línea epipolar $l_e$, y se obtiene como

$$l_e = \mathcal{H}[e_i] \times \mathcal{H}[s_{i,n}(x, y)], \tag{2.2.7}$$

donde $\times$ representa el producto cruz.

En imágenes estéreo, la coordenada $s_{1,n}(x, y)$ se proyecta al plano $I_2(x, y)$, y $s_{2,n}(x, y)$ se proyecta al plano $I_1(x, y)$ como una línea epipolar. Este planteamiento simplifica la búsqueda de correspondencias entre puntos sobre la línea epipolar proyectada. Reduciendo un problema de búsqueda bidimensional a uno unidimensional.

### 2.2.3. Matriz fundamental

La matriz fundamental $F$ es una matriz de $3 \times 3$ que realiza un mapeo del plano imagen $I_i(x, y)$ a la línea epipolar $l_e$ [58]. La matriz fundamental se define como

$$F = \left[\mathcal{H}[e_2]\right]^{\times} K_2 R_2^T R_1 K_1^1, \tag{2.2.8}$$

donde $[\cdot]^{\times}$ es el operador de producto cruz, $\mathcal{H}[e_2]$ son las coordenadas homogéneas del punto epipolar en $I_2(x, y)$, $K_2$ es la matriz de paramétros intrínsecos de $C_2$, $R_2^T$ es la matriz de rotación transpuesta de $C_2$, $R_1$ es la matriz de rotación de $C_1$, y $K_1^{-1}$ es la matriz inversa de paramétros intrínsecos de $C_1$.

Sí asumimos que $C_1$ se encuentra en el origen del sistema coordenado de referencia de la escena, $C_2$ está perfectamente alineada con $C_1$; además, $K_1 = K_2 = K$. En este caso, la Ec. (2.2.8) se convierte en la matriz fundamental de forma canónica para un par de imágenes rectificadas [59], dada por

$$F_c = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{2.2.9}$$

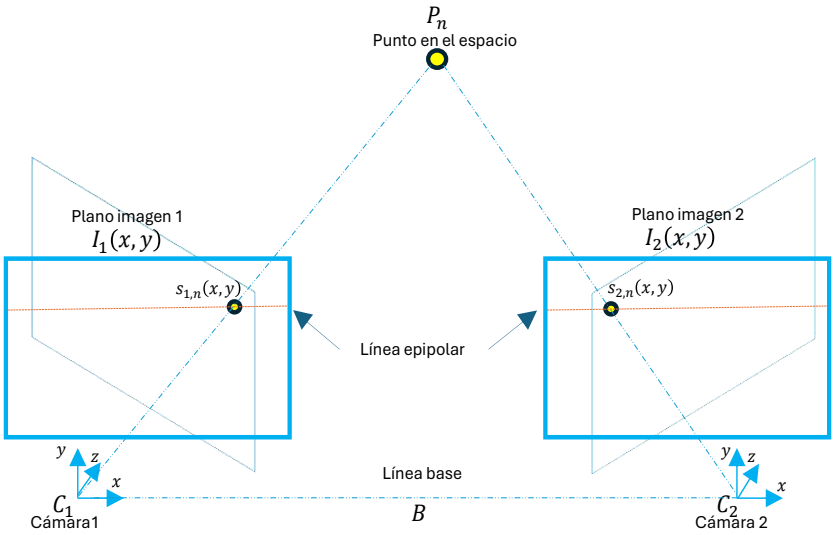



**Figura 2.2.3:** Imágenes estéreo rectificadas. Cada plano $I(x, y)$ es transformado por medio de una matriz $G$, haciendo las imágenes coplanares separadas por una distancia $B$.

Considerando los puntos correspondientes $\{s_{1,n}(x, y) \leftrightarrow s_{2,n}(x, y)\}$, la matriz fundamental $F$ satisface la condición

$$\mathcal{H}[s_{2,n}(x, y)]^T F \mathcal{H}[s_{1,n}(x, y)] = 0. \tag{2.2.10}$$

La matriz fundamental expresa de manera compacta la ubicación de los puntos epipolares y la transformada proyectiva que hay entre las líneas epipolares en un arreglo de cámaras estéreo.

### 2.2.4. Calibración de un sistema de visión estéreo

El proceso de calibración de cámaras estéreo permite estimar los parámetros intrínsecos y extrínsecos del sistema de captura de imágenes [60]. Si se asume que $C_1$ se encuentra en el origen del sistema coordenado de referencia de la escena, los parámetros del par de cámaras estan dados por

$$\begin{aligned} C_1 &= \{K_1\} \\ C_2 &= \{K_2, R_2, t_2\}, \end{aligned} \tag{2.2.11}$$

donde los parámetros $K_1$, $K_2$, $R_2$ y $t_2$ pueden ser estimados utilizando el método propuesto por Zhang et al. [61]. Este método, requiere la captura de un conjunto de imágenes de un patrón plano conocido, considerando distintas posiciones y orientaciones.

### 2.2.5. Rectificación de imágenes estéreo

Un caso particular se presenta cuando las cámaras están perfectamente alineadas, satisfaciendo las condiciones de la Ec. (2.2.9). El proceso de rectificación de imágenes estéreo consiste en encontrar un par de transformadas proyectivas $G_1$ y $G_2$ para generar dos nuevas imágenes, donde los puntos epipolares $e_i$ se encuentren ubicados en el infinito y los planos imagen sea coplanar. Las transformadas proyectivas resultantes deben de

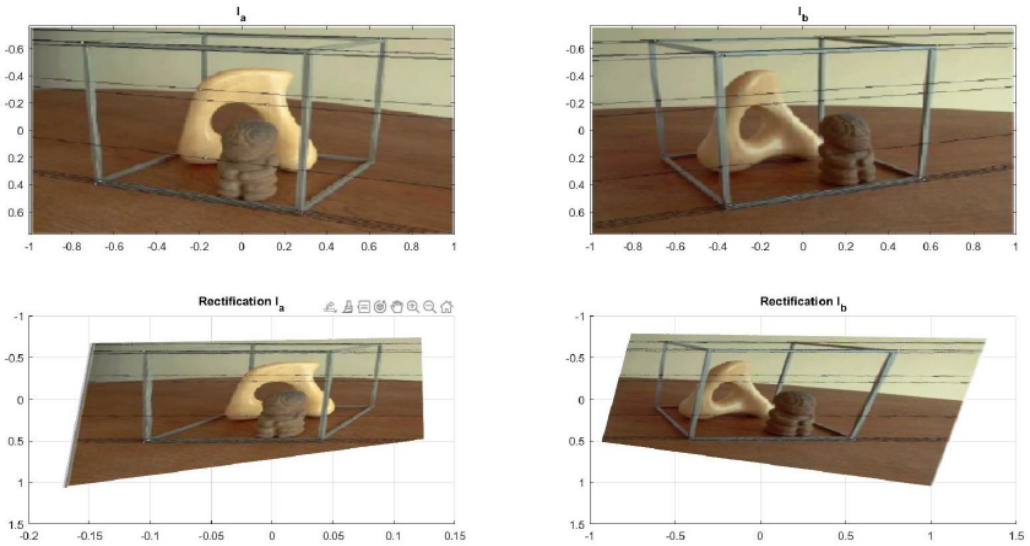



**Figura 2.2.4:** Proceso de rectificación de imágenes estéreo utilizando el método de Hartley et al. [1].

satisfacer la condición descrita por la Ec. (2.2.10). La rectificación de imágenes permite reducir el problema de búsqueda de puntos correspondientes en toda la imagen ($2D$) a solo la línea horizontal ($1D$) del punto $s_{i,n}(x, y)$.

La Fig. 2.2.3 muestra el proceso de rectificación de un par de imágenes estéreo. Los puntos $s_{i,n}(x, y)$ rectificados del plano $I_i(x, y)$ se definen como

$$\begin{aligned} \mathbf{s}_{1,n}(x, y) &= \mathcal{H}^{-1}\left[G_1\mathcal{H}[s_{1,n}(x, y)]\right] \\ \mathbf{s}_{2,n}(x, y) &= \mathcal{H}^{-1}\left[G_2\mathcal{H}[s_{2,n}(x, y)]\right]. \end{aligned} \tag{2.2.12}$$

El proceso de rectificación de imágenes es realizado para todos los puntos de la imagen, como se muestra en la Fig. 2.2.4. Es importante notar, que las imágenes rectificadas son utilizadas para realizar la estimación de profundidad en imágenes estéreo.

### 2.2.6. Estimación de profundidad

Un sistema de visión estéreo está formado por un arreglo de cámaras comúnmente posicionadas de manera horizontal. La estimación de profundidad se obtiene por triangulación a través de establecer la correspondencia de los píxeles en el par de imágenes capturadas. El punto observado $P_n$ se proyecta sobre los planos imagen $I_1(x, y)$ e $I_2(x, y)$. Si el par de imágenes estéreo está rectificado; los píxeles correspondientes se encuentran sobre la misma línea horizontal desplazados por una distancia $d_{12,n}(x, y) = |\mathbf{s}_{1,n}(x) - \mathbf{s}_{2,n}(x)|$ conocida como disparidad, que es inversamente proporcional a la profundidad de $P_n$ [62].

La Fig. 2.2.5 muestra el modelo óptico de un sistema de visión estéreo rectificado desde una vista superior. Se puede observar que la profundidad $Z_n$ del punto observado $P_n$ puede recuperarse utilizando el valor de disparidad $d_{12,n}(x, y)$ entre los puntos proyectados, la longitud focal $f$ de $C_i$, y la línea base $B$ como

$$Z_n = \frac{Bf}{d_{12,n}(x, y)} = \frac{Bf}{|\mathbf{s}_{1,n}(x) - \mathbf{s}_{2,n}(x)|}. \tag{2.2.13}$$

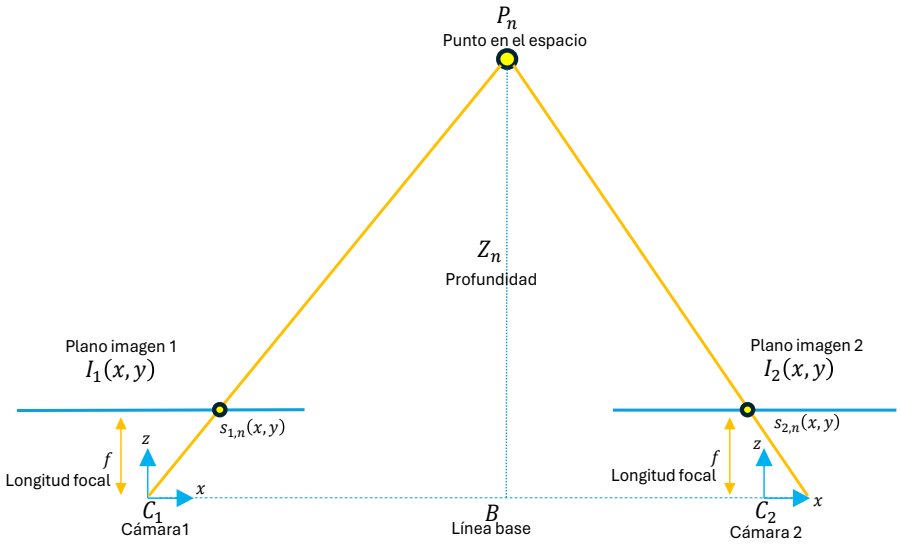



**Figura 2.2.5:** Modelo óptico de un sistema de visión estéreo rectificado.

El mayor reto en la reconstrucción tridimensional utilizando sistemas de visión estéreo es determinar la correspondencia de puntos cuando se presentan perturbaciones en las imágenes como oclusiones, superficies especulares, poca textura, entre otras.

### 2.2.7. Reconstrucción tridimensional basada en triangulación

La reconstrucción tridimensional utilizando sistemas de visión estéreo se puede realizar por medio del método lineal homógeneo de triangulación [57]. Si el par de imágenes estéreo fueron previamente rectificadas, el punto $\mathbf{s}_{1,n}(x,y)$ puede especificarse en términos del punto $\mathbf{s}_{2,n}(x,y)$ como

$$\left[\mathbf{s}_{2,n}(x,y)\right]^T = \left[\mathbf{s}_{1,n}(x,y)\right]^T + \mathcal{H}_0\left[d_{12,n}(x,y)\right], \tag{2.2.14}$$

donde $d_{12,n}(x,y)$ es el valor de disparidad entre los puntos correspondientes de la imagen de referencia $\left(\mathbf{s}_{1,n}(x,y) \leftrightarrow \mathbf{s}_{2,n}(x,y)\right)$, y $\mathcal{H}_0$ es el operador de coordenadas homogéneas con base cero [57]. El punto $P_n$ observado se obtiene utilizando las coordenadas del par de imágenes estéreo como

$$\begin{aligned} \lambda_1\mathcal{H}[s_{1,n}(x,y)] &= \mathcal{C}_1\mathcal{H}[P_n] \\ \lambda_2\mathcal{H}[s_{2,n}(x,y)] &= \mathcal{C}_2\mathcal{H}[P_n], \end{aligned} \tag{2.2.15}$$

donde $\lambda_1$ y $\lambda_2$ son valores escalares desconocidos, y

$$\begin{aligned} \mathcal{C}_1 &= K_1\left[\mathbb{I}_3, \mathbf{0}_3\right] \\ \mathcal{C}_2 &= K_2\left[R^T, -R^T t\right], \end{aligned} \tag{2.2.16}$$

son las matrices cámara de $C_1$ y $C_2$, respectivamente, $\mathbb{I}$, es una matriz identidad $3\times 3$, $R$ es una matriz de rotación, y $t$ es un vector de traslación de $C_2$ con respecto a $C_1$. Nótese que la Ec. (2.2.15) puede rescribirse de la siguiente manera:

$$\begin{bmatrix} \mathcal{C}_1 & \mathcal{H}[\mathbf{s}_{1,n}(x,y)]^T & \mathbf{0}_3 \\ \mathcal{C}_2 & \mathbf{0}_3 & \mathcal{H}[\mathbf{s}_{2,n}(x,y)]^T \end{bmatrix} \begin{bmatrix} \mathcal{H}[P_n] \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \mathbf{0}_6, \tag{2.2.17}$$

donde $\mathbf{0}_6$ y $\mathbf{0}_3$ son vectores de ceros de tamaño $6 \times 1$ y $3 \times 1$, respectivamente. La Ec. (2.2.17) se resuelve para $\begin{bmatrix}\mathcal{H}[P_n] & \lambda_1 & \lambda_2\end{bmatrix}^T$ utilizando el método de descomposición en valores singulares (SVD, por sus siglas en inglés *Singular Value Decomposition* ) [63]. Finalmente, las coordenadas $[X_n, Y_n, Z_n]^T$ del punto observado $P_n$ se pueden obtener como

$$[X_n, Y_n, Z_n]^T = \mathcal{H}^{-1}[P_n]. \tag{2.2.18}$$

### Métodos locales de estimación de disparidad

Numerosos métodos para estimación de correspondencia entre imágenes estéreo han sido propuestos. Estos métodos de estimación pueden clasificarse principalmente por el tipo de información que utilizan, por ejemplo, métodos que utilizan información local de las imágenes, métodos que utilizan información global de las imágenes y métodos basados en aprendizaje profundo [54].

**Cuadro 2.1:** Resumen de técnicas representativas de estimación de disparidad estéreo.

| ***Método de estimación de disparidad estéreo*** | ***Ventajas*** | ***Desventajas*** |
|---|---|---|
| Métodos locales | Computo eficiente<br>Adecuado para tiempo-real<br>Implementación simple | Errores en zonas de baja textura<br>Sensibles al ruido<br>Precisión limitada |
| Métodos globales | Precisión alta en escenas complejas<br>Mejor respuesta ante oclusiones | Difícil de implementar en tiempo-real |
| Métodos basados en aprendizaje profundo | Robusto a oclusiones y ruido<br>Alta precisión en escenas complejas | Requiere gran cantidad de datos para entrenamiento |

Los métodos globales plantean la estimación de disparidad como un problema de minimización de una función de optimización global. Estos métodos producen buenos resultados, pero son computacionalmente costosos, lo que los vuelve poco prácticos para aplicaciones de tiempo real. Por otro lado, los métodos locales utilizan la intensidad de un fragmento predefinido de la imagen para estimar la disparidad de un píxel dado. La ventaja principal de estos métodos es que poseen una complejidad computacional baja. Los métodos basados en aprendizaje profundo son configurados para ajustar hiperpárametros de una función de pérdida comúnmente a través del uso de CNN. La Tabla 2.1 presenta de manera general las ventajas y desventajas de los métodos de estimación de disparidad estéreo.

Los métodos locales de estimación de disparidad estéreo son adecuados para implementaciones en dispositivos con capacidades de paralelismo masivo. Estos métodos locales consisten en cuatro etapas principales: costo de asociación, agregación de costo, selección de disparidad y refinamiento. En la etapa de costo de asociación, se cuantifica la similitud de dos píxeles correspondientes al mismo punto en la escena. La etapa de agregación de costo, minimiza la incertidumbre en la asociación de píxeles. La selección de disparidad, es la selección del menor valor de costo de agregación para cada píxel. Finalmente, en la etapa de refinamiento se reducen los errores de estimación y se recuperan las regiones con oclusiones en la imagen.

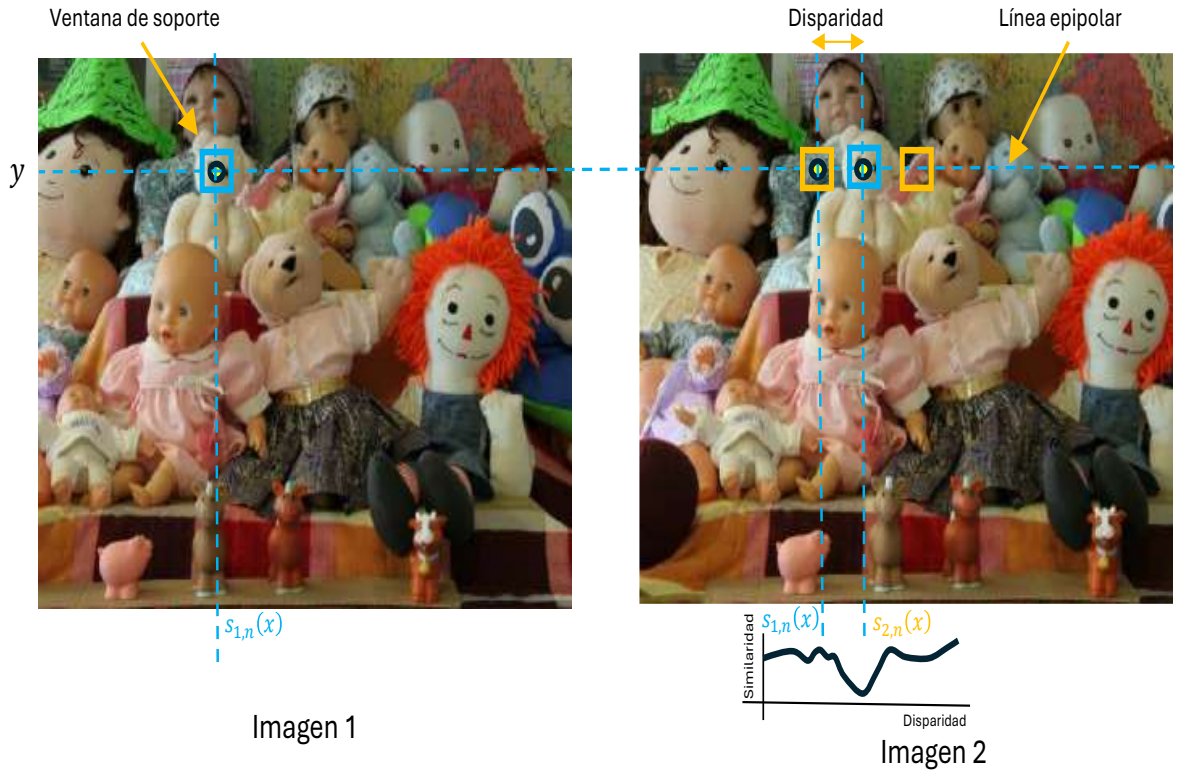



**Figura 2.2.6:** Utilización de ventanas de soporte para estimaciones de disparidad. La función de costo se evalua para cada píxel a lo largo del intervalo de búsqueda.

Distintas funciones de costo de asociación han sido propuestas para resolver el problema de correspondencia utilizando métodos locales estéreo. Entre los métodos más conocidos reportados en la literatura se encuentran el método de suma de diferencias absolutas (SAD, por sus siglas en inglés *Sum of Absolute Differences*), suma de diferencias cuadradas (SSD, por sus siglas en inglés *Sum of Squared Differences*), correlación cruzada normalizada (NCC, por sus siglas en inglés *Normalized Cross Correlation*), transformadas de rango (RT, por sus siglas en inglés *Rank Transform*) y transformada census (CT, por sus siglas en inglés *Census Transform*) [54]. La Fig. 2.2.6 presenta un ejemplo de un método local de estimación de disparidad en imágenes estéreo.

En las secciones siguientes se describen de manera general algunas de las técnicas más conocidas de la literatura científica para el problema de correspondencia local estéreo.

## Estimación de disparidad basada en SAD

Este método consiste en realizar una comparación de similitud entre dos fragmentos de imagen, mediante la suma de sus diferencias absolutas [64]. La agregación de costo usando SAD está dada por

$$SAD(x, y, k) = \sum_{(x,y)\in w} \mid I_1(x, y) - I_2(x - k, y) \mid, \tag{2.2.19}$$

donde $k$ es el valor de disparidad, $I_1(x, y)$ e $I_2(x, y)$ son la imagen izquierda y derecha, respectivamente, del par de imágenes estéreo, y $w$ especifica la región de soporte.

El mapa de disparidad se obtiene al seleccionar el valor de $d$ que minimiza la función $SAD(x, y, d)$ como a continuación:

$$d(x, y) = \arg\min_{k}\{SAD(x, y, k)\}, \quad k \in [k_{min}, k_{max}], \tag{2.2.20}$$

donde $[k_{min}, k_{max}]$ especifica el rango de valores de disparidad.

Este método se ha utilizado en implementaciones de tiempo real debido a su baja complejidad computacional. Sin embargo, la calidad del mapa de disparidad resultante con este enfoque puede verse afectada por la presencia de ruido en los límites del objeto y regiones sin textura.

## Estimación de disparidad basada en SSD

Este método consiste en realizar la suma de las diferencias al cuadrado de los valores de intensidad de los píxeles en la ventana de soporte $w$ alrededor del píxel de interés, como [65]

$$SSD(x, y, k) = \sum_{(x,y)\in w} \mid I_1(x, y) - I_2(x - k, y) \mid^2 . \tag{2.2.21}$$

El mapa de disparidad se obtiene sustituyendo la Ec.(2.2.21) en la Ec.(2.2.20). Este método tiene una alta sensibilidad al ruido y a cambios en la iluminación de la escena.

## Estimación de disparidad basada en NCC

Este método determina la correspondencia entre dos ventanas de la imagen utilizando la correlación cruzada normalizada, dada por [66]

$$NCC(x, y, k) = \frac{\sum_{(x,y)\in w} I_1(x, y)I_2(x - k, y)}{\sqrt{\sum_{(x,y)\in w} I_1^2(x, y) \sum_{(x,y)\in w} I_2^2(x - k, y)}}. \tag{2.2.22}$$

El mapa de disparidad se obtiene al seleccionar el valor de $d$ que maximiza la función $NCC(x, y, k)$ como a continución

$$d(x, y) = \arg\max_{k}\{NCC(x, y, k)\}, \quad k \in [k_{min}, k_{max}]. \tag{2.2.23}$$

Este método es robusto a cambios de intensidad y de contraste. Sin embargo, tiende a difuminar regiones discontinuas.

## Estimación de disparidad basada en transformada Census

Este método realiza una transformación de intensidad no paramétrica de las imágenes de entrada, formando cadenas de bits para cada píxel de las imágenes de entrada [67]. La disparidad se estima utilizando la distancia de Hamming entre las cadenas de *bit* generadas por la transformada Census en las ventanas correspondientes del par de imágenes estéreo, como a continuación:

$$CT(x, y, k) = \bigotimes_{(x,y)\in w} \xi(I_1(x, y), I_2(x - k, y)), \tag{2.2.24}$$

donde $\otimes$ denota el operador de concatenación, y $\xi$ representa la transformación de las ventanas de soporte definida como

$$\xi(I_1(x,y), I_2(x-k,y)) = \begin{cases} 1, & \text{si } I_1(x,y) > I_1(x-k,y) \\ 0, & \text{de lo contrario.} \end{cases} \tag{2.2.25}$$

La transformada Census proporciona una mayor robustez ante cambios de iluminación y a discontinuidades en los valores de disparidad.

### Estimación de disparidad basada en ponderación adaptativa de la región de soporte

Este método, conocido como región de soporte adaptativa (ASW por sus siglas en inglés *Adaptive Support Weight*) utiliza una ventana adaptativa basada en la distancia de color y ubicación espacial de los elementos de la ventana deslizante para realizar la estimación del mapa de disparidad [68]. La ventana adaptativa se construye como

$$w(x,y) = \exp\left[-\left(\frac{w_p(x,y)}{\gamma_p} + \frac{w_q(x,y)}{\gamma_q}\right)\right], \tag{2.2.26}$$

donde $w_p(x,y)$ cuantifica las distancias de color entre cada elemento de la ventana de referencia y el elemento central en el espacio de color CIELAB, $w_q(x,y)$ contiene las distancias espaciales entre cada elemento de la ventana de referencia y el elemento central, y $\gamma_p$,$\gamma_q$ son parámetros de escala.

La agregación de costos se calcula como

$$ASW(x,y,k) = \frac{\sum w_r(x,y) w_f(x-k,y) \left|I_1(x_c,y_c) - I_2(x_c,y_c)\right|}{\sum w_r(x,y) w_f(x-k,y)}, \tag{2.2.27}$$

donde $w_r(x,y)$ y $w_f(x,y)$ representan las ventanas adaptativas construidas utilizando la Ec. (2.2.26) a partir de la información de las imágenes $I_1(x,y)$ e $I_2(x,y)$, respectivamente. El mapa de disparidad estimado $d(x,y)$ se obtiene al sustituir la Ec. (2.2.27) en la Ec. (2.2.20). Este método presenta un buen desempeño en los límites de los objetos y una alta precisión, debido a la asignación de pesos dependiente de la similitud de información.

## 2.2.8. Visión estéreo multilínea base

La mayoría de métodos de visión estéreo en el estado del arte se basan en visión binocular, manteniendo una línea base $B$ fija. Una línea base larga proporciona mayor precisión en la estimación de la disparidad que una línea base corta para el mismo algoritmo de estimación de disparidad [69]. Por otra parte, una línea base larga incrementa la complejidad computacional e incertidumbre en el proceso de estimación de disparidad.

Consideremos el arreglo mostrado en la Fig. 2.2.7. Este arreglo está conformado por $m$ cámaras estéreo rectificadas denotadas por $C_1, C_2, \ldots, C_m$, respectivamente. En esta configuración, la imagen de referencia $\mathbf{I}_1(x,y)$ es capturada por medio de $C_1$, mientras que el resto de las imágenes $\mathbf{I}_m(x,y)$ para $m = 2, \ldots, m$ son capturadas por sus respectivas cámaras $C_m$. Además, $B_m$ es la línea base del par de cámaras $\{C_{m-1}, C_m\}$.

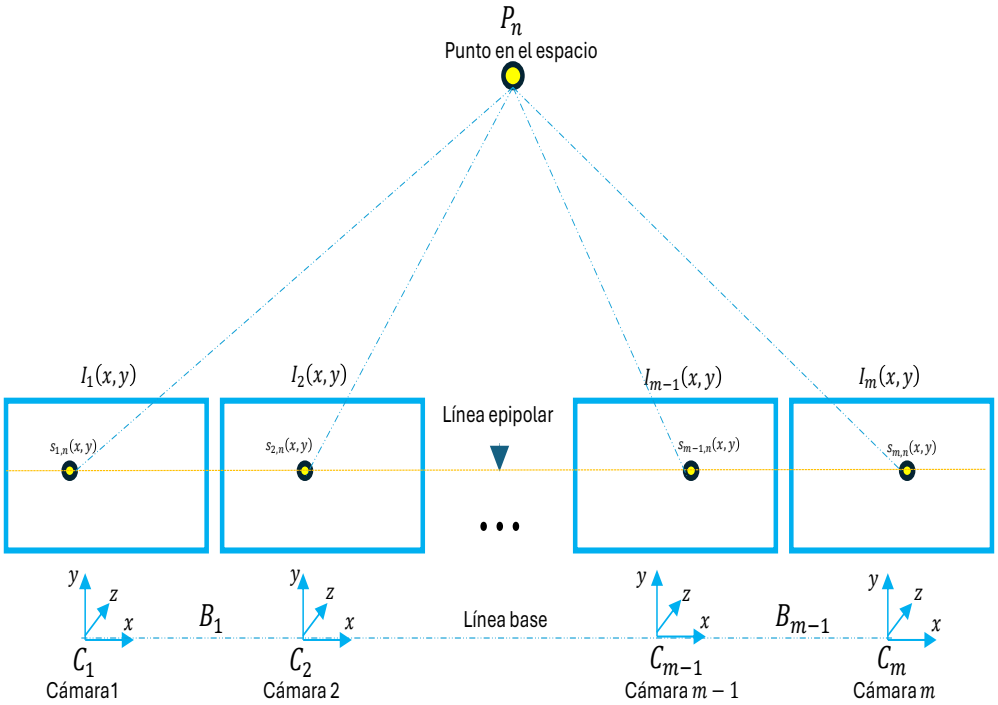



**Figura 2.2.7:** Sistema multi-ocular estéreo.

Sea $d_{12,n}(x,y)$ el mapa de disparidad estimado entre las cámaras $\{C_1, C_m\}$ con la línea base más corta, es decir, cuando $m = 2$. Tenga en cuenta que si el arreglo de cámaras está correctamente rectificado, una buena predicción del mapa de disparidad $d_{1m+1,n}(x,y)$ para el siguiente par de cámaras está dado por

$$d_{1m+1,n}(x,y) = r_{m-1} d_{1m,n}(x,y) \pm \delta, \tag{2.2.28}$$

donde

$$r_{m-1} = \frac{B_{m+1}}{B_m} \tag{2.2.29}$$

es la i-ésima relación de la línea base y $\delta$ es un valor de error de estimación. La profundidad del punto $P_n$ observado, se puede recuperar utilizando la Ec. (2.2.13) con los valores de $d_{1m+1,n}(x,y)$.

La línea base tiene un impacto fundamental en la precisión de la estimación de la profundidad. Una línea base larga mejora la precisión de la estimación. Sin embargo, incrementa el error en la estimación debido a las oclusiones generadas por la posición de la cámara.

## 2.3. Puntos característicos del rostro

Comúnmente, un conjunto de $n$ puntos característicos faciales es representado por un vector de coordenadas $L_n = [x_1, y_1, x_2, y_2, \ldots, x_n, y_n]$, donde $(x_i, y_i)$ es la *i-ésima* coordenada del punto característico detectado en la imagen, como se presenta en la Fig. 2.3.1.

La detección de puntos característicos faciales en imágenes presenta un reto, debido a los efectos de variaciones de las expresiones faciales, cambios en la posición de la cabeza, perturbaciones en la iluminación u oclusiones de la región de interés [45].

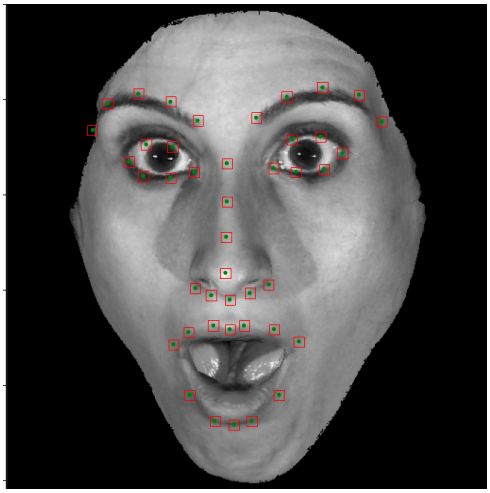


**Figura 2.3.1:** Puntos de interés del rostro ubicados en regiones donde ocurren movimientos musculares como las cejas, contorno de ojos, nariz y boca.

Los métodos de detección de puntos característicos faciales se pueden dividir en tres grupos según la taxonomía propuesta por Cootes et al. [70], presentada en la Tabla 2.2. Los métodos más relevantes reportados en la literatura científica, se describen a continuación.

**Cuadro 2.2:** Taxonomía de los métodos de detección de puntos caraterísticos faciales.

| ***Método*** | ***Ventajas*** | ***Desventajas*** |
|---|---|---|
| Holístico | Modelan la apariencia y forma facial de manera global. | Dificultades para ajustarse a variaciones del rostro. |
| | Procesamiento computacional rápido. | Dificultades para ajustarse en la presencia de oclusiones. |
| Basado en regresión | Modelan la apariencia facial de manera local. | Procesamiento computacional lento. |
| | Robusto a cambios de iluminación. | Compensación entre robustez y precisión. |
| | Robusto a oclusiones parciales. | |
| Con restricciones locales | Capturan de manera implícita la información de la apariencia y forma facial. | Baja precisión ante oclusiones. |
| | Buen desempeño. | Método iterativo que puede quedar ciclado en mínimos locales. |
| | Procesamiento computacional rápido. | Sensible al método de detección facial. |

### Características locales binarias (LBF)

El método de características locales binarias (LBF, por sus siglas en inglés *Local Binary Features*) es un método basado en regresión que estima una proyección lineal para cada punto característico facial utilizando un conjunto de características locales binarias

discriminativas [71]. Este método está compuesto por dos etapas: entrenamiento y prueba. Durante la etapa de entrenamiento, se construye una forma facial $F_s$ de manera iterativa utilizando un método de regresión definido como

$$\Delta F_s^i = L_r^i \psi^i \left( I, F_s^{i-1} \right), \tag{2.3.1}$$

donde $I$ es una matriz que contiene los elementos de la imagen de entradad, $F_s^{i-1}$ es la forma facial estimada en la iteración pasada, $\psi^i(\cdot)$ es una función de mapeo de características, y $L_r^i$ es una matriz de regresión lineal.

Para cada punto característico se obtienen características binarias $\theta_n^i$. Por lo tanto, una función de mapeo de características locales puede formarse concatenando las características para cada iteración como $\psi^i = [\theta_1^i, \theta_2^i, \ldots, \theta_n^i]$.

Para determinar $\psi^i$, la técnica de regresión de bosques aleatorios [72] es aplicada a cada característica binaria combinada con las muestras de entrenamiento. Las características binarias locales obtenidas son utilizadas para construir una matriz de regresión $L^i$ que regulariza las dimensiones de las características como

$$L^i = \underset{L^i}{\operatorname{argmin}} \sum_{p}^{N} ||\Delta \bar{F}_p^i - L^i \psi^i \left( I, F_p^{i-1} \right) ||_2^2, + \lambda ||L^i||_2^2, \tag{2.3.2}$$

donde $\lambda$ es un vector de paramétros de regularización.

## Modelo de apariencia activa (AAM)

El método de modelo de apariencia activa (AAM, por sus siglas en inglés *Active Appearance Model*), es un método holístico que utiliza un modelo estadístico obtenido a partir de las variaciones de la forma y textura faciales de las imágenes de entrenamiento, etiquetadas con los correspondientes puntos de referencia [73]. Los puntos característicos faciales son estimados a través de la comparación entre el modelo generado y el rostro de la imagen de entrada. Este modelo está dado por

$$\begin{aligned} \mathbf{x} = &\quad \bar{\mathbf{x}} + \mathbf{P}_x \alpha \\ \mathbf{s} = &\quad \bar{\mathbf{s}} + \mathbf{P}_s \alpha, \end{aligned} \tag{2.3.3}$$

donde $\mathbf{x}$ y $\mathbf{s}$ representan el vector de puntos característicos y vector de textura, respectivamente, $\bar{\mathbf{x}}$ y $\bar{\mathbf{s}}$ son la media de la forma y la media de la textura de la imagen de entrenamiento combinada, respectivamente, $\mathbf{P}_x$ y $\mathbf{P}_s$ son matrices que describen la variación de las imágenes de entrenamiento en términos de ubicación del punto de referencia y textura de la imagen, respectivamente, y $\alpha$ es un vector escalar que controla el modelo de apariencia.

## Árboles de regresión (Dlib)

Este método basado en regresión desempeña una alineación facial utilizando una cascada de funciones de regresión, la cual predice el vector de forma facial utilizando estimaciones previas y un subconjunto de píxeles de rostro en la imagen como [74]

$$\hat{F}_s^{(t+1)} = \hat{F}_s^{(t)} + r_{f_t}(I, \hat{F}_s^{(t)}), \tag{2.3.4}$$

donde $\hat{F}_s^{(t)}$ es la estimación actual de la forma facial en la imagen $I$, y $r_{ft}(\cdot)$ representa la cascada de funciones de regresión. A partir de los datos de entrenamiento se genera una imagen facial $I_i$, una forma inicial $\hat{F}_{s_i^{(0)}}$, y un paso de actualización objetivo $\Delta F_{s_i^{(0)}}$, que son actualizados en cada iteración como

$$\hat{F}_{si}^{\,(t+1)} = \hat{F}_{si}^{\,(t)} + r_t(I_i, \hat{F}_{si}^{\,(t)}) \tag{2.3.5}$$

$$\Delta F_{s_i}^{(t+1)} = F_{s_i} - \hat{F}_{si}^{\,(t+1)}. \tag{2.3.6}$$

### Método de descenso supervisado (SDM)

El método de descenso supervisado (SDM, por sus siglas en inglés *Supervised Decent Method*) es un método detector de puntos característicos y de alineación facial, dividido en dos etapas: entrenamiento y prueba [75].

Durante la fase de entrenamiento, el método SDM aprende una sucesión de direcciones descendentes para minimizar la diferencia entre la forma facial estimada y los rasgos reales. Las direcciones descendentes obtenidas se utilizan para predecir la forma facial de manera iterativa. El paso del punto de referencia desde la localización actual $(x_k, y_k)$ con respecto a la localización verdadera $(x_g, y_g)$ se calcula como

$$\Delta(x_k, y_k) = (x_g, y_g) - (x_k, y_k). \tag{2.3.7}$$

Después, utilizando el método de Newton para estimar la matriz Hessiana, la Ec. (2.3.7) puede escribirse como

$$\Delta(x_k, y_k) = R_k \Theta_k + d_k, \tag{2.3.8}$$

donde $\Theta_K$ es la forma estimada facial actual, $R_k$ es la dirección descendiente, y $d_k$ es un término de sesgo. La meta es estimar los párametros $R_k$ y $d_k$ como

$$\underset{R_K, d_k}{\operatorname{argmin}} \sum_{(x_k, y_k)^i} ||\Delta(x_k, y_k)^i - R_k \Theta_k^i - d_k||^2, \tag{2.3.9}$$

donde $i$ y $k$ definen los índices de muestra y de iteración, respectivamente. Finalmente, $R_k$ y $b_k$ son usados para actualizar las coordenadas de los puntos faciales como

$$(x_{k+1}, y_{k+1}) = (x_k, y_k) + R_k \Theta_k + d_k. \tag{2.3.10}$$

Este procedimiento es iterativo hasta que los puntos de referencia converjan a la posición real.

## 2.4. Aprendizaje profundo

El aprendizaje profundo es una rama de la inteligencia artificial que agrupa un conjunto de algoritmos capaces de desempeñar tareas que regularmente requieren de inteligencia humana, como percepción visual, reconocimiento de patrones o toma de decisiones. A lo largo de los años, las técnicas de aprendizaje profundo han demostrado un desempeño sobresaliente en la literatura científica para tareas de procesamiento de imágenes y visión por computadora.

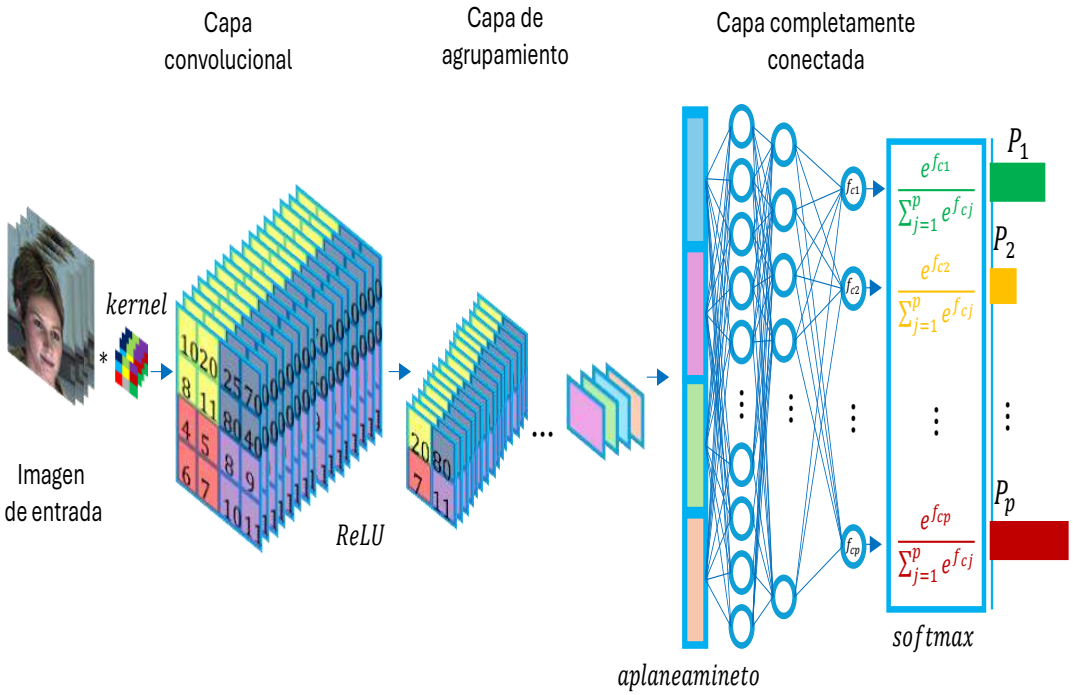



**Figura 2.4.1:** Arquitectura convencional de una CNN para la tarea de clasificación de imágenes.

En los útimos años las técnicas de aprendizaje profundo han tenido gran popularidad, debido a que son capaces de integrar las tareas de extracción de características y clasificación para desempeñarse de manera conjunta. Este enfoque permite realizar una predicción directa de cada emoción, por medio de representaciones complejas obtenidas de imágenes de entrada. Sin embargo, este enfoque presenta dos principales retos para los sistemas FER: un sobreentrenamiento derivado de una basta cantidad de imágenes de entrenamiento e información del rostro ausente derivado de oclusiones, cambios en la pose o iluminación de las imágenes de entrada.

Hoy en día, se han reportado exitosas configuraciones de modelos de aprendizaje profundo, por ejemplo, las redes neuronales convolucionales (CNN) [46], redes de creencias profundas (DBN, por sus siglas en inglés *Deep Belief Networks*) [76], redes neuronales recurrentes (RNN, por sus siglas en inglés *Recurrent Neural Network*) [77], entre otros. Las CNN son una de las técnicas mayormente utilizadas en el área de aprendizaje profundo, son simples arquitecturas de redes neuronales que usan convolución de matrices en al menos alguna de sus capas [78]. La Fig. 2.4.1 presenta una arquitectura completa de una red neuronal convolucional utilizada en la tarea de clasificación de expresiones del rostro. En este enfoque, la información se agrupa como arreglos multidimensionales denominados tensor.

### 2.4.1. Redes neuronales convolucionales (CNN)

Las CNN son una arquitectura de aprendizaje profundo, ampliamente utilizadas para tareas de clasificación utilizando imágenes [46]. Las CNN están formadas por tres capas

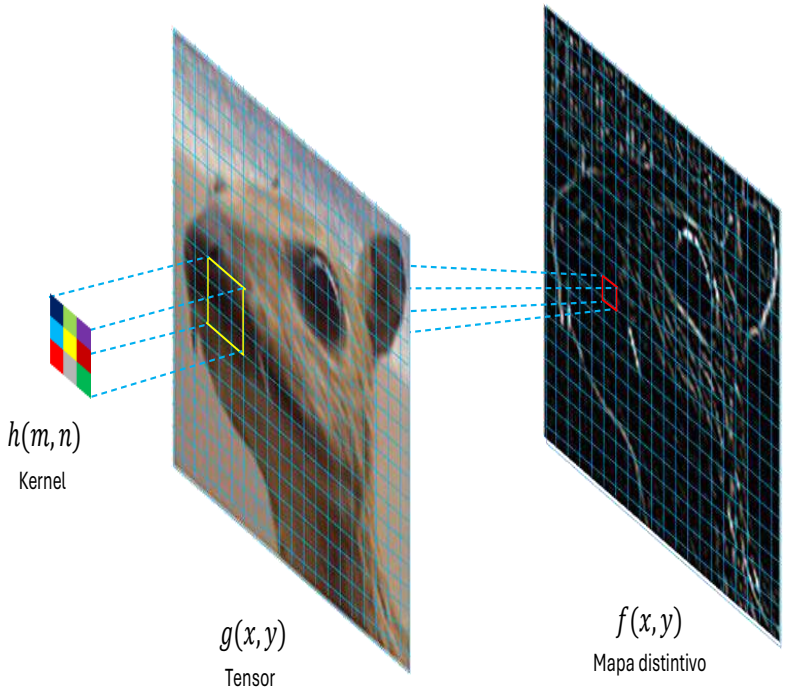



**Figura 2.4.2:** Ilustración de la operación de convolución en una imagen digital.

principales: capa convolucional, capa de agrupación, y capa completamente conectada. Una arquitectura común de CNN está formada por varias capas convolucionales y capas de agrupación apiladas, seguida de una o varias capas completamente conectadas. La transformación de la información a través de estas capas es conocida como propagación hacia adelante.

La capa convolucional realiza la operación lineal de convolución, como se describe a continuación [79]

$$f(x,y) = (g * h)(x,y) = \sum_m \sum_n g(x-m, y-n)h(m,n), \tag{2.4.1}$$

donde $*$ es el operador de convolución, $h(m,n)$ es una matriz de datos llamada *kernel*, $g(x,y)$ es la matriz de entrada comúnmente denominada como tensor, y $f(x,y)$ es el mapa distintivo de características del tensor. Las CNN utilizan una secuencia de capas convolucionales $f(x,y)$ concatenadas, donde la entrada de cada capa es la salida de la capa anterior, siendo la primera entrada la imagen inicial.

La Fig. 2.4.2 muestra un ejemplo del proceso de convolución a una imagen, donde el resultado es una detección de bordes del objeto. Un problema común durante el proceso de convolución se presenta al realizar la operación en los bordes del tensor. Para solucionar este problema, se realiza un incremento en el tamaño de filas y columnas del tensor denominado rellenado de ceros; la idea es poder convolucionar el centro del kernel con los bordes para matener el tamaño del tensor de entrada [80].

Los mapas de distintivos resultantes del proceso de convolución regularmente pasan por una función de activación no lineal, con la finalidad de abstraer características no

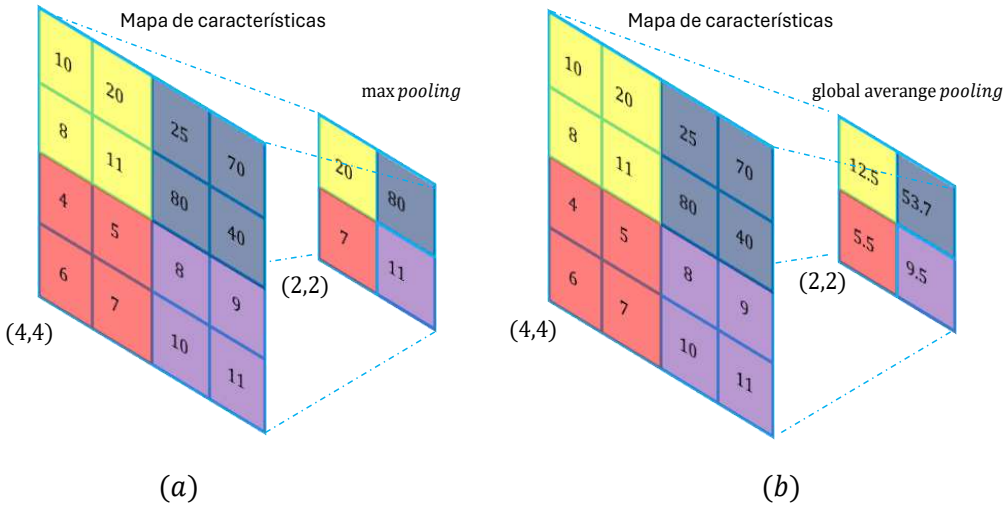



**Figura 2.4.3:** Métodos de agrupación utilizados en arquitecturas de CNN. (a) *Max pooling* toma el valor más alto de la región de interés. (b) *Global averange pooling* calcula el valor medio de la región de interés.

lineales como

$$f^l(x,y) = a_f((g * h)^l(x,y) + b^l(x,y)), \tag{2.4.2}$$

donde $l = 1 \dots L$ es el número de capas, $a_f(\cdot)$ es la función de activación, y $b^l(x,y)$ es un valor de sesgo. Las funciones de activación comúnmente utilizadas en arquitecturas CNN son la función *Sigmoide*, tangente hiperbólica y unidada lineal rectificada $(ReLU)$ [80].

El tamaño de los mapas de características generados en las capas convolucionales es reducido en la capa de agrupación. En esta capa, una función de agrupación es utilizada para disminuir parámetros en zonas determinadas con valores estadísticos cercanos a la posición de valores del tensor a reducir. El objetivo de la capa de agrupación es hacer la representación de tensores invariantes al desplazamiento y cambios sutiles de la imagen de entrada, manteniendo las características más relevantes y condensando la resolución de los tensores. En la literaturas, las técnicas de agrupación más utilizadas son el *max pooling* y *global average pooling* [81]; la Fig. 2.4.3 muestra un ejemplo de ambos métodos.

Finalmente, en la capa completamente conectada se realiza una evaluación de las características extraídas por las capas anteriores para asignarlas a una clase específica basada en una función de probabilidad. Los mapas de características resultantes son transformados a un vector unidimensional, donde cada neurona recibe una combinación del vector generado para aprender combinaciones complejas de patrones de alto nivel relacionados con la salida. Por último, para desempeñar la tarea de clasificación multiclase se utiliza la función de activación *softmax*, donde cada neurona de la capa de salida representa una probabilidad de corresponder a una clase [78]. Sea $\mathbf{F}_c = [f_{c1}, f_{c2}, \dots, f_{cp}]$ el vector de salida, donde cada elemento $f_{cp}$ representa una neurona de la última capa de la red. La función *softmax* convierte los valores del vector $\mathbf{F}_c$ en probabilidades como

$$P_i = \frac{e^{f_{ci}}}{\sum_{j=1}^{p} e^{f_{cj}}}, \tag{2.4.3}$$

donde $e$ es el número de Euler.

### Redes residuales (ResNet)

El uso de varias capas convolucionales puede incrementar la precisión de la red neuronal. Sin embargo, el valor resultante del gradiente puede ser un valor muy grande o un valor igual a cero, lo cual genera un incremento en la tasa de error del entrenamiento y prueba.

Para evitar este problema He et al. [82] propone una nueva arquitectura denominada red residual (ResNet, por sus siglas en inglés *Residual Network*). Esta arquitectura resuelve el problema de degradación de las redes neuronales profundas, utilizando bloques residuales que permiten el flujo directo de información usando saltos en las conexiones, disminuyendo el problema de desvanecimiento de gradiente. La Fig. 2.4.4 presenta la estructura del bloque residual.

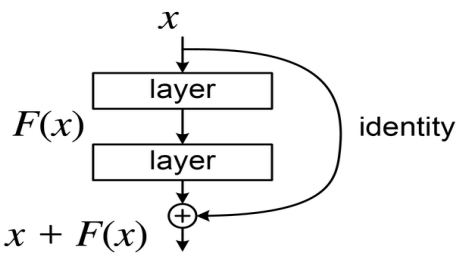



**Figura 2.4.4:** Bloque residual.

Capítulo 3

# Método propuesto para reconocimiento de expresiones usando visión multi-ocular

En este capítulo se presenta el método propuesto para reconocimiento de expresiones faciales utilizando visión multi-ocular. En términos generales, el método propuesto consiste en las siguientes tres etapas principales:

- Captura y rectificación de imágenes multi-oculares del rostro en una escena.
- Estimación de mapas de profundidad de la escena capturada utilizando el enfoque de multilínea base ajustable.
- Extracción y clasificación de características faciales tridimensionales.

La Fig. 3.0.1 presenta un diagrama a bloques del método propuesto. En las siguientes secciones se explica a detalle cada una de la etapas del método propuesto.

## 3.1. Rectificación multi-ocular

En esta sección se presenta la metodología propuesta para realizar el proceso de rectificación multi-ocular estéreo (STROMI, por sus siglas en inglés *STereo-Rectification for Optimized multi-ocular Images*) de una escena. El método de propuesto, plantea la tarea de rectificación multi-ocular como un problema de búsqueda de un conjunto de transformadas proyectivas que minimicen la distorsión inducida por la proyección del plano imagen y que cumpla con las restricciones epipolares. El método de optimización por enjambre de partículas (PSO, por sus siglas en inglés Particle Swarm Optimization) es utilizado para encontrar un vector que contenga los parámetros para construir las matrices de transformación proyectivas de las imágenes de entrada.

### 3.1.1. Método de rectificación estéreo multi-ocular (STROMI)

Sea $\{s_{1,n}(x,y) \leftrightarrow s_{2,n}(x,y)\}$ un conjunto de $n$ puntos correspondientes de un par de imágenes estéreo $I_1(x,y)$ e $I_2(x,y)$ no rectificadas. El conjunto de puntos cumple con la restricción epipolar descrita por la Ec. (2.2.10) dada por

$$\mathcal{H}[s_{1,n}(x,y)]^T G_1^T F_c G_2 \mathcal{H}[s_{2,n}(x,y)] = 0, \tag{3.1.1}$$

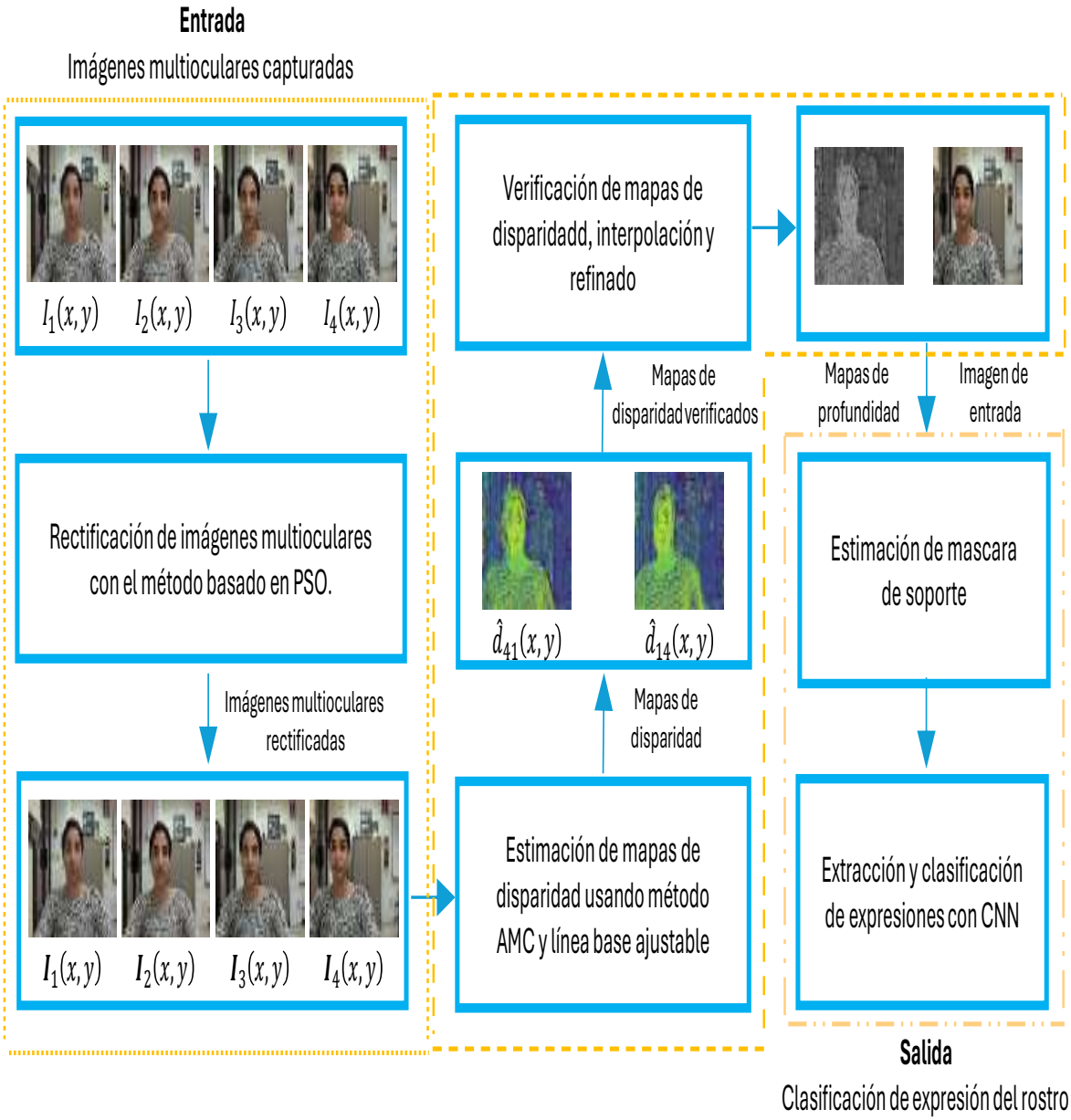



**Figura 3.0.1:** Diagrama de bloques del método propuesto para clasificación de expresiones del rostro utilizando imágenes multi-oculares.

donde para el caso de imágenes rectificadas, la matriz fundamental toma la forma canónica $F_C$, y $G_1$ y $G_2$ representa un par de transformadas proyectivas.

Se requiere encontrar el valor de las matrices $G_1$ y $G_2$ que cumplan con la restricción descrita en la Ec. (3.1.1) y generen la menor distorsión posible a los puntos transformados. La matriz $\{G_i|i = 1, 2\}$ es conocida como matriz homografía, la cual contiene los parámetros intrínsecos y extrínsecos de $C_i$, definida como [57]

$$G = \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix} = K \begin{bmatrix} \bar{r}_1, & \bar{r}_2, & -R^T t \end{bmatrix}, \tag{3.1.2}$$

donde $\bar{r}_1^T$ y $\bar{r}_2^T$ son el primer y segundo renglón de la matriz $R$, respectivamente. La matriz $R$ puede definirse por los ángulos de Euler $(\alpha, \beta, \gamma)$ como

$$R = R(\gamma)R(\beta)R(\alpha) = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}. \tag{3.1.3}$$

La matriz homografía se puede construir utilizando nueve parámetros. Sin embargo, debido a que es invariante a la escala, solo tenemos ocho grados de libertad. De acuerdo con el análisis de Diaz et al. [83], la matriz homografía se puede parametrizar como

$$G = \begin{bmatrix} k_{11} & k_{12} & k_{13} \\ 0 & k_{22} & k_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{21} & 0 \\ r_{12} & r_{22} & 0 \\ r_{13} & r_{23} & 1 \end{bmatrix}, \tag{3.1.4}$$

donde $k_{11}$ y $k_{22}$ son la longitud focal de la cámara, la cual induce un escalamiento en la imagen, $k_{12}$ es un valor de oblicuidad del píxel, $k_{13}$ y $k_{23}$ son los puntos principales de la imagen que inducen un desplazamiento en las direcciones $x$ y $y$, respectivamente. Por lo tanto, utilizando esta parametrización se cubren los ocho grados de libertad de la matriz $G$.

Sea $p = [p_1, p_2, \ldots, p_w]^T \in \Omega \subseteq \mathbb{R}^w$ un vector de tamaño $w \times 1$ que contiene los ocho paramétros necesarios para construir la matriz homografía utilizando la Ec. (3.1.4), para las matrices de rectificación $G_1, G_2 : \Omega \to \mathbb{R}$. La restricción epipolar para el caso rectificado, puede ser cuantificada para el conjunto de puntos correspondientes utilizando las Ecs. (3.1.1) y (3.1.4), como

$$J(p) = \sum_{i=1}^{n} \mathcal{H}[s_{1,n}(x,y)]^T G_1(p)^T F_c G_2(p) \mathcal{H}[s_{2,n}(x,y)], \tag{3.1.5}$$

donde $p$ es el vector que contiene los paramétros de las matrices de rectificación.

La Ec. (3.1.5) describe la función que satisface la rectificación de un par de imágenes estéreo. Sin embargo, existen múltiples combinaciones de matrices homografía que pueden cumplir con las restricciones descritas [59]. Por lo tanto, es necesario asegurar que las transformaciones proyectivas aplicadas a las imágenes generen la menor distorsión posible.

Sea $c_1$ el punto superior izquierdo, $c_2$ el punto superior derecho, $c_3$ el punto inferior derecho, $c_4$ el punto inferior izquierdo y $c_c$ el punto central de una imagen de referencia libre de degradación. Se puede describir la línea horizontal entre los puntos $c_1 - c_2$ y $c_3 - c_4$ en coordenadas homogéneas utilizando la Ec. (2.2.7), como

$$\begin{aligned} l_u &= \mathcal{H}[c_1] \times \mathcal{H}[c_2] \\ l_d &= \mathcal{H}[c_3] \times \mathcal{H}[c_4], \end{aligned} \tag{3.1.6}$$

donde $l_u$ y $l_d$ son paralelas.

La proyección de las líneas $l_u$ y $l_d$, generada por la matriz $G$ esta dada por

$$\begin{aligned} l_u^p &= (G^{-1})^T l_u \\ l_d^p &= (G^{-1})^T l_d. \end{aligned} \tag{3.1.7}$$

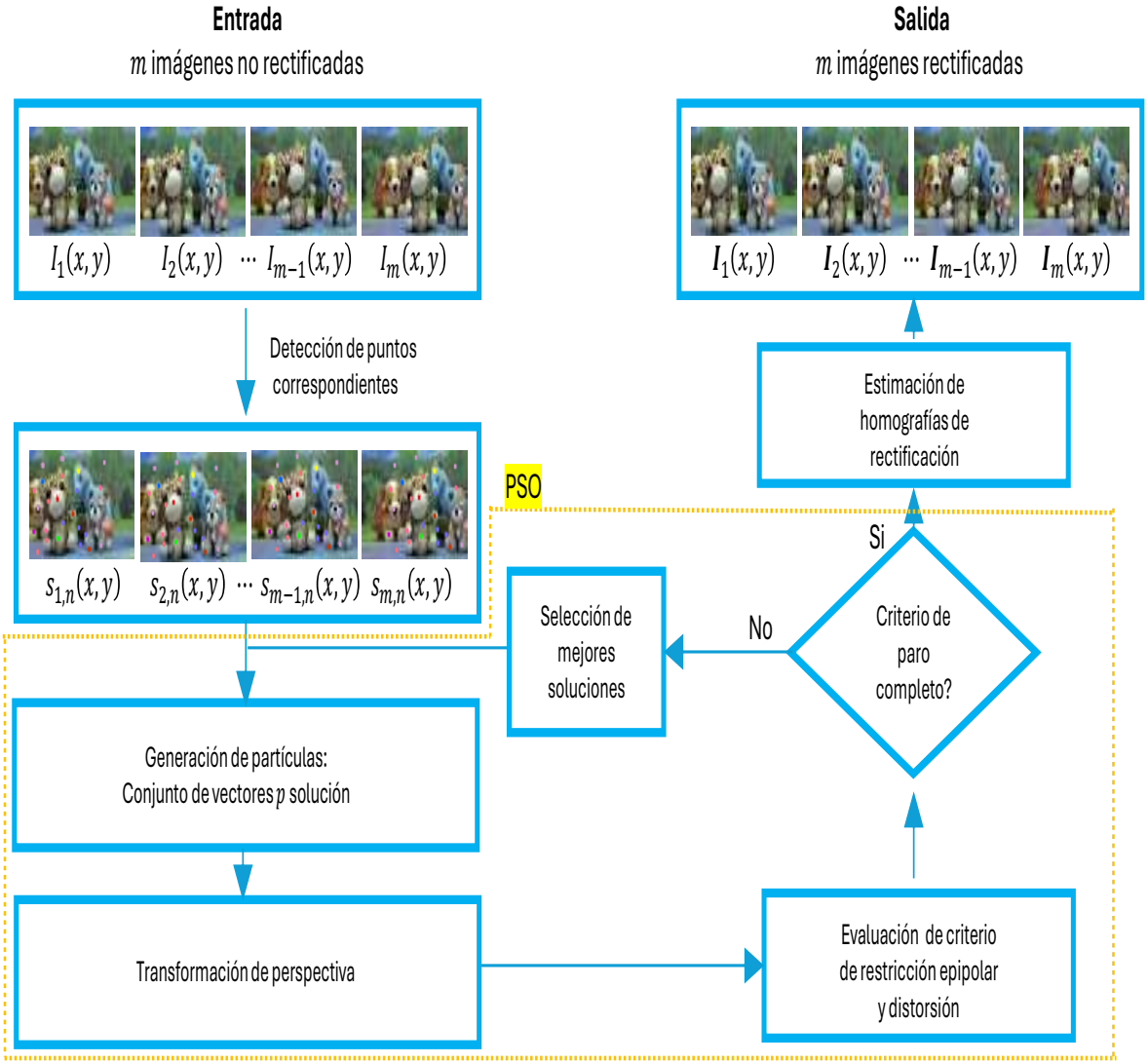



**Figura 3.1.1:** Diagrama a bloques del algoritmo STROMI propuesto para rectificación estéreo multi-ocular.

La proyección del punto central $c_c$, por la matriz $G$ se define como

$$c_c^p = G\mathcal{H}[c_c]. \tag{3.1.8}$$

Utilizando las Ecs. (3.1.7) y (3.1.8), se puede determinar el error de proyección como

$$e_p = |c_c - c_c^p|_2 + l_{d3}^p + l_{u3}^p, \tag{3.1.9}$$

donde $l_{d3}^p$ y $l_{u3}^p$, son el tercer elemento del vector proyectado de la línea $l_d$ y $l_p$, respectivamente.

Utilizando las Ecs. (3.1.5) y (3.1.9) se define la función objetivo como

$$\tau(p) = \xi\left(J(p)\right) + (1-\xi)\left(e_{1p}(p) + e_{2p}(p)\right), \tag{3.1.10}$$

donde $\xi$ es un valor escalar que regula el peso de los criterios a evaluar, $e_{1p}$ y $e_{2p}$ son los errores de proyección de $C1$ y $C_2$, respectivamente.

La Ec. (3.1.10) describe la función objetivo que permite construir el par de matrices homografía de rectificación. Debido a que el proceso de rectificación de imágenes se realiza para pares de imágenes, es necesario utilizar una imagen como referencia $I_1(x, y)$ e ir rectificando con su par $I_m(x, y) : m = 2, \ldots, m$, respectivamente, para obtener una aproximación de imágenes rectificadas, descritas como

$$\tau(p) = \sum_{j=2}^{m} \Big[ \xi \left( J(p) \right) + (1 - \xi) \left( e_{1p}(p) + e_{jp}(p) \right) \Big] \tag{3.1.11}$$

Los paramétros del vector $p$ para construir las homografías de rectificación se buscan iterativamente utilizando el método PSO [84], minimizando la función objetivo descrita por la Ec. (3.1.11).

La Fig. 3.1.1 presenta un diagrama a bloques del método propuesto para rectificación multi-ocular. Es importante mencionar que no es necesario conocer los parámetros intrínsecos y extrínsecos de la cámara para poder realizar el proceso de rectificación. Sin embargo, si se conocen los parámetros, éstos pueden ser utilizados para acotar el área de búsqueda del método.

## 3.2. Estimación de disparidad en imágenes estéreo

Como se ha mencionado en el capítulo 2.2.7, existen diferentes enfoques que presentan un buen desempeño para el problema de asociación estéreo. Sin embargo, aún se requiere explorar alternativas para mejorar la precisión y robustez obtenidas.

Se propone un método local robusto para la etapa de asociación estéreo basado en correlación morfológica adaptativa, en donde se utiliza un nuevo criterio de optimización llamado relación binaria de disimilitud a coincidencia (BDMR, por sus siglas en inglés *Binary Dissimilarity-to-Matching Ratio*). El método es diseñado para mejorar la precisión en la estimación en los bordes y regiones homogéneas de la imagen.

### 3.2.1. Asociación estéreo utilizando correlación morfologica adaptativa

Sean $w_1(x, y)$ y $w_2(x, y)$ ventanas de soporte de tamaño $S_w \times S_w$, con coordenadas centrales en el punto de interés $p(x_0, y_0)$ de las imágenes estéreo rectificadas $\mathbf{I}_1(x, y)$ e $\mathbf{I}_2(x, y)$, respectivamente. Estas ventanas se pueden representar por una descomposición binaria [62, 85] como

$$w_i(x, y) = \sum_{q=q_0}^{q_n} B_{iq}(x, y), \tag{3.2.1}$$

donde $q_0 = \text{mín}\{w_i(x, y)\}, q_N = \text{máx}\{w_i(x, y)\}$, y $B_{iq}(x, y)$ es una imagen binaria de $w_i(x, y)$ definida como

$$B_{iq}(x, y) = \begin{cases} 1, & \text{si } w_i(x, y) \geq q \\ 0, & \text{de lo contrario} \end{cases}. \tag{3.2.2}$$

Considerando las restricciones de la geometría epipolar en imágenes estéreo, la $BDMR$ para las ventanas $w_1(x,y)$ y $w_2(x,y)$, se define como

$$BDMR(k) = \frac{\sum_{q\in q_n}\sum_{x\in S_w}\sum_{y\in S_w}|B_{1q}(x,y) - B_{2q}(x-k,y)|}{\sum_{q\in q_n}\sum_{x\in S_w}\sum_{y\in S_w}|B_{1q}(x,y) + B_{2q}(x-k,y) - 1|}, \tag{3.2.3}$$

donde $k$ se define dentro del rango de búsqueda de disparidad.

Notese que el valor de $BDMR(k) = 0$ cuando $w_1(x,y) = w_2(x-k,y)$, y $BDMR(k) = \infty$ si $w_1(x,y) \neq w_2(x-k,y)$. Se requiere generar una medida de costo de coincidencia estéreo que minimice el valor de $BDMR$. Utilizando las propiedades del valor absoluto, la correlación no lineal que minimiza la Ec. (3.2.3) esta dada por

$$BDMR(k) = \frac{\Phi_{B_1} - \Phi_{B_2}(k) - \frac{2}{(q_n-q_0)s_w^2}\sum_{q\in q_n}\sum_{x\in S_w}\sum_{y\in S_w} MIN\{B_{1q}(x,y), B_{2q}(x-k,y)\}}{1 + \Phi_{B_1} - \Phi_{B_2}(k) - \frac{2}{(q_n-q_0)s_w^2}\sum_{q\in q_n}\sum_{x\in S_w}\sum_{y\in S_w} MAX\{B_{1q}(x,y), B_{2q}(x-k,y)\}}, \tag{3.2.4}$$

donde

$$\begin{aligned}\Phi_{B_1} &= \frac{1}{(q_n-q_0)S_w^2}\sum_{q\in q_n}\sum_{x\in S_w}\sum_{y\in S_w} B_{1q}(x,y),\\ \Phi_{B_2}(k) &= \frac{1}{(q_n-q_0)S_w^2}\sum_{q\in q_n}\sum_{x\in S_w}\sum_{y\in S_w} B_{2q}(x-k,y),\end{aligned} \tag{3.2.5}$$

Al maximixar la Ec. (3.2.4) obtenemos el mínimo valor de $BDMR$ como

$$C(k) = \frac{\sum_{x\in S_w}\sum_{y\in S_w} MIN\{\widehat{w}_1(x,y), \widehat{w}_2(x-k,y)\}}{\frac{1}{QS_w^2} + \sum_{x\in S_w}\sum_{y\in S_w} MAX\{\widehat{w}_1(x,y), \widehat{w}_2(x-k,y)\}}, \tag{3.2.6}$$

donde

$$\begin{aligned}\widehat{w}_1(x,y) &= \sum_{q\in Q} B_{1(q+p_{\Delta q})}(x,y),\\ \widehat{w}_2(x-k,y) &= \sum_{q\in Q} B_{2(q+p_{\Delta q})}(x-k,y),\end{aligned} \tag{3.2.7}$$

son ventanas pre-procesadas de las imágenes estéreo de entrada utilizando una descomposición binaria adaptativa, con una cuantización dada por

$$p_{\Delta_q} = \frac{3\sigma_{w1}}{Q}, \tag{3.2.8}$$

donde $Q = (q_n - q_0)$ es el número de niveles de cuantización y $\sigma_{w1}$ es la desviación estándar de $w_1(u,v)$ con respecto a $p(x_0,y_0)$.

El proceso de asociación estereo utilizando las Ecs. (3.2.7) y (3.2.8) es realizado para el punto de referencia $p(x_0,y_0)$ en la imagen $\mathbf{I}_1(x,y)$. Finalmente, el valor de disparidad es estimado como

$$\widehat{d}_{1j,n}(x_0,y_0) = \arg\max_k\{C(k)\}, \quad k \in [k_{min}, k_{max}]. \tag{3.2.9}$$

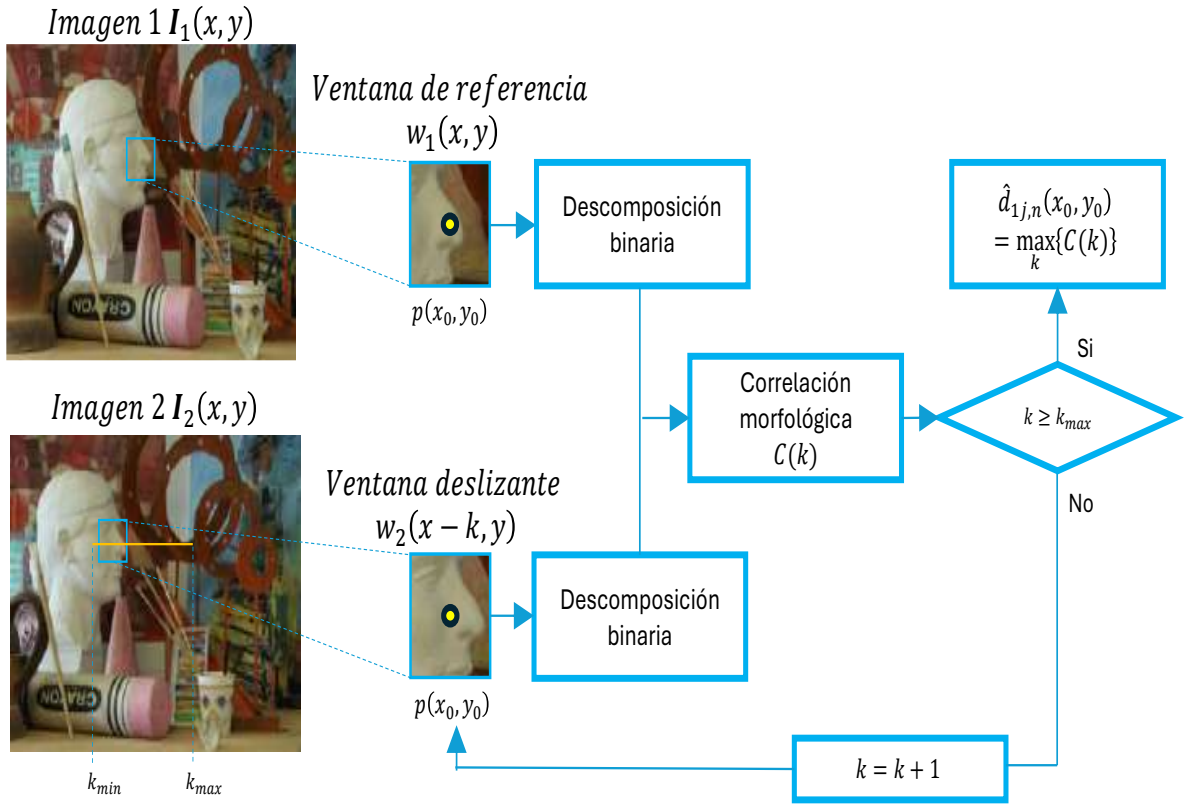



**Figura 3.2.1:** Diagrama a bloques del método propuesto para estimación de disparidad. Para cada píxel del par de imágenes se realiza el proceso de descomposición binaria y correlación morfológica con la ventana deslizante en $k$ píxeles. La disparidad estimada corresponde al valor de correlación máximo obtenido.

Una ventana adaptativa es propuesta para robustecer la respuesta del proceso de estimación. Primero, la ventana de soporte $w_1(x,y)$ de tamaño $S_w \times S_w$ es construida, donde $S_w = 2s+1$, y $s$ es calculada de manera adaptativa como

$$s = (S-1)\exp -\alpha \frac{\mathbf{I}_1(x_0,y_0)}{\sigma_S^2}, \tag{3.2.10}$$

donde $\alpha$ es un valor escalar, $S$ es un valor escalar que define el tamaño máximo de la ventana de soporte, y $\sigma_S^2$ es la desviación estándar de los valores de intensidad de $w_1(x,y)$ con respecto a $p(x_0,y_0)$. La Fig. 3.2.1 presenta el diagrama a bloques del método propuesto.

## 3.3. Método de post-procesamiento de mapa de disparidad

El mapa de disparidad se puede obtener con el método descrito en la sección 3.2 para un par de imágenes estéreo. Sin embargo, debido a la presencia de oclusiones, baja textura en los objetos, u objetos repetidos en la escena, el mapa de disparidad estimado puede contener errores de estimación. La finalidad de la etapa de postprocesamiento es

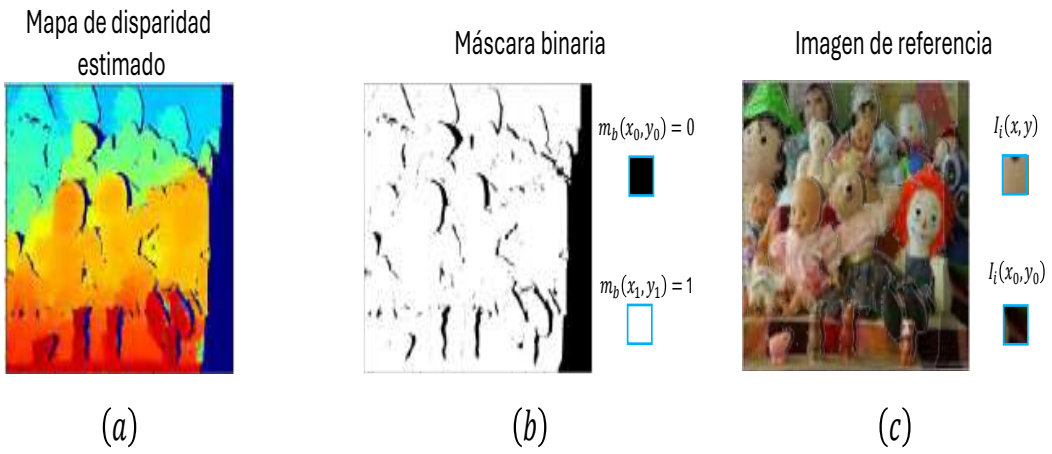



**Figura 3.3.1:** Proceso de postprocesamiento para reemplazar el valor de disparidad estimado incorrectamente con valores de disparidad verificados. El método utiliza información de (a) mapa de disparidad estimado validado, (b) máscara binaria, y (c) la imagen de referencia.

determinar la fiabilidad de cada punto asociado estimado y descartar los valores estimados erróneamente.

Consideremos una máscara binaria $m_b(x, y)$, la cual tendrá valores de 1 en los píxeles con correspondencia válida y valores de 0 para las asociaciones erróneas [86]. La asociación de píxeles es verificada de forma efectiva utilizando la informacion de disparidad estimada del la imagen izquierda a derecha y de derecha a izquierda, respectivamente. Esta la verificación, puede realizarse como

$$m_b(x,y) = \begin{cases} 1, & \text{si} |\widehat{d}_{12,n}(x,y) - \widehat{d}_{21,n}(x - \widehat{d}_{12,n}(x,y), y)| < \Delta_\delta, \\ 0, & \text{de otra manera,} \end{cases} \tag{3.3.1}$$

donde $\Delta_\delta$ es un valor escalar que pondera la diferencia de disparidad permitida, comúnmente $\Delta_\delta \leq 3$. Esta simple verificación permite encontrar puntos asociados incorrectos ocasionados por oclusiones o cualquier perturbación en la imagen. Los puntos mal asociados deben de ser remplazado con valores estimados de manera correcta, por lo cual proponemos un método de postprocesamiento para rellenado de huecos de disparidad basado en una función bayesiana (DHB, por sus siglas en inglés *Disparity Hole-filling using Bayesian function*).

### 3.3.1. Método propuesto de interpolación de disparidad basada en una función Bayesiana (DHB)

Implementando el método de consistencia estéreo descrito en la Ec. (3.3.1) para el mapa de disparidad estimado $\widehat{d}_{12,n}(x, y)$, podemos observar que el resultado puede dividirse en una imagen con dos clases: $m_b(x_1, y_1) = 1$ donde la disparidad ha sido estimada de manera correcta y $m_b(x_0, y_0) = 0$ donde hay errores de estimación.

Nuestro interés es remplazar los valores de las coordenadas de $m_b(x_0, y_0)$ con información de valores de disparidad de $m_b(x_1, y_1)$. Para desarrollar esta tarea, consideremos la

probabilidad *a priori* de que un punto verificado con coordenadas $(x, y)$ pueda ser utilizado para remplazar el valor erróneo en la coordenada $(x_0, y_0)$. Esta probabilidad, puede especificarse como

$$P(x,y) = \frac{1}{\sigma_c^2\sqrt{2\pi}} \exp\left[-\frac{(x-x_0)^2 + (y-y_0)^2}{2\sigma_c^2}\right], \tag{3.3.2}$$

donde se asume una distribución normal con varianza $\sigma_c^2$. Adicionalmente, la función de densidad de probabilidad de que un punto en la imagen con valor de intensidad $I_i(x, y)$ tenga un valor de disparidad similar en las coordenadas $(x_0, y_0)$, está dada por

$$P(I_1(x,y)|\widehat{d}(x_0,y_0)) = \frac{1}{\sigma_t^2\sqrt{2\pi}} \exp\left[-\frac{(I_i(x,y) - I_i(x_0,y_0))^2}{2\sigma_t^2}\right], \tag{3.3.3}$$

donde $\sigma_t^2$ es la varianza del valor de intensidad. De acuerdo con la teoría de Bayes, la probabilidad *a posteriori* de que un punto en la imagen con disparidad $\widehat{d}_{12,n}(x, y)$ e intensidad $I_i(x, y)$ puede remplazar el valor de la disparidad desconocida $m_b(x_0, y_0)$ dado $I_i(x_0, y_0)$ es

$$P(\widehat{d}_{12,n}(x,y)|I(x_0,y_0)) = \frac{P(I_i(x,y)|\widehat{d}_{12,n}(x_0,y_0))P(x,y)}{P(I_i(x,y))}, \tag{3.3.4}$$

donde $P(I_i(x, y))$ es la función de densidad de probabilidad *a priori* del valor de intensidad $I_i(x, y)$. El valor de disparidad en las coordenadas $(x_1, y_1)$ con mayor probabilidad de ser correcto, puede obtenerse como

$$(\widehat{x}_1, \widehat{y}_1) = \underset{(x,y)}{\operatorname{argmax}} \left\{ P(\widehat{d}(x,y)|I_i(x_0,y_0)) \right\}. \tag{3.3.5}$$

Al sustituir los valores de las funciones de densidad de probabilidad y aplicando la función de logaritmo en la Ec. (3.3.5), se obtiene

$$(\widehat{x}_1, \widehat{y}_1) = \underset{(x_1,y_1)}{\operatorname{argmax}} \left\{ \frac{(x-x_0)^2 + (y-y_0)^2}{2\sigma_c^2} + \frac{(I_i(x,y) - I_i(x_0,y_0))^2}{2\sigma_t^2} \right\}. \tag{3.3.6}$$

Es importante notar, que al aplicar el estimador de la Ec. (3.3.6) a los elementos con estimación errónea en el mapa de disparidad inicialmente estimado, podemos recuperar la información faltante y mejorar la precisión en la estimación. La Fig. 3.3.1 presenta las imágenes de la información utilizada para interpolar los valores del mapa de disparidad verificado.

## 3.4. Método de estimación de disparidad usando multilínea base

En este trabajo de tesis, se propone un enfoque de línea base ajustable para un sistema multi-ocular estéreo (BAMS, por sus siglas en inglés *Baseline Adjustable multi-ocular Stereo*) para estimación de disparidad y reconstrucción tridimensional. El enfoque propuesto permite incrementar la resolución de la información tridimensional estimada, disminuyendo la probabilidad de estimaciones erróneas.

### 3.4.1. Estimación de mapas de disparidad utilizando estéreo multi-ocular de línea base ajustable (BAMS)

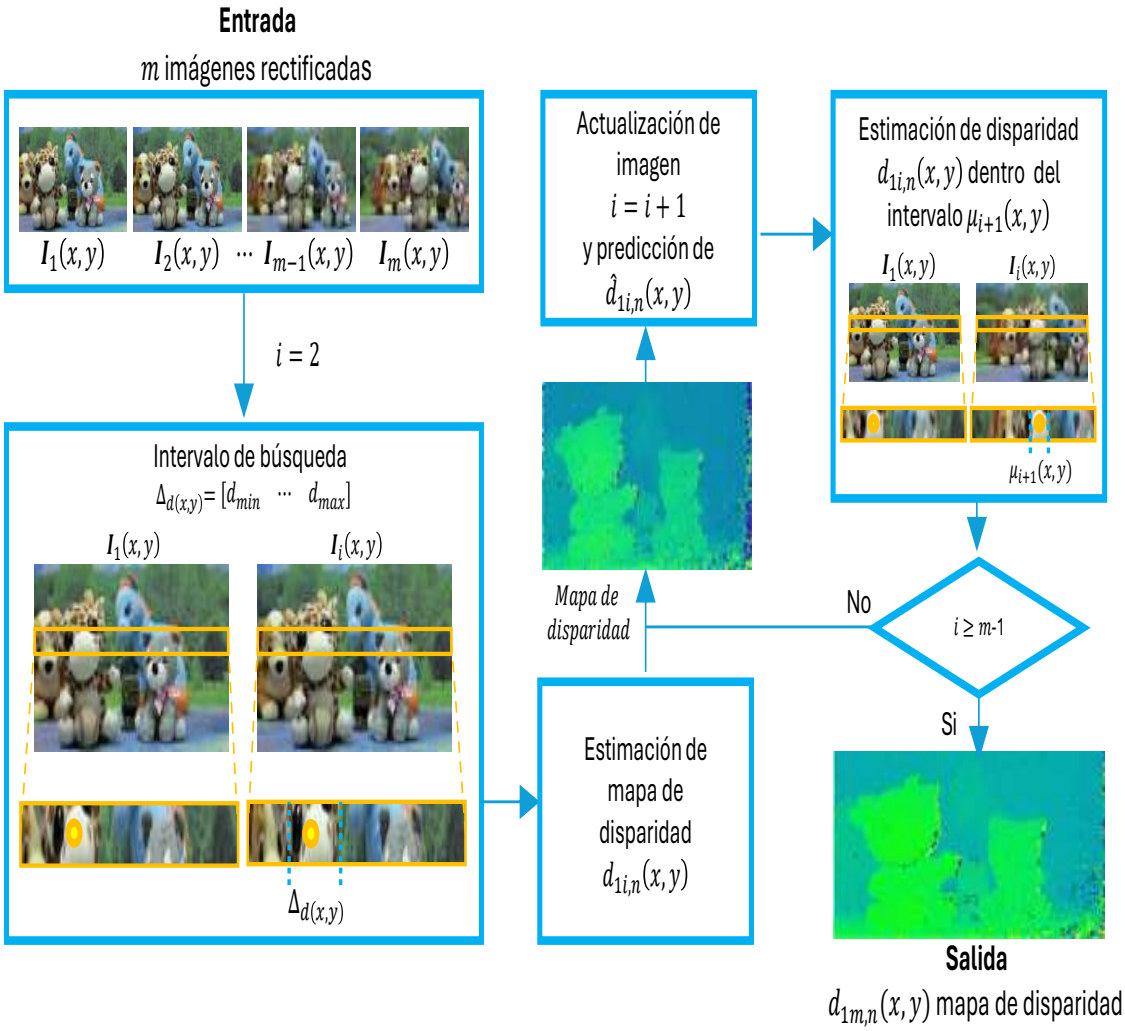



**Figura 3.4.1:** Diagrama a bloques del método de estimación de disparidad utilizando BAMS. Él método puede ser utilizado para múltiples imágenes estéreo. Inicialmente, se requiere la estimación completa del mapa de disparidad con menor distancia de línea base. Los mapas subsecuentes se estiman utilizando un intervalo acotado por las disparidades previamente estimadas.

El objetivo es encontrar un rango de búsqueda $\delta$ reducido, que permita realizar la búsqueda de los valores de disparidad en un intervalo confiable para reducir los errores de estimación. Considerando las cámaras $C_1$ y $C_{i+1}$ para $i = 1$, la estimación del mapa de disparidad $d_i(x, y)$ se realiza en el intervalo de búsqueda

$$\Delta_{d_i(x,y)} = [d_{min}, d_{min+1}, \ldots, d_{max}] \,. \tag{3.4.1}$$

Para el valor estimado de disparidad $d_i(x, y)$, la predicción *a priori* de la disparidad entre las cámaras $C_1$ y $C_{i+2}$ es

$$\hat{d}_{i+1}(x, y) = d_i(x, y) \left( \frac{B_{i+1}}{B_i} \right) . \tag{3.4.2}$$

Sea $\hat{d}_{i+1}(x,y)$ una variable aleatoria con función de distribución de probabilidad $\mathcal{N}(0,1)$, definida como

$$P(z) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} \exp\left(\frac{z^2}{2}\right) dz, \tag{3.4.3}$$

donde

$$z = \frac{d_{i+1}(x,y) - \mu_d}{\sigma_d^2}, \tag{3.4.4}$$

$\mu_d$ es el valor de disparidad a estimar con la mayor probabilidad, y $\sigma_d^2$ es la varianza. El valor de $\mu_d$ se encuenta en el intervalo

$$P(\hat{d}_{i+1}(x,y) + \alpha\sigma_d \geq \mu_d \geq \hat{d}_{i+1}(x,y) - \alpha\sigma_d), \tag{3.4.5}$$

donde $\alpha$ es un valor escalar que regula el intervalo de confianza.

Sea $\Delta_{d(x,y)} = [d_{min}, d_{min+1}, \ldots, \hat{d}_{max}]$ el intervalo completo de búsqueda de algún par de cámaras. Para un valor $\hat{d}_i(x,y)$ previamente estimado y pronosticado $\hat{d}_{i+1}(x,y)$ se puede actualizar el intervalo de búsqueda como

$$\Delta_{d(x,y)} = [d_{min}, d_{min+1}, \ldots, \hat{d}_{i+1}(x,y)], \tag{3.4.6}$$

donde $\hat{d}_{i+1}(x,y) \geq d_{min}$, $\mu_{\Delta d} = \frac{d_{min}+\hat{d}_{i+1}(x,y)}{2}$ y $\sigma_d^2 = \frac{1}{\hat{d}_{i+1}(x,y)-d_{min}} \sum_{i=1}^{\hat{d}_{i+1}} (\Delta_d(i) - \mu_{\Delta d})$.

El intervalo de búsqueda para $\mu_{i+1}(x,y)$, se define como

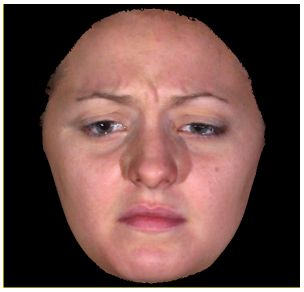
$$[\hat{d}_{i+1}(x,y) - max\{\alpha\sigma_d, d_{min}\}, \hat{d}_{i+1}(x,y) + min\{\alpha\sigma_d, d_{max}\}]. \tag{3.4.7}$$

Utilizando la Ec. (3.4.7), podemos limitar la búsqueda de disparidad de las cámaras con mayor línea base a tan solo unos cuantos píxeles, definidos por la estimación de la disparidad con menor línea base. La Fig. 3.4.1 presenta el diagrama a bloques del método propuesto para estimación de disparidad utilizando multilínea base.

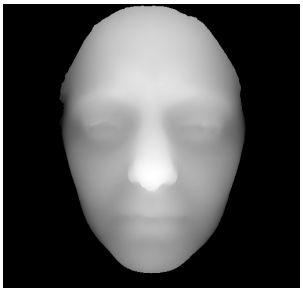
## 3.5. Extracción de características y clasificación de expresiones del rostro basado en CNN

La robustez de las CNN en la tarea de clasificación es elevada siempre y cuando los datos de prueba sean similares a los datos de entrenamiento. Sin embargo, la tarea de clasificación se vuelve más compleja cuando las imágenes presentan cambios en la pose. Una solución a este problema es el uso de información adicional generada a partir de transformaciones afines, rotaciones, reflexiones, y escalamientos de las imágenes de originales [87]. Sin embargo, esta solución no es suficiente para tener una mejora significativa y evitar el problema del sobre ajuste [88]. El uso de la información tridimensional del rostro proporciona mayor robustez ante los cambios de pose e iluminación [47]. La información tridimensional del rostro se puede recuperar utilizando imágenes, por medio de métodos de correspondencia estéreo [89].

En esta sección, se presenta un método de clasificación de expresiones faciales basado en redes neuronales convolucionales. El enfoque propuesto utiliza imágenes $2D$ del rostro y mapas de profundidad $3D$ como entradas. Se incorpora un bloque residual que emplea regiones de soporte definidas mediante una descomposición binaria de puntos de interés facial, sustituyendo la etapa de salto convencional. Este diseño refuerza las características $3D$ extradas en las capas convolucionales para mejorar la robustez del modelo.



(a)



(b)

**Figura 3.5.1:** Información obtenida de un sistema de digitalización multi-ocular. (a) Imagen de entrada en espacio de color RGB. (b) Mapa de profundidad en escala de grises.

### 3.5.1. Método propuesto para clasificación de expresiones del rostro utilizando CNN

Sea $I_t(x, y, 3)$ una imagen en el espacio de color RGB e $I_p(x, y)$ un mapa de profundidad del rostro obtenido de una digitalización, como se muestra en la Fig. 3.5.1. El objetivo es utilizar la combinación de información de textura (2D) e información geométrica (3D) del rostro como información de entrada a una CNN para extraer y clasificar expresiones del rostro.

La información del mapa de profundidad $I_p(x, y)$ es preprocesada para remover discontinuidades y ruido utilizando el método propuesto por Sui et al. [90]. Primero, se remueven los valores atípicos del mapa de profundidad sí se cumple con

$$\max_{(x',y')\epsilon N(x,y)} |I_p(x, y) - I_p(x', y')| > 0.6\sigma_N, \tag{3.5.1}$$

donde $N(x, y)$ es una ventana de soporte, $I_p(x, y)$ es el elemento central de $N(x, y)$, $I_p(x', y')$ son los elementos de $N(x, y)$ alrededor del elemento central, y $\sigma_N$ es el valor de desviación estándar de $N(x, y)$. Posteriormente, se realiza un rellenado de información ausente utilizando interpolación bicúbica considerando información de 16 píxeles alrededor del elemento ausente. Finalmente, se elimina el ruido del mapa utilizando un filtro gaussiano como [91]

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{3.5.2}$$

con una ventana de tamaño $3 \times 3$ y $\sigma = 1$.

Se utilizan áreas clave para reforzar la información de las características detectadas, y las características detectadas del rostro, donde se generan movimientos al realizar gesticulaciones, como lo son el contorno de los ojos, la nariz, los labios y las cejas. Primero, se realiza la detección de 51 puntos característicos del rostro en $I_t(x, y, 3)$ utilizando el método Dlib [74], descrito en la sección 2.3. Estos puntos característicos, se utilizan para construir máscaras de soporte $\mathcal{W}_L$ de tamaño $7 \times 7$ alrededor de los puntos detectados por medio de descomposición binaria umbralizada [83], obtenidas como

$$\mathcal{W}_L(L) = \sum_{q \in Q} B_{\Delta_q}(x, y), \tag{3.5.3}$$

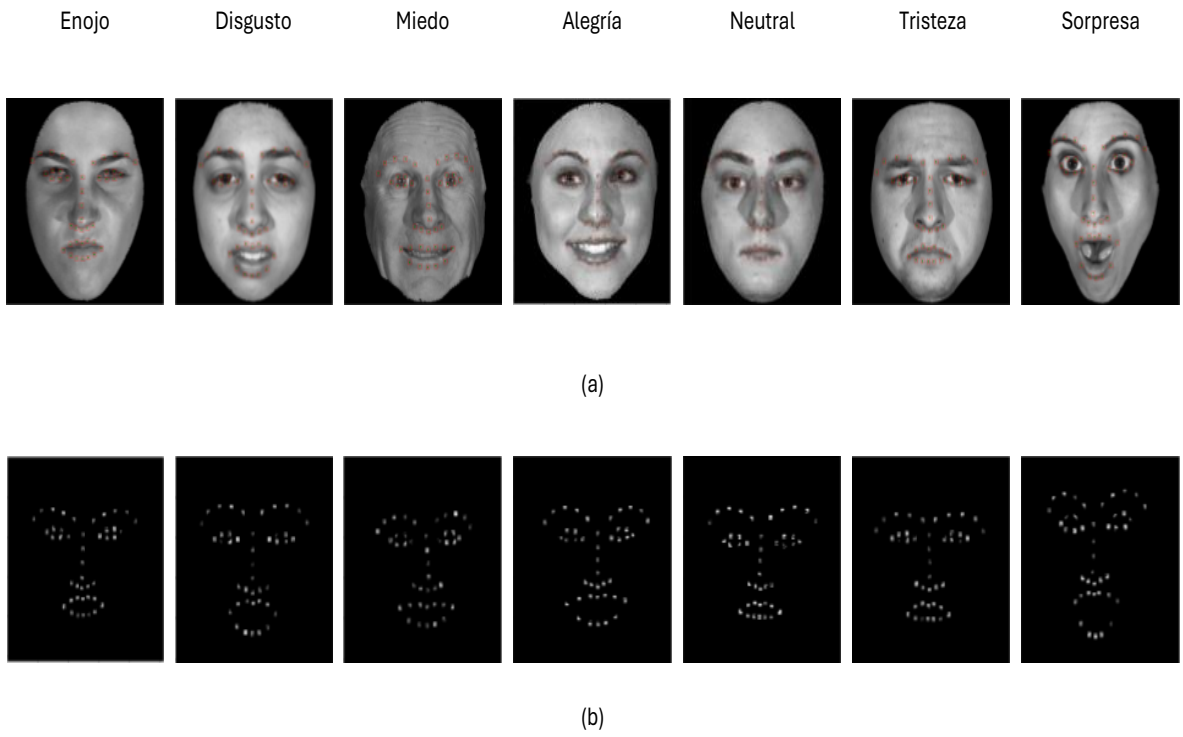



**Figura 3.5.2:** Imágenes de expresiones del rostro. (a) detección de puntos característicos del rostro (puntos rojos) para cada expresión del rostro. (b) Ventanas de soporte alrededor de puntos característicos del rostro.

donde $L$ es el punto característico detectado, $\Delta_q = \frac{3\sigma_{\mathcal{W}_L}}{Q}$ es la cuantización , $Q$ es el número de niveles de cuantización, $\sigma_{\mathcal{W}_L}$ es la desviación estándar de la región de interés. La Fig. 3.5.2, presenta los puntos característicos detectados del rostro y las máscaras de soporte.

La arquitectura propuesta fusiona características $2D + 3D$ para clasificación de expresiones del rostro. Además, esta arquitectura se compone por dos bloques de extracción de características, como se muestra en la Fig. 3.5.3.

Para extraer las características $2D$ de $I_t(x, y, 3)$, el primer bloque utiliza la arquitectura de CNN VGG16 propuesta por Symonya et al. [92]. Esta arquitectura está compuesta por 13 capas convolucionales con filtros de tamaño $3 \times 3$ activadas por la función $ReLU$, 5 capas de agrupación con la función *max pooling*, y 3 capas completamente conectadas. Sin embargo, para el proceso de extracción solo son utilizadas las capas convolucionales y de agrupación. El tamaño de los tensores de entrada y salida es de $(224 \times 224 \times 3)$ y $(7 \times 7 \times 512)$, respectivamente. El número de filtros utilizados por las primeras tres capas convoluciones es de 64, 128, 256, respectivamente, y las últimas dos capas utilizan 512 filtros.

En el segundo bloque se realiza la extracción de características de $I_p(x, y)$, con una red convolucional de 6 capas. Los mapas de profundidad contienen información más suavizada que la información de textura bidimensional [90], por lo cual el tamaño de los primeros dos filtros convolucionales son de $7 \times 7$ y $5 \times 5$, respectivamente. El tensor resultante es utilizado como entrada de un bloque residual [82], donde remplazamos la entrada de

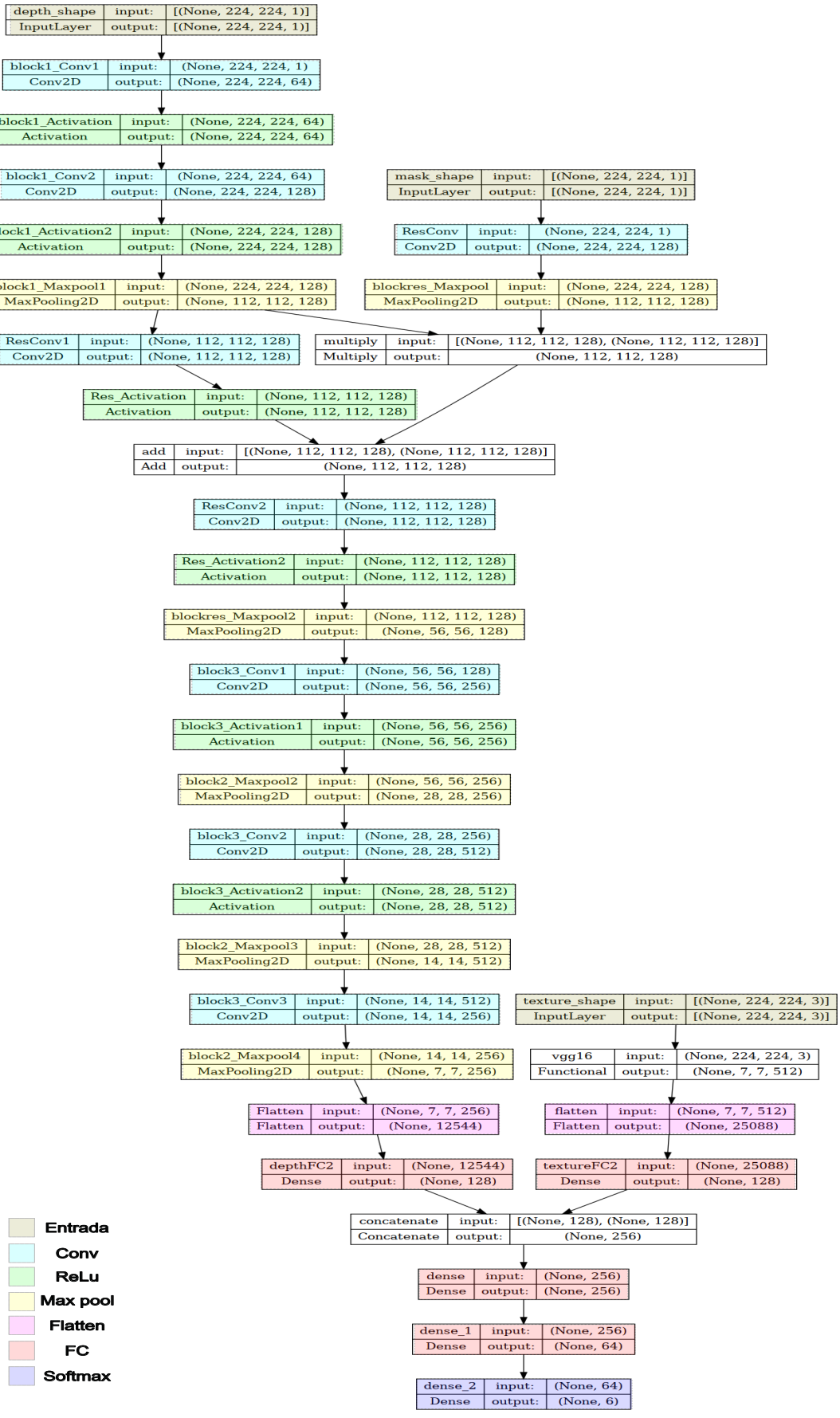



**Figura 3.5.3:** Arquitectura propuesta de clasificación de expresiones faciales.

salto de conexión por la multiplicación de la máscara $\mathcal{W}_L$ para robustecer los pesos en las zonas de interés [93]. El tensor resultante del bloque residual es convolucionado por filtros de tamaño $(3 \times 3)$ y activados por la función ReLU. El tensor de entrada es de tamaño $(224, 224, 1)$, mientras que la salida es de $(7 \times 7 \times 256)$.

Los tensores resultantes de los bloques de extracción de características son transformados en vectores en la capa completamente conectada $F_{c2}$ y concantenados. El vector resultante es reducido en las capas completamente conectadas $F_{c3}$ y $F_{c4}$ a un vector de tamaño $64 \times 1$. Por último, las características del vector resultante $F_{c4}$ son clasificadas utilizando la función *softmax*, para esta arquitectura se utilizarón 6 clases de emociones.

Capítulo 4

# Experimentos y resultados

En esta sección se presentan y se discuten los resultados obtenidos al implementar los métodos propuestos descritos en la sección 3. Los experimentos descritos en este capítulo fueron implementados utilizando el lenguaje de programación Python 3.10.7, en una computadora personal con un procesador *Intel Core* I5, con 16 GB de memoria RAM y sistema operativo *Windows* 10. Cada método fue evaluado y comparado con métodos reportados en la literatura científica, utilizando imágenes de diferentes bases de datos para evaluar diferentes aspectos del desempeño y funcionalidad del enfoque abordado.

## 4.1. Evaluación del método STROMI

En esta sección se presentan los resultados del desempeño del método propuesto de rectificación descrito en la sección 3.1. Para propósitos de evaluación, se utilizaron 40 imágenes binoculares de la base de datos *Media CommLab Real Stereo* (MCL-RS) [94] e IRIA Syntim [95]. Estas bases de datos contienen imágenes estéreo no rectificadas de ambientes interiores y exteriores con diferentes niveles de rotación, desplazamiento y ángulo de visión como se muestra en la Fig. 4.1.1. Los resultados obtenidos son comparados con aquellos obtenidos con los siguientes dos métodos existentes de rectificación estéreo: Fusiello et al. [96] y rectificación de fase estéreo (SPR, por sus siglas en inglés *Stereo Phase Rectification*) [97].

La precisión del método propuesto se mide en términos del error vertical $err_v$ [96], definido como

$$err_v = \frac{1}{N}\sum_{q}^{N} |\mathbf{s}_{1,q}(y) - \mathbf{s}_{2,q}(y)|, \tag{4.1.1}$$

donde $\mathbf{s}_{1,q}(y)$ y $\mathbf{s}_{2,q}(y)$ son las ordenadas $y$ de los puntos correspondientes transformados por las homografías $G_1$ y $G_2$, respectivamente, y $N$ es el número de puntos correspondientes. Adicionalmente, la distorsión de las imágenes rectificadas se cuantifica en términos del ángulo de ortogonalidad $\theta$ y relación de aspecto *Asp.Rat.* definido por Mallone et al. [98].

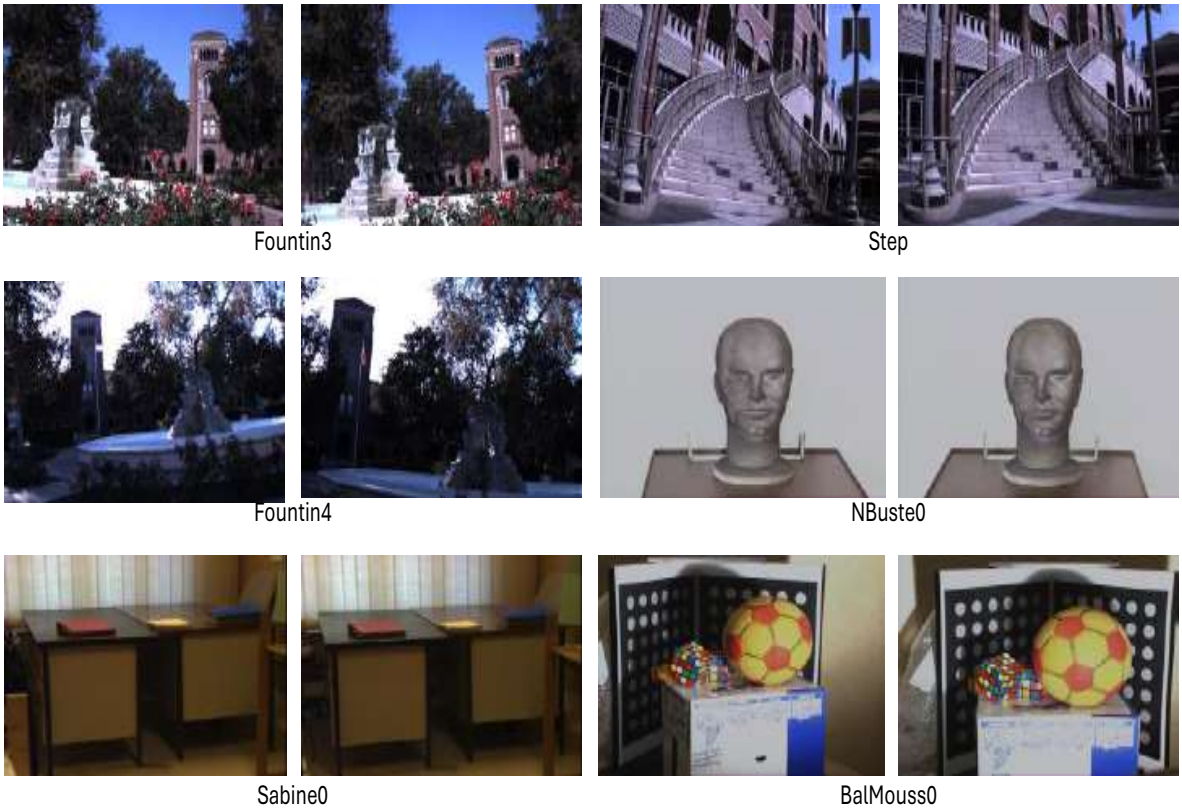



**Figura 4.1.1:** Ejemplos de imágenes estéro no rectificadas de la base de datos INRIA Sytim y MCL-RS, mostrando ambientes interiores y exteriores, diferentes niveles de iluminación, y pose de cámaras.

### 4.1.1. Evaluación del método STROMI utilizando dos cámaras

En el primer experimento, un conjunto de puntos correspondientes del par de imágenes estéreo $\{s_{1,n}(x,y) \leftrightarrow s_{2,n}(x,y)\}$, es estimado por medio del algoritmo SIFT [99]. A continuación, los puntos estimados son filtrados por medio del método de consenso de muestra aleatoria (RANSAC, por sus siglas en inglés *Random Sample Consensus*) [100] para eliminar valores atípicos. Posteriormente, se genera una población de partículas de forma aleatoria y con distribución uniforme, que representan soluciones candidatas de los paramétros necesarios para generar el par de transformadas proyectivas $G_1, G_2$ de rectificación. Cada solución candidata se evalua utilizando la Ec. (3.1.10) considerando los puntos correspondientes filtrados. Los valores resultantes del proceso de evaluación, permiten actualizar la posición y velocidad de las partículas de manera iterativa hasta cumplir el criterio de paro. Finalmente, la mejor solución candidata $p$ es seleccionada para construir el par de matrices de rectificación $G_1$ y $G_2$. Las matrices obtenidas son utilizadas en las imágenes estéreo como lo indica la Ec. (2.2.12).

La configuración de los parámetros del método propuesto es la siguiente: número de puntos correspondientes $n = 100$, población de partículas $N_P = 500$, coeficiente inercial $w = 0.9$, coeficientes de aceleración cognitiva y social $c_1 = 0.5$, $c_2 = 0.3$, respectivamente. Los coeficientes $c_1$ y $c_2$ fueron seleccionados para converger al menor valor de $err_v$, como lo muestra la Tabla 4.1.

**Cuadro 4.1:** Valores de $err_v$ obtenidos en la rectififcación de imágenes estéreo al realizar variaciones en los coeficientes $c_1$ y $c_2$ del método PSO. Los valores tienen influencia en la convergencia del método a una solución candidata.

| | | $C_1$ | | | | |
|---|---|---|---|---|---|---|
| | | *0.5* | *1.0* | *1.5* | *2.0* | *2.5* |
| $C_2$ | *0.3* | 0.26 | 1.49 | 5.18 | 6.58 | 9.97 |
| | *0.5* | 0.36 | 0.80 | 5.21 | 6.85 | 7.4 |
| | *1.0* | 2.84 | 6.42 | 6.22 | 7.28 | 6.82 |
| | *1.5* | 5.3 | 6.11 | 8.71 | 6.8 | 7.58 |
| | *2.0* | 7.41 | 6.19 | 6.88 | 8.22 | 8.9 |
| | *2.5* | 8.32 | 9.12 | 10.76 | 10.87 | 11.4 |

La Fig. 4.1.2 presenta imágenes rectificadas de los métodos evaluados. Los resultados estadísticos obtenidos de las 40 imágenes se presentan en la Tabla 4.2 y la Fig. 4.1.3. Los resultados obtenidos muestran que el método propuesto genera la menor distorsión y la mayor precisión durante el proceso de rectificación; además, presenta una desviación estándar baja para ambas bases de datos en comparación con los métodos existentes evaluados. El método de Fusiello presenta un buen desempeño para ambas bases de datos en términos de $err_v$, por el contrario, genera mayor distorsión a las imágenes. El método SPR presenta el peor desempeño de los métodos considerados en términos de $err_v$, sin embargo, induce una baja distorsión.

**Cuadro 4.2:** Resultados estadísticos de rectificación (media y desviación estándar) de los métodos evaluados en términos de $err_v$ (en píxeles), $\theta$ (en grados) y Asp. Rat. utilizando imágenes de la base de datos MCL-RS y Syntim.

| **Rectificación** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Fusiello et al.** | | **SPR** | | **Propuesto** | |
| **Base de datos** | **Medida de desempeño** | **Media** | **St. Dev.** | **Media** | **St. Dev.** | **Media** | **St. Dev.** |
| MCL-RS | $err_v$ | 0.84 | 0.25 | 0.86 | 0.31 | **0.68** | 0.14 |
| | $\theta(90°)$ | 95.4 | 6.83 | 87.9 | 11.9 | **90.05** | 0.08 |
| | Asp. Rat (1) | 1.2 | 0.16 | 1.01 | 0.05 | **0.99** | 0.002 |
| Syntim | $err_v$ | 0.79 | 1.28 | 1.15 | 0.61 | **0.78** | 0.51 |
| | $\theta(90°)$ | 92.7 | 6.25 | 86 | 4.84 | **89.9** | 0.54 |
| | Asp. Rat. (1) | 1.04 | 0.09 | 0.96 | 0.05 | **0.99** | 0.008 |

Es importante notar, que el método propuesto penaliza la distancia entre los puntos centrales de la imagen de entrada y la imagen rectificada, manteniendo la horizontalidad en los extremos superior e inferior de la imagen cumpliendo la restricción epipolar para el caso rectificado.

Por último, los resultados demuestran que el método propuesto es robusto a cambios severos de pose de la cámara, iluminación y textura en ambientes exteriores e interiores. Sin embargo, este método está planteado para el modelo de cámara *pinhole*, por lo que puede presentar resultados inconsistentes al utilizar cámaras con distorsión radial.

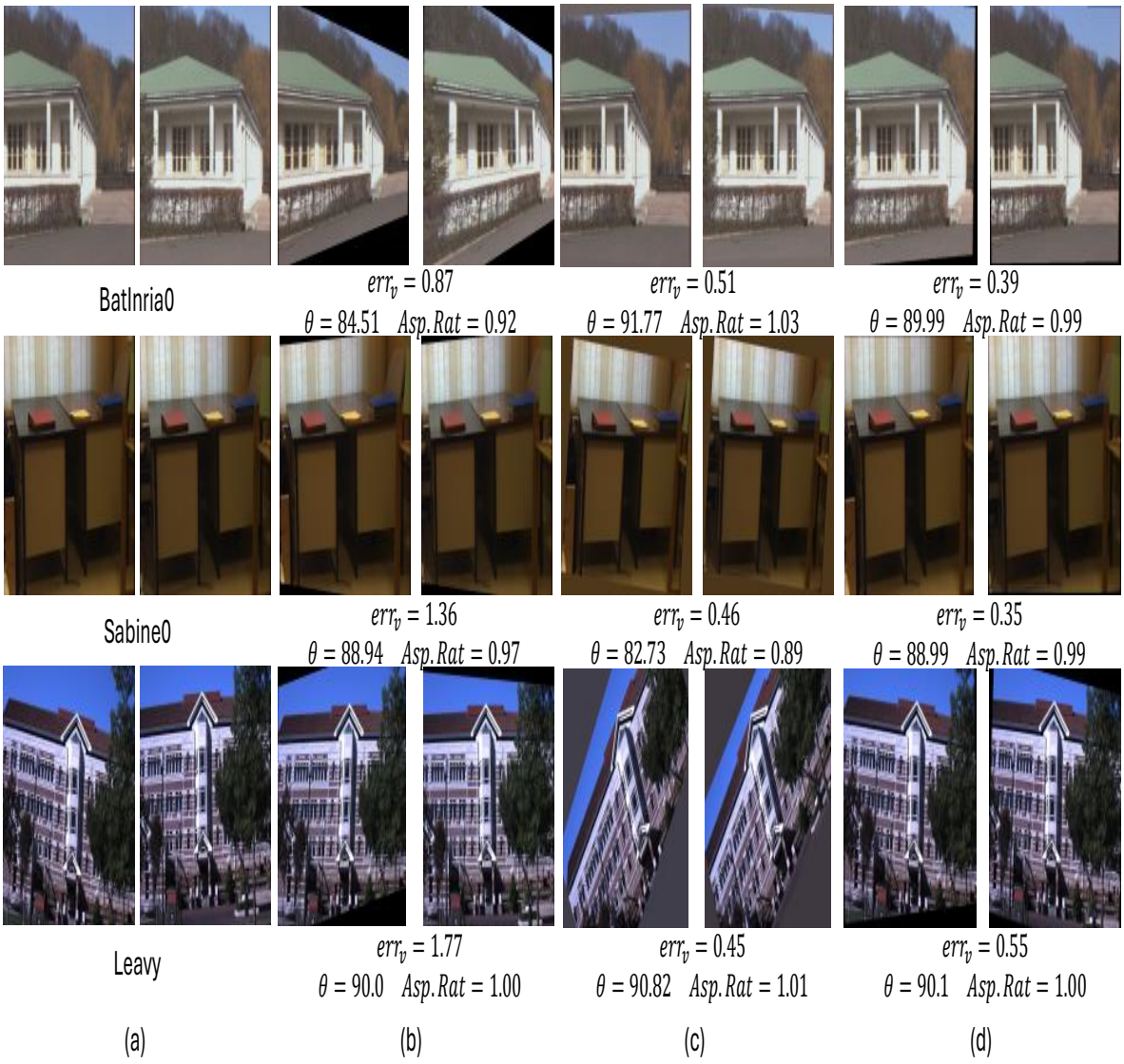



**Figura 4.1.2:** Resultados de rectificación en imágenes de las bases de datos Sytim y MCL-RS en términos de $err_v$, $\theta$ y $Asp.Rat$. (a) Imágenes estéreo no rectificadas. Imágenes rectificadas con el método: (b) SPR, (c) Fusiello et al., (d) propuesto.

### 4.1.2. Evaluación del método STROMI para múltiples cámaras

En este experimento, el método propuesto se evalúa para el problema de rectificación utilizando cuatro cámaras. En esta configuración, se realizan los pasos descritos en la sección 4.1.1 para 25 imágenes de escenas reales capturadas con la plataforma experimental mostrada en la Fig. 4.1.4.

La plataforma experimental está formada por un arreglo multi-ocular de cuatro cámaras 4-Lane MIPI CSI-2 e-CAM130A, con lentes de longitud focal de 7 mm, alineadas de manera horizontal con una separación de línea base igual a 45 mm. Las cámaras son controladas por una tarjeta de desarrollo Jetson Xavier AGX. La Fig. 4.1.4 muestra la plataforma experimental utilizada para captura de imágenes multi-oculares.

Las imágenes capturadas son imágenes a color RGB con una resolución de $1920 \times 1080$

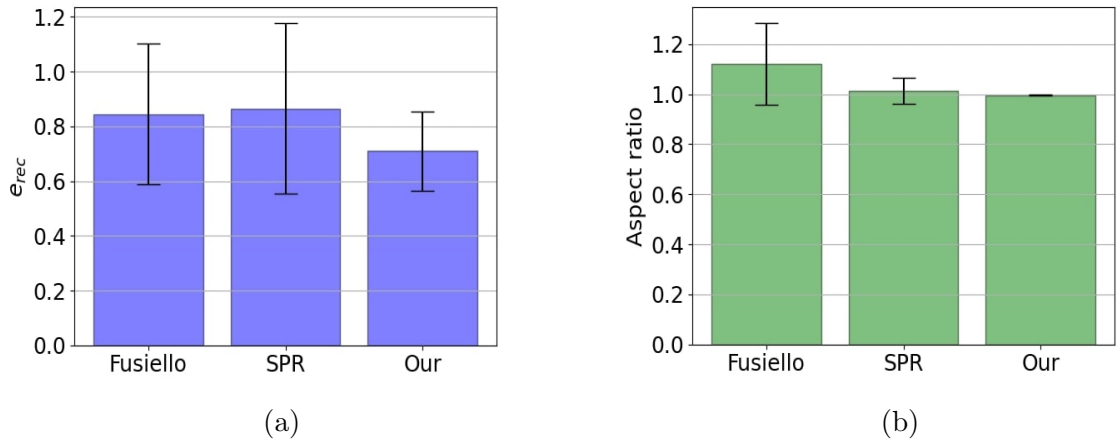
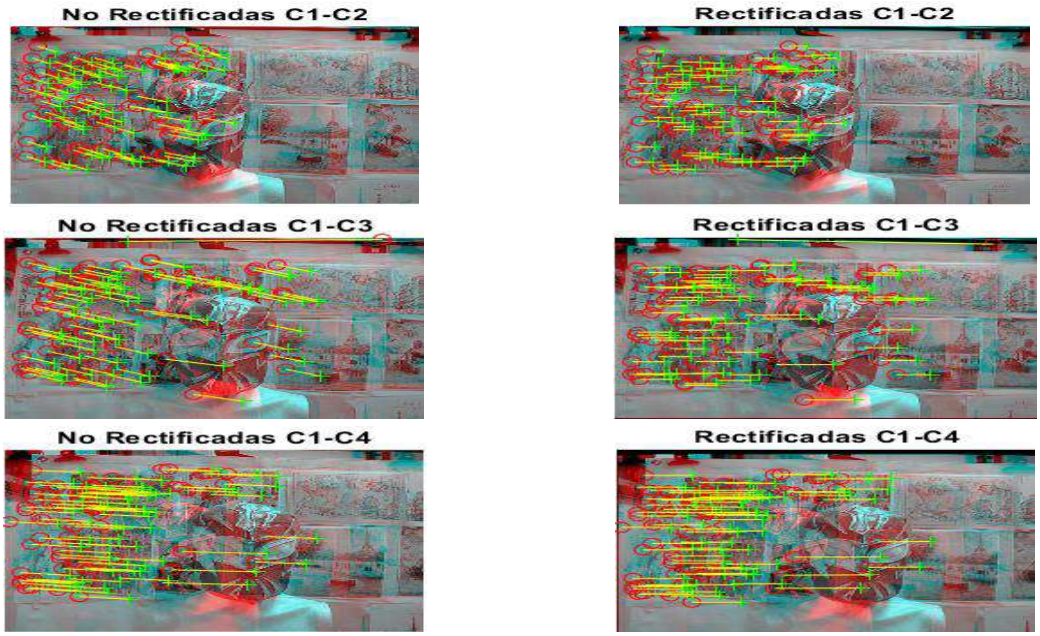



**Figura 4.1.3:** Resultados estadísticos de rectificación (valor esperado y desviación estándar) utilizando imágenes de las bases de datos Sytim y MCL-RS, en términos de (a) $err_v$, (b) $Asp.Rat.$.

**Figura 4.1.4:** Plataforma experimental construida, formada por cuatro cámaras separadas horizontalmente a una distancia de 45 mm y controladas por la tarjeta de desarrollo Jetson Xavier AGX.

píxeles. La Fig. 4.1.5-(a) muestra el proceso de detección de puntos correspondientes en las imágenes capturadas utilizando el método SIFT. Podemos observar la diferencia en las ordenadas de los puntos correspondientes al superponer el par de imágenes con respecto a la imagen de referencia (imagen 1). La Fig. 4.1.5-(b) muestra el resultado al transformar cada una de las imágenes capturadas con su correspondiente homografía de rectificación. Los puntos correspondientes entre el par de imágenes se encuentran a la misma altura al implementar la rectificación.

El desempeño del método propuesto fue evaluado utilizando las mismas medidas de desempeño utilizadas en la sección 4.1.1. Los resultados fueron comparados con el método de rectificación multi-ocular reportado por Yang. La Fig. 4.1.6 (a) muestra un ejemplo de cuatro imágenes capturadas con la plataforma experimental multi-ocular. La Fig. 4.1.6 (b) presenta los resultados de rectificación de las imágenes capturadas con el método de Yang, mientras la Fig. 4.1.6 (c) presenta el resultado con el método propuesto. De los resultados obtenidos, podemos notar que el método propuesto produce menor distorsión en las imágenes rectificadas.




(a) (b)

**Figura 4.1.5:** Resultados de rectificación de cuatro imágenes multi-oculares de una escena real. Los puntos correspondientes (círculos rojos) son estimados utilizando el método SIFT, al superponer el par de imágenes estéreo observamos el desplazamiento vertical que existe entre las imágenes (líneas amarillas). (a) Imágenes no rectificadas , (b) imágenes rectificadas.

La Tabla 4.3 presenta los resultados estadísticos (media y desviación estándar) de 25 imágenes rectificadas con respecto a la cámara 1. Podemos observar que el método de Yang decrementa el error al incrementar la separación de línea base. Sin embargo, el error obtenido es superior al método propuesto, el cual genera una mejora de 92.5 % y 83.2 % para la distancia de línea base corta y la distancia de línea base mayor, respectivamente, utilizadas en esta configuración. Por último, el método propuesto induce menor distorsión en comparación con el método de rectificación de Yang.

## 4.2. Evaluación del método propuesto para estimación de mapas de disparidad

En esta sección se presentan los resultados del desempeño del método propuesto descrito en el capítulo 3.2 para la estimación de mapas de disparidad. Para propósitos de evaluación se utilizan 25 imágenes estéreo de la base de datos Middlebury [64, 66, 102]. Esta base de datos contiene varias imágenes estéreo rectificadas y sin distorsión radial, con una distancia de línea base de 160 mm y una longitud focal de 3740 píxeles.

Los resultados obtenidos son comparados con dos variantes recientes del método de transformada census, denominada transformada census ponderada mejorada (WCT, por sus siglas en inglés *Weighted Census Transform*) [103] y el algoritmo AD-Census (ADCT,

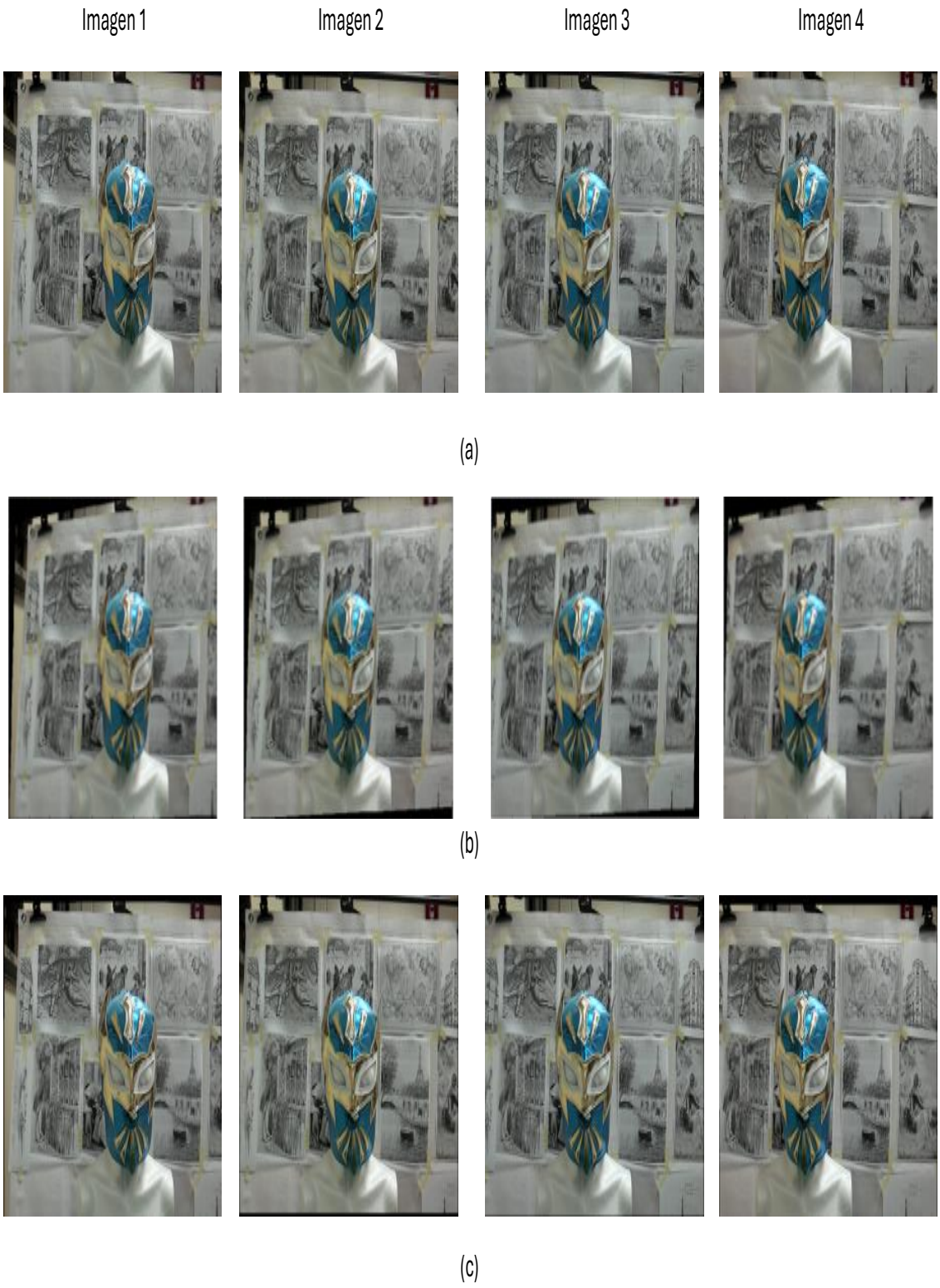



**Figura 4.1.6:** (a) Imágenes multi-oculares capturadas. Imágenes mutioculares rectificadas con el método: (b) Yang, (c) método propuesto.

**Cuadro 4.3:** Resultados estadísticos de rectificación (media y desviación estándar) de los métodos evaluados en términos de $err_v$ (en píxeles), $\theta$ (en grados) y Asp. Rat. utilizando imágenes capturadas.

| Rectificación multi-ocular | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Cámara 1-2 | | Cámara 1-3 | | Cámara 1-4 | |
| **Método** | **Medida de desempeño** | **Media** | **St. Dev.** | **Media** | **St. Dev.** | **Media** | **St. Dev.** |
| Yang et al. [101] | $err_v$ | 4.45 | 0.08 | 4.22 | 0.15 | 3.10 | 0.20 |
| | $\theta(90°)$ | 95.68 | 0.03 | 95.64 | 0.07 | 95.62 | 0.08 |
| | Asp. Rat. (1) | 1.01 | 0.005 | 1.09 | 0.001 | 1.01 | 0.001 |
| Propuesto | $err_v$ | 0.33 | 0.40 | 0.42 | 0.23 | 0.52 | 0.50 |
| | $\theta(90°)$ | 90.01 | 0.03 | 89.99 | 0.01 | 90.01 | 0.05 |
| | Asp. Rat. (1) | 1.01 | 0.01 | 0.99 | 0.02 | 1.02 | 0.01 |

por sus siglas en inglés Absolute Difference Census Transform) [104].

### 4.2.1. Medidas de desempeño para la evaluación de estimación de disparidad

El desempeño del método propuesto para estimación de disparidad es analizado en términos del error absoluto medio (MAE, por sus siglas en inglés *Mean Absolute Error*), y relación señal a ruido máxima (PSNR, por sus siglas en inglés *Peak Signal-to-Noise Ratio*). Estas medidas de desempeño se utilizan comparando los mapas de disparidad de referencia de la base de datos y los mapas de disparidad estimados.

El MAE puede calcularse como

$$MAE = \frac{1}{N} \sum_{x,y} |d_{GT}(x,y) - d_{est}(x,y)|, \tag{4.2.1}$$

donde $d_{GT}(x,y)$ es el valor de disparidad de referencia, $d_{est}(x,y)$ es la disparidad estimada, y $N$ es el número de elementos de la imagen procesada.

La $PSNR$ mide la calidad de la imagen basada en la diferencia de los píxeles de dos imágenes. La $PSNR$ está dada por

$$PSNR = 20 \log_{10} \frac{Max_f}{\sqrt{MSE}}, \tag{4.2.2}$$

donde $Max_f$ es el valor máximo de disparidad y MSE es el valor error cuadrático medio, definido como

$$MSE = \frac{1}{N} \sum_{x,y} ||d_{GT}(x,y) - d_{est}(x,y)||^2. \tag{4.2.3}$$

### 4.2.2. Evaluación del método propuesto de estimación de disparidad

En un primer experimento, se realizó una variación de los tamaños de la ventana de sorporte $S_w$ y el número de niveles de cuantización $Q$. La Fig. 4.2.1 presenta los resultados

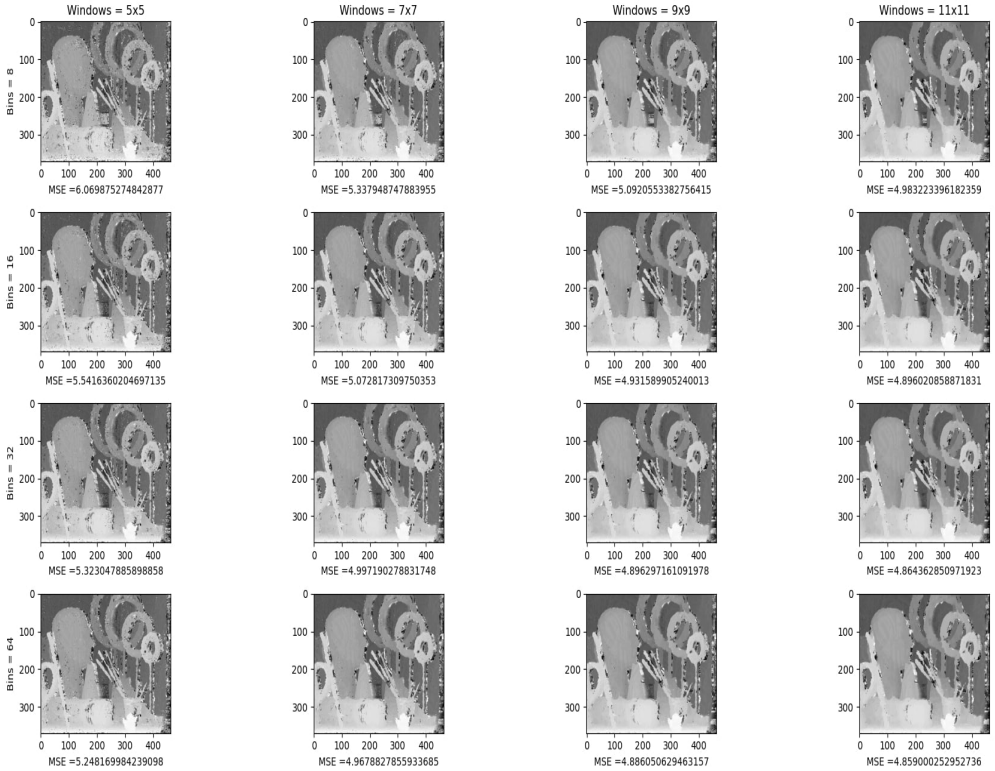



**Figura 4.2.1:** Mapas de disparidad estimado con el método propuesto utilizando diferentes tamaños de ventana y niveles de cuantización.

obtenidos utilizando diferentes combinaciones de tamaños de $S_w$ y niveles de $Q$ para estimar mapas de disparidad. Podemos observar que la calidad del mapa de disparidad estimado incrementa con el tamaño de ventana. Sin embargo, la nitidez de los contornos de los objetos reduce. Adicionalmente, la respuesta del estimador mejora con valores de $Q \geq 8$.

La Tabla 4.4 muestra el valor esperado (representado por el símbolo de barra superior $\overline{\{\}}$) y la desviación estándar (denotada por $\sigma$) del valor de $PSNR$ en 25 imágenes estéreo al realizar variaciones en $S_w$ y $Q$. De acuerdo a los valores presentados en la Tabla 4.4, un mayor valor de $Q$ permite disminuir el tamaño de $S_w$ manteniendo la calidad en la estimación del mapa de disparidad.

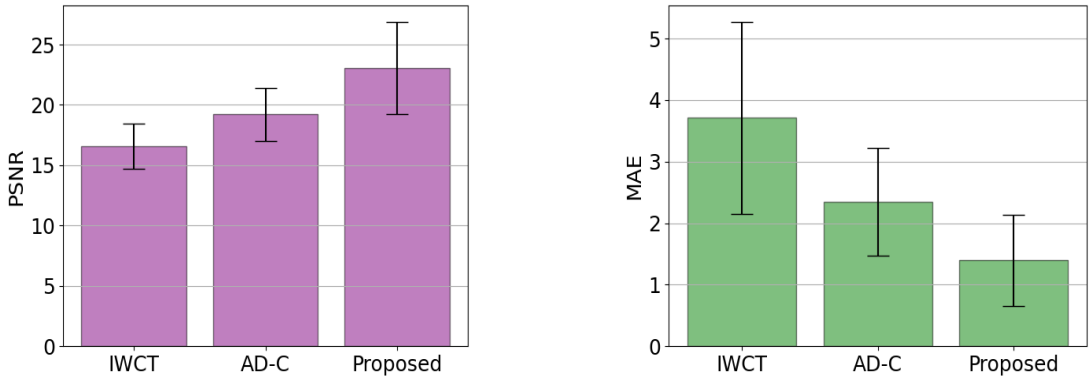
En el segundo experimento, se evaluó el método propuesto en zonas no ocluidas de las imágenes estéreo, utilizando los siguientes parámetros: $s_0 = 5$, $Q = 8$, $\epsilon_v = 1.5$ y $\beta_0 = 2.5s_0$. Los resultados obtenidos son comparados con los métodos IWCT y AD-Census. La Fig. 4.2.3 muestra la estimación de los mapas de disparidad en zonas no ocluidas utilizando los métodos AMC, WCT y ADCT. Además, podemos notar que el método propuesto es capaz de estimar con alta precisión la disparidad en regiones de la imagen con información de intensidad homogénea y en los bordes de los objetos de la escena. Los resultados obtenidos revelan que el desempeño más bajo lo tiene el algoritmo IWCT. El algoritmo ADCT realiza una buena estimación de disparidad en los bordes de los objetos. Sin embargo, el desempeño mostrado es inferior al presentado en el método propuesto.

El valor esperado y la desviación estándar del MAE y PSNR de 25 imágenes estéreo

**Cuadro 4.4:** Valores obtenidos de $\overline{PSNR}$ y $\sigma_{PSNR}$ en estimación de disparidad con el método propuesto, variando los parámetros $Q$ y $S_w$.

| **Bins** | $S_w$ | $\overline{PSNR}$ | $\sigma_{PSNR}$ |
|---|---|---|---|
| 4 | 5 | 22.696 | 5.903 |
| | 7 | 23.679 | 5.577 |
| | 9 | 24.170 | 5.507 |
| | 11 | 24.477 | 5.509 |
| | 13 | 24.642 | 5.513 |
| 8 | 5 | 23.178 | 5.870 |
| | 7 | 23.992 | 5.715 |
| | 9 | 24.393 | 5.599 |
| | 11 | 24.677 | 5.484 |
| | 13 | 24.846 | 5.482 |
| 16 | 5 | 23.914 | 5.879 |
| | 7 | 24.456 | 5.691 |
| | 9 | 24.741 | 5.576 |
| | 11 | 24.949 | 5.526 |
| | 13 | 24.999 | 5.542 |
| 32 | 5 | 24.161 | 5.742 |
| | 7 | 24.626 | 5.634 |
| | 9 | 24.926 | 5.485 |
| | 11 | 25.111 | 5.465 |
| | 13 | 25.169 | 5.447 |

se presentan en la Fig. 4.2.2. Los problemas de oclusiones y regiones con baja textura en imágenes estéreo son un gran reto para los métodos de estimación de disparidad. Los resultados mostrados en las Figs. 4.2.3 y 4.2.2, demuestran que el método propuesto es robusto y preciso para determinar los puntos correspondientes en imágenes estéreo.




(a) (b)

**Figura 4.2.2:** Resultados estadísticos (valor esperado y desviación estándar) de la estimación de la disparidad en zonas no ocluidas en términos de (a) $PSNR$, (b) $MAE$.

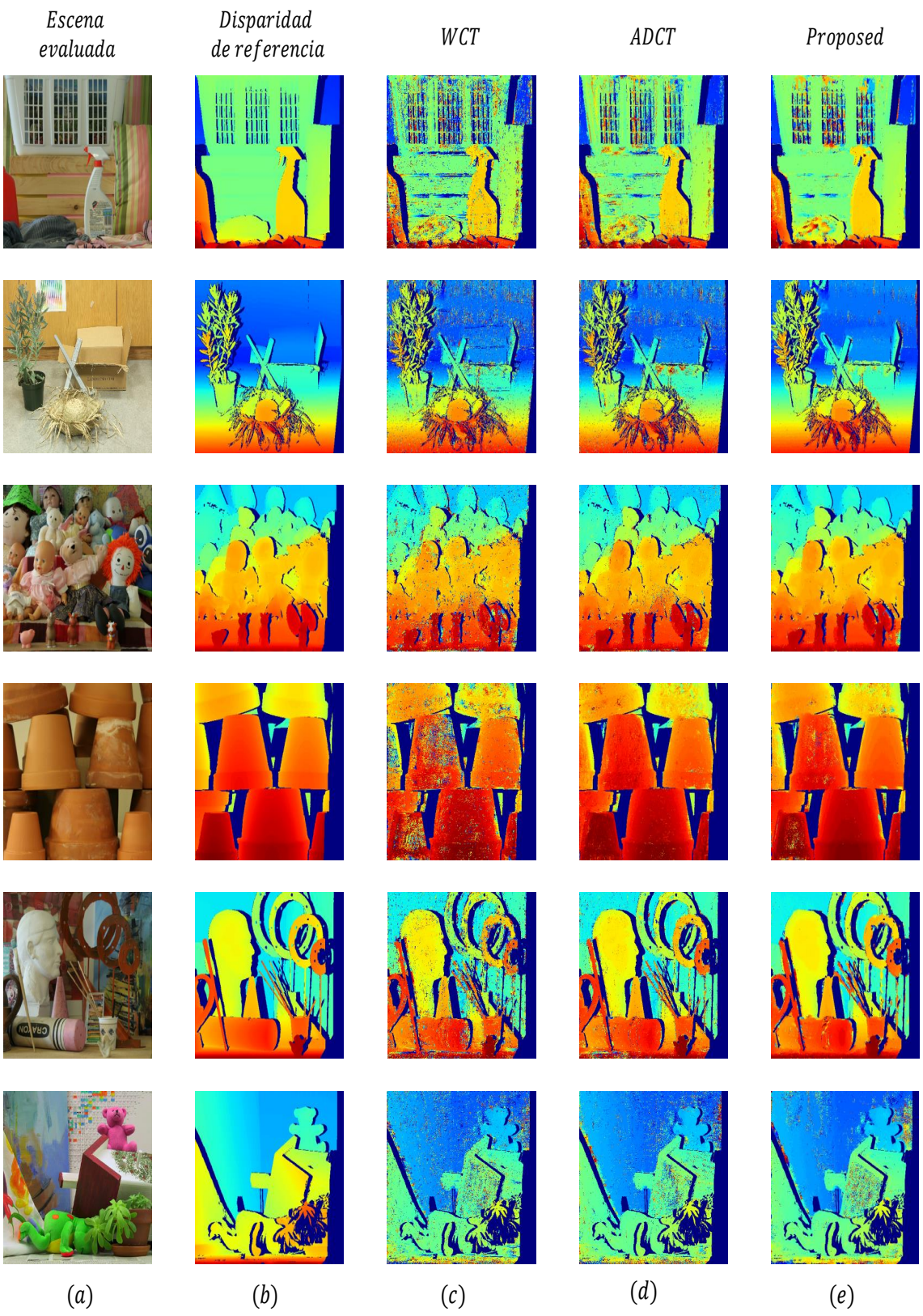


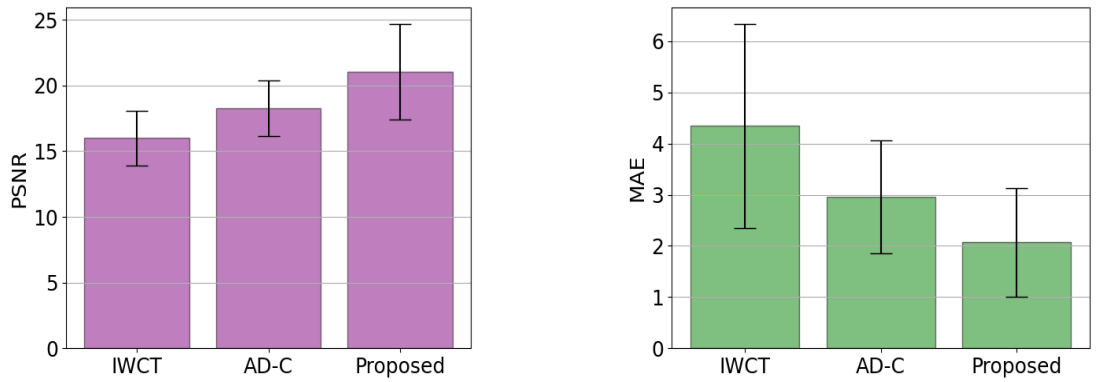

**Figura 4.2.3:** Resultados de estimación de disparidad en imágenes estéreo de la base de datos Middlebury en regiones no ocluidas. (a) Imagen evaluada. (b) Mapa de disparidad de referencia. Mapa de disparidad estimado con el método: (c) IWCT, (d) ADCT, (e) propuesto.

## 4.3. Evaluación del método propuesto de postprocesamiento de mapas de disparidad

En esta sección se presentan los resultados del desempeño del método de post-procesamiento descrito en la sección 3.3.1. Para propósitos de evaluación se utilizaron los mapas de disparidad estimados en zonas no ocluidas con los métodos AMC, IWCT y ADCT presentados en la sección 4.2.2. Los resultados son comparados utilizando los criterios de desempeño descritos en la sección 4.2.1.

### 4.3.1. Evaluación del método de rellenado de huecos basado en función Bayesiana

En este experimento se valida la precisión del método de interpolación de disparidad erróneamente estimada. La Fig. 4.3.2 muestra el resultado de implementar el método de post-procesamiento descrito en la sección 3.3 a los mapas de disparidad estimados en regiones no ocluidas de la Fig. 4.2.3 (zonas color azul oscuro). El método es capaz de recuperar la información como se muestra en la Fig. 4.3.2 para todos los métodos evaluados.




(a) (b)

**Figura 4.3.1:** Resultados estadísticos (valor esperado y desviación estándar) de estimación de disparidad realizadas con el método de post-procesamiento propuesto en términos de (a) $PSNR$, (b) $MAE$.

El valor esperado y la desviación estándar del MAE y PSNR del procesamiento de las 25 imágenes estéreo se presentan en la Fig. 4.3.1. Los resultados estadísticos demuestran que el método propuesto de estimación de disparidad, mantiene el mejor resultado en las evaluaciones realizadas. El método propuesto de post-procesamiento permite recuperar de manera exitosa los valores de disparidad en zonas ocluidas. Sin embargo, la calidad de los valores de disparidad recuperados depende de píxeles vecinos del elemento a recuperar. Por lo tanto, valores erróneos contribuirán a errores en la estimación de disparidades en zonas ocluidas. Podemos observar en la Fig. 4.2.2, los métodos con bajo desempeño en la estimación del mapa de disparidad, presentan una tendencia similar al implementar el método de post-procesamiento propuesto.

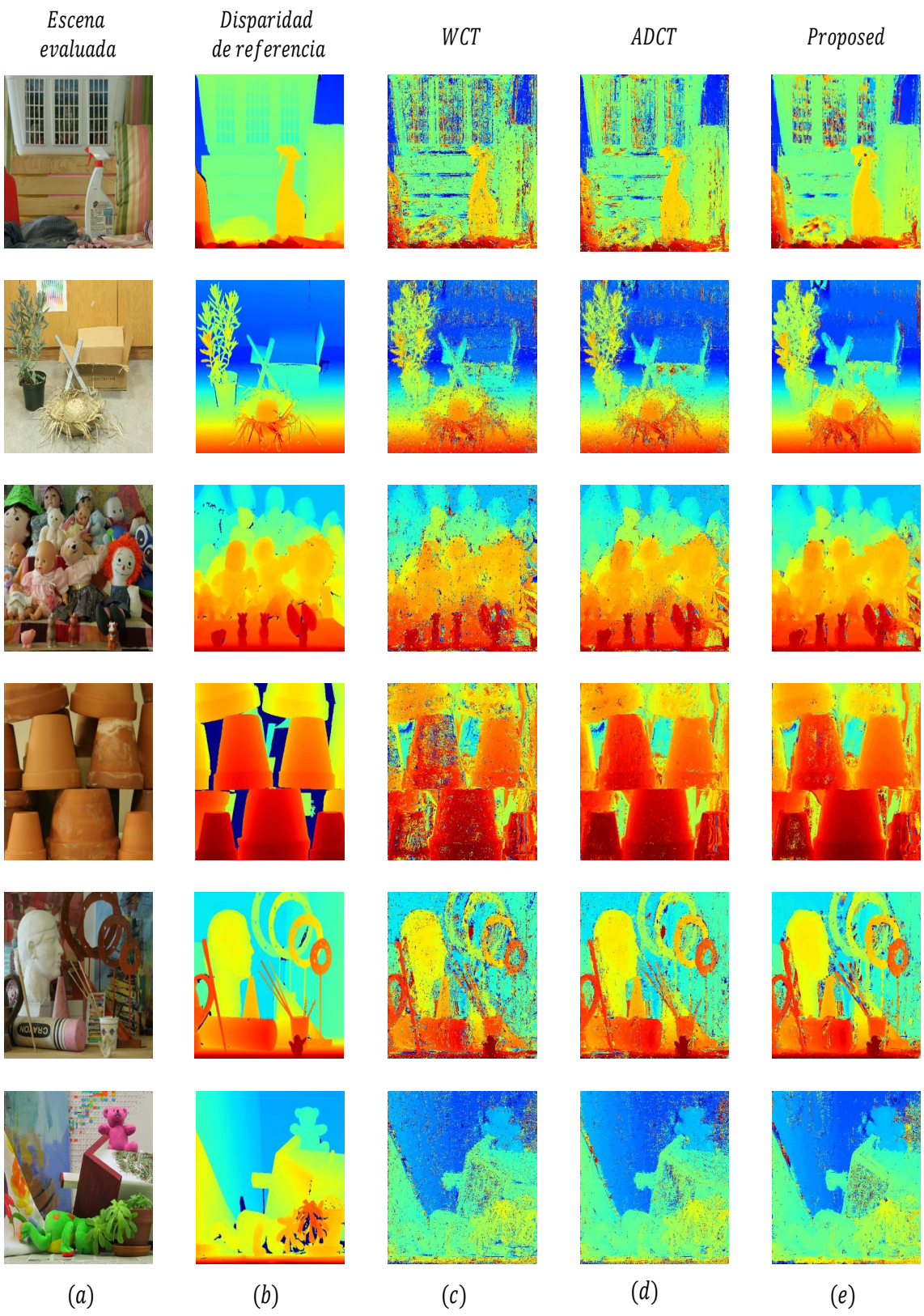



**Figura 4.3.2:** Resultados del método de postprocesamiento de disparidad en imágenes estéreo de la base de datos de Middlebury. (a) Imagen evaluada. (b) Mapa de disparidad de referencia. Mapa de disparidad postprocesado del método: (c) WCT, (d) ADCT, (e) propuesto.

## 4.4. Validación del método propuesto en escenas experimentales

En este experimento se realiza la validación de los métodos evaluados en las secciones 4.2.2 y 4.3 utilizando escenas reales capturadas en laboratorio. Los mapas de disparidad obtenidos, son utilizados para reconstruir la escena tridimensional haciendo uso del método lineal homogéneo descrito en la sección 2.2.7. Las imágenes fueron capturadas utilizando una tarjeta de desarrollo GPU NVIDIA Jetson TX1 y el módulo de dos cámaras de visión estéreo Li-Jetson-IMX274-Dual, con distancia de línea base de 50 cm. Las imágenes capturadas fueron rectificadas con el método de Zhang et al [61].

La Fig. 4.4.1 muestra el proceso de estimación del mapa de disparidad para una escena con objetos a diferentes profundidades. El mapa refinado de la Fig. 4.4.1 (e) es obtenido utilizando el método de Ma et al. [105].

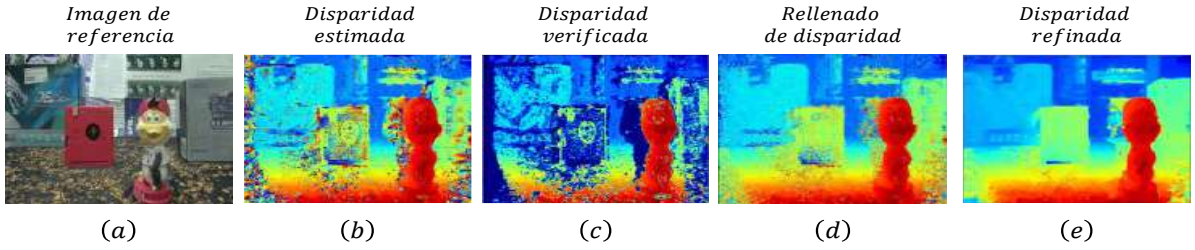



**Figura 4.4.1:** Resultado de estimación de disparidad de una escena real. (a) Imagen estéreo de referencia. (b) Disparidad estimada con el método AMC. (c) Verificación cruzada de valores de disparidad. (d) Mapa de disparidad obtenido utilizando el método de post-procesamiento propuesto. (e) Mapa de disparidad refinado.

En la Fig. 4.4.2 se presenta la reconstrucción tridimensional de la escena, utilizando el mapa de disparidad refinado mostrado en la Fig. 4.4.1 (e) y segregando la información del fondo de la escena.

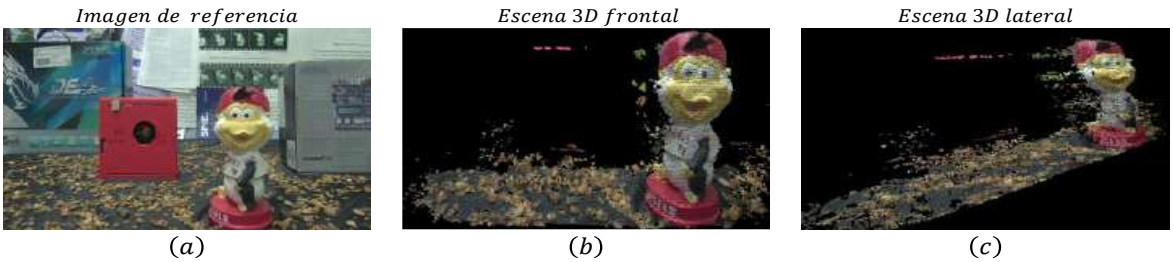
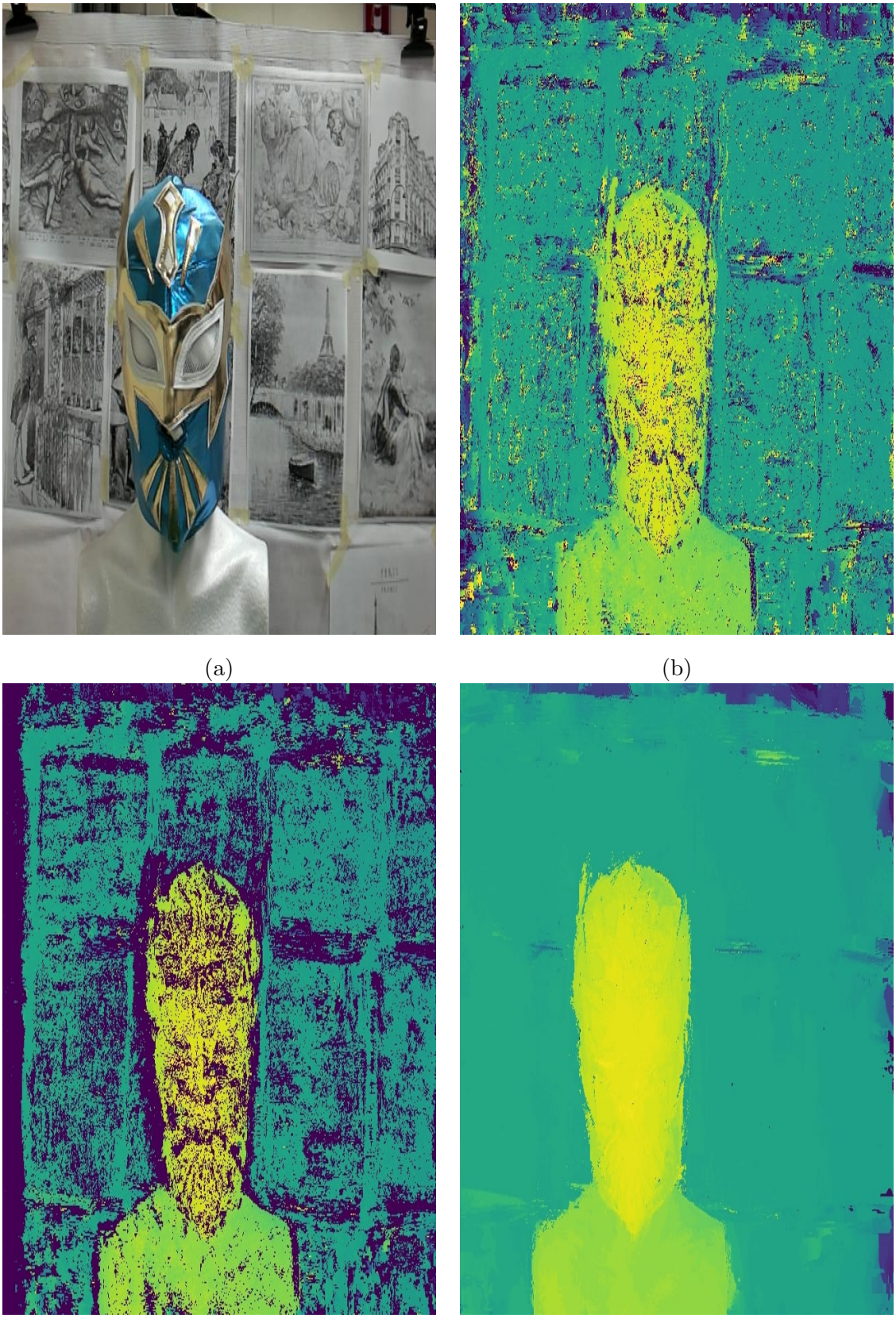



**Figura 4.4.2:** Resultado de la digitalización tridimensional de una escena real. Vista: (a) frontal, (b) lateral.

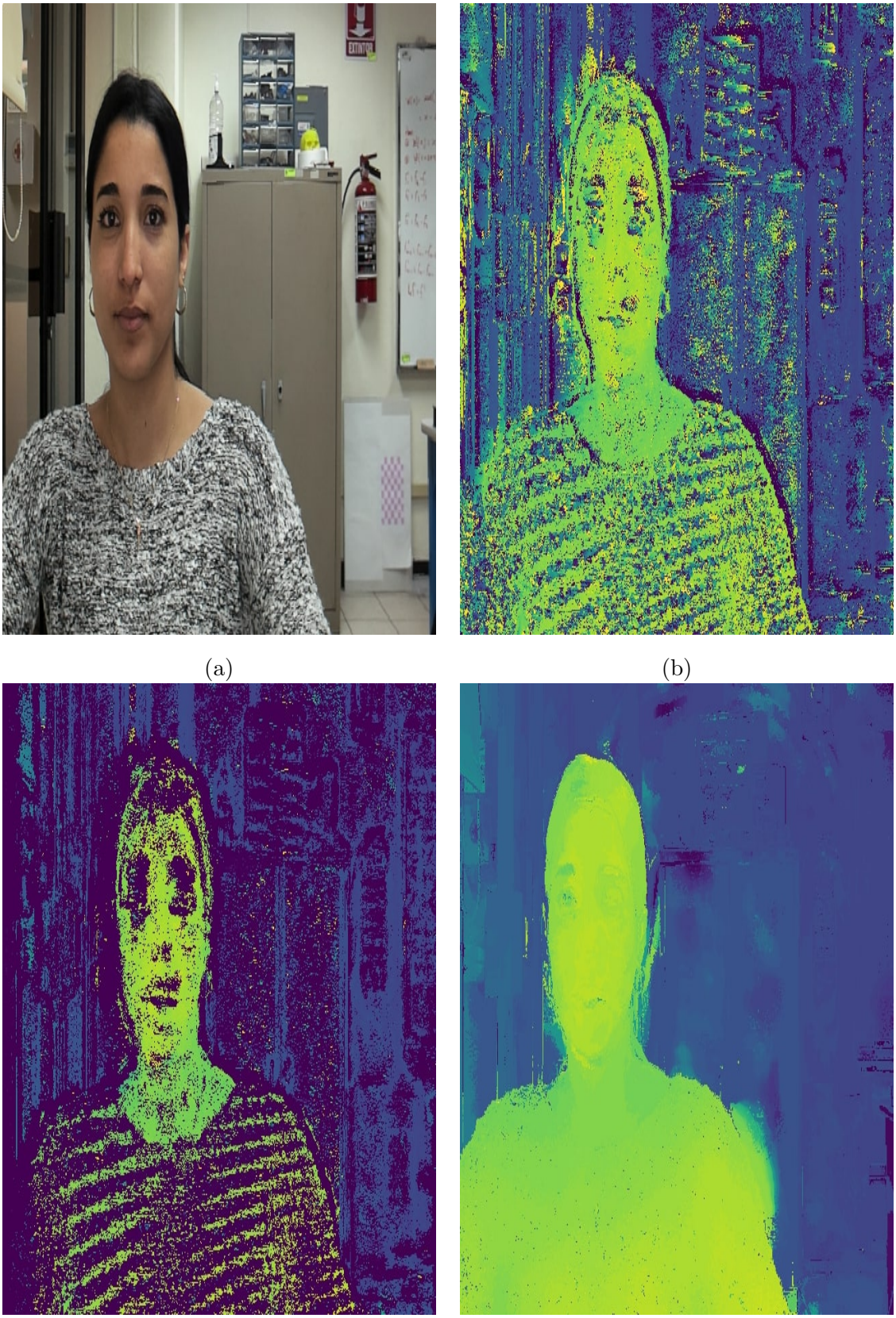
El proceso de adquisición de información del rostro es una etapa fundamental en los sistemas de FER. En este experimento se evaluaron dos escenas; la primera escena está formada por un modelo de rostro con un fondo distintivo, y la segunda escena presenta el rostro de una persona en un ambiente común. Las Figs. 4.4.3 y 4.4.4, muestra el proceso de estimación del mapa de disparidad.



(a) (b)

(c) (d)

**Figura 4.4.3:** Resultados del mapa de disparidad estimado de una escena construida: (a) imagen evaluada, (b) mapa de disparidad estimado, (c) mapa de disparidad verificado, (d) mapa de disparidad refinado.

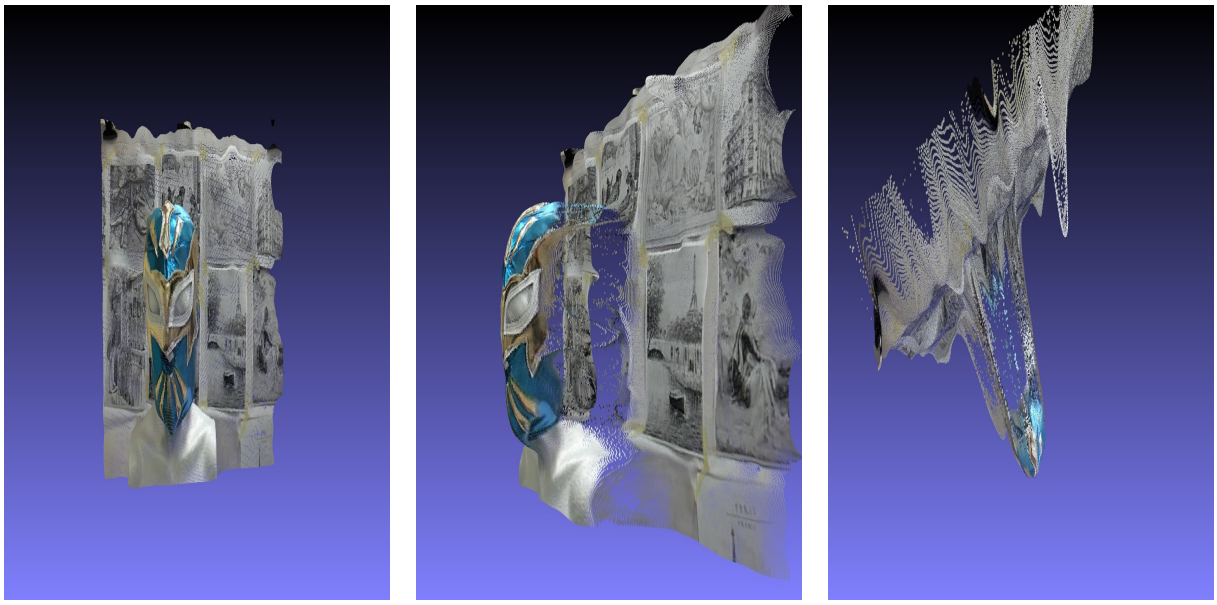


(a) (b)

(c) (d)

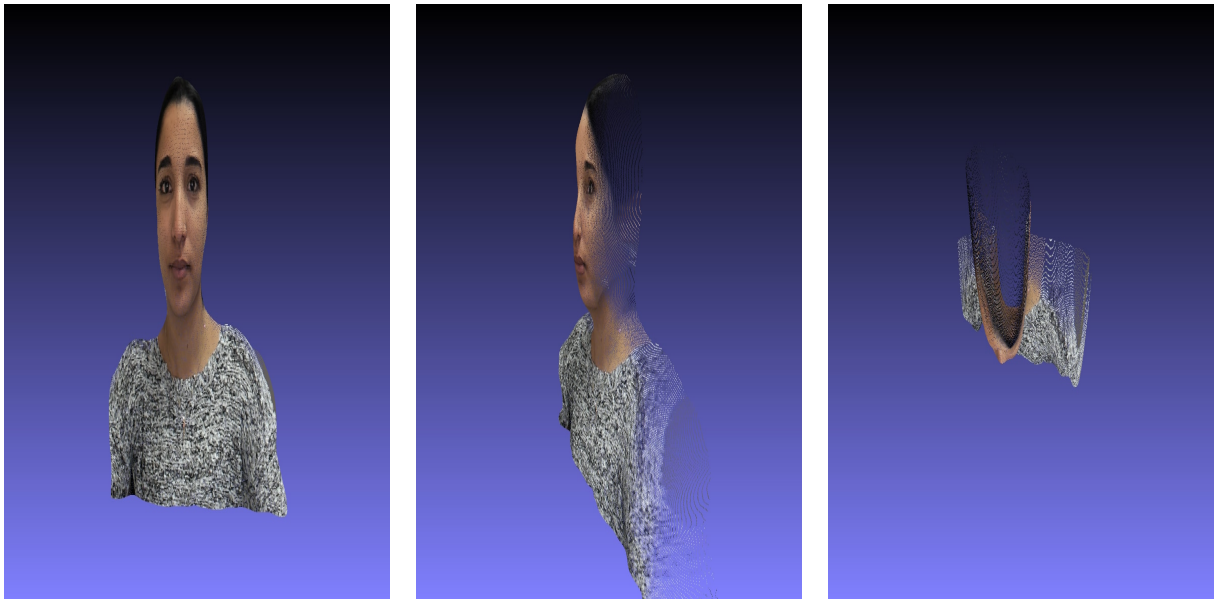
**Figura 4.4.4:** Resultados del mapa de disparidad estimado del rostro en una persona: (a) imagen de referencia, (b) mapa de disparidad estimado, (c) mapa de disparidad verificado, (d) mapa de disparidad refinado.

Las Figs. 4.4.5 y 4.4.6 muestran la reconstrucción tridimensional del par de imágenes estéreo utilizando el mapa de disparidad refinado.



(a) (b) (c)

**Figura 4.4.5:** Resultado de la reconstrucción de una escena real construida. Vista: (a) frontal, (b) lateral, (c) superior.



(a) (b) (c)

**Figura 4.4.6:** Resultado de la digitalización de un rostro real. Vista: (a) frontal, (b) lateral, (c) superior.

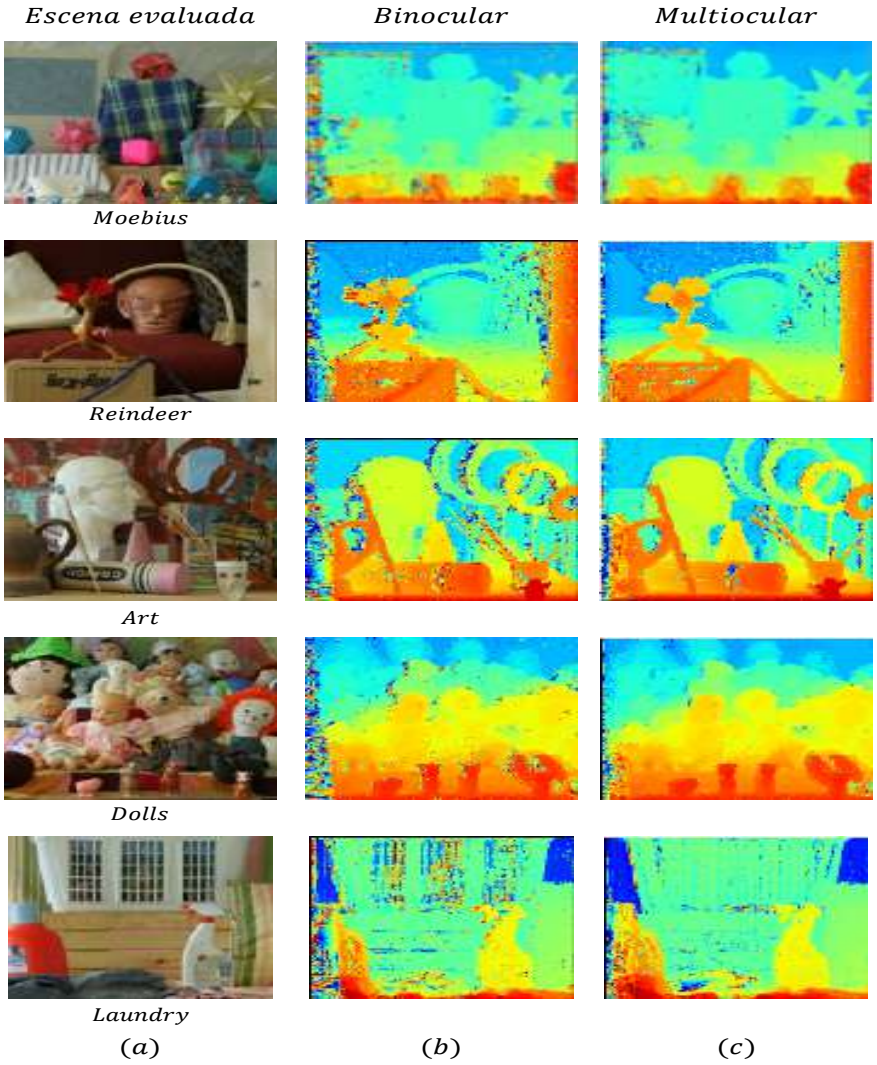
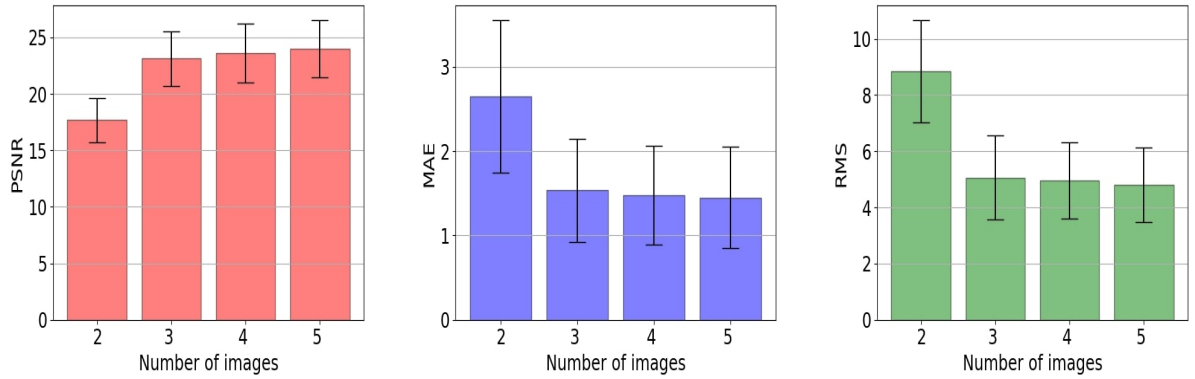



**Figura 4.5.1:** Resultados de estimación de disparidad en imágenes de la base de datos Middlebury. (a) Imagen de referencia. Mapas de disparidad estimado con el enfoque: (b) binocular (160 mm de línea base), (c) multi-ocular (distancia de línea base de 80 y 160 mm).

El resultado de la digitalización muestra una buena calidad en la información para el modelo del rostro. En la reconstrucción del rostro de una persona, se puede observar que se recupera la forma de los rasgos faciales. No obstante, es importante mencionar que los detalles de la textura de la piel y ligeras arrugas no son captados debido a que los niveles de disparidad son muy pequeños; esta información es útil para detectar microexpresiones.

El modelo tridimensional del rostro está límitado al campo de visión del arreglo estéreo. Por lo tanto, la combinación de un arreglo estéreo con separación de cámaras variables permitirá recuperar mayor información espacial y mejorar la resolución de la profundidad.




(a) (b) (c)

**Figura 4.5.2:** Resultados estadísticos (media y desviación estándar) de estimación de disparidad, obtenidos con el método propuesto multilínea base para estimación de intervalo de búsqueda y el método estéreo binocular en zonas no ocluidas: (a) PSNR, (b) MAE, (c) RMS.

## 4.5. Evaluación del método de visión estéreo multilínea base

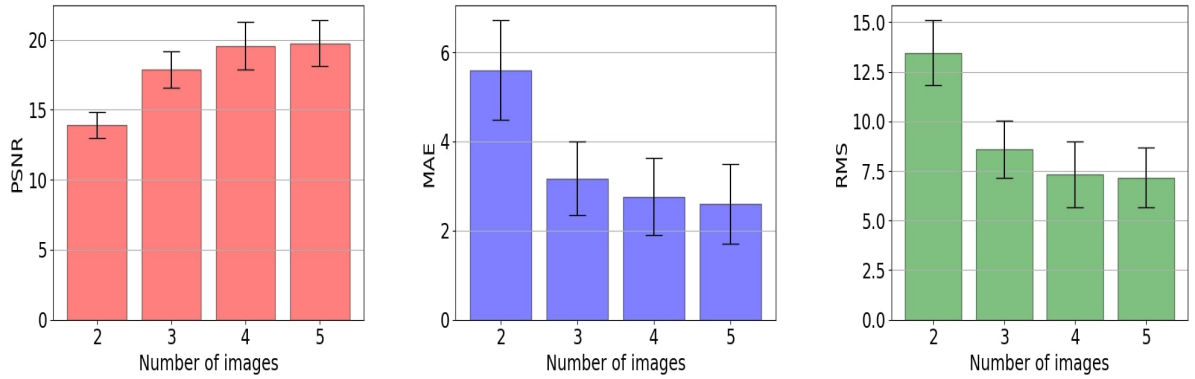
En este experimento, se realizó la evaluación del método presentado en la sección 3.4 utilizando 25 imágenes de la base de datos Middlebury [64, 66, 102], con configuraciones de 2, 3, 4 y 5 imágenes estéreo con una distancia de línea base igual a 40 mm entre cada cámara. Es decir, la separación mayor es de 160 mm entre las imágenes 1 y 5. Mientras que la separación menor es de 40 mm entre las imágnes 1 y 2.

La Fig. 4.5.1 (b) presenta el resultado del mapa de disparidad estimado con el método estéreo binocular, imagen 1-5. La Fig. 4.5.1 (c) muestra el resultado del mapa de disparidad estimado utilizando el método multilínea base propuesto en las imágenes 1-3-5. Podemos observar que la estimación de los mapas de disparidad mejora al utilizar más de dos cámaras, principalmente en los contornos de los objetos en la escena. Es posible disminuir las estimaciones erróneas de disparidad con una línea base estrecha, y utilizar el resultado como guía de búsqueda para distancias de línea base mayores.

Los resultados estadísticos obtenidos de las 25 imágenes estéreo se presentan en las Figs. 4.5.2 y 4.5.3, para zonas no ocluidas y todas las zonas, respectivamente, utilizando las medidas de desempeño descritas en la sección 4.2.1.

Los errores de estimación disminuyen con el enfoque multilínea base ajustable propuesto, comparado con el caso binocular utilizando el mismo método de estimación de disparidad. El uso de más cámaras permite disminuir los errores e incrementar la calidad de la estimación de disparidad. Sin embargo, de los resultados obtenidos notamos que no hay una diferencia significativa entre el uso de 4 y 5 cámaras.

En el segundo experimento, realizamos la comparación del método propuesto con el método reportado en la literatura por Li et al. [106] para estimación de disparidad con multilínea base. Las medidas de desempeño utilizadas son las descritas en la sección 4.2.1. La Tabla 4.5 presenta los resultados estadísticos obtenidos para las 25 imágenes de la base de datos Middlebury [64, 66, 102] en zonas no ocluidas y todas las zonas. Para el caso de




(a) (b) (c)

**Figura 4.5.3:** Resultados estadísticos (media y desviación estándar) de estimación de disparidad, obtenidos con el método propuesto multilínea base para estimación de intervalo de búsqueda y el método estéreo binocular en todas las zonas: (a) PSNR, (b) MAE, (c) RMS.

evaluación del método de Li et al.; se utilizó la configuración de imágenes 1-3-5.

**Cuadro 4.5:** Resultados estadísticos de la estimación de mapas de disparidad utilizando el enfoque multi-ocular.

| | | Asociación estéreo Zonas no ocluidas | | | | | | Asociación estéreo Todos los punto | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MAE | | RMS | | PSNR | | MAE | | RMS | | PSNR | |
| Método | Número de imágenes | Media | St. Dev | Media | St. Dev | Mean | St. Dev | Media | St. Dev | Media | St. Dev | Media | St. Dev |
| Propuesto | 2 | 2.64 | 1.80 | 8.83 | 3.64 | 17.67 | 3.96 | 5.60 | 2.22 | 13.45 | 3.27 | 13.90 | 1.85 |
| | 3 | 1.53 | 1.22 | 5.06 | 2.99 | 23.13 | 4.87 | 3.16 | 1.64 | 8.59 | 2.86 | 17.87 | 2.55 |
| | 4 | 1.47 | 1.16 | 4.96 | 2.73 | 23.64 | 5.21 | 2.76 | 1.72 | 7.33 | 3.31 | 19.55 | 3.38 |
| | 5 | 1.44 | 1.19 | 4.81 | 2.65 | 23.98 | 5.04 | 2.59 | 1.80 | 7.15 | 3.01 | 19.75 | 3.2 |
| Li et al. [106] | 3 | 1.91 | 1.47 | 6.66 | 3.43 | 20.66 | 4.05 | 3.78 | 1.97 | 10.69 | 3.30 | 15.81 | 2.27 |

El método propuesto presenta mejores resultados comparado con el método de Li et al [106]. Los intervalos de búsqueda están definidos por los valores estadísticos del valor de disparidad, lo que genera mayor robustez aumentando el intervalo de búsqueda en valores de disparidad alta y disminuyéndolo en valores de disparidad bajos.

### 4.5.1. Validación del método en una escena experimental

En este experimento se realizó la validación del método descrito en la sección 3.4, utilizando la plataforma experimental mostrada en la Fig. 4.1.4 para 25 escenas de laboratorio. Primeramente, se capturaron cuatro imágenes de una escena observada. Las imágenes fueron rectificadas utilizando el método de rectificación descrito en la sección 3.1, como se muestra en las Figs. 4.5.4 y 4.5.5.

Después, el mapa de disparidad denso es estimado entre las imágenes 1 y 2 utilizando el método AMC. El mapa de disparidad resultante es utilizado para realizar la predicción

de disparidad entre las imágenes 1 y 3 con el método de línea base ajustable, y también para generar el intervalo de búsqueda. Este proceso se repite hasta llegar al mapa de disparidad de las imágnes 1 y 4. El mapa estimado entre las imágenes 1 y 4 se procesa para validar los valores asociados correctos y erróneos. Finalmente, se interpolan los valores faltantes utilizando el método descrito en la sección 3.3.1. Los resultados se presentan en las Figs. 4.5.6-(c) y 4.5.7-(c). En la Fig. 4.5.7-(a) se presenta el resultado de la estimación de disparidad entre las imágenes con mayor separación de línea base utilizando el método estétero binocular. Se puede observar que la disparidad estimada presenta bastantes errores; esto es ocasionado por las oclusiones generadas al separar las cámaras. La distancia de separación induce zonas que no son observadas en ambas imágenes. Estos errores de estimación son reducidos considerablemente con el uso del método multilínea base ajustable. La estimación de disparidad entre imágenes con la línea base angosta se permite reducir zonas ocluidas. Esta estimación inicial es crucial para predecir los valores del intervalo de búsqueda para estimaciones con mayor separación de línea base, como se muestra en la Fig. 4.5.6-(b).

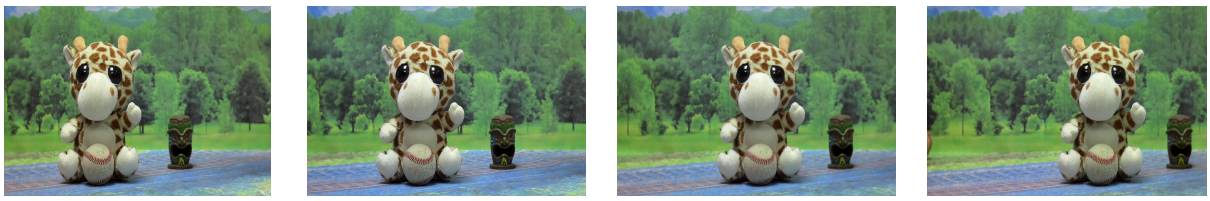


Imagen rectificada 1 Imagen rectificada 2 Imagen rectificada 3 Imagen rectificada 4

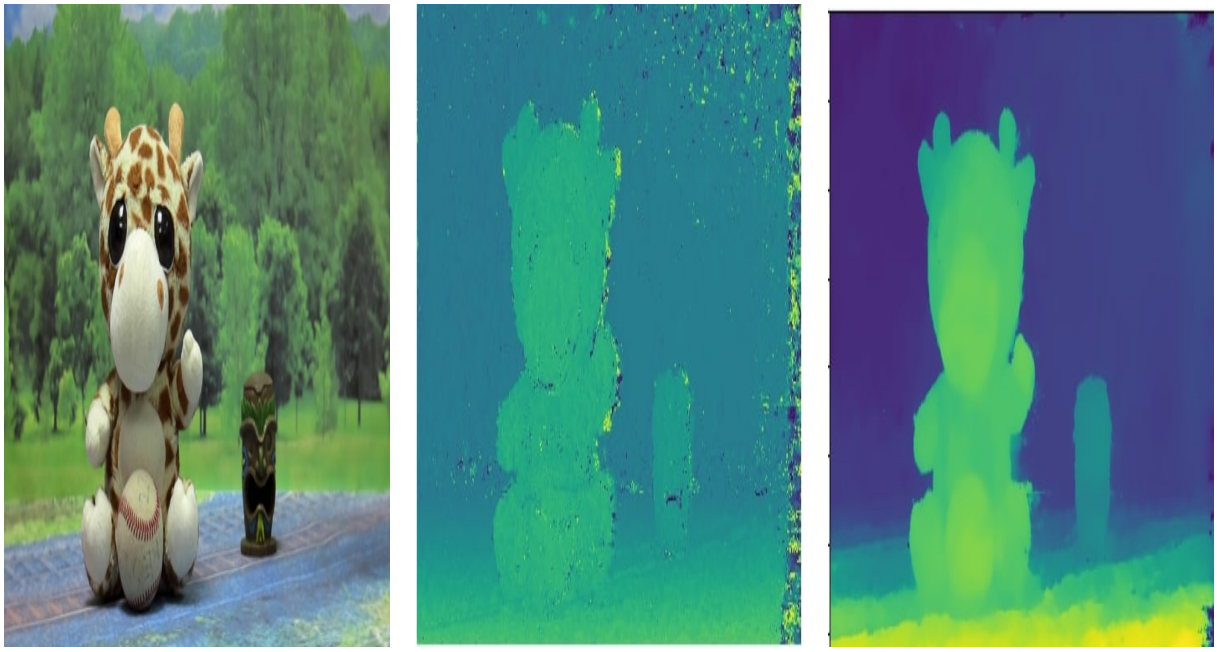
**Figura 4.5.4:** Imágenes capturadas y rectificadas con el método STROMI de una escena experimental.

Imagen rectificada 1 Imagen rectificada 2 Imagen rectificada 3 Imagen rectififcada 4

**Figura 4.5.5:** Imágenes rectificadas con el método STROMI de una escena experimental.

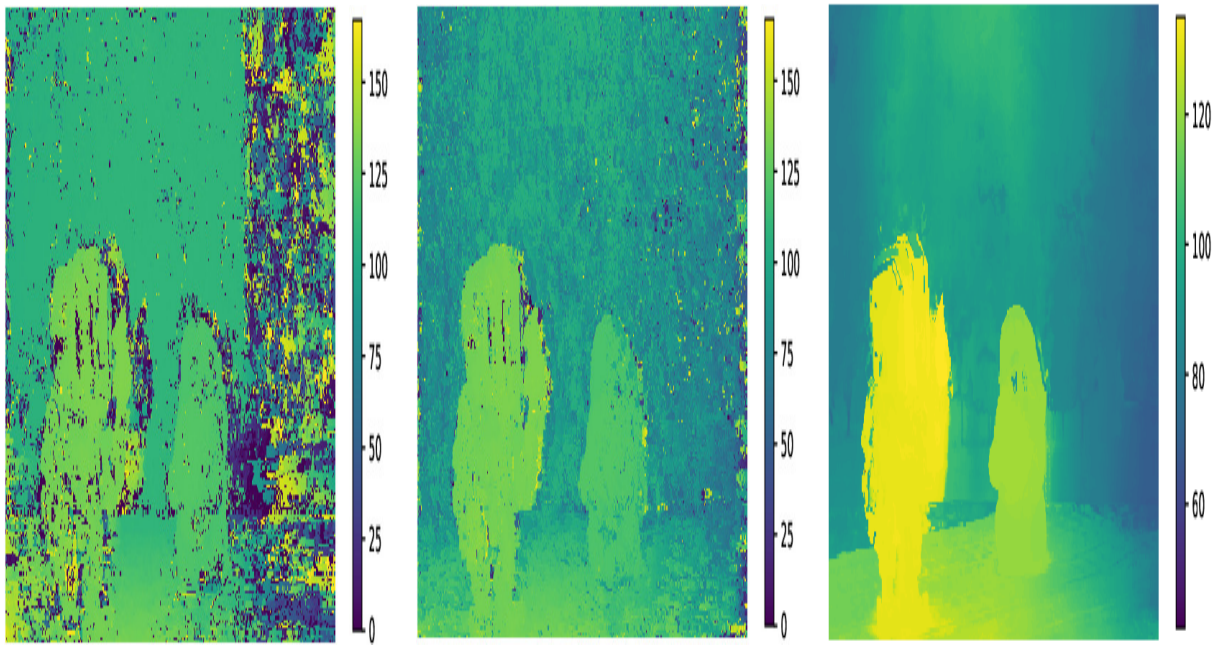
Por último, la escena es reconstruida utilizando el método lineal homogéneo descrito en la sección 2.2.7, el resultado se presenta en la Fig. 4.5.8. Para cuantificar los errores en la precisión de la reconstrucción, se realiza la medición del error de reproyección utilizando la matriz de paramétros intrínsecos obtenida por medio del método lineal directo de transformación (DLT, por sus siglas en inglés *Direct Linear Transform*) [1]. La Fig. 4.5.10 muestra el error de reproyección con valor medio de $\mu_{\text{err}} = (-1.05, 34.46) \times 10^{-6}$ mm y una desviación estándar $\sigma_{err} = 0.165$ mm.

Considerando que el error nominal de los sistemas de visión estéreo está en el orden de los milímetros [107], los resultados obtenidos confirman que el método propuesto utilizando el enfoque de línea base ajustable es robusto, preciso y confiable para reconstrucción $3D$ de una escena observada.



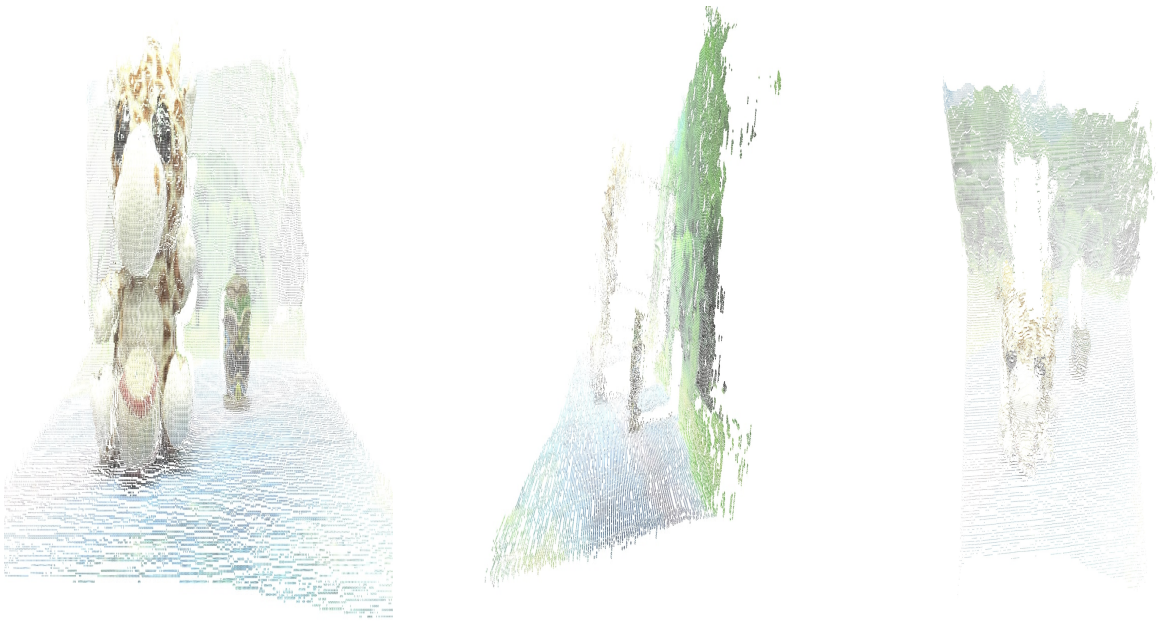
(a) (b) (c)

**Figura 4.5.6:** Estimación de disparidad de una escena real usando el método propuesto. (a) Imagen de referencia. (b) Mapa de disparidad estimado de las imágenes 1 y 4. (c) Mapa de disparidad postprocesado.




(a) (b) (c)

**Figura 4.5.7:** Mapa de disparidad estimado entre las imágenes 1 y 4 utilizando el enfoque: (a) binocular y (b) multi-ocular. (c) Mapa de disparidad del método multi-ocular refinado.



(a) (b) (c)

**Figura 4.5.8:** Nube de puntos espaciales de la reconstrucción 3D de una escena de la vista: (a) frontal. (b) lateral. (c) Superior.

(a) (b) (c)

**Figura 4.5.9:** Nube de puntos espaciales de la reconstrucción 3D de una escena de la vista (a) frontal. (b) lateral. (c) Superior.

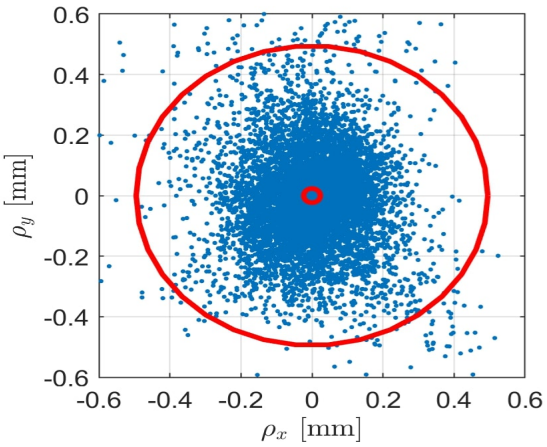



**Figura 4.5.10:** Errores de reproyección obtenidos, estimando los parámetros intrínsecos con el método DLT.

## 4.6. Evaluación del método de clasificación propuesto utilizando CNN

En esta sección se presentan los resultados del desempeño del método propuesto de extracción y clasificación de expresiones del rostro descrito en la sección 3.5. Para propósitos de evaluación, se utilizaron 2400 imágenes de la base de datos *Binghamton University 3D Facial Expression* (BU-3DFE) [108]. Esta base de datos contiene imágenes del rostro de 100 sujetos, con seis expresiones faciales: alegría (HA), tristeza (SA), ira (AN), disgusto (DI) y miedo (FE), como se presenta en la Fig. 4.6.1. Los rostros están divididos en 56 imágenes femeninas y 44 imágenes masculinas, pertenecientes a distintos grupos étnicos, con un rango de edad entre 18 a 70 años. Adicionalmente, la base de datos proporciona la información tridimensional de todas las imágenes.


**Figura 4.6.1:** Imágenes de expresiones del rostro de la base de datos BU-3DFE.

Las imágenes de la base de datos son divididas en imágenes de entrenamiento e imágenes de prueba, el tamaño de las imágenes es de $(224, 224)$. El 80 % de las imágenes son utilizadas para la etapa de entrenamiento mientras que el 20 % restante se utilizan para la etapa de prueba. El entrenamiento se realiza durante 100 épocas, con un lote de 32

muestras y utilizando el optimizador Adam [109] con una tasa de aprendizaje de 0.0001. La Fig. 4.6.2 muestra las curvas de entrenamiento del modelo propuesto.

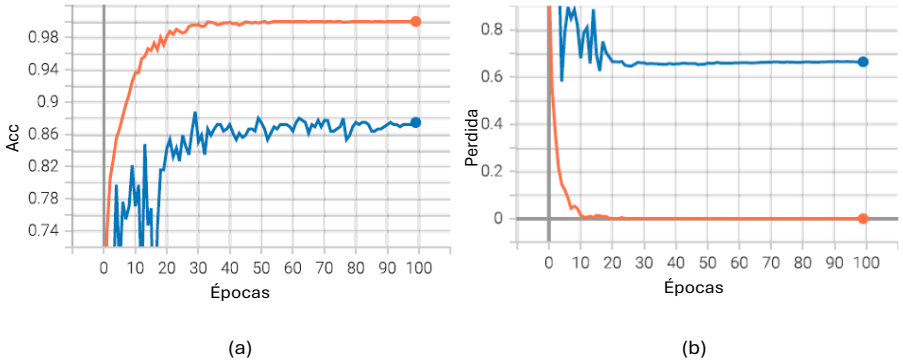



**Figura 4.6.2:** Curvas de entrenamiento (naranja) y validación (azul) de la arquitectura CNN propuesta. (a) Precisión del clasificador. (b) Función de pérdida.

### 4.6.1. Medidas de desempeño para la evaluación de método propuesto para clasificación de expresiones faciales

La evaluación del clasificador propuesto es analizada en términos de la precisión ($Acc$), donde se realiza la medición de las muestras clasificadas de manera correcta, definida como [110]

$$Acc = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}, \tag{4.6.1}$$

donde $T_P$, $T_N$, $F_P$ y $F_N$ son las clasificaciones verdaderas positivas, verdaderas negativas, falsos positivos y falsas negativas, respectivamente. Un mayor valor en la precisión, esta relacionada con un mejor rendimiento en la tarea de clasificación.

El clasificador fue comparado con diferentes métodos de CNN reportados en la literatura, como se muestra en la Tabla 4.6. El método propuesto presenta una mejor precisión en la clasificación de expresiones.

**Cuadro 4.6:** Resultados obtenidos en términos de precisión.

| Método | Acc( %) |
|---|---|
| Mao et al. [111] | 80.47 |
| Lai et al. [112] | 73.13 |
| Zhang et al. [113] | 81.20 |
| Fu et al. [114] | 82.89 |
| Jiang et al. [115] | 83.31 |
| Propuesto | 84.63 |

Finalmente, la Fig. 4.6.3 presenta predicciones del clasificador utilizando imágenes de la base de datos.


**Figura 4.6.3:** Resultados de predicción de expresiones faciales utilizando el método propuesto.

Capítulo 5

# Conclusiones

En este trabajo se presentó el desarrollo de un método de procesamiento de imágenes $2D + 3D$ del rostro para clasificación de expresiones. El método consiste en dos etapas, la etapa de estimación de información espacial del rostro utilizando sistemas de visión multi-ocular y la etapa de clasificación de características del rostro por medio de redes neuronales.

El método propuesto inicia con la captura de múltiples imágenes estéreo de una escena. Las imágenes capturadas son rectificadas utilizando un conjunto de transformadas proyectivas que permitan tener imágenes coplanares con la menor distorsión posible. La búsqueda del mejor conjunto de transformadas proyectivas se realiza a través del uso del algoritmo de enjambre de partículas, donde cada partícula representa un conjunto de transformadas proyectivas candidatas. La evaluación de las soluciones candidatas se realiza utilizando un criterio de evaluación propuesto, el cual minimiza la distorsión geométrica inducida y cumple con las restricciones epipolares. Este método permite realizar la rectificación de $n$ de imágenes estéreo. Los resultados obtenidos al realizar la evaluación del desempeño utilizando imágenes de base de datos, demuestran una mejora en la términos de precisión comparado con dos métodos populares de la literatura. Adicionalmente, el método propuesto fue implementado y evaluado utilizando imágenes capturadas con una plataforma experimental. Los resultados obtenidos fueron comparados con un método similar reportado en la literatura. Los resultados demostraron una mayor adaptabilidad del método propuesto para realizar el proceso de rectificación estéreo en imágenes reales. Sin embargo, el método es susceptible a detección de puntos correspondientes atípicos y distorsión inducida por la lente de la cámara.

A continuación, se realizó el proceso de estimación de información espacial de la escena utilizando $n$ imágenes estéreo rectificadas. Este proceso se realiza por medio de un método de asociación estéreo de línea base ajustable propuesto. Primero, la estimación de mapas de disparidad utilizando un método de asociación estéreo basado en correlación morfológica adaptativa propuesto es utilizado. Un primer mapa de disparidad denso es estimado con un rango de búsqueda predeterminado. Posteriormente, el mapa de disparidad entre la imagen de referencia y la cámara con el segundo menor valor de línea base es inferido, utilizando el mapa de disparidad previamente estimado y la relación de líneas base. Este método ayuda a reducir el rango de búsqueda a unos cuantos píxeles, reduciendo errores en la estimación, evitando operaciones innecesarias e incrementando la información estimada en zonas con oclusión. Este proceso continua iterativamente hasta llegar al par de

imágnes con mayor línea base. Finalmente, el mapa estimado es verificado para eliminar valores erróneos y postprocesado para mejorar la calidad del proceso de estimación. El mapa de disparidad post-procesado es utilizado para realizar la reconstrucción de la escena. De acuerdo con los resultados obtenidos de la evaluación y comparación, el método de estimación de mapas de disparidad basado en correlación morfólogica mejora la respuesta en zonas homógeneas y en los bordes de los objetos de la escena. El método propuesto es robusto y preciso; sin embargo, requiere realizar un ajuste en los parámetros de configuración debido a que existe una relación entre la calidad de la estimación y el número de binarizaciones, comprometiendo el tiempo de respuesta del estimador. Adicionalmente, el planteamiento multilínea base ajustable permite reducir el número de operaciones realizadas en la estimación, y también permite incrementar la resolución de los mapas de disparidad estimados. Por último, el método de post-procesamiento ayuda a rellenar la información mal asociada. Sin embargo, si la calidad del mapa estimado presentaruido excesivo, el método podría utilizar información errónea.

La última etapa está fue diseñada para procesar la información de textura y los mapas de profundidad de la imagen del rostro. Esta información es utilizada como entrada a una red neuronal convolucional con el propósito de extraer características distintivas de cada emoción utilizando información $2D$ y $3D$. Para la extracción de características $2D$, se utiliza un bloque conformado por una red neuronal convolucional preentrenada. La extracción de información $3D$ se realiza por medio de bloques convolucionales, incluyendo un bloque residual, que utiliza máscaras de soporte por medio de descomposición binaria umbralizada en los puntos característicos del rostro en sustitución del salto de conexión. Este bloque permite resaltar zonas de relevancia como ojos, nariz, boca y cejas. Finalmente, ambos bloques se concatenan para tener capas completamente conectadas, y realizar la clasificación de las expresiones del rostro. Los resultados obtenidos muestran una precisión del clasificador del 84.63 %, demostrando que el método propuesto mejora la etapa de clasificación.

# Trabajo futuro

A continuación se presentan algunas áreas de oportunidad para extender y robustecer el trabajo de investigación realizado como trabajo futuro.

- Extender el desarrollo del método propuesto de rectificación estéreo multi-ocular para cámaras con distorsión radial. En el presente trabajo de investigación, se logró desarrollar un método preciso y robusto de rectificación estéreo multi-ocular no calibrado. Sin embargo, se asume un modelo de cámara *pinhole* sin distorsión. El uso de lentes con diferentes longitudes focales inducen aberraciones a las imágenes capturadas, lo que puede generar problemas al realizar la correspondencia de puntos en imágenes y comprometer el desempeño del método propuesto. Una generalización del problema basado en un modelo de cámara con distorsión, permitirá robustecer la respuesta del método para el uso de cualquier dispositivo.

- Explorar diferentes configuraciones de arreglos geométricos de cámaras estéreo multi-ocular. Los resultados presentados en este trabajo de investigación fueron obtenidos utilizando varias cámaras posicionadas de manera horizontal, separadas por una distancia de línea base. Los resultados demuestran que el método propuesto en esta

tesis para estimación de mapas de disparidad y estimación del intervalo de búsqueda de disparidad basada en línea base ajustable, permiten recuperar la información de profundidad de la escena con buena resolución. Sin embargo, la información del ambiente puede ser delimitada por el campo de visión de las cámaras. El uso de diferentes configuraciones geométricas de cámaras permitirá capturar diferentes niveles de detalles de la escena, al ampliar el campo de visión de la escena.

- Explorar el uso de cámaras con combinaciones de valores de longitud focal para el sistema multi-ocular estéreo. El uso de diferentes longitudes focales permitirá capturar la escena en diferentes planos focales, evitando imágenes degradadas por desenfoque. Adicionalmente, incrementará el ángulo de visión. Por lo tanto, el sistema estéreo será capaz de realizar reconstrucciones en zonas ambiguas de la escena o minimizar las zonas ocluidas de la escena.

- Implementar el método de estimación de mapas de disparidad en sistemas de tiempo real. La dinámica de las expresiones del rostro requiere de sistemas de captura y reconstrucción que respondan en tiempo real. Los sistemas de visión estéreo locales son bastante adecuados para realizar reconstrucciones de la escena observada a partir de una sola toma. Además son adecuados para ser implementados en dispositivos de procesamiento paralelo, lo que permite tener una reconstrucción inmediata de la escena observada, y ser utilizados en diferentes aplicaciones de vanguardia.

- Indagar el uso de algoritmos evolutivos para realizar el proceso de refinamiento de mapas de disparidad. La etapa de refinamiento en mapas de disparidad permite corregir la información mal estimada. Estos problemas no se pueden evitar al utilizar configuraciones estéreo, debido a que la posición geométrica de las cámaras genera oclusiones en ciertas zonas de la escena. Adicionalmente, el proceso de rectificación puede inducir una distorsión. Por último, la estimación puede resultar errónea al existir zonas con baja textura o iluminación saturada. Los algoritmos evolutivos permiten plantear la solución de manera iterativa como un problema de minimización adaptativa a la función objetivo. Permitiendo encontrar la función que se adapte de mejor manera y permita minimizar los errores de estimación.

- Explorar el uso de redes neuronales recurrentes para el problema de clasificación de características faciales. Utilizar información adicional como los mapas de vectores normales del rostro. Ajustar los hiperparámetros de entrenamiento de la red neuronal.

# Bibliografía


[1] R. I. Hartley, "Theory and practice of projective rectification," *International Journal of Computer Vision*, vol. 35, pp. 115–127, 2004.

[2] W. F. P. Ekman, "The facial action coding system: A technique for measurement of facial movement," *Consulting Psychologists Press*, 1978.

[3] B. Dominguez-Dager, F. Gomez-Donoso, R. Roig-Vila, F. Escalona, and M. Cazorla, "Holograms for seamless integration of remote students in the classroom," *Virtual Reality*, vol. 28, no. 1, p. 24, 2024.

[4] H.-C. Chu, W. W.-J. Tsai, M.-J. Liao, and Y.-M. Chen, "Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning," *Soft Computing*, vol. 22, pp. 2973–2999, 2018.

[5] N. Bosch, S. K. D'Mello, J. L. Ocumpaugh, R. Baker, and V. J. Shute, "Using video to automatically detect learner affect in computer-enabled classrooms," *ACM Trans. Interact. Intell. Syst.*, vol. 6, pp. 17:1–17:26, 2016.

[6] J. V. Moniaga, A. Chowanda, A. P. Prima, Oscar, and M. D. T. Rizqi, "Facial expression recognition as dynamic game balancing system," *Procedia Computer Science*, vol. 135, pp. 361–368, 2018.

[7] M. T. Akbar, M. N. Ilmi, I. V. Rumayar, J. V. Moniaga, T.-K. Chen, and A. Chowanda, "Enhancing game experience with facial expression recognition as dynamic balancing," *Procedia Computer Science*, 2019.

[8] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," in *ISVC*, 2012.

[9] P. Werner, A. Al-Hamadi, K. Limbrecht, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain assessment with facial activity descriptors," *IEEE Transactions on Affective Computing*, vol. 8, pp. 286–299, 2017.

[10] P. Ekman, "Universals and cultural differences in facial expressions of emotion." 1972.

[11] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 1424–1445, 2000.

[12] R. A. Patil, V. Sahula, and A. S. Mandal, "Automatic recognition of facial expressions in image sequences: A review," *2010 5th International Conference on Industrial and Information Systems*, pp. 408–413, 2010.

[13] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 1548–1568, 2016.

[14] P. A. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, 2001.

[15] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2009.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 vol. 1, 2005.

[17] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, pp. 135–164, 2004.

[18] S. W. Chew, P. Lucey, S. Lucey, J. M. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan, "In the pursuit of effective affective computing: The relationship between features and registration," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, pp. 1006–1016, 2012.

[19] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, "Automatically detecting pain in video through facial action units," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, pp. 664–674, 2011.

[20] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn, "A framework for automated measurement of the intensity of non-posed facial action units," *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 74–80, 2009.

[21] G. Littlewort, M. S. Bartlett, I. R. Fasel, J. M. Susskind, and J. R. Movellan, "Dynamics of facial expression extracted automatically from video," *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 80–80, 2004.

[22] G. Littlewort, J. Whitehill, T. Wu, I. R. Fasel, M. G. Frank, J. R. Movellan, and M. S. Bartlett, "The computer expression recognition toolbox (cert)," *Face and Gesture 2011*, pp. 298–305, 2011.

[23] Y. Koda, Y. Yoshitomi, M. Nakano, and M. Tabuse, "A facial expression recognition for a speaker of a phoneme of vowel using thermal image processing and a speech recognition system," *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 955–960, 2009.

[24] J. Whitehill and C. W. Omlin, "Haar features for facs au recognition," *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 5 pp.–101, 2006.

[25] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, pp. 803–816, 2009.

[26] G. Sandbach, S. Zafeiriou, and M. Pantic, "Local normal binary patterns for 3d facial action unit detection," *2012 19th IEEE International Conference on Image Processing*, pp. 1813–1816, 2012.

[27] C.-K. Tran, T.-H. Ngo, C.-N. Nguyen, and L.-A. Nguyen, "Svm-based face recognition through difference of gaussians and local phase quantization," *International Journal of Computer Theory and Engineering*, vol. 13, pp. 1–8, 2021.

[28] P. Lemaire, M. Ardabilian, L. Chen, and M. Daoudi, "Fully automatic 3d facial expression recognition using differential mean curvature maps and histograms of oriented gradients," *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–7, 2013.

[29] B. Sun, L. Li, X. Wu, T. Zuo, Y. Chen, G. Zhou, J. He, and X. Zhu, "Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild," *Journal on Multimodal User Interfaces*, vol. 10, pp. 125–137, 2015.

[30] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3d facial expression recognition," *Pattern Recognit.*, vol. 44, pp. 1581–1589, 2011.

[31] D. Derkach and F. M. Sukno, "Automatic local shape spectrum analysis for 3d facial expression recognition," *Image Vis. Comput.*, vol. 79, pp. 86–98, 2018.

[32] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, vol. 12, pp. 524–543, 2021.

[33] K. Yurtkan and H. Demirel, "Feature selection for improved 3d facial expression recognition," *Pattern Recognit. Lett.*, vol. 38, pp. 26–33, 2014.

[34] W. Hariri, H. Tabia, N. Farah, A. Benouareth, and D. Declercq, "3d facial expression recognition using kernel methods on riemannian manifold," *Eng. Appl. Artif. Intell.*, vol. 64, pp. 25–32, 2017.

[35] B. B. Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4-d facial expression recognition by learning geometric deformations," *IEEE Transactions on Cybernetics*, vol. 44, pp. 2443–2457, 2014.

[36] L. Tao and B. J. Matuszewski, "Is 2d unlabeled data adequate for recognizing facial expressions?" *IEEE Intelligent Systems*, vol. 31, pp. 19–29, 2016.

[37] Y. Fu, Q. Ruan, G. An, and Y. Jin, "Fast nonnegative tensor factorization based on graph-preserving for 3d facial expression recognition," *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, pp. 292–297, 2016.

[38] M. Jazouli, A. Majda, and A. Zarghili, "A $p recognizer for automatic facial emotion recognition using kinect sensor," *2017 Intelligent Systems and Computer Vision (ISCV)*, pp. 1–5, 2017.

[39] S. Soltanpour, B. Boufama, and Q. M. J. Wu, "A survey of local feature methods for 3d face recognition," *Pattern Recognit.*, vol. 72, pp. 391–406, 2017.

[40] S. Zhou and S. Xiao, "3d face recognition: a survey," *Human-centric Computing and Information Sciences*, vol. 8, pp. 1–27, 2018.

[41] H. Huang, J. Chai, X. Tong, and H.-T. Wu, "Leveraging motion capture and 3d scanning for high-fidelity facial performance acquisition," in *SIGGRAPH 2011*, 2011.

[42] G. Fyffe, K. Nagano, L. Huynh, S. Saito, J. Busch, A. Jones, H. Li, and P. E. Debevec, "Multi-view stereo on consistent face topology," *Computer Graphics Forum*, vol. 36, 2017.

[43] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, 2019.

[44] M. Sajjad, F. U. M. Ullah, M. Ullah, G. Christodoulou, F. A. Cheikh, M. Hijji, K. Muhammad, and J. J. Rodrigues, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal*, vol. 68, pp. 817–840, 2023.

[45] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.

[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[47] E. Ahmed, A. Saint, A. E. R. Shabayek, K. Cherenkova, R. Das, G. Gusev, D. Aouada, and B. Ottersten, "Deep learning advances on different 3d data representations: A survey," *arXiv preprint arXiv:1808.01462*, vol. 1, 2018.

[48] U. Prabhu, J. Heo, and M. Savvides, "Unconstrained pose-invariant face recognition using 3d generic elastic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1952–1961, 2011.

[49] J. Salvi, S. Fernandez, T. Pribanić, and X. Lladó, "A state of the art in structured light patterns for surface profilometry," *Pattern Recognit.*, vol. 43, pp. 2666–2680, 2010.

[50] S. Zhang, "High-speed 3d imaging with digital fringe projection techniques," 2016.

[51] M. W. Takeda and K. Mutoh, "Fourier transform profilometry for the automatic measurement of 3-d object shapes." *Applied optics*, vol. 22 24, p. 3977, 1983.

[52] M. Y. Kim, S. M. Ayaz, J. Park, and Y. J. Roh, "Adaptive 3d sensing system based on variable magnification using stereo vision and structured light," *Optics and Lasers in Engineering*, vol. 55, pp. 113–127, 2014.

[53] W. Kazmi, S. Foix, G. Alenya, and H. J. Andersen, "Indoor and outdoor depth imaging of leaves with time-of-flight and stereo vision sensors: Analysis and comparison," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 88, pp. 128–146, 2014.

[54] R. A. Hamzah and H. Ibrahim, "Literature survey on stereo vision disparity map algorithms," *J. Sensors*, vol. 2016, pp. 8 742 920:1–8 742 920:23, 2016.

[55] M. S. Hamid, N. Abd Manap, R. A. Hamzah, and A. F. Kadmin, "Stereo matching algorithm based on deep learning: A survey," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 5, pp. 1663–1673, 2022.

[56] R. Juarez-Salazar, G. A. Rodriguez-Reveles, S. Esquivel-Hernandez, and V. H. Diaz-Ramirez, "Three-dimensional spatial point computation in fringe projection profilometry," *Optics and Lasers in Engineering*, vol. 164, p. 107482, 2023.

[57] R. Juarez-Salazar and V. H. Díaz-Ramírez, "Operator-based homogeneous coordinates: application in camera document scanning," *Optical Engineering*, vol. 56, 2017.

[58] A. Harltey and A. Zisserman, "Multiple view geometry in computer vision (2. ed.)," 2003.

[59] B. Cyganek and J. P. Siebert, *An introduction to 3D computer vision techniques and algorithms.* John Wiley & Sons, 2011.

[60] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[61] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 1330–1334, 2000.

[62] P. Maragos, "Optimal morphological approaches to image matching and object detection," in *1988 Second International Conference on Computer Vision.* IEEE Computer Society, 1988, pp. 695–696.

[63] V. Klema and A. Laub, "The singular value decomposition: Its computation and some applications," *IEEE Transactions on automatic control*, vol. 25, no. 2, pp. 164–176, 1980.

[64] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, 2004.

[65] A. Fusiello, U. Castellani, and V. Murino, "Relaxing symmetric multiple windows stereo using markov random fields," in *EMMCVPR*, 2001.

[66] H. Hirschmüller and D. Scharstein, "Evaluation of cost functions for stereo matching," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[67] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European conference on computer vision*. Springer, 1994, pp. 151–158.

[68] K. jin Yoon and I.-S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28 4, pp. 650–6, 2006.

[69] J. Wang, C. Peng, M. Li, Y. Li, and S. Du, "The study of stereo matching optimization based on multi-baseline trinocular model," *Multimedia Tools and Applications*, pp. 1–12, 2022.

[70] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5*. Springer, 1998, pp. 484–498.

[71] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1685–1692.

[72] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[73] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[74] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[75] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 532–539.

[76] Y. Hua, J. Guo, and H. Zhao, "Deep belief networks and deep learning," in *Proceedings of 2015 international conference on intelligent computing and internet of things*. IEEE, 2015, pp. 1–4.

[77] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[78] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[79] B. Jähne, *Digital image processing*. Springer Science & Business Media, 2005.

[80] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, pp. 611–629, 2018.

[81] A. Zafar, M. Aamir, N. Mohd Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta, and S. Almotairi, "A comparison of pooling methods for convolutional neural networks," *Applied Sciences*, vol. 12, no. 17, p. 8643, 2022.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[83] V. H. Díaz-Ramírez, R. Juarez-Salazar, J. Zheng, J. E. Hernandez-Beltran, and A. Márquez, "Homography estimation from a single-point correspondence using template matching and particle swarm optimization." *Applied optics*, vol. 61 7, 2022.

[84] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. ieee, 1995, pp. 1942–1948.

[85] P. Garcia-Martinez, C. Ferreira, J. Garcia, and H. H. Arsenault, "Nonlinear rotation-invariant pattern recognition by use of the optical morphological correlation," *Applied Optics*, vol. 39, no. 5, pp. 776–781, 2000.

[86] S. D. Cochran and G. Medioni, "3-d surface description from binocular stereo," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 14, no. 10, pp. 981–994, 1992.

[87] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE, 2018, pp. 117–122.

[88] J. Wang, L. Perez *et al.*, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Networks Vis. Recognit*, vol. 11, no. 2017, pp. 1–8, 2017.

[89] J. Luo, E. S. Ruezga, and J. Davis, "How accurate is passive stereo for 3d face reconstruction?" in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2516–2520.

[90] M. Sui, H. Li, Z. Zhu, and F. Zhao, "Afnet-m: Adaptive fusion network with masks for 2d+ 3d facial expression recognition," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 116–120.

[91] K. R. Castleman, *Digital image processing*. Prentice Hall Press, 1996.

[92] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[93] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 30–40.

[94] F. Isgro and E. Trucco, "Projective rectification without epipolar geometry," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 1. IEEE, 1999, pp. 94–99.

[95] H. Ko, H. S. Shim, O. Choi, and C.-C. J. Kuo, "Robust uncalibrated stereo rectification with constrained geometric distortions (usr-cgd)," *Image and Vision Computing*, vol. 60, pp. 98–114, 2017.

[96] A. Fusiello and L. Irsara, "Quasi-euclidean epipolar rectification of uncalibrated images," *Machine Vision and Applications*, vol. 22, pp. 663–670, 2011.

[97] R. Juarez-Salazar, O. I. Rios-Orellana, and V. H. Diaz-Ramirez, "Stereo-phase rectification for metric profilometry with two calibrated cameras and one uncalibrated projector," *Applied Optics*, vol. 61, no. 21, pp. 6097–6109, 2022.

[98] J. Mallon and P. F. Whelan, "Projective rectification from the fundamental matrix," *Image and Vision Computing*, vol. 23, no. 7, pp. 643–650, 2005.

[99] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[100] H. K. Sangappa and K. Ramakrishnan, "A probabilistic analysis of a common ransac heuristic," *Machine Vision and Applications*, vol. 30, pp. 71–89, 2019.

[101] J. Y. Z. D. F. Guo and H. Wang, "Multiview image rectification algorithm for parallel camera arrays," *Journal of Electronic Imaging*, vol. 23, no. 3, pp. 1–33, 2013.

[102] D. Scharstein and C. J. Pal, "Learning conditional random fields for stereo," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[103] Y. Hou, C. Liu, B. An, and Y. Liu, "Stereo matching algorithm based on improved census transform and texture filtering," *Optik*, vol. 249, p. 168186, 2022.

[104] Y. Wang, M. Gu, Y. Zhu, G. Chen, Z. Xu, and Y. Guo, "Improvement of ad-census algorithm based on stereo vision," *Sensors*, vol. 22, no. 18, p. 6933, 2022.

[105] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu, "Constant time weighted median filtering for stereo matching and beyond," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 49–56.

[106] J. Li, Z. Li, and H. Zhao, "Efficient global stereo-matching method of general images under a long baseline based on baseline estimation." *Applied optics*, vol. 60 27, pp. 8248–8257, 2021.

[107] T.-M. Wang and Z.-C. Shih, "Measurement and analysis of depth resolution using active stereo cameras," *IEEE Sensors Journal*, vol. 21, no. 7, pp. 9218–9230, 2021.

[108] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 2006, pp. 211–216.

[109] P. K. Diederik, "Adam: A method for stochastic optimization," *(No Title)*, 2014.

[110] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[111] Q. Mao, Q. Rao, Y. Yu, and M. Dong, "Hierarchical bayesian theme models for multipose facial expression recognition," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 861–873, 2016.

[112] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 263–270.

[113] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3359–3368.

[114] Y. Fu, Q. Ruan, Z. Luo, Y. Jin, G. An, and J. Wan, "Ferlrtc: 2d+ 3d facial expression recognition via low-rank tensor completion," *Signal Processing*, vol. 161, pp. 74–88, 2019.

[115] Y. Jiang and Q. Ruan, "Multi-feature tensor neighborhood preserving embedding for 3d facial expression recognition," *IEEE Access*, vol. 9, pp. 106 303–106 316, 2021.