www.komputersapiens.org



KOMPUTER SAPIENS

Revista de Divulgación de la Sociedad Mexicana de Inteligencia Artificial

Año 17
Volumen 1
enero-abril 2025

Videovigilancia predictiva

Obteniendo la geometría de objetos: renderizado

Del cerebro a la máquina: sinapsis digitales

Visión artificial e inclusión: Traductor de señas

Detección de enfermedades foliales en maíz con IA

Localización basada en visión computacional

Visión artificial y la robótica: Robots que ven

Aprendizaje profundo para la localización aérea

Detección de residuos sólidos urbanos reciclables

ISSN 2007-0691

9 772007 069007

©Komputer Sapiens, Año XVII Volumen I, enero-abril, es una publicación cuatrimestral de la Sociedad Mexicana de Inteligencia Artificial, A.C., con domicilio en Ezequiel Montes 56 s/n, Fracc. los Pilares, Metepec, Edo. de México, C.P. 52159, México, http://www.komputersapiens.org, correo electrónico: editorial@komputersapiens.org, tel. +52 (833)357.48.20 ext. 3024, fax +52 (833) 215.85.44. Impresa por Sistemas y Diseños de México S.A. de C.V., calle Aragón No. 190, colonia Álamos, delegación Benito Juárez, México D.F., C.P. 03400, México, se terminó de editar el 6 de enero de 2025.
Reserva de derechos al uso exclusivo número 04-2009-111110040200-102 otorgado por el Instituto Nacional de Derechos de Autor. ISSN 2007-0691.
Los artículos y columnas firmados son responsabilidad exclusiva de los autores y no reflejan necesariamente los puntos de vista de la Sociedad Mexicana de Inteligencia Artificial. La mención de empresas o productos específicos en las páginas de Komputer Sapiens no implica su respaldo por la Sociedad Mexicana de Inteligencia Artificial.
Queda estrictamente prohibida la reproducción total o parcial por cualquier medio, de la información aquí contenida sin autorización por escrito de los editores.
Komputer Sapiens es una revista de divulgación en idioma español de temas relacionados con la inteligencia artificial.
Creada en LaTeX, con la clase **papertex** disponible en el repositorio *CTAN*: Comprehensive TeX Archive Network, http://www.ctan.org/
Indizada en el IRMDCT de CONACYT y en Latindex.

Directorio SMIA

Presidenta	María Lourdes Martínez Villaseñor
Vicepresidenta	Hiram Eredín Ponce Espinosa
Secretario	Roberto Antonio Vázquez Espinosa de los Monteros
Tesorero	Gilberto Ochoa Ruiz
Vocales:	Gustavo Arroyo Figueroa
	Noé Alejandro Castro Sánchez
	Hugo Jair Escalante Balderas
	Sofia Natalia Galicia Haro
	Antonio Marín Hernández
	Bella Citlali Martínez Seis
	Iris Iddaly Méndez Gurrola
	Leobardo Morales Tiburcio
	César Jesús Núñez Prado
	Iván Alejandro Olvera Rodríguez
	Obdulia Pichardo Lagunas
	David Eduardo Pinto Avendaño
	Victor Josué Ruiz Martínez

Komputer Sapiens

Directora general	María Lourdes Martínez Villaseñor
Editora en jefe	Karina Mariela Figueroa Mora
Editor científico	Héctor Gabriel Acosta Mesa
Editores asociados	Marco Antonio Aguirre Lam
	Laura Cruz Reyes
Editores invitados	Gilberto Ochoa Ruiz
Coordinadora de producción	Viridiana Mena Gómez
e-Tlakuilo	Karina Figueroa
	Rosa María Valdovinos
Estado del IArte	Reyna Carolina Medina Ramírez
	Jorge Rafael Gutiérrez Pulido
Sakbe	Claudia G. Gómez Santillán
	Laura Cruz Reyes
	Laura Gómez Cruz
IA & Educación	María Yasmín Hernández Pérez
	J. Julieta Noguez Monroy
	María Lucía Barrón Estrada
Deskubriendo Konocimiento	Alejandro Guerra Hernández
	Gildardo Sánchez Ante
Corrección de estilo	Guillermo de Jesús Hoyos Rivera
	Luis Miguel García Velázquez
	Marcela Quiroz Castellanos
	Nelson Rangel Valdez
	Nicandro Cruz Ramírez
	Sonia Lilia Mestizo Gutíerrez
	Susana Aurora Velasco Montiel

Directores Fundadores

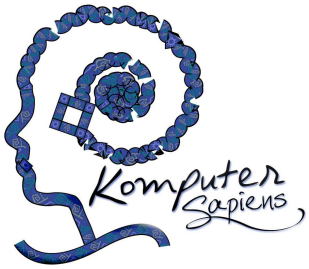
Carlos Alberto Reyes García
Ángel Kuri Morales

Comité Editorial

Ildar Batyrshin
Oscar Herrera Alcántara
Jesús Favela Vara
Sofía Natalia Galicia Haro
Miguel González Mendoza
Raúl Monroy Borja
Eduardo F. Morales Manzanares
Leonardo Garrido Luna
Carlos Alberto Reyes García
Antonio Sánchez Aguilar
Luis Enrique Sucar Succar
Ángel Kuri Morales
José A. Martínez Flores
Juan Manuel Ahuactzin Larios
Manuel Montes y Gómez
Ofelia Cervantes Villagómez
Alexander Gelbukh
Grigori Sidorov
Laura Cruz Reyes
Ramon Brena Pinero
Juan Humberto Sossa Azuela
Félix Castro

Árbitros

Adriana Laura López Lobato
Aldo Márquez-Grajales
Andrés Espinal
Angel Sánchez García
Arturo Cerón López
Carlos-Alberto López-Herrera
Cleber Zanchettin
David Herrera-Sánchez
Emmanuel Ovalle-Magallanes
Estefania Talavera
Fernando Arámbula
Francisco López-Tiro
Gildardo Sánchez-Ante
Héctor Acosta-Mesa
Homero Vladimir Ríos Figueroa
Hugo Jair Escalante
Ivan Reyes-Amezcua
Jorge Cruz-Duarte
Jorge Valverde-Rebaza
José Luis Reyes
Josimar Chire Saire
Karen Sánchez
Karina Figueroa Mora
Karina Gutiérrez-Fragoso
Luis Luevano
Nancy Pérez-Castro
Salvador Hinojosa
Sergio Hernández Méndez
Sonia Mestizo-Gutiérrez



Contenido

Del Cerebro a la Máquina: Sinapsis Digitales con Redes Neuronales de Impulso
Pág. 10 ⇒Valeria Carrillo Cabrera y Edgar Alejandro Guerrero Arroyo

Visión artificial e inclusión: creando un traductor de señas con Teachable Machine
Pág. 15 ⇒ Marco Antonio Aguilar Cortés, Miguel Aguilar Cortés y Luis Ángel Martínez Cabrera

Obteniendo la geometría de objetos en una imagen usando renderizado diferencial
Pág. 20 ⇒ José Abraham Ramos Badillo, Luis Eduardo Gamboa Guzmán y Karina Mariela Figueroa Mora

Odometría Visual con IA: Innovando las Tecnologías de Localización
Pág. 25 ⇒ Leonardo Herrera

Videovigilancia predictiva: anticipándose al robo
Pág. 30 ⇒ L. Enrique Morales Márquez, Iván Olmos Pineda y J. Arturo Olvera López

La Visión Artificial y la Robótica: Robots que Ven
Pág. 34 ⇒ JRodrigo Juárez-Villalva, Efrén Castro-Bobadilla, Hugo Yáñez-Badillo, Laura Cleofas-Sánchez, J. Nabor Balderas-Gutiérrez y Francisco Beltrán-Carbajal

Detección de enfermedades foliares en maíz mediante aprendizaje profundo
Pág. 40 ⇒ Víctor A. González Huitrón, Jesús A. Medrano Hermosillo, Abraham E. Rodríguez Mata, José A. Almaráz Damián y Pedro Celestino Ramírez

Detección de residuos sólidos urbanos reciclables con Visión Artificial
Pág. 45 ⇒ Josué Garatachia Mejía, Erubiel Urbina Gómez, Hector A. Montes Venegas, Rosa María Valdovinos Rosas, Marcelo Romero Huertas

Aprendizaje Profundo para Localización Aérea
Pág. 50 ⇒ Aldrich Alfredo Cabrera-Ponce, Manuel Isidro Martin Ortíz y Jose Martínez-Carranza

Columnas

Sapiens Piensa. Editorial pág. 2

Sapiens Invitados. Editorial pág. 4

e-Tlakuilo pág. 5

Estado del IArte pág. 7

Sakbe pág. 8

IA & Educación pág. 55

Deskubriendo Konocimiento pág. 57

Sapiens Piensa

Gilberto Ochoa Ruiz

La inteligencia artificial (IA), y en particular la visión computacional (VC) tienen cada vez mayor impacto en nuestras vidas, y de acuerdo con estudios de la OMPI, es de lejos una de las áreas de investigación de la IA con mayor cantidad de publicaciones, patentes y desarrollos comerciales. La VC se ocupa de la teoría detrás de los sistemas artificiales que extraen información de las imágenes y que pueden tomar muchas formas: como secuencias de video, vistas de múltiples cámaras, datos multidimensionales de un escáner 3D, nubes de puntos 3D de sensores LiDaR o dispositivos de escaneo médico. La disciplina tecnológica de la visión artificial busca aplicar sus teorías y modelos a la construcción de sistemas de visión artificial. La VC también involucra la resolución de tareas a partir de la combinación de información semántica (clasificación, localización y detección de objetos) con información geométrica, aunque pueden desarrollarse aplicaciones haciendo uso de éstas de forma independiente. Las subdisciplinas de la VC incluyen la reconstrucción de escenas, la detección y objetos de objetos, detección de eventos, reconocimiento de actividades, seguimiento de video, reconocimiento de objetos, estimación de pose 3D, aprendizaje, indexación, estimación de movimiento, servo-control visual, modelado de escenas 3D y restauración de imágenes, entre muchas otras.

En este número de Komputer Sapiens, se presentan 9 artículos que abordan temas relacionados a la VC entre las que se encuentran: Explicabilidad, Aprendizaje Automático, Interacción Hombre-Máquina, Medicina, Reconocimiento de Señas y Computación Afectiva, entre otras. Los artículos fueron seleccionados cuidadosamente por el consejo editorial designado con base tres evaluaciones realizadas por las personas del Comité Técnico conformado por expertos del área. Las personas que realizaron la revisión de cada trabajo examinaron aspectos como el tema a abordar (visión computacional), la relevancia, la calidad de la exposición del tema, la legibilidad y el enfoque a la divulgación.

El primer artículo, titulado "**Videovigilancia predictiva: anticipándose al robo**" aborda cómo la visión computacional puede ayudar en predecir eventos delictivos. El artículo describe algunos métodos para este propósito y su funcionamiento, así como algunos de los retos a enfrentar en esta tarea. Una de estas limitaciones está relacionada con el hecho que la mayor parte de las soluciones sólo se enfocan en ciertos tipos de delitos, usando las mismas bases de datos, por lo que los autores concluyen que una área de investigación consiste en incrementar las cantidades de datos existentes para incluir una mayor variedad de delitos. Por ejemplo, la inclusión de imágenes o casos en la vía pública, es decir, entornos menos menos controlados; es de suma importancia para obtener modelos más generales, capaces de predecir de manera adecuada la ocurrencia de los delitos.

Otro tema de vanguardia es abordado por el artículo "**Del Cerebro a la Máquina: Sinapsis Digitales con Redes Neuronales de Impulso**", el cual explora el potencial de las Redes Neuronales de Impulsos (Spiking Neural Networks, SNNs) para tareas de clasificación y regresión, centrándose en la comparación con redes neuronales tradicionales. Las SNNs representan un enfoque de tipo bioinspirado, que emula mejor el comportamiento neuronal biológico. Este enfoque les confiere ventajas significativas sobre métodos de aprendizaje profundo en términos de eficiencia energética y capacidad para procesar información temporal. Los autores presentan diversos experimentos para demostrar los beneficios de las SNN en tareas de clasificación. El artículo concluye que, si bien las SNNs presentan desafíos inherentes, como la necesidad de ajustar cuidadosamente los parámetros temporales y la sensibilidad a las condiciones de entrenamiento, su potencial para aplicaciones que requieren procesamiento temporal y eficiencia energética es prometedor.

Otra aplicación de impacto (social en este caso) es presentada en el trabajo "**Visión artificial e inclusión: creando un traductor de señas con Teachable Machine**" que aborda la creación de un modelo de traducción de señas utilizando Teachable Machine, una plataforma desarrollada por Google que permite entrenar modelos de Aprendizaje Automático de manera accesible y sin necesidad de conocimientos avanzados en programación. Se partió de un conjunto de imágenes sin fondo para entrenar el modelo, permitiendo la identificación y correcta ejecución de las señas de la Lengua de Señas Mexicana (LSM). El modelo se integró en una aplicación web, donde los usuarios pudieron realizar ejercicios frente a la cámara, obteniendo retroalimentación visual sobre la precisión de sus señas. Los resultados reportados por los autores fueron positivos, demostrando que la herramienta ayuda a los usuarios a identificar y aprender señas de manera efectiva. El trabajo futuro de los autores tomará en cuenta comentarios de los usuarios de este primer prototipo, tales como la incorporación de más señas y la integración de elementos de gamificación para hacer la experiencia más interactiva. Además, se identificó el desafío de incluir señas que requieren movimiento.

El siguiente artículo en esta colección, titulado "**Detección de enfermedades foliares en maíz median-**

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

te CNN y ViT: un breve análisis comparativo" presenta un análisis comparativo de modelos de aprendizaje profundo para la detección de enfermedades foliares en maíz utilizando el conjunto de datos Plantvillage. Los autores presentan resultados del entrenamiento y evaluación de seis modelos distintos de redes neuronales profundas, incluyendo arquitecturas basadas en Convolutional Neural Networks (CNN) y Vision Transformers (ViT). El estudio compara el desempeño de estos modelos en términos de tiempos de entrenamiento y métricas de calidad como precisión, recall, accuracy y F-score. Los resultados obtenidos revelan diferencias en la eficiencia de las distintas arquitecturas, proporcionando una visión integral sobre cuál de estos enfoques ofrece el mejor rendimiento en la detección de enfermedades en las hojas de maíz. Este análisis busca guiar futuras investigaciones y aplicaciones prácticas en la agricultura de precisión mediante el uso de técnicas avanzadas de visión artificial

El enfoque del artículo "**Localización Basada en Visión Computacional**" es en localización, misma que es crucial para la navegación autónoma y muchas otras aplicaciones. Se abordan técnicas de localización que prescinden del uso de visión computacional, también la localización basada en ésta y dicha técnica. La localización basada en visión computacional utiliza como dispositivo central la cámara de vídeo que es montada en un vehículo de interés para determinar la localización del vehículo a través de secuencias de imágenes capturadas. Diversas características clave (extraídas de imágenes consecutivas capturadas por la cámara) son analizadas de forma eficiente a través de una microcomputadora a bordo del vehículo para determinar el movimiento realizado por la cámara y así la localización de dicho vehículo. A esta tecnología se le conoce como odometría visual y se ha empleado para navegar vehículos en Marte donde no existe GPS. Similar a los sensores inerciales, la odometría visual es también independiente y compacta, donde lo único necesario para su operación es la cámara y una microcomputadora, lo cual la hace sumamente interesante en aplicaciones como robótica y realidad aumentada.

El penúltimo artículo de la colección, titulado "**Obteniendo la geometría de objetos en una imagen usando renderizado diferencial**" aborda un tema entre la visión computacional y las gráficas por computadora. Efectivamente, el renderizado de imágenes fotorrealistas ha tenido un gran desarrollo en años recientes y su implementación la podemos ver en muchos lugares como en películas animadas, videojuegos, diseño asistido por computadora, etc. El renderizado consiste en obtener una imagen 2D a partir de una escena 3D considerando la naturaleza de la luz. Ahora bien, el problema inverso consiste en reconstruir la escena 3D a partir de imágenes 2D. Esta área de estudio se llama renderizado inverso, la cual aborda grandes problemas de la visión computacional con un enfoque diferente, que consiste en entender cómo los objetos de una escena interactúan entre sí considerando la naturaleza de la luz. Algunas de sus aplicaciones son la reconstrucción de objetos en una escena, análisis de iluminación en una imagen, la optimización de la luz natural como parte de un proceso de diseño arquitectónico, entre otros. En este artículo se explica qué es un renderizador tradicional para después adentrarse en esta novedosa área de estudio del renderizado inverso, específicamente el método de renderizado diferencial.

En el último trabajo de esta colección "**La Visión Artificial y la Robótica: Robots que Ven**" se presentan las principales generalidades de la visión computacional y su relación con el desarrollo de sistemas robóticos. Se inicia con una breve reseña histórica de los sistemas de visión por computadora, destacando la relación que ésta tiene con la visión de máquina y la visión artificial. Posteriormente, se destacan algunos puntos importantes en el desarrollo de sistemas avanzados de navegación para robots móviles, donde se hace especial énfasis en la importancia de la visión computacional para mejorar la autonomía de estos sistemas que siguen revolucionando al mundo desde su aparición. Por último, se presentan dos casos de estudio relacionados con dos sistemas desarrollados por estudiantes de la carrera de ingeniería mecatrónica, donde se emplea la visión computacional con diferentes perspectivas.

Invitamos a todos los lectores, colegas, estudiantes y personas interesadas, a sumergirse en cada uno de los artículos de su interés. Y agradecemos a todos los autores que contribuyeron en el envío de artículos, así como al comité científico por concretar esta edición.✷

SOBRE EL AUTOR



Gilberto Ochoa Ruiz es miembro del Sistema Nacional de Investigadores (Nivel I), de la Red Tematica CONACYT en Inteligencia Computacional Aplicada (RedICA), la Sociedad Mexicana de Inteligencia Artificial (SMIA), de la Asociación Mexicana de Computación (AMEXCOMP), entre otras. Es además investigador asociado a tres programas Erasmus Mundus de la Unión Europea, el primero con el Master in Computer Vision and Robotics (Vibot) y al Master in Medical Imaging and Applications (MAIA+) y al Master in Computational Color and Spectral Imaging (COSI).

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Sapiens Invitados

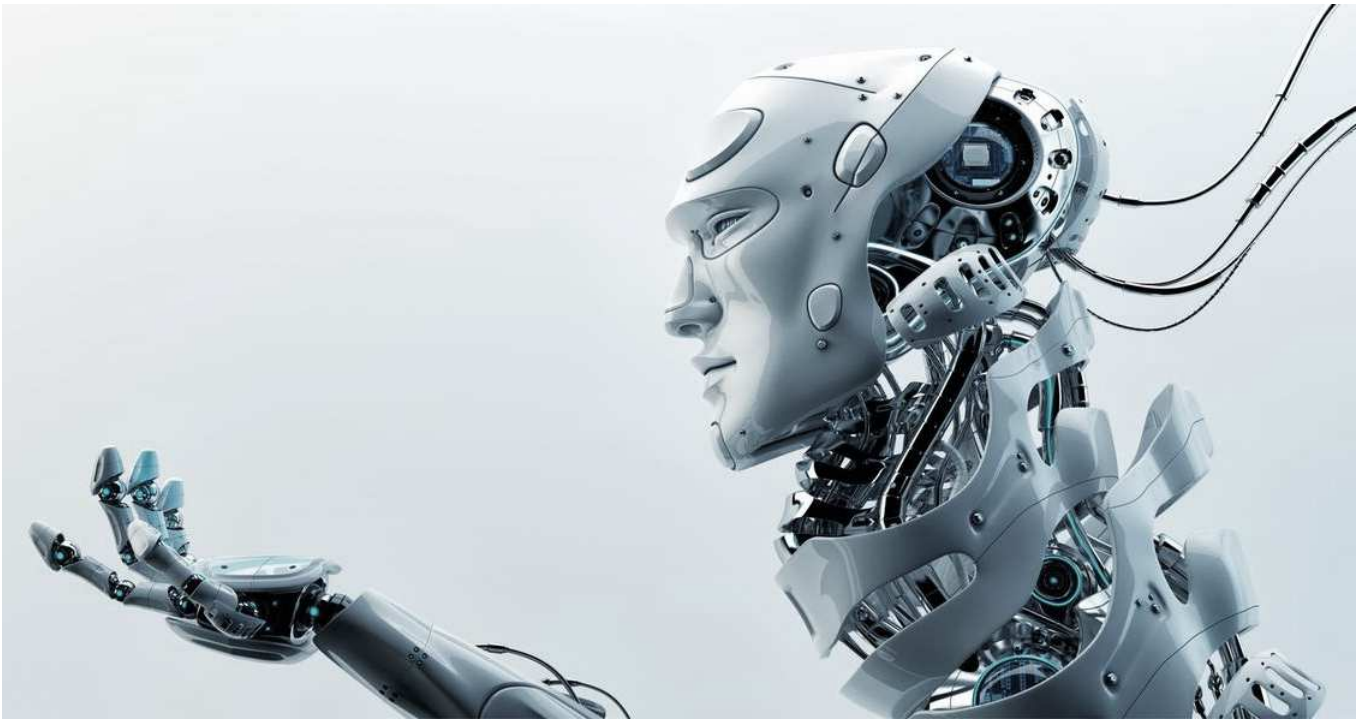
Karina Mariela Figueroa Mora

Para este volumen de Komputer Sapiens se presentan dos artículos invitados.

El primero fue seleccionado como uno de los mejores artículos del 1er Coloquio de Mujeres en Computación llevado a cabo durante la Asamblea anual de la Academia Mexicana de Computación. El artículo titulado Detección de residuos sólidos urbanos reciclables con visión artificial presenta el resultado de utilizar la Inteligencia Artificial (aprendizaje automático y redes neuronales) para la selección y separación de residuos para su reciclaje. Los residuos representan parte del problema de contaminación de nuestro tiempo. Para todos es sabido que la separación y reutilización de los mismos es una pequeña contribución al problema. Este artículo encaja muy bien en este volumen por tratarse de un uso de la visión computacional.

El segundo artículo invitado fue sometido y aprobado para el volumen 3 del año 16 sin embargo, no fue publicado en su momento. En este artículo titulado Aprendizaje profundo para localización aérea es otro uso de la visión computacional, como describen los autores, para establecer una posición de referencia en lo que se espera la restauración de la señal del sistema GPS. Esto es, a partir de imágenes capturadas se extrae información que permiten la estimación de la posición mientras el dron sigue su trayectoria de vuelo.

Invitamos a nuestros lectores a leer estos artículos que amplian el panorama de los alcances de la visión computacional.✷



© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

e-Tlakuilo: Cartas de nuestros lectores

Karina Figueroa y Rosa María Valdovinos
etlakuilo@komputersapiens.org

Los autores agradecemos el interés de los lectores en nuestro artículo **"La evaluación de habilidades en cirujanos laparoscópicos: La inteligencia artificial al rescate"**. A continuación, respondemos las preguntas que amablemente nos hicieron llegar.

La evaluación de habilidades en cirujanos laparoscópicos: La inteligencia artificial al rescate

Pregunta: Oscar Eduardo Franco Juan

¿Cuáles son los resultados preliminares obtenidos con las RNN y las SVM?

Nuestro artículo presenta solamente el procesamiento inicial de las señales de EMG y de aceleración, así como una interfaz simple en Python. Los modelos en RNN y SVM son propuesta de lo que se podría hacer con estas señales. Pero aún no se realiza.
Sin embargo, considerando el estado del arte al respecto, se espera una precisión de al menos 80 %, según lo que reportan en (Lam et al., 2022), quien realizó un artículo de revisión sobre técnicas para evaluar habilidades de cirujanos usando ML. Si bien, no son las mismas condiciones ni mediciones, esperaríamos obtener al menos algo similar o superior.
Lam, K., Chen, J., Wang, Z., Iqbal, F. M., Darzi, A., Lo, B., Sanjay Purkayastha, & Kinross, J. M. (2022). Machine learning for technical skill assessment in surgery: a systematic review. Npj Digital Medicine, 5(1). `https://doi.org/10.1038/s41746-022-00566-0`.

Si bien, el conjunto de datos es sumamente amplio en cuestión de señales por operador, sólo se están evaluando a 34 operadores del sistema quirúrgico Da Vinci. ¿Los resultados obtenidos podrían estar sesgados por este reducido espacio muestral?

Efectivamente, nuestra muestra actual de 34 participantes (10 expertos en cirugía laparoscópica, 10 personal médico y 14 personas de la población general) es una muestra pequeña en términos de participantes. El tamaño de sujetos muy posiblemente podría causar sesgos. Considerando la regla empírica de 10 veces el número de características por clase y si se plantean, por ejemplo, 10 características por sujeto y tres clases (Expertos, no expertos y novatos), necesitaríamos al menos 30 sujetos por clase, por lo que requeriría al menos 90 sujetos en total.
Es importante mencionar que en el artículo aún no presentamos resultados obtenidos de modelos, solo es el procesamiento inicial de las señales. Por lo que si quisiéramos entrenar modelos robustos, requerimos un mayor número de sujetos, los cuales necesitamos considerar como trabajo futuro.

¿Cuál es el costo computacional del proyecto?

Dado el estado del proyecto, aún no podemos cuantificar el costo computacional si quisiéramos considerar el tiempo de entrenamiento, predicción o tiempo total. Ya que aún no implementamos ningún modelo.
Sin embargo, para el procesamiento de las señales por sujeto, considerando:

- 6 sensores por participante
- Señales EMG a 2000 Hz
- Señales de aceleración a 148 Hz
- 3 minutos por tarea de entrenamiento

El procesamiento actual incluyendo filtrado digital para eliminar ruido e interpolación para datos faltantes, le toma menos de 1 segundo a una computadora con procesador i7, 16MB de RAM y SSD.

¿Cuál fue el mayor reto durante el desarrollo de esta primera parte del proyecto?

El principal reto en este tipo de trabajos es la obtención de datos. La logística para reunir a los participantes, conseguir el tiempo para el uso de la consola de entrenamiento del robot Davinci, etc. fue sin duda lo más complicado.
La segunda tarea más complicada será sin duda asegurar la calidad de las señales, ya que las señales EMG sueles ser muy sensibles a los artefactos de movimiento, por lo que se deben revisar cuidadosamente con el fin de no introducir datos "basura" al modelo.

¿Bajo que parámetros se evalúa que las características extraídas de las señales son las apropiadas para someter a entrenamiento?

La selección de características extraídas de las señales se basa en lo reportado en el estado del arte cuando se usa este tipo de señales (EMG y Aceleración). Generalmente para este tipo de señales se consideran parámetros relacionados con la energía de la señal y su comportamiento

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

frecuencial. Que en este caso de estudio indicarían de alguna manera si los movimientos de los participantes fueron mayores o menores al promedio, considerando que todos realizan las mismas tareas.
Por ejemplo, se esperaría que un cirujano experto realice la tarea con movimientos más estables, cortos y precisos, comparados con los novatos, lo que se reflejaría en señales con menor energía y tiempos más cortos.
Sin embargo, en el estado actual del proyecto, aún no se llega a evaluar ciales son las señales más adecuadas, dado que aún no se implementa algún modelo.

¿De qué manera se evaluarían los resultados de los modelos para verificar que dichos sean precisos?

Si bien aún no se han implementado los modelos de IA, se planea evaluar los resultados mediante:

- Validación cruzada k-fold
- Comparación con evaluaciones de expertos
- Evaluación longitudinal de la progresión de habilidades

De manera personal considero que otra aplicación además de evaluar en tiempo real al cirujano, es poder evaluarlo durante su formación dando como resultado un "egreso" que haga valer que el cirujano está capacitado para realizar laparoscopías de manera eficaz.

Agradecemos su pregunta, efectivamente, un posible uso del proyecto planteado es precisamente que sea útil para evaluar en tiempo real el progreso de los cirujanos. De esta manera, una vez que se determinen que combinación de sensores (EMG, acelerómetros, colocación y número) es el más adecuado, así como el tipo de características de las señales y se cuente con el modelo entrenado más adecuado. Se podría implementar para evaluar y proporcionar un puntaje a los cirujanos en tiempo real mientras realizan su entrenamiento, y quizá hasta podría proporcionar sugerencias a los cirujanos para mejorar su técnica y asemejarse más a un experto.



© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Estado del IArte

Jorge Rafael Gutiérrez Pulido (@jrgpulido) y Reyna Carolina Medina Ramírez
estadoiarte@komputersapiens.org

IA++. Hacia un compromiso de uso razonable

En 2024 hemos sido testigos de la eclosión del uso de algoritmos de Inteligencia Artificial (IA) para realizar todo tipo de actividades, desde aquellas de la vida cotidiana hasta la exploración espacial. Los usuarios comunes tienen acceso, incluso sin solicitarlo, a las capacidades que ofrece la IA en aplicaciones tales como Whatsapp, Facebook, Instagram, Office por mencionar sólo algunos ejemplos. El uso que hacemos de dichas aplicaciones, permite que sigan aprendiendo y mejoran sus algoritmos. Lo que la mayoría de estos usuarios desconoce es que estas actividades - resumir un documento, generar una presentación o crear un sticker - tienen un costo energético y por ende un impacto al medio ambiente. El problema no es el acceso o la facilidad del uso de este tipo de aplicaciones para el público en general, ni las aplicaciones *per se*; sino la falta de compromiso entre los consumidores, los productores de tecnología IA y las reglas que rigen a ambos. En columnas anteriores hemos abordado el impacto al medio ambiente y costo energético que ocasiona el uso indiscriminado de las tecnologías[1], así como la necesidad de establecer criterios éticos y regular el desarrollo de tecnologías derivadas de algoritmos de IA [2].

El uso razonable de aplicaciones de IA depende de un compromiso entre tres actores: el usuario, las empresas y el gobierno. Como usuarios de aplicaciones debemos educarnos en el uso de la tecnología por convicción y no por facilidad, debemos evitar el abuso que se hace de estas tecnologías para realizar tareas humanamente sencillas. En lo que respecta a las empresas tanto desarrolladoras como consumidoras, Marvin Minsky[3] en el premio BBVA 2013 destacaba que se pueden crear máquinas tan inteligentes como las personas, teniendo los recursos económicos y el capital humano dedicado a la IA. Once años después de esta declaración, somos testigos de la lucha masiva entre las empresas que implementan los últimos avances en IA generativa. La revista Forbes[4] lanzó su sexto IA50 que reconoce a las empresas privadas de IA prometedoras, entre las que encontramos *OpenAI*, *Anthropic*, *Databricks* entre otras. También se destaca *Mistral*, considerado la competencia de *OpenIA*, un startup que desarrolla modelos de IA de código abierto y en su mayoría gratuitos que requieren menos procesamiento computacional. Mistral lanzó "*LeChat*" una alternativa que requiere menos procesamiento que *ChatGPT*. Entre las empresas que incorporan IA a sus aplicaciones existentes tenemos a *Google Workspace* y *Microsoft Office* como ejemplos. Por otro lado, las empresas consumidoras de tecnología IA deben mantener un equilibrio, entre los ingresos económicos y discernir entre utilizar la IA para potencializar a su capital humano o tomar la decisión de sustituirlos. Finalmente, el pasado 21 de marzo la Asamblea General de la ONU adoptó una resolución para regular los desarrollos tecnológicos en el área de IA. En particular, pide a los estados miembros: *que se abstengan de utilizar sistemas de inteligencia artificial que no pueden funcionar de conformidad con las normas internacionales de derechos humanos o los pongan en riesgo*[5].

Como personas pertenecientes a una sociedad, desde esta columna les invitamos a reflexionar y educarnos en el consumo razonable de aplicaciones con IA. Discriminar las tareas de nuestra vida cotidiana en las cuales es importante o crucial utilizar la IA; ser conscientes que toda consulta conlleva un costo energético y por ende un impacto al medio ambiente. El avance de la IA la podemos comparar con el sol. Siempre estará ahí, y así como hemos podido aprovecharlo, también hemos aprendido a protegernos, por lo que podríamos actuar de la misma forma con el uso de la IA.✷

[1] Hacia una IA sustentable - IArte - Año 14 (2022) número 3
[2] Ética para la IA - IArte - Año 15 (2023) número 3
[3] https://shorturl.at/51kZL
[4] https://www.forbes.com.mx/lista-forbes-ia-50-estas-son-las-principales-empresas-emergentes-de-inteligencia-artificial/
[5] https://news.un.org/es/story/2024/03/1528511

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Sakbe

Claudia Guadalupe Gómez Santillán, Laura Cruz Reyes y Laura Gómez Cruz
sakbe@komputersapiens.org

Visión Computacional

La visión computacional es un área de conocimiento enfocada en llevar a cabo algoritmos para que las computadoras puedan generar, interpretar y descifrar imágenes, fotos o videos, lo más similar a como lo haría el ser humano. La inteligencia artificial (IA) aplicada a la visión computacional ha logrado un incremento rápido en los últimos años, este 2024 ha sido un año importante en el desarrollo de estas tecnologías permitiendo a las máquinas aprender a reconocer patrones y objetos, analizar imágenes y tomar decisiones basadas en la información visual.

Algunos institutos o universidades que ofrecen cursos, maestrías o diplomados a distancia en Visión Computacional son:

La Universidad Católica de San Pablo ⇒ mediante su Departamento de Ciencia de la Computación, ofrece el Diplomado en Visión por Computador con Machine Learning y Deep Learning (https://cs.ucsp.edu.pe/postgrado/diplomado-vision-computador-con-machine-learning-deep-learning/), está dirigido a estudiantes o profesionales que necesitan una especialización en visión artificial dirigida al mercado laboral, la industria, la investigación o al emprendimiento. La plataforma de cursos virtuales **Udemy** ofrece una gran cantidad de cursos dentro de los cuales se pueden encontrar cursos básicos y avanzados en Python para visión artificial, procesamiento digital de imágenes con Python, Computer Vision con Python - Visión artificial con Python, entre otros (https://www.udemy.com/es/topic/computer-vision/). También la **UNAM** a través de su **Posgrado en Ciencias e Ingeniería de la Computación** cuenta con el Laboratorio Avanzado de Procesamiento de Imágenes donde se desarrollan métodos y algoritmos para la visión computacional y el procesamiento digital de imágenes usando modelos matemáticos de Inteligencia Artificial, así como propiedades y características inspiradas en el Sistema de Visión Humano, dando la oportunidad de investigación y desarrollo para estudiantes de licenciatura, maestría y doctorado en esta área de estudio (https://lapi.fi-p.unam.mx/).

https://cs.ucsp.edu.pe/postgrado/diplomado-vision-computador-con-machine-learning-deep-learning/, https://www.udemy.com/es/topic/computer-vision/ y https://lapi.fi-p.unam.mx/

Apple Vision Pro ⇒ Dentro de las aplicaciones de visión computacional novedosas que surgieron en este año esta **Apple Vision Pro** que es un visor de realidad virtual que te permite transformar cualquier habitación en tu propio cine personal o te ofrece espacio ilimitado para organizar todo lo que necesites a tu alrededor, incorporando flujos de trabajo de tu Mac mediante *Mac Virtual Display*, además de que se puede usar como una

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

cámara 3D de Apple, permite capturar fotos espaciales y vídeos espaciales en 3D, con audio espacial envolvente (https://www.apple.com/apple-vision-pro/). La visión computacional también ha impactado positivamente en la industria; según la compañía Bosh las áreas en las que la visión computacional moldeará y cambiará el futuro del sector industrial son: en el control de calidad e inspección, optimización de la cadena de suministro, mantenimiento predictivo, colaboración máquina humano, entre otras; "En Bosch, como pioneros en Industria 4.0, se tiene una visión computacional como un pilar fundamental de la Industria 4.0 y estamos comprometidos a integrar esta tecnología para crear fábricas más inteligentes, seguras y eficientes", destaca Juan Antonio Relaño Pinilla, responsable de tecnología e innovación de Robert Bosch España (https://www.automaticaeinstrumentacion.com/texto-diario/mostrar/5048984/bosch-identifica-seis-areas-clave-donde-vision-computacional-transformara-industria-40, https://revistabyte.es/actualidad-it/industria-vision/).

https://www.apple.com/apple-vision-pro/, https://www.automaticaeinstrumentacion.com/texto-diario/mostrar/5048984/bosch-identifica-seis-areas-clave-donde-vision-computacional-transformara-industria-40 y https://revistabyte.es/actualidad-it/industria-vision/



Imagen tomada de: https://blog-es.lac.tdsynnex.com/que-es-la-vision-computacional/.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Del Cerebro a la Máquina: Sinapsis Digitales con Redes Neuronales de Impulso

Valeria Carrillo Cabrera y Edgar Alejandro Guerrero Arroyo

Primeros impulsos

Los humanos siempre hemos tenido curiosidad y asombro sobre el funcionamiento de nuestro cerebro, por su misteriosa capacidad de almacenar recuerdos, tomar decisiones y procesar la información externa. La prueba de la afirmación anterior eres tú, leyendo este artículo. Investigadores han dado su máximo para satisfacer esa curiosidad y entender mejor este órgano prodigioso y, aunque es seguro que falta mucho por descubrir, en los últimos años se han dado pasos gigantes hacia esa meta.

Una de las innovaciones más destacadas son las redes neuronales (NN: Neural Networks), un modelo de aprendizaje automático que intenta emular el proceso cerebral, pues consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para resolver problemas muy interesantes, como visión por computadora y reconocimiento de voz.

Las NN se ejecutan en supercomputadoras centralizadas que requieren muchos recursos computacionales y energéticos debido a su complejidad y al procesamiento intensivo de datos, Figura 1. Sin embargo, el cerebro humano no sigue un procedimiento centralizado: está compuesto por miles de millones de neuronas en redes interconectadas donde cada neurona realiza operaciones locales, y la información se procesa de manera distribuida en todo el cerebro. Además, es enérgicamente más eficiente, pues realizar billones de operaciones por segundo consume solo 20 vatios, lo mismo que una bombilla, mientras que las supercomputadoras requieren miles de vatios, comparables a la energía de todo un edificio.

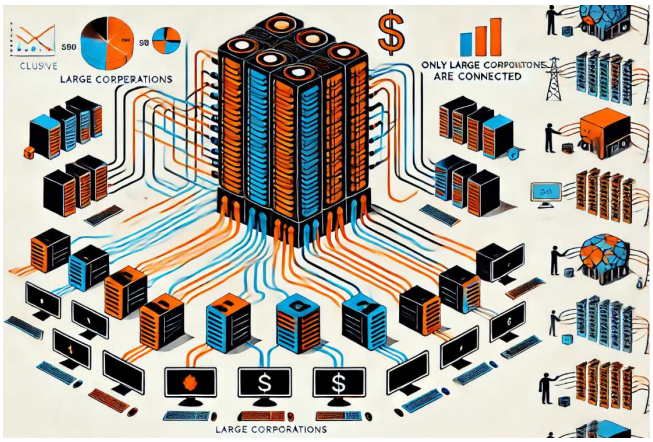


Figura 1. El cómputo centralizado y recursos energéticos costosos es práctico solo para las grandes compañías.

Es entonces donde entran en juego las redes neuronales por impulso (SNN: Spiking Neural Networks), que reflejan mejor la realidad biológica utilizando energía para activar neuronas solo cuando se recibe un estímulo.

Las investigaciones que sentaron las bases para el modelo SNN y describieron cómo una neurona biológica responde a los estímulos eléctricos son atribuidas al francés Louis Édouard Lapicque quien, en 1907, estimuló fibras nerviosas en el anca de una rana para observar cómo la duración y amplitud de la corriente afectaba la contracción muscular. El neurocientífico concluyó que estas dinámicas se asemejaban a un circuito de filtro de paso bajo. En español, esto significa que se permite el paso de señales de baja frecuencia mientras se atenúan o bloquean las de alta frecuencia.

Esta idea evolucionó a lo que hoy en día llaman modelo de neuronas con fuga de integración y disparo (LIF: Leaky Integrate-and-Fire model), el cual es una simplificación del modelo de neurona propuesto en 1952 por Hodgking y Huxley [1]. Una simplificación era necesaria, pues el trabajo de estos fisiólogos es un modelo complejo sobre los mecanismos iónicos base de la propagación de los potenciales de acción, señales eléctricas breves que permiten a las neuronas transmitir información. Como dato cultural, Hodgkin y Huxley centraron su estudio en el axón gigante del calamar, un modelo ideal debido a su gran tamaño (de 0.5 mm de diámetro aprox.), que controla parte de su sistema de propulsión de agua. Por este aporte a la comprensión de la fisiología neuronal se les otorgó el *Premio Nobel de Fisiología o Medicina* en 1963.

Anatomía de las neuronas por impulso

A través de las investigaciones sabemos que, cuando una neurona recibe un estímulo de una fuente externa o de otra neurona, se abren los canales iónicos en la membrana de la neurona que permiten que ciertos iones de sodio (partículas cargadas) entren rápidamente en la célula. Este flujo provoca un cambio en el voltaje de la membrana de la neurona, el potencial de membrana. Este proceso se denomina despolarización. Además, la membrana tiene una tendencia natural a permitir que la carga eléctrica escape a través de ella, como una "fuga".

Si con todo y la fuga el cambio en el voltaje alcanza un umbral específico, se genera entonces un potencial de acción, un "disparo" eléctrico que se propaga a lo largo del axón de la neurona hacia su terminal. Al llegar a

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

la terminal, este impulso provoca la liberación de neurotransmisores que mandan la señal a la siguiente neurona. Después, la neurona vuelve a su estado de reposo o periodo refractario, preparándose para recibir otro estímulo.

En un intento por replicar esas funciones cerebrales, recurrimos al modelo LIF, que de manera similar a las redes neuronales clásicas hace una suma ponderada de los datos de entrada, excepto que: **a)** la información de entrada serán impulsos y no datos etiquetados o numéricos como los que usan las neuronas clásicas y **b)** en lugar de procesar la suma inmediatamente con una función de activación, la neurona LIF integrará esta adición a lo largo del tiempo, si esta integración supera un cierto umbral entonces la neurona LIF envía un impulso.

La idea es definir el potencial de membrana como una función P_t del tiempo y de los impulsos de entrada. Y si contamos el tiempo de manera discreta, es decir, dividimos en instantes de tiempo, entonces cada evento, como la llegada de un impulso, se registra en un momento específico dentro de estos pasos. Imaginemos un arreglo de números [0...0 1 0...0] donde el 1 indica un impulso y su posición indica el instante de tiempo, t. Este enfoque facilita el procesamiento digital, ya que operar con ceros y unos es sencillo en general. Para simular el efecto de fuga, el potencial de membrana disminuye por un valor constante, β, en cada paso de tiempo. A través de la Figura 2 podemos imaginar los procesos de cada modelo.

La neurona LIF no se preocupa por cómo se ve el impulso que envía, sino por cuándo lo envía. Esto es importante porque las neuronas en nuestro cerebro también funcionan así, activándose en momentos específicos. Para imitar bien el cerebro, los datos también deben ser temporales. El uso de estos datos secuenciales nos recuerda a otras redes que en su momento acapararon la atención, las redes recurrentes (RNN: Recurrent Neural Networks).

Las RNN se presentaron como un tipo especial de redes con conexiones retroalimentadas, es decir, permiten que la salida de algunas neuronas afecte la entrada de las mismas, lo que significa que pueden mantener y utilizar información sobre eventos pasados en la secuencia de datos de entrada. Pero mantengámoslo simple, imagina que leemos un cuento, cada que pasamos la página queremos recordar lo que pasó en la anterior para entender bien la historia presente. Estas redes son capaces de retener información y utilizarla en momentos posteriores.

Conceptualmente, las SNN se pueden pensar como una generalización de las RNN en el dominio de los impulsos, Figura 3. Las SNN tienen la capacidad de procesar información temporal de manera más fina y biológicamente plausible al utilizar el tiempo exacto de los impulsos para transmitir información, mientras que las RNN utilizan un estado oculto continuo para mantener la información temporal, más detalles en la referencia [2].

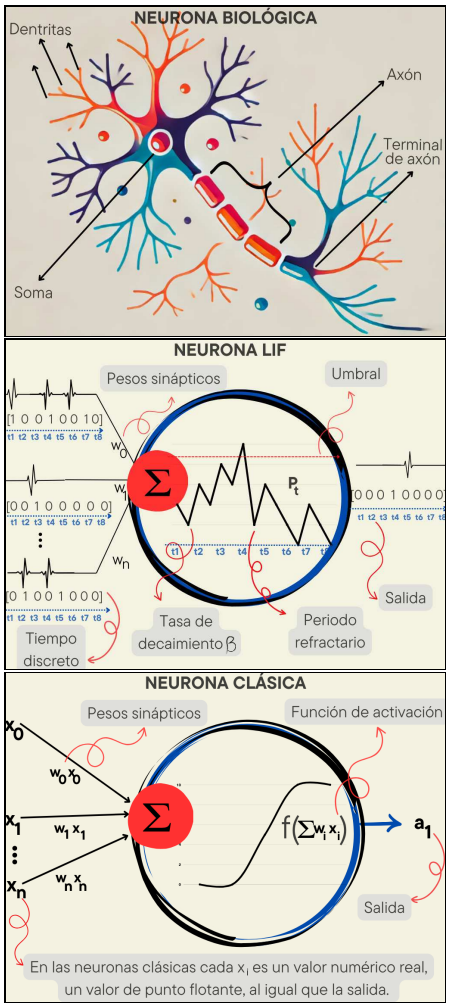


Figura 2. Comparación de los componentes clave de cada modelo de neurona. En la neurona LIF los impulsos suceden en ciertos instantes de tiempo que contribuyen al potencial, P_t, y cuando pasa el umbral en $t = 4$ se genera un impulso y se activa el periodo refractario.

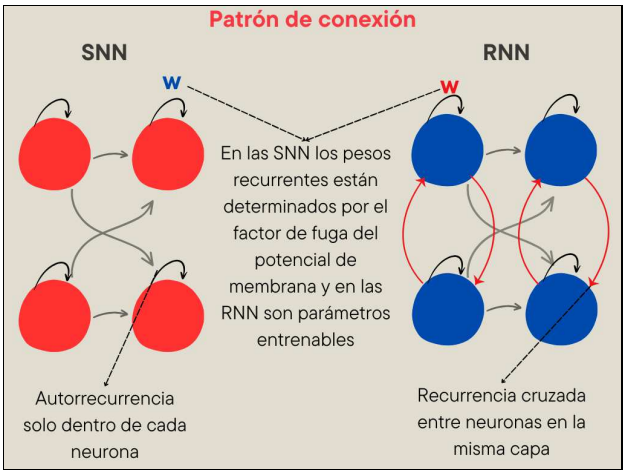


Figura 3. Patrón de conexión SNN vs RNN.

Codificación temporal: Frecuencia y Latencia

Todo el proceso de despolarización comienza con información externa traducida al lenguaje de las neuronas biológicas: impulsos eléctricos. Entonces el objetivo es traducir los datos que se quieren procesar al lenguaje de las neuronas LIF. Y, aunque el código neuronal biológico aún no se comprende completamente, se han propuesto métodos para imitarlo como la **codificación en tasas**.

Imagina que las neuronas son luces intermitentes. La codificación en tasas se enfoca en cuántas veces parpadean en un período de tiempo. Al mirar algo brillante, las neuronas parpadean más. Si consideramos, por ejemplo, un píxel blanco de alguna imagen (tiene un valor de 255), tendrá una alta tasa de impulsos, mientras que un píxel negro (con un valor de 0) no genera impulsos, y los grises tienen tasas proporcionales.

Otra forma de codificación es la **codificación en latencia**, que se enfoca en el momento preciso de los impulsos. Está motivada en la idea de que estímulos importantes tienden a provocar disparos antes en las neuronas. En el caso de los píxeles, un píxel blanco provoca disparos antes que uno gris o uno negro.

Así, mientras la codificación en tasas se centra en la cantidad de impulsos, la codificación en latencia se enfoca en el momento de los impulsos. La Figura 4 muestra una representación en frecuencia y latencia de una imagen del MNIST, un conjunto de datos de dígitos escritos a mano. Para los curiosos, los detalles de estos procesos están en la referencia [3].

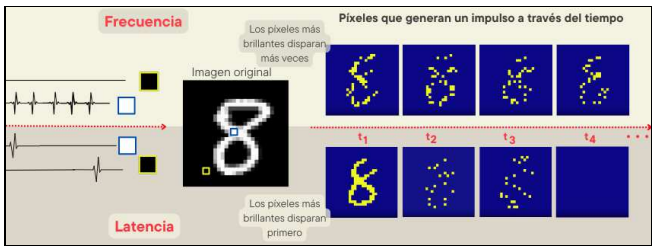


Figura 4. Codificación en frecuencia vs latencia de una imagen de un 8 escrito a mano en escala de grises.

Clasificación a pulso

Con la anatomía de las neuronas LIF, y la codificación adecuada de la información, se puede escalar este modelo a una red profunda, es decir, una red neuronal con múltiples capas ocultas entre la capa de entrada y la de salida. La presencia de estas capas permite que la red capture características y patrones en los datos. Todo lo anterior suena complejo y difícil, sin embargo, existen librerías, como Pytorch y snnTorch, diseñados específicamente para construir y entrenar redes neuronales de impulso, esto al menos le quita lo difícil.

La parte compleja habrá que entenderla, o al menos intentarlo. Supongamos que queremos crear una red completamente conectada para realizar una tarea de clasificación del conjunto MNIST, de tres capas con dimensiones 784-N-10, la primera capa con 784 neuronas de entrada, pues es la cantidad de píxeles de una imagen del MNIST (28×28), la segunda con N neuronas y la tercera 10, correspondientes a los posibles dígitos. Realizamos:

1. **Definición de parámetros** para cada capa de la red como la cantidad de neuronas de entrada, capas ocultas, neuronas de salida y tasa de decaimiento del potencial. Las capas se inicializan usando las funciones de PyTorch para definir las conexiones y snnTorch para definir las neuronas LIF.

2. **Inicialización de estados y variables ocultas** como potencial de membrana y pesos sinápticos.

3. **Creación de trenes de picos de entrada** usando la codificación temporal para convertir la información de las imágenes en un formato que el modelo LIF entienda. Por ejemplo, para la codificación en tasas se genera una secuencia de impulsos que se distribuyen a lo largo de, digamos, 200 pasos de tiempo para las 784 neuronas de entrada, donde los valores de los píxeles determinan la posibilidad de que una neurona dispare un impulso.

4. **Proceso de simulación** en cada paso de tiempo, la entrada actual se multiplica por los pesos de la primera capa para calcular la corriente post sináptica, la cual se suma al potencial de membrana, que se actualiza con base a la entrada y el decaimiento. Si el potencial excede el umbral, la neurona dispara un impulso, e inicia el periodo refractario. Este proceso se repite en la segunda capa con las salidas de la primera capa como entradas.

5. **Conteo de impulsos** total generados por cada una de las 10 neuronas de salida.

6. **Clasificación.** La clasificación pronosticada es aquella que corresponde a la neurona de salida que generó el mayor número (o tasa) de impulsos. Por ejemplo, si la neurona que representa la clase “8” generó más impulsos que cualquier otra neurona, entonces la red clasifica la imagen como un “8”.

7. **Entrenamiento**. Se determina el desempeño del modelo con una función de error y se ajustan los pesos repitiendo los pasos del 4 al 7 hasta minimizar el error.

8. **Clasificación final.** El modelo bien entrenado genera las predicciones correctas. Véase la Figura 5.

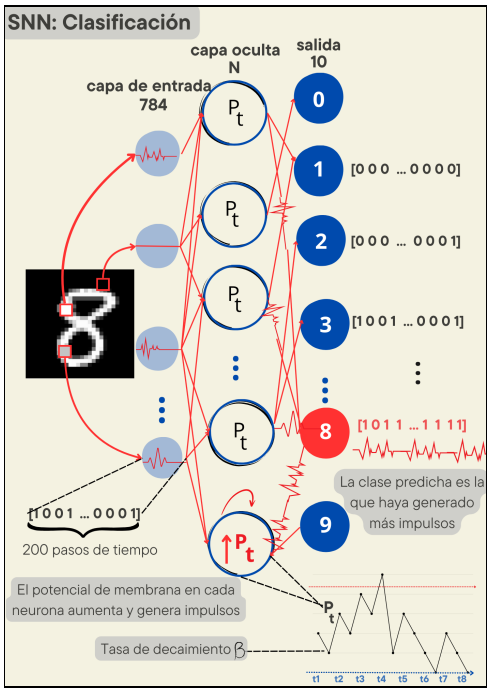


Figura 5. Diagrama del proceso de clasificación de una SNN.

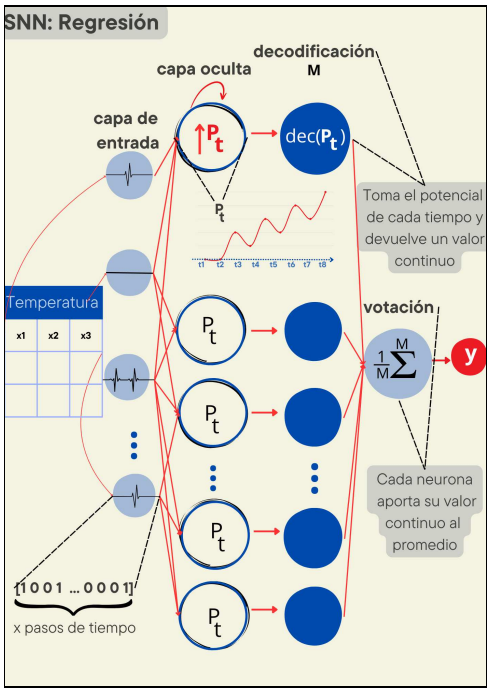


Figura 6. Diagrama del proceso de regresión de una SNN.

Regresión a pulso

En contraste con la clasificación donde se busca asignar una categoría, en tareas de regresión la salida es un número real. Por ejemplo, supongamos que queremos pronosticar el precio de una casa basado en sus características.

El proceso que sigue una SNN para resolver un problema de regresión solo difiere en un par de pasos de los mencionados en la sección anterior [4], véase la Figura 6.

1. **Capas de decodificación.** En lugar de contar los pulsos, se utilizan capas de decodificación que transforman los potenciales de membrana acumulados en cada paso de tiempo en valores continuos.

2. **Capa de votación de población.** Cada neurona de la capa de decodificación aporta su estimación del valor continuo basado en su potencial de membrana. La capa de votación toma todas las estimaciones y las promedia, proporcionando un único pronóstico final, y.

Entonces, después de pasar los datos por las capas intermedias, los potenciales de membrana se convierten en estimaciones de precio en la capa de decodificación. Luego, la capa de votación promedia estas estimaciones para obtener una predicción final del precio.

El código fuente y los ejemplos de redes por impulso se pueden encontrar en el repositorio de GitHub [5].

SNN vs NN

El modelo SNN es más eficiente energéticamente que el modelo NN pues, debido a su naturaleza basada en eventos, solo consumen energía cuando generan un impulso, mientras las redes tradicionales requieren un consumo de energía constante. Algunos estudios han aplicado SNN en tareas de detección de objetos, mostrando resultados comparativos con las NN pero con un potencial de energía mucho menor [6].

Las SNN son ideales para procesar información temporal debido a su capacidad para manejar datos secuenciales y depender de eventos discretos. Esto las hace adecuadas para tareas como el reconocimiento de patrones temporales y el procesamiento de señales biológicas. En cambio, las NN están diseñadas para procesar datos de manera paralela, siendo eficientes para tareas donde cada entrada se procesa de forma independiente, como la traducción automática de idiomas.

Aunque las NN llevan más tiempo siendo investigadas y tenemos más acceso a estudios, bibliotecas y herramientas para utilizarlas, la comunidad sigue desarrollando softwares que facilitan el uso de SNN con el objetivo de apostar por un cómputo descentralizado.

Conclusiones

Exploramos cómo las SNN introducen elementos biológicamente inspirados, como el potencial de membrana y los disparos neuronales, para procesar información de manera similar al cerebro humano. Además, discutimos cómo se utilizan en tareas específicas, como la clasifica-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ción y la regresión, y cómo difiere su enfoque del de las redes neuronales clásicas.

De todo lo anterior, podemos concluir que la elección de utilizar el modelo de SNN depende del contexto de la aplicación, los requisitos específicos de la tarea y las limitaciones tecnológicas y energéticas, pues este modelo ofrece ventajas en términos de eficiencia energética y procesamiento temporal.

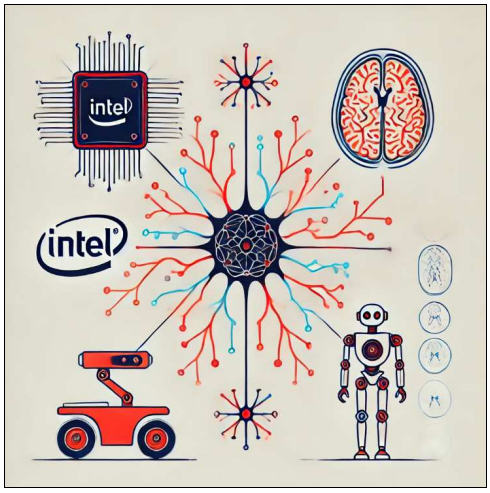


Figura 7. Aplicaciones destacadas de las SNN.

Entre las aplicaciones en las que ha destacado el modelo protagonista de este artículo se puede mencionar el chip neuromórfico Loihi de Intel, diseñado para emular el funcionamiento del cerebro humano. Utiliza una arquitectura de red neuronal por impulso para realizar tareas de forma eficiente. También los robots móviles, con sus capacidades continuas de observación y acción de alta dimensión, se vieron obligados a evolucionar hacia modelos de redes neuronales por impulso para enfrentar las deficiencias inherentes a sus limitados recursos energéticos.

Además de humanizar chips y evolucionar robots, este modelo ha asombrado por sus aplicaciones médicas. Existen investigaciones exitosas sobre las ventajas que se pueden tener al utilizarlas en clasificación de señales y diagnóstico de enfermedades usando señales del corazón y del cerebro [7]. Esto por supuesto mejora la eficiencia y precisión de los doctores al detectar afecciones de manera temprana, Figura 7.✷

REFERENCIAS

1. Hodgkin, A. y Huxley, F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4), 500–44.

2. Weihua, H., YuJie, W., Lei, D., Guoqi, L., Haoyu, W., Yang, T., Wei, D., Wenhui, W. y Yuan, X. (2005). Comparing SNNs and RNNs on Neuromorphic Vision Datasets: Similarities and Differences. *Frontiers in Neuroscience*, 16. Department of Precision Instrument, Tsinghua University, Beijing 100084, China. Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106, USA. Center for Brain Inspired Computing Research, Tsinghua University, Beijing 100084, China. Lab of Cognitive Neuroscience, THBI, Tsinghua University, Beijing 100084, China.

3. Jason, K., Max, W., Emre, N., Xinxin, W., Gregor, L., Girish, D., Mohammed, B., Doo, S. y Wei, D. (2023). Training Spiking Neural Networks Using Lessons From Deep Learning. *Proceedings of the IEEE*, 111(9).

4. Jason, K., Max, W., Emre, N., Xinxin, W., Gregor, L., Girish, D., Mohammed, B., Doo, S. y Wei, D.(2023). Training Spiking Neural Networks Using Lessons From Deep Learning. Proceedings of the IEEE, 111(9).

5. GitHub. (2024). *mnist-snn-demo*. Disponible en: `https://github.com/vccboom/mnist-snn-demo`. Accesado en: 2024-08-14.

6. Kashu, Y., Viet-Khoa, V., Darshan, B., y Ngan, L.(2022). Spiking Neural Networks and Their Applications: A Review. *Brain Sci.*, 12(7), 863.

7. Li, X., Zhang, X., Yi, X., Liu, D., Wang, H., Zhang, B., Zhang, B., Zhao, D. y Wang, L. (2023). Review of medical data analysis based on spiking neural networks. *Procedia Computer Science*, 221, 1527-1538.

SOBRE LOS AUTORES



Valeria Carrillo Cabrera es estudiante de la Licenciatura en Matemáticas en el Centro Universitario de Ciencias Exactas e Ingenierías, CUCEI, de la Universidad de Guadalajara. Actualmente se encuentra trabajando en su proyecto de tesis dirigido por el Dr. Edgar Alejandro Arroyo Guerrero, con el tema de Análisis comparativo entre las redes neuronales clásicas y las redes neuronales por impulso. Sus áreas de interés son la inteligencia artificial y la ciencia de datos.



Edgar Alejandro Guerrero Arroyo obtuvo el grado de doctor en el Centro de Investigación en Matemáticas, CIMAT, en el 2016. Actualmente es profesor de tiempo completo en el Departamento de Matemáticas del Centro Universitario de Ciencias Exactas e Ingenierías, CUCEI, de la Universidad de Guadalajara. Sus líneas de investigación se centran en modelación numérica de ecuaciones diferenciales, optimización numérica y ciencia de datos.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Visión artificial e inclusión: creando un traductor de señas con Teachable Machine

Marco Antonio Aguilar Cortés, Miguel Aguilar Cortés y Luis Ángel Martínez Cabrera

Introducción

La discapacidad es una condición que puede surgir en cualquier etapa de la vida, desde la gestación hasta la vejez. Se presenta en limitaciones que pueden afectar el funcionamiento intelectual, motriz, sensorial (como la vista y el oído) y la capacidad de adaptación a diferentes entornos, como el hogar, la escuela y la comunidad [1]. Según datos del Instituto Nacional de Estadística y Geografía (INEGI), en México, 20 millones 838 mil 108 personas viven con alguna discapacidad o limitación o condición mental, lo que representa un impacto considerable no solo en sus vidas, sino también en las de sus familias [2]. Entre este grupo, el 21.9 % enfrenta problemas auditivos y el 15 % tiene dificultades para hablar o comunicarse [3]. Estas barreras de comunicación limitan su calidad de vida, pues afectan su capacidad para participar plenamente en actividades cotidianas como el trabajo, la educación e incluso la interacción familiar.

En respuesta a estos desafíos, la tecnología ha abierto nuevas posibilidades para mejorar la comunicación de las personas con discapacidad auditiva y del habla. Una de las áreas más significativas es la visión artificial, que, entre otras cosas, ha permitido traducir señas a un formato comprensible, sin embargo, el desarrollo de estas soluciones suele requerir conocimientos avanzados en programación, lo que dificulta su implementación para la mayoría de las personas.

Este artículo presenta una manera accesible de crear un traductor de señas utilizando visión artificial, sin necesidad de contar con habilidades en desarrollo de software. A través de la plataforma Teachable Machine, es posible entrenar un modelo de inteligencia artificial utilizando un conjunto de datos y aplicarlo en lenguajes de programación como JavaScript. Los resultados permiten identificar y traducir las primeras letras del abecedario a través de señas. Con esto, crece la posibilidad de generar modelos que permitan identificar más señas facilitando la comprensión de las necesidades de las personas con discapacidad auditiva y del habla, y acercando a la sociedad a un entorno más inclusivo que permita ofrecer una mejor calidad de vida.

Investigaciones previas

En la actualidad, se han desarrollado diversas herramientas tecnológicas que han permitido un avance significativo en la identificación de señas, con el objetivo de fomentar un entorno más inclusivo y facilitar la comunicación con personas que utilizan la Lengua de Señas. Gracias a la inteligencia artificial y la tecnología de predicción de imágenes, es posible reconocer una seña y comprender su significado, lo que representa un primer paso importante hacia el aprendizaje de la Lengua de Señas Mexicana (LSM).

Un ejemplo es el uso de redes neuronales convolucionales [4], donde, a través de una página web, se recopiló una gran cantidad de imágenes para entrenar un modelo que permite la predicción utilizando el lenguaje de programación Python. Los resultados mostraron que el 80 % de las letras fueron identificadas, sin embargo, el proceso de entrenamiento es complejo, lo que resalta la necesidad de simplificarlo y facilitar su implementación en sitios web o dispositivos móviles para hacerlo más accesible e interactivo para el usuario.

En la investigación de [5], se emplearon herramientas como MediaPipe y OpenCV para extraer marcas clave de las manos, el cuerpo y el rostro en tiempo real. Con estos datos, se generó un conjunto de entrenamiento para un modelo basado en redes neuronales convolucionales. Los resultados indicaron una precisión del 94.9 % en el reconocimiento de 75 palabras aisladas y del 94.1 % en la identificación de 20 frases de prueba en un contexto médico.

Por su parte, en [6] se desarrolló un sistema para el reconocimiento automático del abecedario dactilológico, el cual es un sistema alfabético utilizado por personas sordociegas para comunicarse mediante señas en la Lengua de Señas Mexicana. Este sistema se basó en una red neuronal recurrente, logrando una similitud del 80.74 %. Sin embargo, algunas letras fueron difíciles de diferenciar debido a sus rasgos visuales similares, lo que afectó el resultado final.

En [7], se utilizó MediaPipe para la detección de señales y la extracción de texturas de cada imagen, como parte del proceso de aprendizaje. Posteriormente, se compararon dos algoritmos: Support Vector Machine (SVM), un algoritmo de aprendizaje supervisado utilizado en tareas de clasificación y regresión, y Random Forest, un algoritmo de machine learning que también se usa para problemas de clasificación y regresión. Los resultados mostraron que Random Forest fue más preciso, alcanzando más del 90 % de similitud en la predicción de las señas.

Por otro lado, en [8] implementó un algoritmo de aprendizaje supervisado para interpretar señas y tradu-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

cirlas al alfabeto urdu, utilizado principalmente en Pakistán e India. Este proyecto se integró en una aplicación móvil que, al capturar una imagen de la seña, genera el carácter correspondiente en urdu, lo que ha sido de gran ayuda para la comunidad sorda en Pakistán.

Asimismo, en [9] utilizó redes neuronales convolucionales para traducir señas del lenguaje de señas estadounidense a un formato comprensible, alcanzando un 99 % de precisión gracias al uso de un gran volumen de imágenes para entrenar el modelo con una cámara RGB.

De acuerdo con el INEGI, en México hay 20,838,108 personas con algún tipo de discapacidad, limitación o condición mental.

Todos los proyectos mencionados dependen de procesos complejos para el entrenamiento de modelos, lo que dificulta la implementación de soluciones por parte de quienes no tienen conocimientos en programación. Por ello, es necesario explorar nuevas maneras de crear modelos accesibles, de modo que su uso pueda extenderse a un mayor número de personas con discapacidades auditivas y del habla.

Obtención y procesamiento de imágenes

Para entrenar un modelo capaz de predecir una seña, es fundamental contar con un conjunto de imágenes para cada elemento. Mientras mayor sea el número de imágenes, más preciso será el modelo resultante. En este caso, se utilizó un conjunto de datos proveniente de una investigación previa, la cual también aplica visión por computadora para la traducción del lenguaje de señas [10]. Este conjunto de datos incluye 300 imágenes de las 21 señas del alfabeto del LSM, sin embargo, cada elemento tiene un fondo verde que podría afectar la precisión de las predicciones. Para resolver este problema, se empleó el lenguaje de programación Python que permitió procesar los elementos y eliminar el fondo verde, mejorando así la exactitud del modelo. En la Figura 1 se muestra el antes y el después del procesamiento de cada una de las imágenes.



Figura 1. Procesamiento del fondo de imagen.

Entrenamiento del Modelo

Una vez que se generaron las imágenes sin fondo, el siguiente paso fue entrenar el modelo. La plataforma empleada fue Teachable Machine, que, según el sitio oficial, es una herramienta web desarrollada por Google que permite crear modelos de aprendizaje automático de manera rápida y sencilla, accesible para cualquier persona. Con Teachable Machine, es posible entrenar un ordenador para que reconozca imágenes, sonidos y posturas sin escribir código. Para acceder a la plataforma, la dirección es: `https://teachablemachine.withgoogle.com/`.

Una vez entrenado el modelo, este se puede integrar fácilmente en proyectos, sitios web, aplicaciones y más, lo que ofrece una amplia gama de posibilidades creativas y prácticas.

El proceso comienza con subir las imágenes, después se asignan a la clase correspondiente, y se procede al entrenamiento del modelo. La Figura 2 muestra el procedimiento para cargar las imágenes de cada una de las señas.

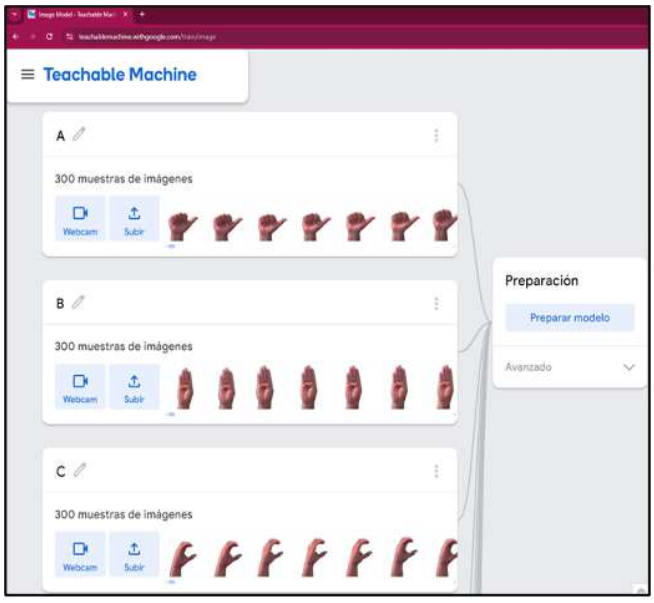


Figura 2. Carga de imágenes a Teachable Machine.

Al finalizar la preparación del modelo, la plataforma ofrece la opción de realizar pruebas utilizando una webcam o un archivo. En la Figura 3 se muestra un ejemplo de la vista previa.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Teachable Machine es una plataforma web que permite crear modelos de aprendizaje automático de forma rápida y sencilla.

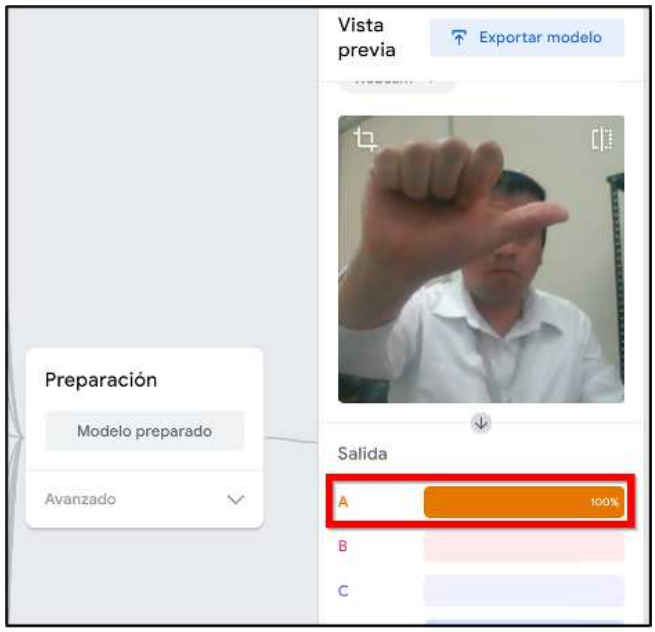


Figura 3. Pruebas de reconocimiento de seña.

Como se puede apreciar, Teachable Machine permite realizar predicciones precisas de señas, ajustándose automáticamente si el usuario cambia de seña o realiza movimientos. Además, ofrece la opción de generar modelos compatibles con aplicaciones web y móviles. En esta investigación, se utilizó un modelo integrado en una aplicación web, con instrucciones detalladas proporcionadas por la plataforma para facilitar su integración.

Implementación

Con el modelo listo, se procede a integrarlo en una aplicación web utilizando JavaScript. La función principal de esta aplicación es permitir que, a través de la cámara web, los usuarios realicen una seña y visualicen la probabilidad de que estén representando correctamente la letra correspondiente. Esta retroalimentación se presenta de manera visual, mediante una barra de progreso que indica si la seña está siendo ejecutada correctamente.

El propósito principal de desarrollar e implementar la aplicación es ayudar a los usuarios a perfeccionar sus señas, brindando una retroalimentación inmediata y accesible. Aunque la funcionalidad es similar a la vista previa de Teachable Machine, esta aplicación web permite que cualquier persona acceda a ella sin necesidad de tener una cuenta en la plataforma. Esto significa que el proyecto puede ser publicado en un sitio web y estar disponible desde un navegador web con conexión a internet. El único requisito para que la cámara funcione correctamente es utilizar un Protocolo Seguro de Transferencia de Hipertexto (HTTPS). Las pruebas del sistema se llevaron a cabo en un entorno local.

Resultados

Tras desarrollar la aplicación web e integrar el modelo de traducción de señas, se invitó a dos expertos en Lengua de Señas y a trece alumnos del Instituto Tecnológico Superior de Teziutlán a realizar ejercicios frente a la cámara. Los expertos evaluaron la efectividad de la aplicación, ya que cuentan con el conocimiento adecuado para realizar las señas correctamente. Por otro lado, los alumnos, al no estar familiarizados con las señas, permitieron evaluar la capacidad del modelo para detectar errores y guiar en la corrección.

Con la colaboración de los expertos, se verificó que el modelo no solo ayudaba a identificar las señas, sino que también brindaba una retroalimentación del porcentaje de similitud. Los resultados generales fueron positivos, con una alta aceptación de la aplicación, destacando su utilidad y facilidad de uso. Sin embargo, los expertos también propusieron algunas mejoras para enriquecer la experiencia, como añadir más señas y optimizar la retroalimentación. En las Figuras 4 y 5 se presenta un ejemplo visual de la retroalimentación proporcionada por la aplicación.



Figura 4. Letra no identificada.

Una de las recomendaciones más destacadas fue registrar el progreso de un usuario en una base de datos y así implementar técnicas de gamificación, es decir, recompensar al usuario en función de objetivos alcanzados, por ejemplo, acumulación de puntos, regalos, niveles y demás elementos que fomenten el aprendizaje del lenguaje de señas. Esto no solo haría la experiencia más

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

interactiva y motivadora, sino que también fomentaría el aprendizaje continuo del lenguaje de señas.



Figura 5. Seña identificada.

Además, se sugirió ampliar el conjunto de señas disponibles en la aplicación. Un punto importante por mencionar es que algunas letras del abecedario requieren movimientos específicos, lo que complicó su integración en el modelo entrenado actual. Este desafío fue identificado como una oportunidad para futuros trabajos, donde se explorará la posibilidad de incorporar señas que impliquen movimiento, asegurando así una cobertura más completa del lenguaje de señas.

Por otro lado, los alumnos valoraron la plataforma como una herramienta innovadora. Aunque señalaron la existencia de aplicaciones similares, destacaron que estas no ofrecen un porcentaje de similitud en las señas. En la Figura 6 se presentan las evaluaciones realizadas por los alumnos, que abarcan aspectos como la eficacia, el diseño y el nivel de satisfacción con la plataforma.

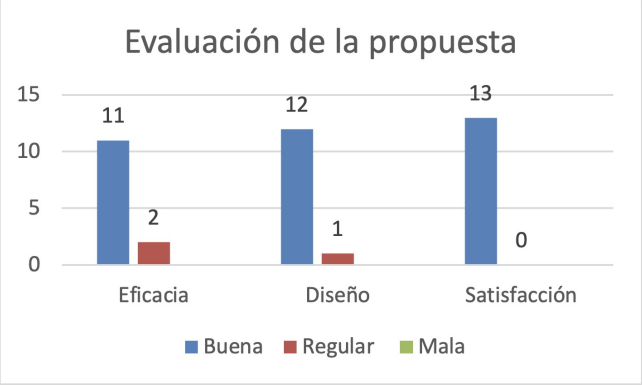


Figura 6. Evaluación por parte de los alumnos.

Aunque la aplicación cumplió con los objetivos iniciales y mostró ser una herramienta efectiva para el aprendizaje de señas, se identificaron áreas clave para su desarrollo futuro. Integrar más señas, mejorar la interactividad a través de la gamificación, y abordar el reto de las señas que requieren movimiento serán los próximos pasos para hacer de esta herramienta un recurso aún más valioso para la inclusión de personas con discapacidad auditiva en la sociedad.

Conclusiones

El desarrollo del modelo de traducción de señas a través de visión artificial presenta muchos desafíos. Factores como la cantidad y calidad de las imágenes utilizadas durante el entrenamiento influyen directamente en la precisión de las predicciones. Aunque es posible alcanzar probabilidades altas, es importante tener en cuenta que la precisión del modelo no será del 100 %, ya que las variaciones en las imágenes o en las condiciones de captura pueden afectar el resultado.

Además, la calidad de la cámara utilizada es un aspecto importante. Un dispositivo con baja resolución puede disminuir la probabilidad de éxito en las predicciones, ya que los detalles necesarios para identificar correctamente las señas pueden no ser capturados con suficiente claridad. Por lo tanto, se recomienda utilizar una cámara de buena resolución para maximizar la eficacia del modelo.

En este proyecto, el conjunto de datos de imágenes utilizado fue tomado de una fuente preexistente, lo cual facilitó el proceso inicial de desarrollo, sin embargo, no se descarta la posibilidad de generar un dataset propio, con características específicas, para realizar otras pruebas y mejorar la precisión del modelo en situaciones particulares.

A pesar de estas consideraciones, el modelo desarrollado a través de una plataforma gráfica como Teachable Machine cumple con los objetivos planteados. Los resultados obtenidos indican que, a pesar de ser un enfoque simplificado, el modelo es efectivo y demuestra que la visión artificial tiene un gran potencial para mejorar la inclusión de personas sordomudas en diversos entornos. Este proyecto se continuará mejorando para que la población del municipio de Teziutlán, Puebla y sus alrededores conozcan más la Lengua de Señas Mexicana. Con avances adicionales en esta tecnología, es posible que en el futuro se logren herramientas aún más robustas que faciliten la comunicación y la integración de personas con discapacidad auditiva y del habla en la sociedad.✷

REFERENCIAS

1. CONAFE. (2010). Discapacidad auditiva. Guía didáctica para la inclusión en educación inicial y básica. Recuperado el 27 de Mayo de 2024, de https://www.gob.mx/cms/uploads/attachment/file/106806/discapacidad-auditiva.pdf.

2. INEGI. (2020). Censo de Población y Vivienda. Obtenido de https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=Discapacidad_Discapacidad_01_29827fe7-b1cd-4bd2-81d6-9d08bda47df8&idrt=151&opc=t.

3. D' Artigues, K. (2023). El problema son las barreras, no las personas con discapacidad. Obtenido de IBERO: https://ibero.mx/prensa/el-problema-son-las-barreras-no-las-personas-

con-discapacidad-katia-dartigues#:~:text=Personas%20con%20discapacidad%2C%20la%20minor%C3%ADa,son%20la%20minor%C3%ADa%20m%C3%A1s%20grande.

4. Aguilar-Cortés, M. A., Aguilar-Cortés, M., Robles-Calderón, J., Pérez-López, A. y Vicenteño-Rivera, H. (2022). Interpretación de letras en lenguaje de señas a través de imágenes utilizando redes neuronales convolucionales. *Coloquio de investigación multidisciplinaria*, 10(1), SC-41 - SC-49.

5. Ramírez.Sánchez, J. E., Anguiano-Rodríguez, A. y González-Mendoza, M. (2021). Interpretación en tiempo real de lengua de señas mexicana con CNN y HMM. *Research in Computing Science*, 150(6), 91 - 102.

6. Mejía-Pérez, K. y Córdova Esparza, D. (2023). Reconocimiento de la Lengua de Señas Mexicana mediante redes neuronales. *Nthe*, (41), 27 - 33.

7. Maciel-Cataño, E. J., Estrada-Gutiérrez, J. C., Flores-Jiménez, A., Cano-Lara, M. y Jiménez-Rodríguez, M. (2024). Detección del abecedario de Lengua de Señas Mexicanas (LSM) usando MediaPipe, SVM y Random Forest. *Transregiones*,(7), 73 - 84.

8. Razzaq, A., Baig, I. A., Hussain, A., Shahid, S. y Rehman, T.-u. (2021). Dataset of Pakistan Sign Language and Automatic Recognition of Hand Configuration of Urdu Alphabet through Machine Learning. *Data in Brief*, 36(107021).

9. Sharma, S. y Singh, S. (2021). Vision-based hand gesture recognition using deep learning for the interpretation of sign language. *Expert Systems With Applications*, 182, 115657.

10. Mancilla-Morales, E., Vázquez-Aparicio, O., Arguijo, P. y Melendez-Armenta, R. (2019). Traducción del lenguaje de señas usando visión artificial. *Research in Computing Science*, 148(8), 79 - 89.

SOBRE LOS AUTORES



Marco Antonio Aguilar Cortés trabaja en el Tecnológico Nacional de México / Instituto Tecnológico Superior de Teziutlán. Es Doctor en Investigación Educativa, además de contar con una Maestría en Sistemas Computacionales y otra en Pedagogía. A lo largo de su trayectoria, ha participado activamente en el desarrollo de software para diversas instituciones tanto públicas como privadas. Ha desempeñado el rol de asesor en numerosos eventos académicos de relevancia regional y nacional. Cuenta con el reconocimiento de Perfil Deseable y forma parte de un Cuerpo Académico en Consolidación. Además, es Candidato a Investigador por el Sistema Nacional de Investigadoras e Investigadores (SNII).



Miguel Aguilar Cortés es Docente investigador adscrito al área de Ingeniería en Sistemas Computacionales del Tecnológico Nacional de México/ Instituto Tecnológico Superior de Teziutlán. Es Maestro en Sistemas Computacionales e Ingeniero en Mecatrónica; ha sido asesor de diferentes proyectos de investigación con aplicación a nivel nacional, participando en eventos científico-tecnológico y en redacción de artículos de investigación. Cuenta con Perfil Deseable.



Luis Ángel Martínez Cabrera es Docente investigador adscrito al área de Ingeniería en Mecatrónica del Tecnológico Nacional de México/ Instituto Tecnológico Superior de Teziutlán es Ingeniero en Mecatrónica con especialidad en Automatización y Manufactura por el Instituto Tecnológico Superior de Zacapoaxtla (2009), y Maestro en Sistemas Computacionales con especialidad en Redes y Seguridad Informática por el Instituto Tecnológico Superior de Teziutlán (2020). Como investigador, ha liderado proyectos financiados por el TecNM, contribuyendo con artículos científicos en las áreas de automatización, robótica móvil y manufactura. Su trayectoria se fortalece con certificaciones nacionales e internacionales en diseño CAD, lo que le permite integrar soluciones tecnológicas innovadoras en sus investigaciones y proyectos.

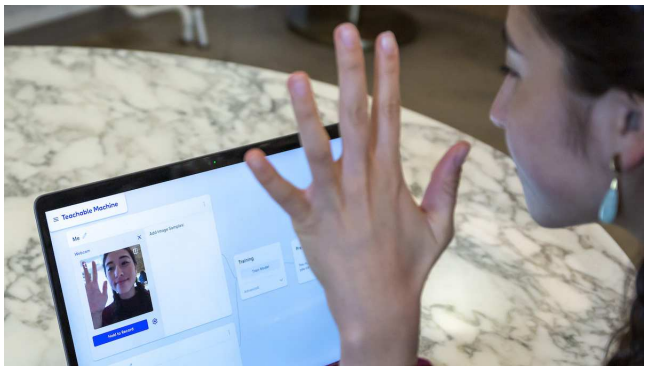


Imagen tomada de:
https://teachablemachine.withgoogle.com/.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Obteniendo la geometría de objetos en una imagen usando renderizado diferencial

José Abraham Ramos Badillo, Luis Eduardo Gamboa Guzman y Karina Mariela Figueroa Mora

El renderizado fotorealista consiste en generar imágenes con aspecto real y atractivo. En la Figura 1 mostramos la geometría de una habitación, observa cómo se logra transformar a objetos con aspecto realista después de pasar por un proceso de renderizado. La implementación de éste permite disfrutar películas animadas, videojuegos, diseño asistido por computadora, etc.

Formalmente, el renderizado consiste en obtener una imagen 2D a partir de la descripción geométrica de una escena tridimensional. Esto incluye detallar la forma y posición de cada objeto, donde cada objeto tiene un material asociado. Finalmente, considerando la naturaleza de la luz y su interacción con la escena, se logran crear imágenes tan reales como si hubieran sido tomadas con una cámara fotográfica real (Figura 1).

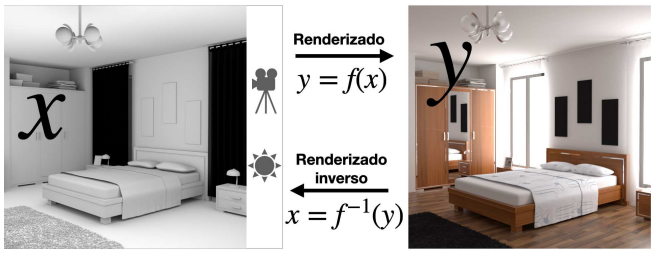


Figura 1. El renderizador toma los parámetros de una escena (izquierda) y crea una imagen muy realista (derecha). El renderizador inverso toma la imagen real y aproxima cómo eran los parámetros que se usaron para generar esa foto (izquierda).

El renderizado consiste en obtener una imagen 2D a partir de la descripción de una escena tridimensional.

Pensemos en el proceso de renderizado como la evaluación de una expresión $y = f(x)$, donde x es un vector que contiene la descripción completa de la escena: geometría y posición de los objetos, tipo de material, tipos de fuentes de luz y cámara. La función f denota el renderizador y, finalmente y es la imagen 2D renderizada. Aquí surge una pregunta: ¿es posible ir en sentido contrario evaluando una función inversa de f para obtener una descripción de la escena a partir de una imagen?, en términos formales, ¿es posible hacer $x = f^{-1}(y)$? (Figura 1). Como es de esperarse, la reconstrucción de escenas 3D a partir de información 2D es un proceso extremadamente complejo. Esta área de estudio se llama renderizado inverso (RI), la cual aborda grandes problemas de la visión computacional con un enfoque diferente, que consiste en entender cómo los objetos de una escena interactúan entre sí.

Existen algunas estrategias para atacar el problema RI, por ejemplo, el método de renderizado diferencial (RD) [1]. Algunas de sus aplicaciones son la reconstrucción de geometrías tridimensionales a partir de imágenes 2D, reconstrucción de posturas y formas de manos humanas, reconstrucción facial, fabricación de materiales translúcidos, creación de cáusticos, creación de un modelo 3D con fotografías, la optimización de la luz natural como parte de un proceso de diseño arquitectónico, análisis de iluminación en una escena, análisis por síntesis, entre otros.

En este artículo, explicaremos qué es un renderizador para después adentrarnos en esta novedosa área de estudio del renderizado inverso, específicamente el método de renderizado diferencial. Finalmente, se presentará un ejemplo de su aplicación en la reconstrucción de geometrías a partir de imágenes 2D.

Renderizado

Antes de adentrarnos en el problema inverso, centrémonos por un momento en el problema del renderizado. Existen muchos métodos para realizar esta tarea, uno de ellos es conocido como Path Tracing (PT) [2], el cual es una técnica ampliamente utilizada para generar imágenes fotorealistas, gracias a que simulan el transporte de luz, permitiendo tener en cuenta aspectos como la iluminación global y la dispersión de la luz en los diferentes tipos de materiales, a través de una escena tridimensional. Ahora bien, es necesario entender cómo funciona un renderizador para explicar la idea detrás de un Renderizador Diferencial (RD).

Comencemos revisando el comportamiento de la luz en alguna superficie arbitraria. Imagina que estás observando un punto en específico en la superficie de una tetera. La luz que se aleja de esa superficie hacia los ojos

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

surge debido a una de dos posibles situaciones (Figura 2):

1. La superficie emite luz (radiancia), por ejemplo una fuente de luz o una superficie en llamas.
2. La luz podría provenir directa o indirectamente de algún otro lugar y luego reflejarse en una superficie.

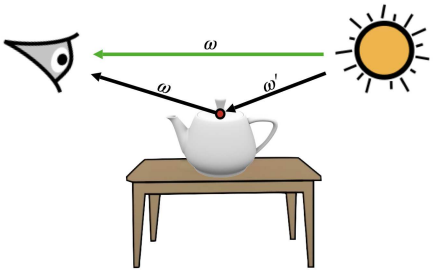


Figura 2. Al observar un punto que emite luz hacia la cámara (flecha verde), a esta luz la conocemos como radiancia emitida. El camino con flechas negras denota la radiancia saliente del punto rojo, que se calcula como la suma de toda la luz reflejada desde diferentes direcciones (radiancia incidente).

Ahora bien, obtener el valor de color observado, esto es, la radiancia saliente de la superficie, implica resolver complicadas ecuaciones matemáticas que describen el comportamiento de la luz. Desafortunadamente, estas ecuaciones no tienen una solución exacta (analítica), por lo que es necesario recurrir a métodos numéricos como Monte Carlo Path Tracing [2]. Aún en este caso, estimar una solución es computacionalmente difícil.

Comúnmente el proceso de renderizado de un PT comienza con la emisión de rayos desde una cámara virtual hacia la escena tridimensional. Esto para garantizar que los caminos de luz generados tengan una correspondencia directa con los píxeles de la imagen (Figura 3).

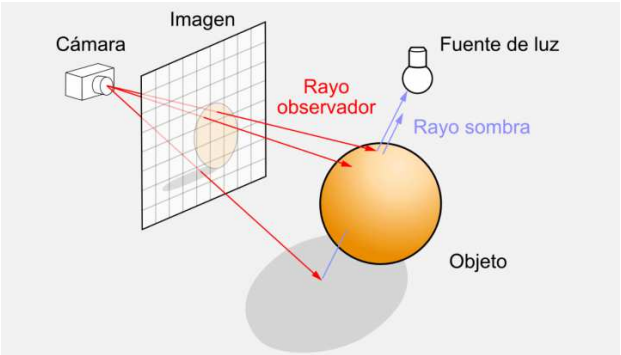


Figura 3. La cámara emite rayos desde su origen a través de una cuadrícula de píxeles (plano de la imagen). El color final de cada píxel se obtiene sumando la contribución de todos los rayos de luz que salen de ese píxel.

Estos rayos interactúan con los objetos en su camino, reflejándose y refractándose según las propiedades de los materiales de dichos objetos, permitiendo simular con precisión cómo se comporta la luz en el mundo real. Por ejemplo, si la luz golpea un espejo, se refleja. Pero si la luz golpea un cristal, puede reflejarse o atravesar el cristal con una dirección determinada; a este fenómeno le llamaremos refracción (Figura 4).

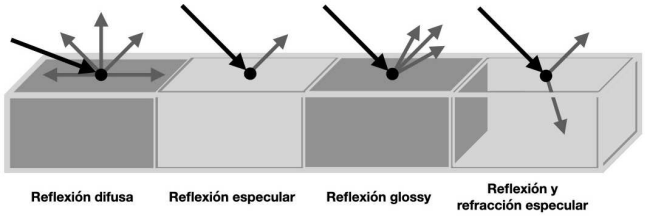


Figura 4. La forma en que se refleja depende del material: los materiales difusos, como el yeso, dispersan la luz uniformemente, mientras que los materiales glossy, como los metales pulidos, reflejan los objetos a su alrededor de manera similar a la reflexión especular de un espejo, pero con menos claridad.

En un PT, el camino normalmente termina cuando se encuentra una fuente de luz o se ha alcanzado un número máximo de rebotes, como se muestra en la Figura 5. En cada intersección, se estima la intensidad de la luz resolviendo una ecuación especial que determina el comportamiento de la luz. A esta ecuación le llamaremos la ecuación de renderizado o ecuación de transporte de luz [2]. Este proceso se repite múltiples veces para cada píxel de la imagen, acumulando los resultados para generar la imagen final. No bromeabamos cuando dijimos que es computacionalmente difícil.

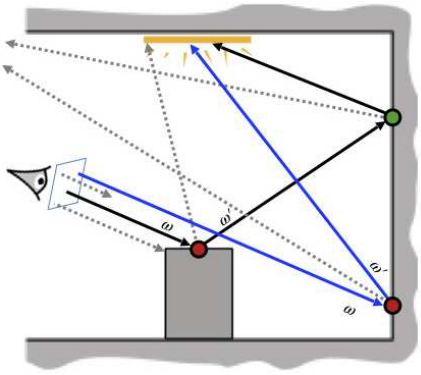


Figura 5. Los rayos salen de la cámara y viajan por la escena. En cada intersección, se calcula una nueva dirección para un rayo y se aplica la ecuación de renderizado, creando un camino de luz que terminará mediante un método de paro.

Renderizado Diferencial

Imagina que estás pintando un cuadro y dispones de una fotografía real que deseas plasmar en tu pintura. El renderizado diferencial se encarga de ajustar la pintura para

que se asemeje lo más posible a la fotografía. En primer lugar, se mide la diferencia entre la pintura y la fotografía (error). Luego, se modifican los colores y detalles en la pintura para mejorarla, como si estuviera probando con diferentes pinceles y tonalidades hasta lograr que la pintura se parezca más a la fotografía. Este proceso de ajuste y prueba se asemeja a un juego cuyo objetivo es perfeccionar la pintura mientras el error disminuye.

Tomando en cuenta cómo funciona un renderizador diferencial, es importante mencionar que el comportamiento de la luz será ligeramente diferente de lo que hemos visto anteriormente. Esta luz especial (ajustable) no puede ser observada de la manera normal. Para verla, necesitaremos gafas especiales, similares a las que se utilizan en el cine para ver películas en 3D. Sin embargo, en este caso, las gafas no crearán una sensación de profundidad tridimensional, sino que nos permitirán observar esta luz especial en función de algún parámetro de escena a modificar, como la posición de una fuente de luz, el color de un objeto, la textura, entre otros.

Existen dos formas principales de generar luz especial que cambia respecto a algún parámetro de la escena: mediante un emisor o una reflexión con parámetros ajustables. Esto significa que son elementos o parámetros que queremos ajustar para que la imagen se asemeje a la imagen de referencia. Ejemplos de estos parámetros ajustables son la posición de una fuente de luz, el color de un objeto o la geometría de algún elemento en la escena. Una vez creada esta luz especial, notemos que se dispersa de manera similar a la luz ordinaria en objetos con parámetros no ajustables. A diferencia de los parámetros no ajustables, que permanecen constantes, los ajustables pueden ser modificados. Por ejemplo, si en una escena queremos modificar la posición del sol porque la imagen original proyecta sombras en diferentes direcciones, solo la posición del sol sería un parámetro ajustable. Los demás elementos de la escena permanecerían inalterados, ya que representan fielmente sus valores.

El renderizado diferencial es una técnica avanzada, que se utiliza para ajustar y optimizar modelos 3D basándose en cómo cambia la imagen renderizada con respecto a los parámetros de escena. Esta técnica es particularmente útil en el campo del renderizado inverso, donde el objetivo es mejorar la precisión y la calidad de las reconstrucciones tridimensionales [3]. Pero, ¿cómo funciona exactamente? A continuación vamos a verlo en detalle.

El primer paso en el proceso de renderizado diferencial es generar una aproximación de una imagen a partir de un modelo 3D. Este modelo puede ser cualquier cosa, desde una figura de acción hasta un edificio complejo. El renderizador toma el modelo 3D, (aplica tipos de texturas, materiales, geometrías, luces, cámaras, etc) y produce una imagen 2D (Figura 6).

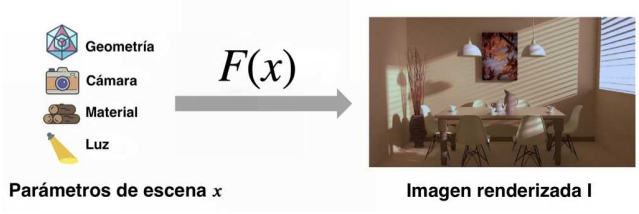


Figura 6. Escena renderizada de un comedor iluminada por el sol desde fuera de la ventana. En este caso, el parámetro que se desea obtener mediante la optimización es la posición del sol.

El renderizado inverso consiste en obtener la descripción de una escena partiendo de una imagen 2D.

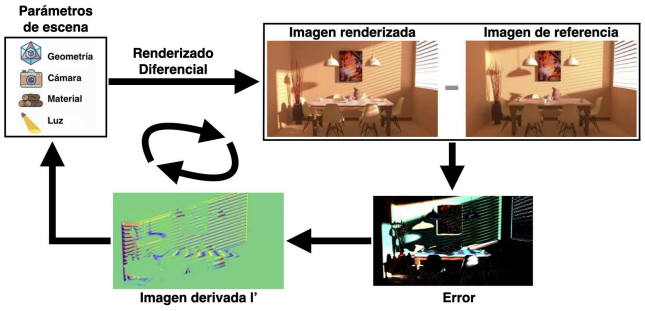


Figura 7. Las sombras de los objetos en la imagen cambian dependiendo de la posición del sol. El error es la diferencia de las imágenes (renderizada y de referencia). La razón de cambio (derivada) del error ayuda a cambiar los parámetros en este caso, la posición del sol.

Una vez que se ha generado la imagen, se compara con una imagen de referencia o foto tomada del mundo real de la cual se quiere hacer coincidir. Esta comparación es fundamental porque permite medir qué tan bien el modelo 3D se asemeja a la realidad. El objetivo es minimizar la diferencia entre la imagen renderizada y la imagen de referencia (Figura 7).

El error calculado nos ayuda a identificar en qué aspectos el modelo 3D no coincide con la imagen de referencia, ya sea en términos de forma, color, textura o iluminación. Cuanto menor sea el error, más precisa será la reconstrucción del modelo 3D de la escena (Figura 7).

Con el error calculado, el siguiente paso es ajustar el modelo 3D para reducir esta diferencia. Aquí es donde el renderizado diferencial demuestra su potencial. Al utili-

zar técnicas matemáticas avanzadas, como el cálculo de gradientes se pueden realizar ajustes en los parámetros de la escena (como la posición de las luces o cámaras, las texturas o la geometría del modelo) para minimizar el error [4]. Por ejemplo, la razón de cambio entre cada imagen calculada indica en qué dirección deben realizarse estos ajustes (Figura 7).

La optimización no se realiza en un único paso. El proceso se repite varias veces, generando nuevas imágenes, calculando el error y ajustando el modelo 3D. Con cada iteración, el modelo 3D se ajusta y refina hasta que la imagen generada se asemeje lo más posible a la imagen de referencia. Este ciclo iterativo permite un refinamiento continuo y una mejora progresiva en la precisión del modelo.

Un ejemplo aplicado a la geometría

Un ejemplo de cómo obtener parámetros de geometría utilizando el motor de renderizado Mitsuba 3 [5,6], se puede ver en la Figura 8(a), donde la idea es reconstruir un changuito. Como paso inicial, se crea una escena que contenga una aproximación burda de este objeto, en este caso una esfera (Figura 8(b)) y aplicando renderizado diferencial se optimiza la escena para que se parezca cada vez más a la imagen de referencia. La optimización intenta hacer coincidir con la geometría del changuito (Figuras 8(c) y 9).

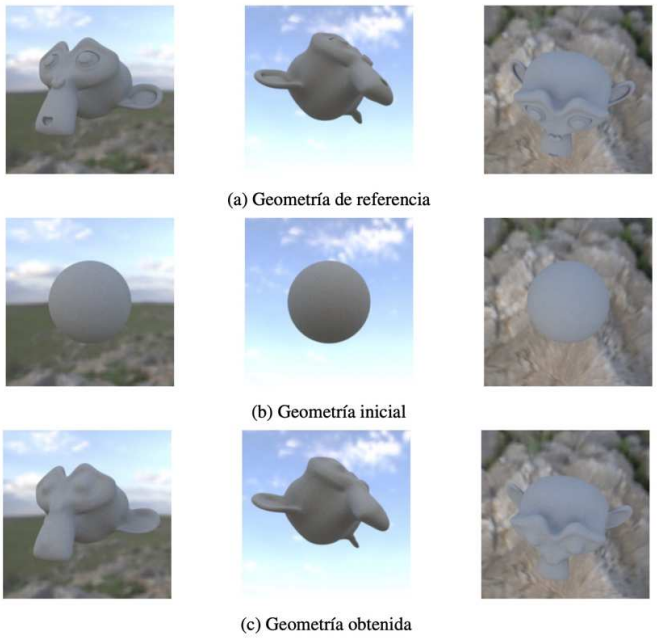


Figura 8. Se muestran varios ángulos de la foto de un changuito con el objetivo de obtener su representación tridimensional. Es necesario contar con diferentes ángulos para considerar toda la información de la geometría objetivo. El renglón superior (a) muestra la geometría de referencia, el (b) presenta una aproximación inicial de esta geometría y el (c) muestra la reconstrucción final de la geometría obtenida.

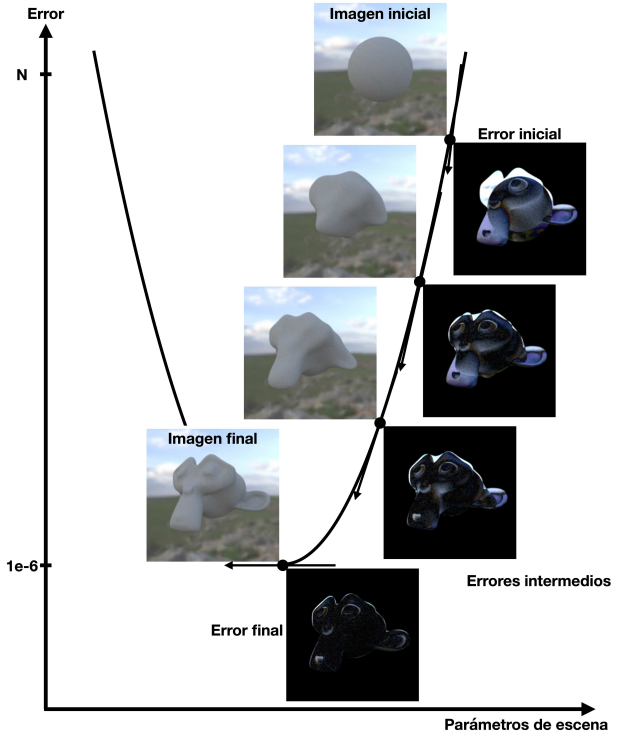


Figura 9. El error disminuye en cada paso, lo que significa que el modelo del changuito se va acercando más a la geometría objetivo. Al final, después de muchas iteraciones, el modelo 3D que se obtiene coincide casi perfectamente con el de las imágenes originales. Es como si tuvieras una bola de plastilina y mirando las fotos del changuito, la fueras moldeando hasta que te quede igual a la de las fotos.

Conclusiones

El renderizado diferencial está en constante evolución y sus aplicaciones prometen revolucionar muchos campos de la investigación. Esta poderosa herramienta está transformando cómo se crean y optimizan las imágenes en gráficos por computadora. Al hacer que el proceso de renderizado sea diferencial, se abren nuevas posibilidades para la integración de algoritmos de aprendizaje automático, permitiendo avances significativos en la calidad y eficiencia de las imágenes generadas por computadora. Además, el permitir la reconstrucción de escenas 3D a partir de imágenes 2D, ofrece un potencial inmenso para la creación de aplicaciones avanzadas en realidad aumentada, cine, animación, robótica y más. A medida que la tecnología continúa evolucionando, el renderizado inverso seguirá desempeñando un papel crucial en el futuro de los gráficos por computadora y la visión computacional.✷

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

REFERENCIAS

1. Loper, M.M. y Black, M.J. (2014). OpenDR: An Approximate Differentiable Renderer. *Computer Vision - ECCV 2014*. Lecture Notes in Computer Science. 8695. Springer, Cham. https://doi.org/10.1007/978-3-319-10584-0_11.
2. Kajiya, J. T. (1986). The rendering equation. *SIGGRAPH Comput. Graph.*, 20(4), 143-150. https://doi.org/10.1145/15886.15902
3. Kato, H., Beker, D., Morariu, M., Ando, T., Matsuoka, T., Kehl, W. y Gaidon, A. (2020). Differentiable Rendering: A Survey. ArXiv, abs/2006.12057.
4. Zhang, C., Miller, B., Yan, K., Gkioulekas, I. y Zhao, S. (2020). Path-Space Differentiable Rendering. *ACM Trans. Graph.*, 39(4), 143:1-143:19.
5. Jakob, W., Speierer, S., Roussel, N. y Vicini, D. (2022). Dr.Jit: A Just-In-Time Compiler for Differentiable Rendering. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 41(4). https://doi.org/10.1145/3528223.3530099
6. Jakob, W., Speierer, S., Roussel, N., Nimier-David, M., Vicini, D., Zeltner, T., Nicolet B., Crespo, M., Leroy, V. y Zhang, Z. (2022). Mitsuba 3 renderer. https://mitsuba-renderer.org.
7. Rhodin, H., Robertini, N., Richardt, C., Seidel, H.-P. y Theobalt, C. (2016). A Versatile Scene Model with Differentiable Visibility Applied to Generative Pose Estimation. https://arxiv.org/abs/1602.03725.

SOBRE LOS AUTORES



José Abraham Ramos Badillo es Ingeniero en Computación por la Universidad Michoacana de San Nicolás de Hidalgo y actualmente cursa la Maestría en Ciencias en la misma universidad. Sus intereses en investigación incluyen el renderizado fotorealista basado en física, la optimización y los métodos numéricos.



Luis Eduardo Gamboa Guzman es Doctor en Ciencias Computacionales por parte de la Universidad de Montreal, Canadá. Realizó sus estudios de Maestría en Ciencias en la UMSNH, y de Ingeniero en Sistemas Computacionales en el Instituto Tecnológico Nacional campus Morelia. Actualmente es profesor investigador en la Facultad de Ingeniería Eléctrica de la Universidad Michoacana de San Nicolás de Hidalgo y miembro del Sistema Nacional de Investigadores. Su investigación abarca el renderizado fotorealista basado en física, el procesamiento digital de señales y los métodos numéricos.



Karina Mariela Figueroa Mora es Doctora en Ciencias de la Computación por la Universidad de Chile, Ingeniera Electricista y Maestra en Ingeniería Eléctrica opción Sistemas Computacionales por la Facultad de Ingeniería Eléctrica de la Universidad Michoacana de San Nicolás de Hidalgo. Profesora e Investigadora de la Facultad de Ciencias Físico-Matemáticas UMSNH. Pertenece al Sistema Nacional de Investigadoras, perfil PRODEP. Actualmente está enfocada en recuperación de Información y desarrollo de software educativo para la enseñanza de matemáticas.

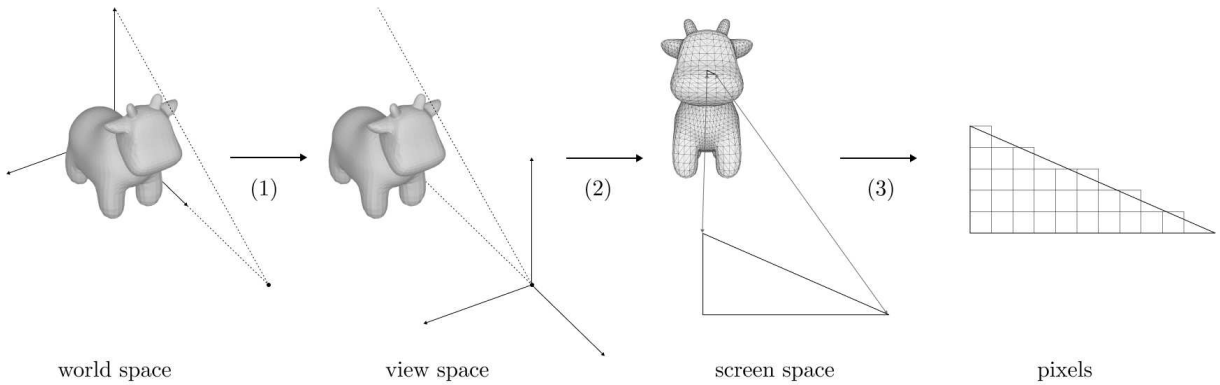


Imagen tomada de: https://andrewkchan.dev/posts/diff-render.html?form=MG0AV3.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Odometría Visual con IA: Innovando las Tecnologías de Localización

Leonardo Herrera

Introducción

Imagina viajar en un avión que te lleva de forma segura de una ciudad a otra, y luego te enteras de que el avión en el que volaste era un avión sin piloto, es decir, autónomo. Esto parecería sorprendente, pero en la actualidad estos vuelos ya han sido realizados por la compañía Airbus. De forma análoga, taxis autónomos han sido desplegados en ciudades como San Francisco y Phoenix, donde el usuario puede solicitar servicio de taxi a través de una aplicación de teléfono móvil y este llega a la ubicación del usuario y lo conduce hacia su destino sin conductor local o remoto. Waymo es una de las compañías pioneras en desarrollar este tipo de vehículos que dan servicio de taxi en dichas ciudades.

La arquitectura que dota de autonomía a estos vehículos -ya sean aviones, carros, satélites, barcos, entre otros- es comúnmente conocida en la literatura como navegación autónoma y se muestra en la Figura 1. Los módulos que la constituyen son: módulo de percepción, de planeación, y de control. El módulo de percepción toma información de sensores para determinar el entorno del vehículo. Entre las funciones de este módulo está la de determinar la localización del vehículo, la localización de otros vehículos, obstaculos, y todo lo relacionado con determinar el medio en el que se desempeña el vehículo. El módulo de planeación considera la información del módulo de percepción para generar trayectorias factibles y seguras que pueda seguir el vehículo para alcanzar un destino deseado sin caer en situaciones de riesgo. Finalmente, el módulo de control habilita los actuadores del vehículo, turbinas en el caso de un avión, para dirigirlo a través de la trayectoria generada por el módulo de planeación a su destino.



Figura 1. Arquitectura de Navegación Autónoma.

El enfoque de este artículo es en el módulo de percepción y particularmente en el submódulo de localización, mismo que es crucial para la navegación autónoma y para otras aplicaciones que se abordarán más adelante. Para entrar en contexto, se inicia el documento con un capítulo que expone algunas técnicas de localización que prescinden del uso de visión computacional. Posteriormente se sigue con el capítulo que aborda la localización basada en visión computacional. Después se expone el capítulo de aplicaciones potenciales que puede tener la localización basada en visión computacional. Finalmente, se presentan las conclusiones.

Localización que prescinde de visión computacional

Un sistema paradigmático y ubicuo de localización es el GPS (Global Positioning System). En la actualidad, cada teléfono inteligente cuenta con antena receptora de señales de GPS que permiten ubicar el teléfono en cualquier parte del planeta. El Sistema GPS inicialmente fue desarrollado para fines militares por el Departamento de Defensa de los Estados Unidos. Sin embargo, esta tecnología fue liberada a plenitud para usos civiles en el año 2000.

El GPS consta de tres componentes cruciales: antenas, constelación de satélites, y estaciones terrestres. Las antenas se encuentran integradas en los teléfonos inteligentes, los satélites en órbita terrestre media -MEO (Medium Earth Orbit)-, y las estaciones terrestres en diversas localidades alrededor del planeta: Colorado, Hawaii, Ecuador, Uruguay, Diego Garcia, Guam, Nueva Zelanda, entre otras.

El primer lanzamiento de satélite GPS fue realizado en 1978 y subsecuentes lanzamientos se llevaron a cabo para completar la constelación de 24 satélites en 1993. En la actualidad la constelación es redundante, ya que cuenta con 31 satélites, de los cuales 24 son suficientes para su operación, y los restantes son respaldos que entran en acción en caso de requerimiento. Las estaciones terrestres son las encargadas de mantener la constelación en armonía, que los satélites orbiten de acuerdo con un almanaque que dicta su ubicación en función del tiempo. Las estaciones terrestres, por ejemplo, monitorean y controlan los satélites en caso de decaimiento natural causado por atracción de cuerpos como la Tierra, u otros cuerpos celestes. Los satélites envían señales a las antenas del celular, mismas que son procesadas internamente en circuito integrado para producir la localización de la antena. Para una exitosa localización se requiere que al menos cuatro satélites sean visibles para la antena.

El GPS, aunque prominente, no cubre todas las necesidades de navegación. Algunas limitaciones son su baja frecuencia, su funcionamiento exclusivo en espacios abiertos, y su determinación única de posición y velocidad lineales.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Su baja frecuencia de 1 Hz, por ejemplo, no permite la navegación de aeronaves o misiles hipersónicos, ya que estos objetos alcanzan velocidades mayores a cinco veces la velocidad del sonido -mayores a 5000 Km/hora-. Asumiendo 5000 Km/ hora, el GPS solo podría medir la ubicación del objeto cada 1.38 Km, lo cual resultaría catastrófico para el objeto por quedarse tanta distancia sin saber su ubicación. Solución a este problema es dada por técnicas de fusión de datos, donde el GPS y sensores inerciales de alta frecuencia como acelerómetros, giroscopios, y magnetómetros son fusionados para proporcionar una localización de alta frecuencia que cubra la necesidad de navegación de objetos hipersónicos.

Los sensores inerciales pueden localizar a muy alta frecuencia, operar en espacios cerrados por periodos limitados, y proporcionar variables angulares complementarias a la localización proporcionada por el GPS. Los sensores inerciales son independientes y compactos, estos no requieren de componentes adicionales como el GPS que se constituye de antenas, satélites, y estaciones terrestres. Para referencia del lector, la Figura 2 muestra dos sensores inerciales de distintos fabricantes. Los sensores inerciales, sin embargo, tienen una problemática de desviación progresiva de la localización debida a que pequeños errores en las aceleraciones y velocidades angulares que estos miden son proyectados como grandes errores en la localización que estos proporcionan. Esta desviación crece conforme el tiempo transcurre, de tal modo que la localización es de calidad solo para el periodo inicial desde que se inicia la operación del sensor. Por otra parte, la desviación progresiva no existe en el GPS, de tal suerte que ambos sensores, inerciales y GPS son complementarios que al fusionarlos generan una sinergia que permite una localización de alta frecuencia, completa (posición, orientación, velocidad lineal, y variaciones angulares), y que puede operar en espacios cerrados por periodos considerables. Referencias relevantes a fusión GPS y sensores inerciales son [1] y [2].

GPS y Sensores Inerciales: Sinergia para una Localización Superior.

Aunque la sinergia GPS y sensores inerciales puede operar en espacios cerrados por periodos considerables, esta sinergia funciona siempre y cuando haya GPS, aunque intermitente. Localización en túneles, cuevas, o minas, debido al bloqueo permanente de las señales provenientes de los satélites destruiría la sinergia. Para esta limitación, tecnologías como la desarrollada en el Proyecto POINTER (Precision Outdoor and Indoor Navigation and Tracking for Emergency Responders) de la NASA JPL (National Aeronautics and Space Administration-Jet Propulsion Laboratory) han sido introducidas [3].

La tecnología PONTER consiste de tres componentes: receptor, transmisor, y estación. El receptor tiene el tamaño aproximado de un teléfono móvil y es portado por algún usuario. El transmisor que se mantiene fijo en alguna localidad emite campos magnéticos quasiestáticos que pueden penetrar a través de muros y otras estructuras que son imposibles de penetrar por las señales de GPS. El receptor detecta las señales del emisor, calcula su localización, y envía esta información a la estación base que consta de una computadora. Dicho Proyecto ha sido desarrollado para ubicar a primeros respondientes, como bomberos, en edificios en llamas para poderlos rescatar en caso de ser requerido.

La tecnología POINTER puede ser transferida a localización en minas y salvar vidas en caso de derrumbes. El desastre minero de la mina Pasta de Conchos ocurrido en 2006 en Coahuila es un caso emblemático que cautivo al mundo. Según el periódico Los Angeles Times [4], se dieron versiones contradictorias y alejadas sobre la ubicación de los mineros atrapados bajo tierra. Una fuente declaro que se encontraban a una profundidad aproximada de 480 m, mientras que la otra declaro que eran aproximadamente 150 m. Con la tecnología de localización aquí descrita se eliminaría la incertidumbre en ubicarlos y así gestionar un plan adecuado de rescate en función de su ubicación.

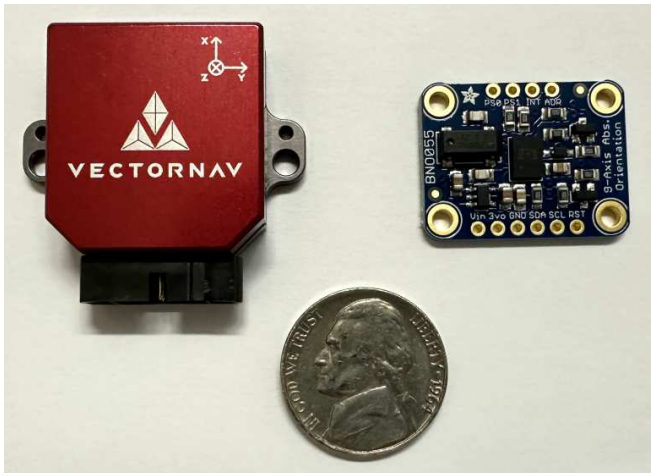


Figura 2. Sensores inerciales. A la izquierda sensor fabricado por la compañía Vectornav. A la derecha sensor fabricado por la compañía Adafruit.

Localización basada en visión computacional

La localización basada en visión computacional utiliza como dispositivo central la cámara de video que es montada en un vehículo de interés para determinar la localización del vehículo a través de secuencias de imágenes

capturadas. A esta tecnología se le conoce como odometría visual [5] y se ha empleado para navegar vehículos en Marte [6], donde no existe GPS. Similar a los sensores inerciales, la odometría visual es también independiente y compacta, donde lo único necesario para su operación es la cámara y una microcomputadora. Un método común de la odometría visual es el basado en características [7]. Este método utiliza algoritmos para extraer puntos clave en imágenes consecutivas. Estos puntos clave permanecen consistentes entre las imágenes debido a que estas se capturan ante pequeños movimientos de la cámara. Asumiendo emparejamiento, es decir, una correspondencia entre los puntos clave de las imágenes consecutivas, se realiza la estimación del movimiento, lo que permite determinar la localización de la cámara. Esto se logra minimizando un error mediante algoritmos como PnP (Perspective-n-Point) y otras técnicas de optimización.

Odometría Visual: Una Proeza para la Localización en Marte.

La odometría visual puede emplear cámaras con uno o dos lentes para la captura de la secuencia de imágenes. Cuando se utiliza cámara con un lente se denomina odometría visual monocular, con dos lentes se denomina odometría visual estereo. El uso de cámara con un lente puede llevar al problema de profundidad, es decir al problema de perdida de una dimensión que produce ambigüedad en la localizacion. El uso de cámara con dos lentes soluciona este problema debido a las distintas percepciones del objeto adquiridas por los lentes. La odometría visual inercial es una extensión de las ya mencionada que emplea y combina sensores inerciales como los mencionados anteriormente para producir una localización mejorada.

La odometría visual ha sido actualizada con los nuevos algoritmos de inteligencia artificial, los cuales permiten una localización cuasidirecta o directa. a) Localización quasidirecta: se emplea típicamente una red neuronal artificial para predecir los puntos clave, en lugar de utilizar los complejos algoritmos matemáticos del esquema tradicional. De esta manera, solo resta estimar la localización a partir de optimizar el error de proyección mediante el esquema tradicional. b) Localización directa: se emplea también de forma típica una red neuronal artificial para predecir la localización de manera directa, sin necesidad de pasar por el proceso tradicional de extracción de puntos, emparejamiento, y optimización. La localizacion directa da lugar a un esquema entrada-salida que mitiga complejidades del esquema tradicional. Como entrada se toma las imagene capturada por la cámara y como salida se produce la localización.

El esquema entrada-salida también está tomando gran relevancia en la arquitectura de percepción, planeación, y control de la Figura 2. Esta arquitectura ha sido el statu quo en la navegación autónoma; sin embargo, está siendo reformulada a un esquema entrada-salida debido a desafíos como:

1.-Tiempo de procesamiento: es el tiempo que se consume de sensado a actuación, es decir el tiempo transcurrido desde que se recibe información de sensores hasta que el vehículo es dirigido por sus actuadores. Este tiempo puede llevar a resultados catastróficos si es lo suficientemente grande para las velocidades ejecutadas por el vehículo.

2.-Ambiente impredecible: es el ambiente incierto en el que se desempeña el vehículo, un vehículo puede experimentas turbulencia en el ambiente que perturbe su movimiento, así como obstáculos dinámicos, tales como otros vehículos que se interpongan en el camino. El no tomar estas incertidumbres en consideración, puede llevar a resultados catastróficos.

3.-Borrosidad en imágenes: borrosidad de las imágenes capturadas por la cámara debido al movimiento del vehículo puede ocasionar una percepción sesgada de la realidad y llevar a resultados inesperados.

Para lidiar con lo anterior, el esquema entrada-salida remplaza los tres módulos descritos por una red neuronal artificial que simplemente se alimenta con los datos de sensores para producir la salida, que son las señales de control que dirigen el vehículo, ver Figura 3. Este esquema de entrada salida es similar al que utilizan de forma natural los humanos al conducir un vehículo. El humano percibe a través de sus sentidos (sensores) el entorno, transmite esta información al cerebro, la procesa, y toma acción sobre el vehículo a través de los actuadores para dirigirlo hacia un destino.

Esta nueva tecnología, inspirada en comportamiento humano, está tomando una gran relevancia. Como prueba de ello, en la prestigiosa conferencia CVPR (Computer Vision and Pattern Recognition) 2024 celebrada en Seattle, se llevó a cabo una competencia sobre carros autónomos manejados por esta tecnología (End-to-End Driving at Scale). El equipo ganador fue la compañía NVIDIA, misma que publicó su contribución en [8].

Desventajas de esta nueva tecnología son, por mencionar algunas, las cantidades masivas de datos requeridos para el entrenamiento de la red; datos que capturen todos los posibles escenarios que el vehículo autónomo pueda experimentar. También la capacidad de cómputo masivo requerido para el entrenamiento de la red. Por último, la perdida de intuición de los módulos tradicionales. Los tres módulos son absorbidos por una red neu-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ronal donde su esencia se pierde. Respecto a este último punto, la localización (ya sea por visión computacional, GPS, Celeste, etc.) resulta completamente reformulada y ajena al esquema tradicional.

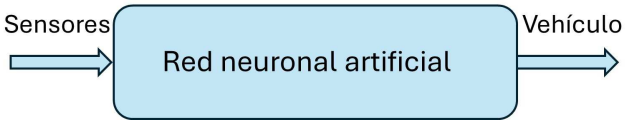


Figura 3. Arquitectura de navegación autónoma con inteligencia artificial de inicio a fin.

Aplicaciones potenciales

Se describen en esta sección dos de las potenciales aplicaciones que puede tener la localización basada en visión computacional.

ABL es un avión Boeing 747-400F modificado para defensa en contra de misiles balísticos. El sistema opera con una nariz laser que permite la desintegración de estos. El proyecto fue desarrollado por la Fuerza Aérea de los Estados Unidos, y exitosamente logro su objetivo en experimentos realizados. Sin embargo, el proyecto fue cancelado en 2011. La localización basada en visión computacional puede jugar un papel crucial en esta tecnología, ya que permitiría la localización automática del misil con respecto al sistema ABL. Dicha localización permitiría, debido al conocimiento de la orientación del misil, determinar un punto vulnerable en el cual se enfoque el láser. Esta es una aplicación interesante debido a que se trata de localizar un objeto remoto no cooperativo sobre el cual no se tiene influencia.



Figura 4. Airborne Laser, creditos a US Missile Defense Agency.

La nave Dragon 2 de Space X utilizada por primera vez en 2020 para transportar astronautas de suelo norteamericano a la EEI es una proeza de la ingeniería. La localización basada en visión computacional puede jugar un rol importante en la etapa de acoplamiento entre la nave y la estación. Para lograr un acoplamiento automático, la localización de la nave con respecto a la estación, producida por la visión computacional, permitiría que el sistema de propulsión de la nave genere la acción de control requerida para alcanzar el acoplamiento.



Figura 5. Acoplamiento entre nave Dragon 2 y Estación Espacial Internacional, creditos a NASA y Space X.

Conclusiones

La localización es un componente fundamental en la navegación, gracias a esta se pueden lograr misiones como la navegación autónoma, la navegación en Marte, los viajes internacionales, y muchas otras. El artículo presentó un breve recorrido por este campo, empezando con la localización que prescinde de visión computacional, luego la basada en visión computacional, y por último, aplicaciones potenciales de la localización basada en visión computacional. Las técnicas aquí descritas solo son una fracción del paquete de técnicas de localización existentes. Como en muchos campos de la ingeniería, la inteligencia artificial no ha estado ajena al campo de la localización. La actualización de los algoritmos de localización por inteligencia artificial es un hecho y traerá nuevas tecnologías como las que ya estamos viviendo con los taxis y aviones autónomos.✳

REFERENCIAS

1. Farrell, J. (2008). Aided navigation: GPS with high rate sensors. *McGraw-Hill, Inc.*
2. Grewal, M. S., Weill, L. R. y Andrews, A. P. (2007). Global positioning systems, inertial navigation, and integration. *John Wiley and Sons.*
3. `https://www.jpl.nasa.gov/news/pointer-seeing-through-walls-to-help-locate-firefighters`. (Sitio web disponible en EEUU).
4. `https://www.latimes.com/archives/la-xpm-2006-feb-20-fg-mexmine20-story.html`. (Sitio web disponible en EEUU).
5. Scaramuzza, D. y Fraundorfer, F. (2011). Visual odometry. *IEEE robotics and automation magazine*, 18(4), 80-92.
6. Cheng, Y., Maimone, M. W. y Matthies, L. (2006). Visual Odometry on the Mars Exploration Rovers-A Tool to Ensure Accurate Driving and Science Imaging. *IEEE Robotics and Automation Magazine*, 13(2), 54-62.
7. Gao, X., and Zhang, T. (2021). Introduction to Visual SLAM: From Theory to Practice. *Springer Nature.*
8. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P. y Zieba, K. (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

SOBRE EL AUTOR



Leonardo Herrera recibió su grado de Doctor en Ciencias en Instrumentación y Control del Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE) en 2018, y los grados de Maestro en Ciencias y Licenciatura en Ingeniería Eléctrica del Instituto Tecnológico de la Laguna en 2011 y 2009, respectivamente. De 2016 a 2017 participó en dos estancias de investigación en NASA LaRC y NASA JPL. También realizó un postdoctorado en el Mechanical and Aerospace Engineering Department de Naval Postgraduate School de 2019 a 2022. Actualmente, desempeña trabajo de investigación en dicha institución. Sus intereses de investigación incluyen Aprendizaje Profundo, Control de Robots, Navegación, y Filtro de Kalman.



© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Videovigilancia predictiva: anticipándose al robo

L. Enrique Morales Márquez, Ivan Olmos Pineda y J. Arturo Olvera López

Los sistemas de video vigilancia surgieron a inicios de los años 40, como método para monitorear actividades militares alemanas. Pocos años después, empezaría su uso comercial el cual evolucionaría al ser dotado de opciones como almacenamiento, captura a color, o incluso, monitoreo por internet y características inteligentes como reconocimiento de personas y movimiento.

La capacidad de registrar eventos en video ha permitido vigilar lo que sucede en vía pública e interiores. Si bien, existe personal para revisar en vivo lo que ocurre en cámara, en muchas ocasiones la cantidad de metrajes que deben atenderse al mismo tiempo excede las capacidades humanas, situación que sucede principalmente en ciudades grandes en donde a pesar de tener centros de monitoreo, un delito puede ocurrir sin que este se atestigüe.

Si los sistemas de vigilancia pudieran detectar de manera automática patrones sospechosos previos a un delito sin la necesidad de ser revisados por personal humano, entonces podrían tomarse medidas para evitar que sucedan.

Características que levantan sospecha

Al andar por la calle, en más de una ocasión las personas han tenido la sensación de que alguien más tiene intención de robarles, esta sospecha suele estar justificada por diversos factores que se analizan al prestar atención por un tiempo. Además, cuando los humanos detectamos alguna de estas características en otra persona, normalmente tomamos medidas para evadirla.

Los comportamientos inusuales suelen estar relacionadas con la apariencia y el actuar del individuo, descriptores como la manera de vestir, el corte de cabello o la complexión tienen una fuerte influencia en el juicio que generan las personas. Por otro lado, existen conductas a las que se les suele prestar atención ante una sospecha, indicadores como la mirada, la postura y el movimiento, así como las acciones e interacción con el entorno, además de algunas otras que van más allá del campo de estudio de la visión por computadora, como el uso de la voz [1].

Se conoce que, en el caso específico de los hurtos en tienda, acciones como observar por mucho tiempo a otros clientes o al personal, evadir a otros individuos, ocultar las manos o interactuar con el personal después del robo y antes de abandonar la tienda, son características importantes que pueden describir a los delincuentes [2].

Si es posible llevar a cabo estas interpretaciones a partir de lo que se observa, entonces ¿por qué no podría hacerlo una computadora usando una cámara de video? La realidad es que se han hecho intentos por lograr este objetivo. Mucho se ha trabajado para poder reconocer en tiempo real la ocurrencia de un delito, sin embargo, la tarea de detectar el comportamiento sospechoso y predecir la posibilidad de que acontezca un crimen es un terreno bastante menos explorado.

Se han obtenido ya algunos resultados que vale la pena revisar para darse cuenta de que, la predicción por computadora de ocurrencia de delitos con solamente *observar* los videos de vigilancia no es una meta de ciencia ficción, un esquema simple para un sistema de predicción para este fin puede observarse en la Figura 1.

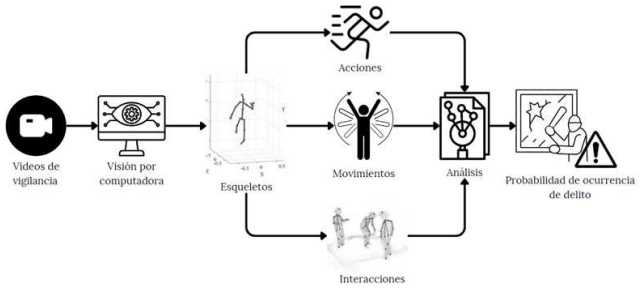


Figura 1. Sistema de predicción de delitos.

Enfoque en el momento previo al robo

Uno de los enfoques más populares en la visión por computadora es el uso de algoritmos que *observan* el contenido de un fotograma de video, este tipo de algoritmos se conocen como "redes convolucionales". Cuando repiten el procedimiento para cada cuadro de un video obteniendo información en función del tiempo se llaman "redes convolucionales tridimensionales" [3]. Se puede pensar en estos modelos de inteligencia como una forma de imitar la manera en que el cerebro capta información. Al tomar una secuencia de imágenes, un análogo computacional del cerebro tomará cada fotograma por separado y la pasará a través de varias etapas en las que se buscan características visuales como formas, bordes o texturas, así, se tiene una idea sobre cómo cambia el contenido de las imágenes conforme avanza el tiempo.

Cuando se tienen todos los datos, otra etapa los combina para obtener características más complejas y determina las acciones u objetos que ha detectado.

Una vez entendido el funcionamiento de estos modelos, describiremos brevemente el método de predicción

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

propuesto en [3]. Este proceso toma los fotogramas de una secuencia de video de vigilancia y los organiza en tres grupos: el primero corresponde al segmento de comportamiento previo, que considera la parte del video con el desarrollo normal de la escena; el segundo es el segmento de comportamiento sospechoso, donde se observan los fotogramas anteriores a un crimen; y, por último, el segmento de evidencia del crimen, que muestra el momento en que se comete el acto delictivo (ver Figura 2).

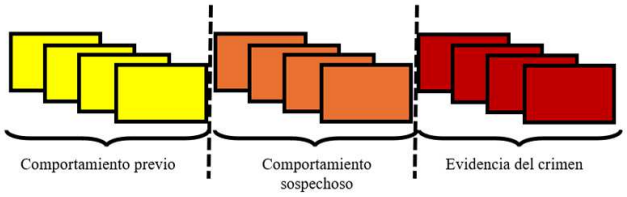


Figura 2. División del video en segmentos.

Este enfoque tiene como objetivo extraer únicamente los momentos en los que pueden existir patrones sospechosos y obtener características que aparezcan únicamente en este tiempo, aprendiéndolas y determinando cuales de ellas anuncian que un robo en tienda está por suceder.

Enfoque en determinación de sospecha

Una manera distinta de anticipar un delito que se apega un poco más al pensamiento humano es la determinación de sospecha. Esto no señala o sugiere que un acto delictivo ocurrirá, simplemente indica que se debe tener atención sobre ciertos individuos en un momento determinado [4]. A partir de videos de los que ya se conoce la existencia de comportamiento sospechoso, se construye un algoritmo de inteligencia muy similar al mencionado en la sección anterior, con la diferencia de tener una cantidad mucho menor de etapas pero que extraen cantidades más grandes de información.

Se analiza una cantidad pequeña de fotogramas del video con el propósito de acelerar el proceso, si se encuentra una cantidad de cuadros consecutivos que puedan considerarse sospechosos con alguna región de particular interés, se marca dicho espacio que puede o no incluir una persona y se envía una notificación. Una reconstrucción del tipo de resultados obtenidos con este sistema se muestra en la Figura 3.

Enfoque en el potencial de ocurrencia de robo

Anteriormente se mencionó que algunos indicadores como la apariencia y el comportamiento tienden a levantar sospechas. Sabiendo que algo inusual está aconteciendo, no sería extraño que en este momento surgiera la pregunta "*¿qué tan sospechoso se ve esto?*".

Una persona vistiendo un pasamontaña dentro de una tienda podría resultar mucho más inquietante que una persona que solamente use un cubrebocas, e incluso, un cubrebocas puede resultar más sospechoso que alguien más que se presenta con el rostro descubierto. De igual modo, un individuo que merodea demasiado alrededor de un punto puede llamar mucho más la atención que otro que se mantiene dentro de una misma zona.

La manera en que se determina *qué tanto* un objeto puede pertenecer a una categoría es una parte fundamental de lo que se conoce como "lógica difusa" que, combinan esos "*qué tanto*" que describen a los objetos o personas usando ciertas reglas para inferir junto con operaciones matemáticas, esto permite determinar una predicción y qué tan seguro se está de ella.

Con esta técnica, se ha analizado el grado de cobertura facial de los individuos, qué tanto merodea una persona dentro de una tienda y el tamaño de la multitud que aparece en un video para determinar un potencial de robo de manera numérica, algo similar a una probabilidad de ocurrencia [5]. Una reconstrucción del resultado de dicho sistema se puede apreciar en la Figura 4.



Figura 3. Cuadros con actividad normal y sospechosa.



Figura 4. Potencial calculado de robo.

Sobre la confiabilidad de estos sistemas

Como en cualquier otra tarea que tenga detrás a la inteligencia artificial, surge la interrogante de si estos avances

son confiables y si se pueden considerar exitosos, y las respuestas son complicadas de obtener.

Es importante recordar que cuando las computadoras aprenden, normalmente comparan sus resultados contra otros que son considerados verdaderos e irrefutables. Sin embargo, se depende totalmente de que esos resultados en teoría verdaderos efectivamente representen a la realidad, por tanto, si casi la totalidad de las predicciones de la máquina empatan con las que se consideran verdaderas, entonces es válido decir que el algoritmo es confiable y exitoso.

Los enfoques actuales buscan determinar la posibilidad de ocurrencia de delitos en tiempo real.

Las aportaciones que se han discutido en este documento reportan un grado de certeza de 86 %, 98 % y 77 % respectivamente, por tanto, se puede inferir que para la tarea de determinar si existe alguna actividad sospechosa, sí se ha tenido éxito, pero condicionado por las limitaciones con las que se ha entrenado a las máquinas. Por otro lado, en cuestión de predecir la ocurrencia de un acto delictivo como tal, se tienen resultados mayormente positivos pero que aún requieren ser mejorados para poder decir que han sido un éxito rotundo.

Como consecuencia, sería recomendable utilizar estos sistemas predictivos como auxiliares y no como el único criterio para tomar una decisión, al menos hasta que no se alcancen resultados mucho más apegados con la realidad.

¿Para quién se diseñan?

Los sistemas de videovigilancia actualmente se encuentran al alcance de la mayoría de las personas, tanto para uso empresarial como para cuidar de los hogares, pero ¿los sistemas de predicción de eventos delictivos son de utilidad para todos? En principio así debería ser, pues cualquier sistema de seguridad tiene como objetivo proteger la integridad física o material de las personas, aunque hay que tomar en cuenta que un sistema pensado para operar en escenarios de exterior y para una gran cantidad de delitos, sería de mayor interés para el departamento de seguridad pública de una ciudad.

Después de todo, un sistema para el hogar no requiere de capacidades para todos los delitos y tampoco poder reconocer escenarios de vía pública con grandes multitudes.

Un sistema para el hogar o comercio bien podría cubrir una cantidad menor de delitos y requerir un aprendizaje menos demandante y más específico, pero un sistema para operar en conjunto con las fuerzas de seguridad pública requiere de una cantidad de escenarios, manejo de multitudes y situaciones que están por encima de lo que puede ocurrir en un entorno más controlado como los interiores.

Además, las notificaciones generadas desencadenarán acciones distintas para cada caso, pues en la calle se pueden enviar elementos de seguridad al sitio requerido, mientras que en una casa podrían llevarse a cabo distintas acciones predefinidas por el usuario. Al final, sin importar en donde se apliquen estas soluciones de inteligencia artificial, se persigue el mismo objetivo que es intentar anticiparse a los delitos y formar parte de las acciones de prevención.

Desafíos para los sistemas de predicción de delitos

Los trabajos de visión por computadora son desafiantes, tanto para los investigadores como para los equipos de cómputo en sí. Por un lado, se requiere de equipos poderosos para aprender patrones de grandes volúmenes de imágenes y videos, por otro lado, los investigadores deben lidiar con otros problemas relacionados con interpretación y manejo de los datos.

Consideremos el caso en el que un par de personas captadas en cámara ejecutan comportamientos muy similares [6], pero una de ellas se encuentra asaltando de manera discreta a la otra. ¿Cómo saber que algo está sucediendo? ¿Y si el sistema califica a ambas personas como posibles delincuentes o concluye que no hay razón para sospechar?

Atender estas interrogantes en primera instancia puede parecer agobiante, no obstante, se debe considerar que de los videos se puede extraer información que a simple vista puede no ser perceptible, además de que, ante una mayor variedad de escenarios, el algoritmo puede aprender a distinguir mejor los comportamientos.

Otro de los desafíos actuales radica en los tipos de delitos que se predicen, actualmente los estudios para detección en tiempo real consideran una gran variedad de ilícitos callejeros, sin embargo, las tareas de predicción se han enfocado mayormente en el hurto en tienda, posiblemente debido a que suelen llevarse a cabo entornos más pequeños y en los que se tiene un mayor control de los escenarios.

Sin embargo, si para el hurto se han podido determinar indicadores para sospechar de una persona, entonces también debería ser posible determinar patrones para otros delitos y bajo esa misma línea, encaminar los esfuerzos en predecir otros actos. Valdría la pena pensar también en la manera en que las personas evolucionan y

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

cambian sus maneras de actuar. Diariamente se pueden escuchar noticias sobre nuevos modus operandi de los delincuentes, por lo que enseñar a las computadoras a predecir crímenes con un determinado patrón puede ser un avance tecnológico con fecha de caducidad, por lo que la constante actualización de los sistemas de predicción y detección representa también un reto a futuro que seguramente permanecerá como una constante ante los cada vez más creativos métodos de la delincuencia.

Conclusiones

En este artículo se abordó el por qué es posible predecir eventos delictivos usando visión por computadora, algunos métodos creados para este propósito y su funcionamiento, así como retos a enfrentar en esta tarea. También se reconoce que los estudios recientes han enfocado sus esfuerzos en algún tipo de delito, por lo que, tomando las mismas bases, se puede pensar a futuro en extender estos avances hacia una cantidad mayor de actos ilícitos incluso en vía pública, es decir, entornos menos controlados para poder llegar a métodos más generales capaces de predecir de manera adecuada la ocurrencia de delitos.

Sin duda es una tarea compleja pero que sigue en desarrollo y que, por la naturaleza evolutiva del comportamiento humano, se mantendrá en actualización constante.✳

REFERENCIAS

1. Reid, S., Coleman, S., Vance, P., et al. (2021). Using Social Signals to Predict Shoplifting: A Transparent Approach to a Sensitive Activity Analysis Problem. *Sensors*, 21(20).
2. Reid, S., Vance, P., Coleman, S., et al. (2021). Visual Social Signals for Shoplifting Prediction. En *PATTERNS 2021: PATTERNS 2021, The Thirteenth International Conference on Pervasive Patterns and Applications*, 37 –42.
3. Martínez-Mascorro, G., Abreu-Pederzini, J., Ortíz-Bayliss, J., et al. (2021). Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks. *Computation*, 9(2).
4. Gandapur, M. (2022). E2E-VSDL: End-to-end video surveillance-based deep learning model to detect and prevent criminal activities. *Image And Vision Computing*, 123.
5. Pouyan, S., Charmi, M., Azarpeyvand, A., et al. (2023). Propounding First Artificial Intelligence Approach for Predicting Robbery Behavior Potential in an Indoor Security Camera. *IEEE Access*, 11, 60471–60489.
6. Martínez-Mascorro, G., Ortiz-Bayliss, J. y Terashima-Marín, H. (2020). Detecting Suspicious Behavior on Surveillance Videos: Dealing with Visual Behavior Similarity between Bystanders and Offenders. En *2020 IEEE ANDESCON*, 1–7.
7. Alarcón, L. (2023). Seguridad pública y videovigilancia: Una óptica criminológica. *Archivos de criminología, seguridad privada y criminalística*, 21(11).

SOBRE LOS AUTORES



Luis Enrique Morales Márquez. Estudiante de doctorado en la Facultad de Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla. Sus intereses de investigación incluyen: visión por computadora, procesamiento de imágenes digitales, aprendizaje máquina, inteligencia artificial, lógica difusa, entre otras.



Ivan Olmos Pineda. Profesor investigador de tiempo completo en la Facultad de Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla. Sus intereses de investigación incluyen: aprendizaje máquina, reconocimiento de patrones, inteligencia artificial, árboles de decisión, entre otras.



José Arturo Olvera López. Profesor investigador de tiempo completo en la Facultad de Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla. Sus intereses de investigación incluyen: reconocimiento de patrones, minería de datos, aprendizaje máquina, análisis de imágenes digitales, biometría, entre otras.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

La Visión Artificial y la Robótica: Robots que ven

Rodrigo Juarez-Villalva, Efren Castro-Bobadilla, Hugo Yañez-Badillo, Laura Cleofas-Sanchez, J. Nabor Balderas-Gutierrez y Francisco Beltran-Carbajal

El funcionamiento del ojo de un mamífero depende de diferentes tipos de células que trabajan en conjunto en la retina. Existen células sensoriales llamadas bastones (que ayudan a distinguir entre la luz y la oscuridad) y conos (que permiten distinguir los colores). Además, otro tipo de células llamadas células interruptoras: como las bipolares y las ganglionares (encargadas de transferir información del ojo al cerebro), procesan información relevante como la nitidez de los bordes y la detección de movimiento [1].

Tanto las células sensoriales como las interruptoras, en conjunto con la potente capacidad del cerebro para interpretar una gran cantidad de información, permiten crear la imagen que se percibe del medio. Cuando es necesario llevar a cabo algún procedimiento médico, como los implantes de retina, el implante reemplaza a las células sensoriales degeneradas y estimula las células interruptoras que aún pueden funcionar.

La visión, sentido de la vista o percepción visual, es el sentido que proporciona la mayor cantidad de información a gran variedad de especies, entre ellos los seres humanos [2], debido a que posibilita la percepción del entorno con gran precisión y a una gran velocidad, permitiendo de esta manera la toma de decisiones y la interacción con el entorno. En paralelo, la visión por computadora constituye un campo de investigación que busca emular la capacidad humana para interpretar imágenes de manera instantánea e identificar objetos utilizando una amplia variedad de algoritmos aplicados en sistemas computacionales.

La visión por computadora es una disciplina que se enfoca en el análisis de imágenes, facilitando que los algoritmos detecten diferentes atributos, rasgos o características de los objetos. Su habilidad para detectar colores, formas, movimientos y profundidades, convierte a la vista en un recurso fundamental para la supervivencia y el desarrollo de habilidades avanzadas, lo que ha promovido el interés de la comunidad científica alrededor del mundo para dotar a las máquinas con esta capacidad [2].

Los conceptos de visión por computadora (o visión computacional), visión de máquina y visión artificial, frecuentemente se usan de manera indistinta, y aunque se relacionan, precisa aclarar que existen algunas particularidades y diferencias. La visión por computadora estudia cómo las computadoras interpretan imágenes. Esto se puede usar para tareas como el reconocimiento facial y la segmentación de imágenes. Dentro de esta, la visión de máquina se concentra en aplicaciones industriales como el control de calidad y la inspección de líneas de producción, como el sistema mostrado en la Figura 1.

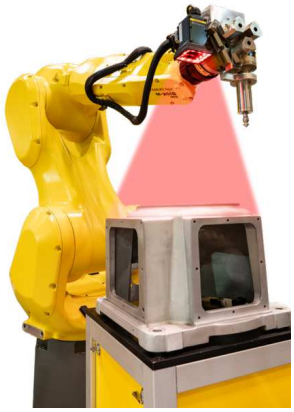


Figura 1. El Robot M-20iB-35S de Fanuc con iRVision durante una aplicación de desbarbado [3].

Por otro lado, la visión artificial es un término más amplio que abarca cualquier sistema que puede *ver*, utilizando inteligencia artificial, donde destacan aplicaciones como los vehículos o robots autónomos, como el que se observa en la Figura 2. Estos sistemas recurren a algoritmos que procesan, analizan y toman decisiones basadas en información capturada por cámaras y sensores.

Por lo tanto, se puede definir formalmente que la visión artificial es una rama de la inteligencia artificial y de las ciencias de la computación que se enfoca en permitir que las máquinas y los sistemas informáticos interpreten y comprendan imágenes del mundo real. En la Figura 2, se observa un robot móvil autónomo creado por Sevensense/ABB. *Cada robot móvil, equipado con visión e inteligencia artificial, escanea una parte única de la nave industrial; en conjunto, estos robots complementan la visión de los demás para formar un mapa completo, lo que les permite trabajar de forma autónoma en un entorno que cambia rápidamente*, [4].

En general, los diferentes enfoques coinciden en que el concepto de visión involucra un proceso computacional que requiere de varios niveles de procesamiento, desde la extracción de las propiedades básicas de las imágenes, hasta la interpretación de patrones complejos para la toma de decisiones.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

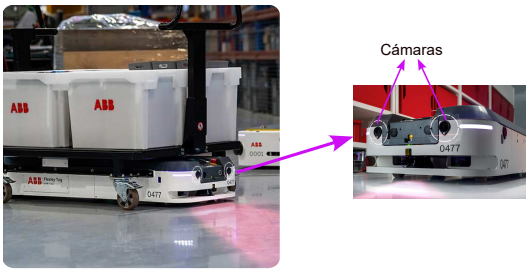


Figura 2. Robot móvil autónomo utilizando localización y mapeo visual simultáneo [4].

Antecedentes

El término de visión por computadora surgió a mediados de 1970s y hacía referencia a un enfoque de percepción visual para imitar la inteligencia humana y dotar a los robots con inteligencia. El origen de la visión por computadora fue el resultado de un intento de emular la inteligencia humana, pero resultó ser más complejo de lo que se esperaba. En [2] se destacan de manera general los principales avances en materia de visión computacional, a partir de su surgimiento y hasta la actualidad.

La visión por computadora es el proceso de capturar objetos del mundo real, realizar preprocesamiento y desarrollar algoritmos para su identificación en tiempo real mediante métodos computacionales.

En los años 70, se comenzó a distinguir la visión por computadora del procesamiento digital de imágenes. La diferencia radicaba en que la visión por computadora tenía la intensión de recuperar la estructura tridimensional del objeto a partir de esas imágenes y utilizarla para comprender completamente el escenario bajo análisis, por lo que se desarrollaron algoritmos para la detección de bordes y etiquetado de líneas para el análisis de cuerpos regulares.

En esa época también se implementaron algoritmos para la representación tridimensional de objetos irregulares y de visión estereoscópica (capacidad del ser humano de integrar imágenes de dos ojos en una sola imagen tridimensional), como en el flujo óptico (detección de movimiento en una secuencia continua de imágenes).

Durante la década de 1980, se lograron numerosos avances en el procesamiento de imágenes gracias al desarrollo de herramientas matemáticas sofisticadas. Se propusieron algunas funciones especiales y colecciones de imágenes para mejorar el análisis. De igual manera, se propusieron técnicas avanzadas para deducir la forma a partir del sombreado, la textura y el enfoque, así como para detectar bordes.

El campo de la visión por computadora experimentó un crecimiento importante en los años 90 en la reconstrucción de estructuras a partir del movimiento. Se plantearon diversos algoritmos de estéreo multivista (donde se utilizan técnicas para reconstruir la forma tridimensional de una escena o un objeto a partir de múltiples imágenes tomadas desde diferentes ángulos) y se mejoraron diversos algoritmos para el reconocimiento facial.

En la década de los 2000's, la visión por computadora se transformó con el auge de la fotografía computacional. La integración del aprendizaje automático (*Machine Learning*) revolucionó el reconocimiento de objetos y escenas, con un enfoque en técnicas basadas en características. La capacidad de aprender automáticamente a partir de una colección masiva de datos revolucionó completamente la visión por computadora. Se desarrollaron algoritmos más eficientes.

Las grandes bases de datos con información etiquetada y los avances en el desarrollo de las unidades de procesamiento gráfico impulsaron el reconocimiento de imágenes en la década de 2010. Durante esta etapa, se intensificó la innovación en la fotografía computacional, el desarrollo de herramientas sofisticadas basadas en el aprendizaje profundo (*Deep Learning*) y la realidad aumentada. Así mismo, la implementación de algoritmos en tiempo real y el uso de sensores especializados, impulsaron el desarrollo de vehículos autónomos terrestres y aéreos (*drones*).

La Visión Artificial en la Robótica

La visión computacional se ha convertido en una tecnología clave en diversas áreas, donde cada aplicación enfrenta desafíos únicos en el manejo y análisis de imágenes y videos [5]. La robótica, los vehículos autónomos, la manufactura, la interpretación de imágenes aéreas y satelitales, el diagnóstico médico mediante imágenes, el reconocimiento biométrico, la elaboración de planos, son algunas de las principales aplicaciones donde se ha aprovechado el potencial de la visión por computadora para abordar problemas sociales en la vida cotidiana [6].

Desde su surgimiento, los robots manipuladores o brazos robóticos, han tenido un impacto significativo en el ámbito social y económico, que se ha visto reflejado principalmente en ambientes industriales, donde juegan un papel fundamental para el progreso de los sectores productivos de cada país.

A medida que se ha expandido la necesidad de realizar tareas en ambientes externos no controlados, repetitivas, complejas o peligrosas, que ponen en riesgo la

integridad humana, ha surgido el desarrollo de los robots móviles para afrontar estos desafíos. Esta clase de robots pueden moverse sobre el suelo, usando ruedas o patas, volar por el aire usando alas fijas o rotores, moverse a través del agua, o navegar sobre ella [7].

La robótica móvil ha adoptado un papel protagónico en el desarrollo de aplicaciones que involucran trabajos subterráneos o submarinos, la desactivación de explosivos, la exploración espacial, las tareas de monitoreo y vigilancia, la búsqueda y rescate, la exploración en zonas comprometidas por desastres naturales, la agricultura de precisión, entre otras. La mayoría de estas aplicaciones se han potenciado considerablemente con la capacidad que tienen algunos robots de *ver* su entorno.

Si el entorno del robot no cambia, un mapa preciso del entorno podría ser suficiente para que el robot pudiese navegar. No obstante, si las condiciones no fueran ideales, este enfoque no funcionaría. Por ejemplo, para un auto autónomo, la existencia de otros vehículos en movimiento, peatones, señales de tráfico y los desperfectos u obras en la carretera, le impedirían navegar de manera satisfactoria solamente utilizando la información proporcionado por un sistema de posicionamiento global (conocido como GPS).

Para resolver este tipo de problemáticas, es necesario *observar* el mundo y planificar las acciones en consecuencia, y aunque esto es relativamente sencillo para los seres humanos, aún no es fácil programar en una máquina, resultando en un gran desafío para la visión artificial (o visión robótica).

Históricamente, los sistemas de visión han sido alimentados con datos provenientes de diversas fuentes, destacando las fuentes visibles e infrarrojas, sensores tridimensionales y diversos dispositivos médicos del tipo escáner como la tomografía computarizada y la resonancia magnética [6]. Para los sistemas robóticos, se han empleado varias fuentes de datos, como sensores de fuerza, sistemas GPS, sonares, odometría, unidades de medición inercial, sensores ópticos del tipo LiDAR (como el que se observa en la Figura 3), y cámaras. Esto ha permitido la integración exitosa de robots en aplicaciones complejas como la agricultura de precisión [8].

La visión artificial ha sido fundamental para el desarrollo de sistemas mecatrónicos avanzados, como los robots, ya que le permite a este tipo de sistemas percibir y entender su entorno, que es crucial para navegar, interaccionar y realizar tareas de manera autónoma. Además, proporciona la precisión necesaria para ejecutar tareas complejas y adaptarse a cambios en tiempo real empleando algoritmos que procesan, analizan y toman decisiones basadas en datos visuales capturados por cámaras y sensores.

La navegación de robots es una disciplina que se encarga de guiar a un robot desde una posición inicial hasta una posición objetivo en un entorno desconocido.

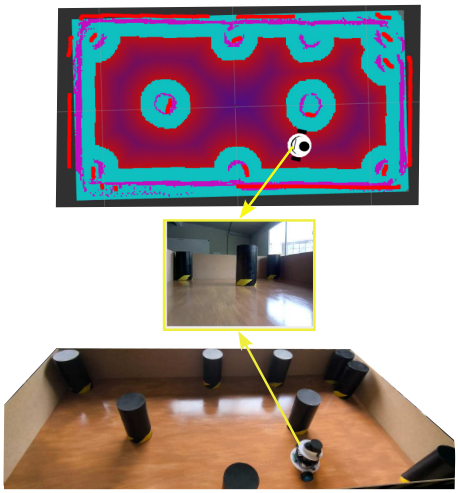


Figura 3. Mapa virtual creado mediante la tecnología LiDAR para la navegación de un robot móvil.

Aunque la navegación parece una tarea simple para los humanos, en realidad es un desafío complejo que requiere de la integración de diversas disciplinas como la robótica, el control, la visión computacional y la inteligencia artificial. La visión es fundamental para potenciar la capacidad de navegación de los robots en entornos complejos de manera autónoma, debido a que facilita la interacción en entornos de colaboración, conformado por humanos y robots colaborativos (o CoBots), lo que es clave en la automatización de procesos modernos mejorando la eficiencia y reduciendo costos, contribuyendo así con la filosofía de las Industrias 4.0 y 5.0.

La visión por computadora permite la reconstrucción del entorno y la creación de mapas en 3D, lo que mejora la navegación y la interacción en entornos complejos y no estructurados. Además, dota a los robots de habilidades cognitivas avanzadas, permitiéndoles no solo *ver*, sino también interpretar y aprender del entorno, lo que es esencial para su autonomía.

Para navegar, los robots dependen de su capacidad de percepción, localización, planificación, evasión de obstáculos y control: la percepción permite al robot interpretar su entorno mediante sensores; la localización determina la posición en el entorno mediante técnicas como

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

la localización y mapeo simultáneo (SLAM); la planificación de rutas ayuda a trazar rutas seguras mediante algoritmos especializados como los arboles de exploración [9]; la evasión de obstáculos emplea visión artificial para detectar y esquivar obstáculos en tiempo real; y el control, interpreta el movimiento planificado ajustando la trayectoria en función del entorno y una posible retroalimentación visual.

La navegación de robots continúa enfrentando una gran variedad de desafíos, como cambios en las condiciones del entorno y la incertidumbre en la percepción, lo que ha promovido el desarrollo de técnicas modernas que han potenciado el desarrollo de sistemas robóticos avanzados para lograr una navegación más autónoma y precisa, y dentro de las que destacan Optical Flow, Visual Odometry, Dense Mapping y el Deep Reinforcement Learning.

Casos de estudio: TTR_{01} y ρSus

La participación de estudiantes de las diferentes áreas de la ingeniería, a nivel licenciatura y posgrado, en el desarrollo de sistemas robóticos avanzados que empleen visión artificial, es fundamental, debido a que se requiere afrontar múltiples desafíos tecnológicos complejos, lo que permite avances en campos como la robótica, la automatización, la medicina, la inteligencia artificial, entre otras, que definirán el futuro tecnológico de nuestro país.

Para ilustrar la aplicación e implementación de la visión artificial para el desarrollo de sistemas mecatrónicos, se presentan dos casos de estudio donde se introducen las principales ideas desarrolladas por estudiantes de Ingeniería Mecatrónica para el reconocimiento de gestos en el control de sistemas aéreos autónomos y la navegación de un robot móvil diferencial utilizando SLAM.

Sistema TTR_{01}: el sistema (**T**urtle **T**est **R**obot, version **01**), que se observa en la Figura 4, es un robot móvil de tipo diferencial, desarrollado (diseño, manufactura y construcción) por los autores, y el cual puede navegar empleando un modelo SLAM gestionado en el ambiente de desarrollo para proyectos de robótica ROS (Robot Operating System).

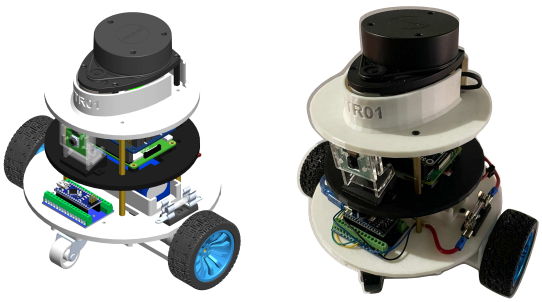


Figura 4. Sistema robótico móvil TTR_{01}.

Este sistema emplea la metodología SLAM para estimar la posición del robot con base en los objetos que detecta. También, emplea mapas cuadriculados del entorno en los que las celdas se representan con dos estados posibles: libre u ocupado, a partir de las mediciones promediadas de un sensor LiDAR, Figura 5.

Cada celda del mapa recibe un valor en una coordenada de la cuadrícula, indicando si está ocupada por un objeto o si está libre de obstáculos. Este proceso permite al algoritmo SLAM construir una representación del entorno en tiempo real (cómo se observa en la Figura 3), ayudando al robot a navegar y evadir obstáculos empleado de manera satisfactoria algoritmos para la planificación de rutas.

El control de está plataforma robótica se logra utilizando odometría (uso de datos de sensores de movimiento en las ruedas para estimar los cambios en su posición) y un algoritmo de control basado en el error de seguimiento de la ruta planificada, y el cual establece que nivel de voltaje debe de aplicarse a los actuadores (motores eléctricos) para disminuir ese error conforme el tiempo transcurre.

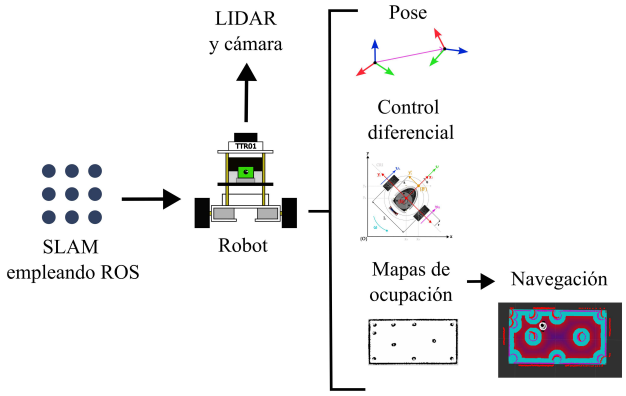


Figura 5. Esquema de funcionamiento del TTR_{01}.

Dentro de los principales elementos que conforman el sistema TTR_{01}, destacan elementos electrónicos para gestionar la energía del vehículo proporcionada por una batería de polímero de litio; un par de actuadores con ruedas para proveer de movimiento a la estructura del robot; una computadora Raspberry Pi4 que cumple la función de *cerebro* para el sistema, ya que en ella se encuentra un algoritmo informático que le permite *pensar*; un sensor LiDAR y una cámara que le permiten al robot *ver* o percibir su entorno, y un chasis manufacturado con impresión 3D.

Sistema ρv$\mathbb{S}$us: este sistema hace referencia a (**ro**tational **s**tructure **s**ystem for **u**nmanned **s**ystems) el cual fue diseñado para incorporar el reconocimiento de gestos corporales en el control de un vehículo aéreo no

tripulado de cuatro rotores. El sistema que se muestra en la Figura 6, posee dos grados de libertad (permite dos clases de movimiento), indicados con las líneas verde y morada, uno de traslación y uno rotacional, ambos respecto al eje vertical.

La plataforma está diseñada para resguardar la integridad del *dron* durante las pruebas de vuelo. Consta de dos cables de acero recubierto, sujetados en los extremos laterales a una placa de aluminio central donde se resguarda el *frame* o estructura principal. En la parte inferior, se ha colocado un amortiguador para reducir el impacto del robot en una posible caída, y una abrazadera tipo **U**, unida con un destorcedor y una bandola, permitiendo que la plataforma cuente con un grado de libertad rotacional ψ, respecto al eje vertical.

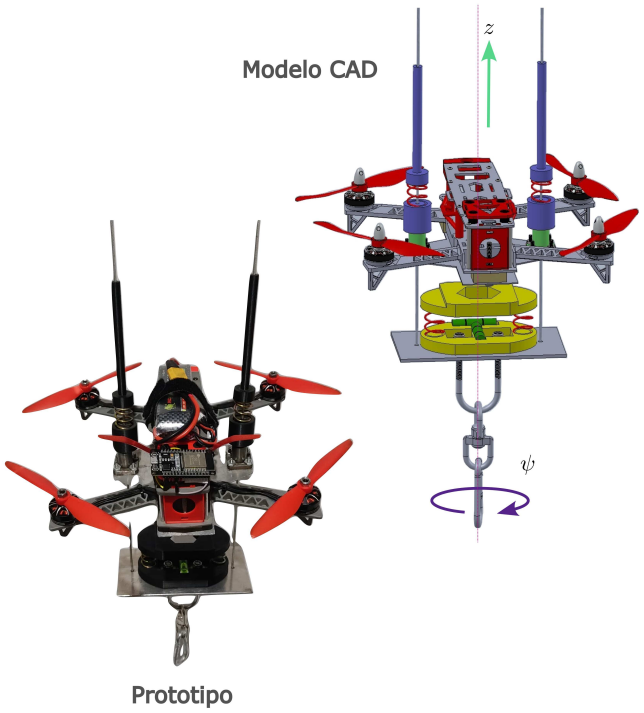


Figura 6. Sistema ρSus con dos grados de libertad.

El sistema toma una fotografía mediante una cámara digital, la cual puede alojarse a bordo de la plataforma. Posteriormente, la fotografía es procesada por el algoritmo de reconocimiento basado en *machine learning*, el cual fue previamente entrenado con una gran cantidad de datos. De manera interna, una red neuronal convolucional, de la clase YoloV5 [10] instalada en una computadora, determina cual es el comando que se le ha asignado a la plataforma, utilizando gestos o señas de las manos, por ejemplo: bajar, subir, girar a la derecha o a la izquierda, etc. La plataforma recibe órdenes desde la unidad central a través de un módulo Wi-Fi/Bluetooth y los cuatro rotores se mueven de acuerdo con el comando establecido por el sistema de reconocimiento, como se observa en la Figura 7.

Reflexión

La navegación autónoma, la interacción hombre-máquina, la manipulación precisa de objetos y la inspección automatizada son algunos ejemplos de áreas de oportunidad para la visión artificial en la robótica, con el objetivo de mejorar la eficiencia y la autonomía en varias aplicaciones tanto industriales como agrícolas, por lo que es imperativo ampliar y redoblar los esfuerzos en tareas de investigación y desarrollo tecnológico en este contexto.✷

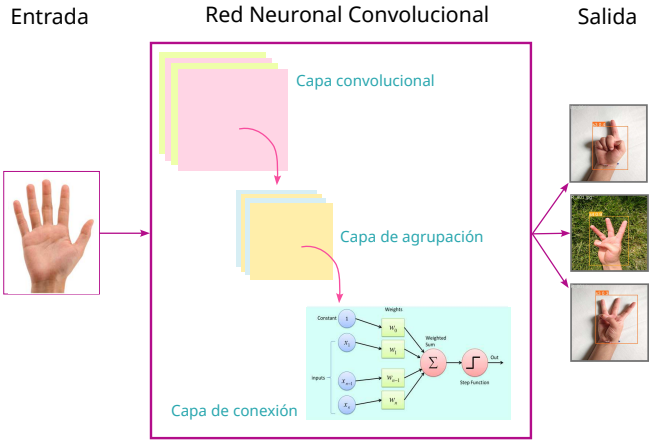


Figura 7. Reconocimiento de gestos del sistema ρSus.

REFERENCIAS

1. Nachtigall, W. y Wisser, A. (2015). Anthropo and Biomedical Technology. En Nachtigall W. y Wisser A. (Eds.) *Bionics by Examples*. Cham, Springer, 171 - 183.

2. Davies, E.R. (2018). Computer Vision: Principles, Algorithms, Applications, Learning. Quinta edición. Cambridge, Academic Press, ISBN 978-0-128-09284-2.

3. Person, J. (2021). Top benefits of integrated robotic machine vision. Recuperado el 15 de agosto de 2024, de `https://www.fanucamerica.com/docs/default-source/articles/fabshop-benefits-of-vision-by-josh-person.pdf`.

4. Demaitre, E. (2024). Sevensense acquisition to add VSLAM smarts to mobile robots, says ABB. Recuperado el 15 de agosto de 2024, de `https://www.therobotreport.com/sevensense-acquisition-adds-vslam-smarts-mobile-robots-says-abb/`.

5. Klette, R. (2014). Concise Computer Vision: An Introduction into Theory and Algorithms. London, Springer-Verlag London, ISBN 978-1-4471-6319-0

6. Davies, E. (2022). The dramatically changing face of computer vision. En Davies E.R. y Turk M. A. (Eds.) *Advanced Methods and Deep Learning in Computer Vision*. Cambridge, Academic Press, 1-91, ISBN 978-0-128-22109-9.

7. Corke. P., Jachimczyk, W. y Pillat, R. (2023). Robotics, Vision and Control: Fundamental Algorithms in MATLAB. Tercera edición. Cham, Springer International Publishing, ISBN 978-3-031-07261-1.

8. Christensen, H. (2016). Part C Sensing and Perception. En Siciliano B. y Khatib, O. (Eds.) *Springer Handbook of Robotics*. Cham, Springer. 713-892.

9. Siegwart, R., Nourbakhsh, I.. y Scaramuzza, D. (2011). Introduction to Autonomous Mobile Robots. Segunda edición. Cambridge, The MIT Press, ISBN 978-0-262-01535-6.

10. Redmon, J., Divvala, S., Girshick, R. y Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection". En *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, EEUU, 779-788.

SOBRE LOS AUTORES



Rodrigo Juarez-Villalva es un estudiante de último semestre de Ingeniería en Mecatrónica del Tecnológico de Estudios Superiores de Tianguistenco, con experiencia en programación y desarrollo de robots manipuladores. Se ha destacado en la aplicación y desarrollo de sistemas mecatrónicos, tiene un particular interés en el aprendizaje automático y las nuevas tecnologías emergentes en el campo de la robótica. Su trabajo de investigación se centra en desarrollar soluciones innovadoras para mejorar la funcionalidad de sistemas complejos con soluciones óptimas.



Efren Castro-Bobadilla es un estudiante que cursa el último semestre de Ingeniería en Mecatrónica en el Tecnológico de Estudios Superiores de Tianguistenco. Cuenta con experiencia en la instalación, mantenimiento y gestión de instalaciones eléctricas industriales. Sus principales intereses son el desarrollo de soluciones mecatrónicas avanzadas para optimizar la funcionalidad de sistemas complejos, así como en las tecnologías de la I4.0, como el aprendizaje automático, la visión artificial y los robots colaborativos.



Hugo Yañez-Badillo es profesor investigador miembro del grupo de investigación Sistemas Mecatrónicos para la I4.0 en el TecNM: Tecnológico de Estudios Superiores de Tianguistenco. Dentro de las áreas de interés en investigación y desarrollo tecnológico destacan el diseño e implementación experimental y mediante simulación de controladores automáticos, sistemas embebidos para sistemas mecatrónicos, estimación de parámetros, sistemas autónomos, sistemas ciber-físicos y algoritmos inteligentes con aplicaciones en sistemas robóticos. Es miembro del Sistema Nacional de Investigadores SNII del Conahcyt, nivel 1.



Laura Cleofas-Sanchez es doctora en Ciencias de la Computación por el Centro de Investigación en Computación del IPN de México. Sus áreas de interés son ciencia de datos, inteligencia artificial, minería de datos, redes neuronales, aprendizaje profundo, ciudades inteligentes, internet de las cosas, mezcla de expertos, algoritmos genéticos, métodos de agrupamiento, aprendizaje automático, reconocimiento de patrones y visión computacional. Actualmente es profesora investigadora miembro del grupo de investigación Sistemas Mecatrónicos para la I4.0 del TecNM: Tecnológico de Estudios Superiores de Santiago Tianguistenco. Ha sido miembro del Sistema Nacional de Investigadores SNII del Conahcyt.



J. Nabor Balderas-Gutierrez es profesor investigador miembro del grupo de investigación Sistemas Mecatrónicos para la I4.0 del TecNM: Tecnológico de Estudios Superiores de Santiago Tianguistenco. Sus áreas de interés son las aplicaciones ambientales de materiales y la electrónica de potencia aplicada para el desarrollo de sistemas mecatrónicos, para la prevención y control de la contaminación del agua. Además, ha demostrando un fuerte compromiso con la mejora del entorno ambiental a través de proyectos de investigación relacionados con el control de contaminantes. Es miembro del Sistema Nacional de Investigadores SNII del Conahcyt, nivel Candidato.



Francisco Beltran-Carbajal es doctor en Ciencias por el Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, CINVESTAV. Actualmente es profesor Titular del Departamento de Energía, en la Universidad Autónoma Metropolitana, Unidad Azcapotzalco. Sus principales intereses de investigación incluyen el control de vibraciones, identificación de sistemas, maquinaria rotativa, mecatrónica y control automático de sistemas de conversión de energía. Es miembro del Sistema Nacional de Investigadores SNII del Conahcyt, nivel 2.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Detección de enfermedades foliares en maíz mediante aprendizaje profundo

Víctor A. Gonzalez Huitrón, Jesús A. Medrano Hermosillo, Abraham E. Rodríguez Mata, José A. Almaráz Damián y Pedro Celestino Ramírez

El maíz, originario de México, es el cultivo que genera más valor económico al país.

Representa la esencia de una cultura milenaria que ha moldeado la identidad del país desde tiempos prehispánicos, siendo un componente fundamental de la dieta mexicana, utilizado en una variedad de platillos que forman parte de la gastronomía nacional y está presente en prácticamente toda la sociedad mexicana. El cultivo de maíz desempeña un papel crucial en la economía agrícola de México, donde se cultivan diversas variedades adaptadas a las condiciones climáticas y geográficas de las distintas regiones.

Sin embargo, la producción de maíz enfrenta múltiples desafíos, entre los cuales las enfermedades foliares son uno de los más preocupantes. Estas enfermedades, causadas por hongos, bacterias y/o virus, pueden reducir drásticamente el rendimiento de las cosechas, comprometiendo no solo la producción de alimentos, sino también la subsistencia de muchas familias que dependen del cultivo. La detección temprana de estas enfermedades es esencial para implementar medidas de control efectivas y mitigar sus efectos. Un enfoque eficiente en la identificación de síntomas puede marcar la diferencia entre una cosecha exitosa y una pérdida significativa de producción.

La importancia de detectar enfermedades en los cultivos de maíz radica en varios factores. En primer lugar, el maíz es un cultivo estratégico para la seguridad alimentaria en México, y cualquier disminución en su producción puede tener repercusiones graves en la disponibilidad de alimentos y en los precios en el mercado. Además, la reducción de pérdidas por enfermedades puede contribuir a mejorar los ingresos de los agricultores, lo que es crucial para el desarrollo económico de las comunidades rurales.

Bajo esta premisa, la tecnología y la innovación juegan un papel fundamental. La inteligencia artificial (IA) se ha convertido en una herramienta prometedora para la detección de enfermedades foliares en diversos cultivos y hortalizas. A través de técnicas de aprendizaje automático y procesamiento digital de imágenes, la visión artificial puede identificar patrones y anomalías en las plantas que podrían pasar desapercibidos para el ojo humano. Los sistemas basados en visión artificial pueden analizar grandes cantidades de datos, procesando información de sensores en cámaras de alta resolución para detectar síntomas de enfermedades en etapas tempranas. Esto permite actuar rápidamente, aplicando tratamientos específicos y minimizando el uso de pesticidas, lo que se traduce en una agricultura de menor impacto a la biodiversidad.

Asimismo, el uso de visión artificial no solo mejora la detección de enfermedades, sino que también optimiza la gestión de cultivos en general. Los algoritmos pueden predecir brotes de enfermedades basándose en condiciones climáticas y otros factores, permitiendo a los agricultores implementar prácticas de manejo integrado de plagas y enfermedades. Esta sinergia entre la tradición agrícola y la innovación tecnológica puede transformar la producción de maíz en México, asegurando su continuidad y preservando nuestra herencia cultural.

En este artículo, exploraremos con mayor detenimiento dos enfoques para el diseño de clasificadores de imágenes mediante aprendizaje profundo, mostrando un análisis comparativo a partir de los resultados obtenidos con diversas arquitecturas basadas en redes neuronales convolucionales y Vision Transformers (CNN y ViT respectivamente, por sus siglas en inglés). Incluyendo resultados de entrenamiento y evaluación para dichos modelos.

Visión artificial y aprendizaje profundo

La visión artificial es una rama de la inteligencia artificial que se enfoca en la capacidad de las computadoras para interpretar y comprender imágenes o vídeos de manera similar a como lo hace el ser humano. Este campo utiliza algoritmos y modelos matemáticos para analizar la información visual y extraer patrones significativos. En el contexto agrícola, la visión artificial se ha vuelto una herramienta poderosa para la detección temprana de enfermedades en cultivos.

Para clasificar imágenes y detectar enfermedades foliares, se utilizan modelos de aprendizaje profundo, el cual es un área del aprendizaje automático que se basa en redes neuronales. Estos modelos son capaces de aprender a partir de grandes cantidades de datos, extrayendo características relevantes de las imágenes sin necesidad de intervención humana. En el caso de las enfermedades foliares en el maíz, estos modelos pueden ser entrenados con imágenes de hojas sanas y enfermas, permitiendo que el sistema reconozca patrones asociados a diferentes enfermedades. La capacidad de identificar síntomas

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

de manera precisa en el menor tiempo posible es fundamental para una gestión efectiva de cultivos y aplicar tratamientos preventivos o correctivos.

Las CNN es uno de los enfoques más utilizados en el aprendizaje profundo para clasificación de imágenes. Las CNN están diseñadas para procesar datos con una estructura de vecinos cercanos por cada pixel (px) o valor en una imagen. Utilizan capas convolucionales que aplican filtros a la imagen para con la finalidad de detectar características como bordes, texturas y formas. A medida que la información fluye a través de estas capas, una CNN puede aprender representaciones jerárquicas, desde características simples (p.e. formas, colores, o texturas en las capas inferiores) hasta patrones más complejos (p.e. características específicas de una clase en las capas superiores). Esto las convierte en una opción competitiva para tareas de clasificación y detección de enfermedades en cultivos.

Por otro lado, un enfoque más reciente en el campo de la visión artificial son los ViT. A diferencia de las CNN, que se basan en la convolución, un Vision Transformer utiliza mecanismos de atención para procesar las imágenes. Este modelo divide la imagen en parches (comúnmente de 64px) y los trata como una secuencia, aplicando atención a cada uno de ellos para aprender las relaciones entre diferentes partes de la imagen. En este contexto podemos entender por *atención* un proceso a través del cual la red neuronal dará mayor ponderación a partes relevantes de la imagen de acuerdo a su etiqueta, agrupando las partes de la imagen que son similares para facilitar la decisión. Esta metodología ha demostrado ser muy eficaz en tareas de clasificación de imágenes, ofreciendo una alternativa en comparación a las CNN.

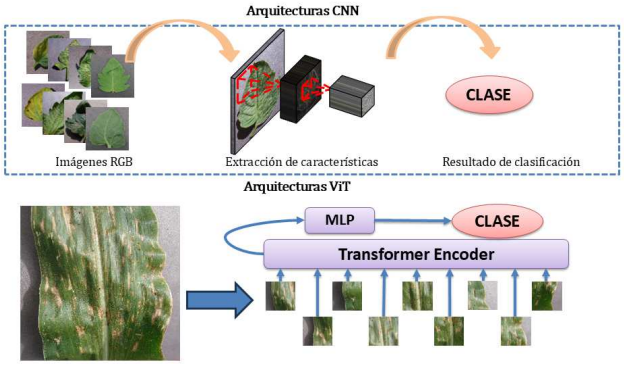


Figura 1. Diagrama de flujo para CNN y ViT.

Cabe destacar que la principal diferencia entre las CNN y los ViT radica en su arquitectura y en cómo procesan la información visual. Las CNN son eficientes para captar patrones locales, mientras que los ViT sobresalen en la comprensión de relaciones globales dentro de la imagen, lo que les permite a estos últimos considerar interacciones entre diferentes áreas de la misma. Ambos enfoques tienen sus cualidades propias, y la elección entre ellos dependerá del tipo de datos, la tarea específica a realizar y los requisitos de implementación. En la Fig. 1 brevemente se muestra la diferencia en el tratamiento de una imagen.

Antecedentes

En trabajos previos se ha establecido que la visión artificial puede brindar resultados satisfactorios; en [1], por ejemplo, se entrenó el dataset *PlantVillage* mediante arquitecturas de CNN de propósito general. El dataset *PlantVillage*, contiene imágenes etiquetadas de diferentes enfermedades foliares, permitiendo a los autores entrenar modelos de CNN con un alto grado de precisión. El estudio demostró que las CNN pueden alcanzar una clasificación precisa de enfermedades con un rendimiento comparable al de expertos humanos, estableciendo un precedente importante.

Otro ejemplo es el trabajo de Gonzalez-Huitron et al. [2] donde se plantea el entrenamiento para una sola hortaliza a partir de un subconjunto de *PlantVillage* enfocado a tomates. Cabe destacar que se establece la implementación en un sistema de bajo consumo y portátil como lo es una Raspberry Pi 4, sin afectar de manera relevante la calidad de los resultados.

Moonwar et al. [4] muestran en su trabajo el entrenamiento realizado con modelos y arquitecturas más recientes como lo son ViT; dicho entrenamiento ha logrado sobrepasar la calidad obtenida por las CNN tradicionales.

Este artículo centra su estudio en la selección de subconjunto de imágenes de maíz de *PlantVillage*. El propósito es establecer un análisis comparativo entre CNN y ViT a partir de resultados observados, a fin de identificar las condiciones generales que guíen en la selección de la mejor de estas técnicas para una aplicación enfocada en la agricultura de precisión. Las cuatro clases del subconjunto son:

- **Cercospora**. Enfermedad foliar causada por el hongo Cercospora zeae-maydis. Afecta principalmente a las hojas de la planta de maíz provocando una reducción significativa en el rendimiento del cultivo.
- **Corn Rust**. Esta enfermedad es causada por varios hongos del género Puccinia, siendo los más comunes Puccinia sorghi y Puccinia polysora. Estos hongos producen pústulas de color naranja a marrón en las hojas del maíz.
- **Northern Leaf Blight**. Es causada por diferentes hongos. Las formas más comunes de Corn Leaf Blight son provocadas por los patógenos Exserohilum turcicum (anteriormente conocido como Hel-

minthosporium turcicum) y Bipolaris maydis, este último es el causante de la enfermedad Northern Leaf Blight específicamente.

El conjunto de datos se compone de 3,232 imágenes RGB en total. De éste, un subconjunto formado por el 70 % de las imágenes fue utilizado para entrenamiento de un clasificador multiclase con 4 salidas (usando 10 % de este subconjunto para validación). El restante 30 % de imágenes fue empleado para la fase de pruebas. En la Fig. 2 se observa una muestra de las imágenes RGB.

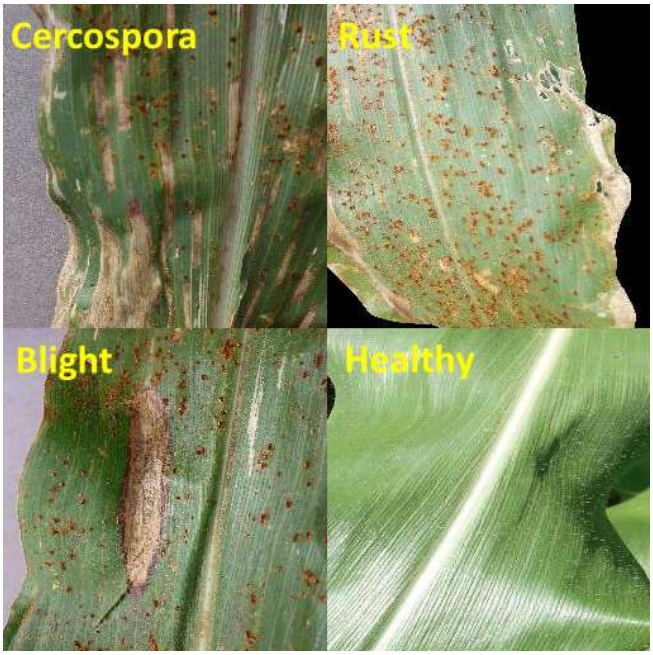


Figura 2. Ejemplo de imágenes RGB en el conjunto de datos.

Los modelos resultantes fueron evaluados con las siguientes medidas de calidad:

- **Accuracy**. Exactitud del modelo, indica el número total de instancias correctamente clasificadas en relación al total de elementos. Podemos interpretarlo como un porcentaje del total de aciertos del modelo, su valor va de 0 a 1.
- **Precision** De manera práctica nos indica la cantidad de errores del tipo Falso Positivo en un clasificador, entre mayor sea su valor, menor es el número de falsos positivos del modelo.
- **Recall**. Indica la cantidad de errores del tipo Falso Negativo, entre más cercano a 1 sea su valor más baja es la cantidad de falsos negativos obtenida.
- **F-Score**. Es el promedio de Precision y Recall, su valor entre más alto sea indica menor presencia de errores, sin embargo, si su valor es menor al de *Accuracy*, significa que el balance entre falsos positivos y falsos negativos presenta una desproporción.

Aportaciones

Se realizó el entrenamiento durante 20 épocas para 6 arquitecturas, 3 enfocadas a CNN, que son GoogleNET, MobileNETv3 y ResNET50, y 3 arquitecturas de ViT, ViT-b-16, ViT-b-32 y Swin-Transfomer. El criterio para la selección de las arquitecturas fue de acuerdo al total de sus parámetros, teniendo modelos relativamente pequeños, medianos y grandes.

Como medidas de calidad se establecieron *Accuracy, Precision, Recall y F-Score* los cuales son de utilidad para una correcta interpretación de los resultados.

En la Fig. 3 se observan resultados con GradCam (Gradient-weighted Class Activation Mapping). Esta técnica ayuda a visualizar y entender qué partes de una imagen contribuyen más a la decisión del modelo, utiliza los gradientes de la capa final de una red con respecto a una clase determinada. Ésto permite identificar qué neuronas de esa capa tienen más impacto en la predicción de la clase seleccionada.

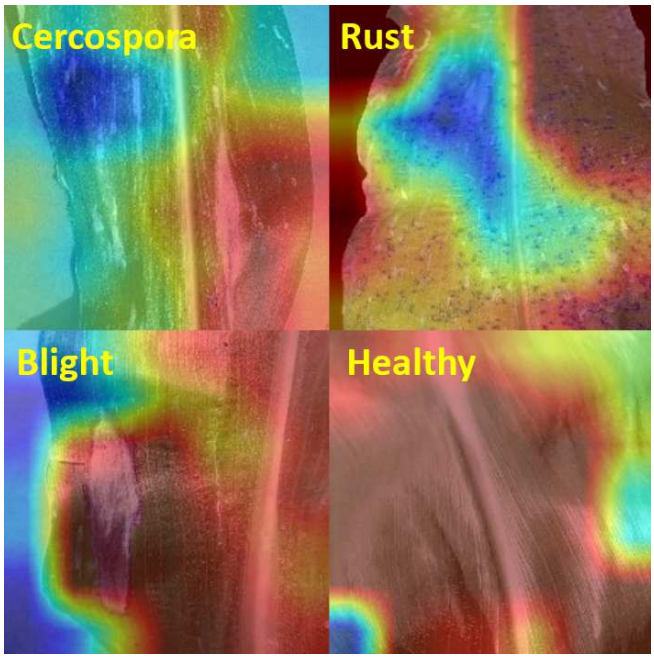


Figura 3. Ejemplo de explicabilidad con ResNet50 mediante GradCam.

En [9] se presentan el proyecto bajo el lenguaje Python y la librería PyTorch para mayor facilidad en la reproducción de los resultados.

Es relevante mencionar que el valor de Accuracy durante el entrenamiento de las arquitecturas fue notoriamente mejor con ViT, ya que se alcanzaban valores más cercanos a 1 en una menor cantidad de épocas. Lo anterior es razonable de acuerdo al trabajo en [3] en donde se establece que estas arquitecturas logran resultados comparables o mejores a CNN con menor cantidad de datos y épocas. Sin embargo el tiempo de entrenamiento de

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

CNN fue notoriamente menor (desde 25 hasta 40 minutos) respecto a ViT (entre 40 hasta 100 min).

Respecto a las medidas de calidad, los valores de *precision* y *recall* por clase, para los datos de prueba, se muestran en la Tabla 1.

Tabla 1. Medidas de calidad en los modelos entrenados.

Modelo	Clase	Precision	Recall	F-score
GoogleNet	Cercos	0.915	0.903	0.909
	Rust	0.994	1.00	0.997
	Blight	0.945	0.952	0.949
	Healthy	1.00	0.994	0.997
ResNet50	Cercos	0.870	0.957	0.911
	Rust	1.00	0.997	0.998
	Blight	0.979	0.935	0.957
	Healthy	0.994	0.997	0.995
MobileNet	Cercos	0.889	0.801	0.843
	Rust	0.868	0.996	0.928
	Blight	0.915	0.940	0.928
	Healthy	1.00	0.904	0.949
ViT-b-32	Cercos	0.915	0.881	0.898
	Rust	1.00	0.997	0.998
	Blight	0.935	0.958	0.947
	Healthy	1.00	1.00	1.00
ViT-b-16	Cercos	0.902	0.972	0.936
	Rust	1.00	1.00	1.00
	Blight	0.986	0.954	0.970
	Healthy	1.00	0.997	0.998
Swin-V2	Cercos	0.883	0.951	0.915
	Rust	1.00	0.997	0.998
	Blight	0.976	0.944	0.960
	Healthy	1.00	1.00	1.00

Para interpretar los resultados de la tabla, es necesario considerar la aplicación y el costo o importancia por cada tipo de error, ya que un Falso Positivo (*Precision*) indicaría que una imagen se ha clasificado como enferma cuando no lo estaba realmente; este tipo de error conlleva la decisión de aplicar medidas preventivas o correctivas en la presencia de alguna anomalía. Por otro lado, un Falso Negativo (*Recall*) nos indicaría que una imagen ha sido clasificada como saludable aun cuando presentaba alguna anomalía. Es de suma importancia la correcta interpretación de los posibles errores de un modelo para determinar sus limitaciones y casos de uso. Tomando lo anterior en cuenta, y observando los resultados de la tabla, en el caso de tener resultados en *Accuracy* similares, el valor de *Recall* es crítico, para que en caso de que el clasificador presente un error, este tenga mayor probabilidad de ser un Falso Positivo. Los valores de *Recall* más elevados especialmente en la clase *Healthy* se observan en las arquitecturas ViT.

Conclusiones

En este breve análisis se presentó la comparativa entre arquitecturas CNN y ViT, detallando los valores de Precison-Recall obtenidos y mostrando las diferencias en el entrenamiento de ambas arquitecturas. De acuerdo a los resultados presentados, las arquitecturas ViT son un candidato ideal cuando se desean modelos confiables; sin embargo, se debe considerar que su uso requiere largos tiempos de entrenamiento y hardware con buena capacidad de cómputo. En contraste, las arquitecturas CNN sacrifican ligeramente el rendimiento a favor de tiempos de entrenamiento relativamente cortos, de su implementación en dispositivos de bajo consumo energético o con especificaciones de hardware reducidas.

Como trabajo a futuro, resta explorar si es posible calibrar la salida de estas arquitecturas para obtener una relación Precision-Recall mas acorde a las necesidades del sector, mediante la obtención y análisis de las curvas Precision-Recall. Además, se propone indagar en tácticas de ajuste de los clasificadores con el fin de armonizar de manera más efectiva la relación entre precisión y recuperación, mejorando así su desempeño en el ámbito agrícola. Cabe destacar que el dataset no fue generado con cultivos en el país, lo cual plantea la necesidad de crear estos conjuntos de datos para el diseño de soluciones con mayor pertinencia en el área.✷

REFERENCIAS

1. Mohanty, S. P., Hughes, D. P. y Salathé, M. (2016). Using Deep Learning for Image-Based Plant Disease Detection. *Frontiers in Plant Science*, 7.
2. Gonzalez-Huitron, V., León-Borges, J. A., Rodriguez-Mata, A. E., Amabilis-Sosa, L. E., Ramírez-Pereda, B. y Rodriguez, H. (2021). Disease detection in tomato leaves via CNN with lightweight architectures implemented in Raspberry Pi 4. *Computers and Electronics in Agriculture*, 181, 105951.
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Others. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv Preprint arXiv:2010. 11929.
4. Moonwar, W., Hasan Mahin, M. R., Nadia, Y., Kibria, F. B., Biswas, A., Chowdhury, F. I. y Rasel, A. A. (2023). Tomato Leaf Disease Classification with Vision Transformer Variants. En *H. Lu, M. Blumenstein, S.-B. Cho, C.-L. Liu, Y. Yagi, & T. Kamiya (Eds.), Pattern Recognition*, 95–107.
5. Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. CoRR, abs/2010.16061. Retrieved from https://arxiv.org/abs/2010.16061
6. Li, Z., Liu, F., Yang, W., Peng, S. y Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12), 6999–7019.
7. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M. y Adam, H. (2019). Searching for MobileNetV3. En *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z. y Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. En *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.
9. Gonzalez-Huitron, V. (2024). VitAndCNNKorn (Version 1.0). Retrieved from `https://github.com/alx3416/ViT-Korn`.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

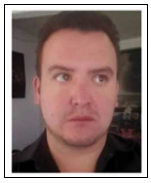
SOBRE LOS AUTORES



Víctor A. González Huitrón Obtuvo el grado de Doctor en Comunicaciones y Electrónica en 2017 por parte del Instituto Politécnico Nacional. Actualmente es miembro del Sistema Nacional de Investigadores y se desempeña como profesor e investigador en el Instituto Tecnológico de Querétaro. Sus intereses de investigación son: Visión por computadora, Inteligencia Artificial y ciencia de datos.



Jesús A. Medrano Hermosillo Profesor investigador del TecNM campus Chihuahua colaborando en la línea de procesamiento de señales eléctricas y electrónicas, específicamente en las áreas de optimización de sistemas inteligentes, redes neuronales de alto orden, convergencia de sistemas dinámicos y control (lineal, no lineal, óptimo). Ha participado en proyectos que involucran sistemas de potencia, robots móviles, quadrotors, electro movilidad e inteligencia artificial.



Abraham E. Rodríguez Mata Investigador SNI nivel 1. Ingeniero Químico experto en el control y modelado de sistemas biotecnológicos y sistemas no lineales incluyendo sistemas electromecánicos. Sus intereses de investigación incluyen control y diseño de observadores y modelos de aprendizaje profundo para sistemas biotecnológicos, mecatrónicos, robótica optimización y control de procesos.



José A. Almaráz Damián Obtuvo su doctorado en Comunicaciones y Electrónica, y su grado de maestría en Ingeniería en Microelectrónica e Ingeniería en Computación por el Instituto Politécnico Nacional (IPN) de México en 2022, 2016 y 2012. Actualmente es Investigador Asociado en el CICESE-UT 3. Sus intereses de investigación incluyen Procesamiento Médico de Señales/Imágenes, Reconocimiento de Patrones, Aprendizaje Automático y Sistemas Embebidos.



Pedro Celestino Ramírez Obtuvo su doctorado en Administración y Gestión Empresarial por la Universidad del Centro del Bajío de Celaya, Gto, y su grado de maestría en Docencia por la Universidad Marista de México e Ingeniería en Sistemas Computacionales por la Universidad Autónoma de Coahuila en 2022, 2008 y 1988. Actualmente es docente de tiempo completo del Tecnológico Nacional de México campus Querétaro. Sus intereses de investigación incluyen investigación educativa e ingeniería de software.

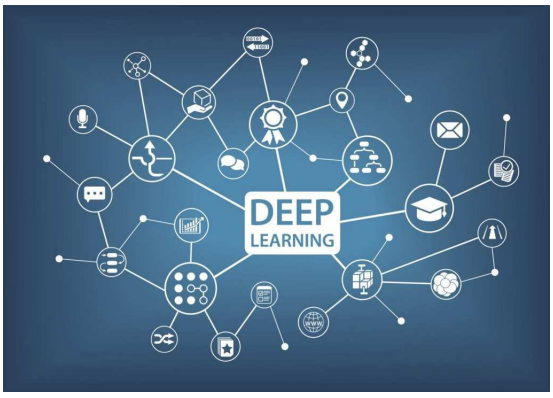


Imagen tomada de:
https://www.initiumsoft.com/blog_initium/aprendizaje-profundo/.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO INVITADO

Detección de residuos sólidos urbanos reciclables con Visión Artificial

Josué Garatachia Mejía, Erubiel Urbina Gómez, Hector A. Montes Venegas, Rosa María Valdovinos Rosas y Marcelo Romero Huertas

La contaminación por basura es uno de los problemas ambientales más serios y crecientes de nuestro tiempo [1]. Este problema se debe a factores como el aumento de la población, el crecimiento urbano y los patrones de consumo de las personas. El incremento constante en la generación de residuos y su inadecuada gestión han llevado a una situación crítica, donde la acumulación de desechos amenaza los ecosistemas y la salud pública.

Una parte importante de la gestión de estos residuos involucra la separación y el posterior reciclaje de los materiales seleccionados. Reciclar estos materiales de manera adecuada generaría beneficios ambientales y económicos significativos al reducir su incineración o vertimiento; adicionalmente, puede disminuir la necesidad de nuevas materias primas y los costos generales de producción asociados.

El proceso de selección y separación de los residuos para su reciclaje varía significativamente en distintas partes del mundo. Las fuentes de influencia de esta variabilidad incluyen los intereses económicos, factores culturales y hábitos de consumo, las necesidades específicas de cada región y las políticas gubernamentales relacionadas con la gestión de residuos [2].

En consecuencia, las tasas de reciclaje varían considerablemente según el país, el tipo de material y su aplicación o valor comercial. A pesar de que existen plantas de selección y separación de residuos en muchas ciudades alrededor del mundo, la tasa de reciclaje mundial es apenas del 9 % [3]. En la Unión Europea en 2020 la tasa de reciclaje fue de 46 %, siendo los envases (64 %), residuos urbanos (49 %) y de residuos electrónicos (39 %) [3].

En México, se generan más de 42 millones de toneladas de residuos sólidos al año, lo que equivale a 175 veces el volumen de la pirámide del Sol de Teotihuacán. Sin embargo, el país tiene una de las tasas de reciclaje general más bajas del mundo con un 6 % [3]. En contraste, en 2023 México recuperó el 63 % de los envases de PET utilizados, principalmente debido a su valor comercial y a las campañas llevadas a cabo por organizaciones como ECOCE que cuenta con 30 plantas recicladoras de PET, además de la planta más grande del continente americano, cuya meta para el 2025 es llegar a una tasa del 70 % y del 80 % para el 2030 [1].

Otro factor que influye en la selección y separación de los residuos es la disponibilidad de tecnología y la infraestructura adecuada para la gestión de residuos. A través de los años los investigadores han explorado diferentes técnicas de separación automática para mejorar la eficiencia general del proceso de reciclaje. La separación se realiza mayoritariamente utilizando métodos mecánicos, sensores y actuadores diversos a lo largo del proceso para detectar los materiales reciclables.

En México existen varias plantas de selección y separación de residuos sólidos urbanos. Una de estas plantas fue inaugurada en la Ciudad de México a finales del 2021 y otra más a finales del 2023. Estas plantas cuentan con un sistema de bandas de separación manual de materiales reciclables donde se segregan mecánicamente residuos orgánicos e inorgánicos, así como diversos materiales reciclables de valor comercial (Figura 1).

La selección manual de residuos reciclables es un proceso lento y costoso, implica complejidades logísticas importantes, además de estar sujeto a errores humanos y riesgos para la salud de los trabajadores al exponerlos continuamente a materiales riesgosos. Estas dificultades resaltan la necesidad de automatizar la selección y separación para mejorar la eficiencia, reducir costos, incrementar la producción y garantizar un proceso más seguro y preciso.



Figura 1. Línea de clasificación manual en la estación de transferencia y selección de residuos de Azcapotzalco en la CDMX[1].

[1] Fuente: https://escuanto.mx/es-cuanto-visita-la-estacion-de-transferencia-y-seleccion-de-residuos-azcapotzalco-que-trae-beneficios-para-cdmx/

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

En los últimos años, se han desarrollado diversos enfoques tecnológicos para abordar la detección de residuos. Entre estos, el uso de técnicas de Inteligencia Artificial ha demostrado ser particularmente efectivas. Estudios recientes han aplicado *Redes Neuronales Convolucionales* para la detección de residuos reciclables como papel aluminio, chatarra de metal, botellas de plástico transparente, tapones de botellas de plástico, entre otros; obteniendo resultados con una precisión de hasta 95 % [4]. Este resultado se debe en gran medida a que los objetos a detectar estaban completamente visibles (pocas oclusiones u obstrucciones) con condiciones de iluminación y escala homogénea de los objetos. Otro estudio para detección de residuos en plantas de reciclaje arroja una precisión promedio del 53 % en su simulación experimental [5]. Los objetos reciclabas presentes en las imágenes se ven afectados en su mayoría por la oclusión de los objetos reciclables de interés.

De manera similar, se han utilizado otros métodos de *Aprendizaje Automático* [6] para detectar únicamente residuos plásticos en imágenes tomadas en entornos urbanos comunes sin variabilidad en escala, forma y fondos con texturas naturales, alcanzando una precisión del 95 % [7].

La segmentación de imágenes con arquitecturas como *U-Net* para separar plásticos del fondo de las imágenes reporta una precisión de hasta un 92 % [8]. Sin embargo, las imágenes utilizadas no representan los retos reales de objetos encontrados en entornos comunes urbanos o en las plantas de selección y separación de residuos sólidos.

Visión artificial para detección de residuos en entornos urbanos

Los métodos de Visión Artificial (VA) son algoritmos ampliamente estudiados que pueden encontrarse implementados en herramientas de libre acceso. El estudio aquí presentado se centra en la detección de residuos mediante la utilización de algoritmos de detección de características y coincidencia de descriptores en imágenes. La implementación se lleva a cabo en Python utilizando la biblioteca *OpenCV* (`https://docs.opencv.org`). El objetivo principal es demostrar la eficiencia de estos métodos en un sistema capaz de detectar residuos reciclables de manera efectiva que pueda ser utilizado en plantas de selección y separación de residuos sólidos urbanos.

Imágenes por analizar

La experimentación utiliza el conjunto de imágenes con anotaciones confiables de acceso libre y gratuito llamado *Trash Annotations in Context* o TACO (`http://tacodataset.org`). Este conjunto contine un total de 1,500 imágenes con 4,784 objetos distribuidos en 60 clases o categorías. Los objetos visibles en cada imagen están etiquetados manualmente de acuerdo con una taxonomía jerárquica establecida por los autores originales. Las etiquetas determinan la localización en la imagen de cada objeto encerrándolo en un recuadro y asignándole una etiqueta de identificación, lo que permite asociar cada punto individual del objeto con la etiqueta asignada.

Del total de imágenes se seleccionaron 970 con 1,790 anotaciones en 23 clases. Estas imágenes corresponden a materiales reciclables, y fueron tomadas en diversos entornos para evaluar sistemas de detección de residuos sólidos urbanos.

Un proceso común en la detección de objetos en imágenes involucra: a) la detección de puntos de interés o puntos clave; b) la extracción de un vector de características en cada punto detectado; y, c) asociar mediante una búsqueda los vectores extraídos para cada objeto de interés dentro de las imágenes que se están analizando. Para tal efecto, es necesario separar los objetos de interés del fondo de las imágenes; posteriormente, se debe extraer una serie de muestras de los objetos que se desean localizar, tal como se muestra en las imágenes de la Figura 2.



Figura 2. Pre-procesamiento de imágenes.

Extracción de características

Para realizar la extracción de características inicialmente se utilizó el algoritmo ORB (*Oriented FAST and Rotated BRIEF*) para la detección de puntos clave y la generación de descriptores en las imágenes [9].

Primeramente, se aplica un detector de esquinas para determinar los puntos clave. Luego se calcula el centroide ponderado por intensidad del área con cada esquina detectada ubicándola en el centro, para después calcular la dirección desde cada esquina hasta el centroide y determinar así su orientación. El conjunto de las orientaciones en cada esquina forma el vector de características final.

ORB es un algoritmo rápido, robusto e invariante a la rotación que resulta útil para tareas como el reconocimiento de objetos y la reconstrucción en 3D. ORB utiliza valores binarios para representar las características de la imagen y se ve menos afectado por el ruido de la

imagen que otros algoritmos, además de que utiliza un mecanismo de compensación de orientación para hacerlo invariante de rotación.

No obstante, ORB no es tan efectivo como otros algoritmos para describir características en imágenes que cambian de ángulo y escala. Los descriptores obtenidos por ORB se comparan utilizando un método de coincidencia por búsqueda exhaustiva. Una muestra de los resultados en las imágenes se ver en la Figura 3.

La generación de residuos sólidos domésticos y su inadecuada gestión es un problema ambiental creciente a nivel mundial.

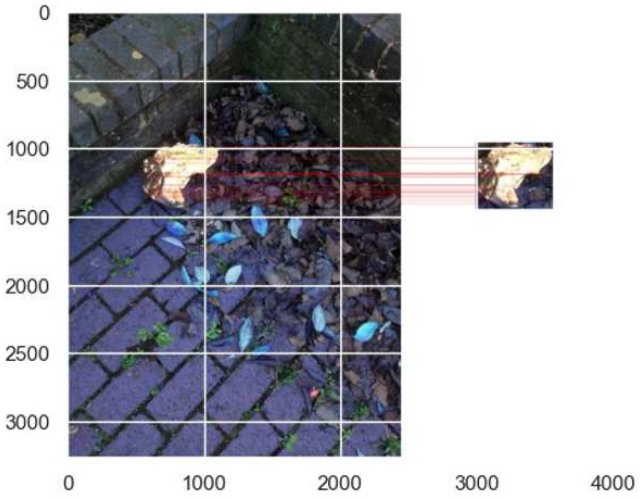


Figura 3. Obtención de los puntos y descriptores de características con el algoritmo ORB.

Adicionalmente, con el fin de examinar las ventajas comparativas de ambos métodos se utilizó el algoritmo SIFT (*Scale-Invariant Feature Transform*) [9], el cual sigue los siguientes pasos:

- *Diferencia de Gaussianas* (D_oG): Se aplica una serie de convoluciones Gaussianas a la imagen original para crear una pirámide de escala.
- *Localización de Puntos Clave*: Los puntos clave se identifican en los espacios de escala mediante la búsqueda de máximos y mínimos en el espacio D_oG.
- *Asignación de Orientación*: Cada punto clave se asigna una orientación basada en el gradiente local de la imagen, utilizando un histograma de 36 contenedores para los 360 grados.
- *Descriptores*: En cada punto clave, se extrae una vecindad de 16x16 puntos, dividida en bloques de 4x4. Para cada bloque se calculan los histogramas de gradientes locales, generando un vector de 128 valores para cada punto clave. Los objetos presentes en la imagen están representados por cada matriz de valores de (N*128), donde N depende de la imagen.

Coincidencia de descriptores

Con el fin de localizar los descriptores calculados para cada objeto en las imágenes examinadas, se utilizaron dos métodos para realizar la correspondencia de los descriptores, Búsqueda Exhaustiva para ORB y FLANN (*Fast Library for Approximate Nearest Neighbors*) [10] para SIFT. En la primera, los descriptores de ORB se comparan directamente uno a uno para encontrar coincidencias válidas. FLANN, por otro lado, contiene algoritmos optimizados para la búsqueda rápida de vecinos más cercanos en grandes conjuntos de datos de alta dimensionalidad. Por cada descriptor en SIFT, FLANN localiza en el conjunto de datos el descriptor que más se asemeje. En ambos casos, se debe superar el 70 % de coincidencias válidas para considerar que un objeto ha sido localizado.

Evaluación y comparación

La efectividad del sistema se evalúa comparando el número de descriptores válidos de SIFT y ORB. Debido a que tanto ORB como SIFT se pueden representar como descriptores binarios, es conveniente realizar su comparación utilizando la distancia de *Hamming*, la cual cuenta la cantidad de elementos diferentes para cadenas binarias.

Desempeño por Categoría

El estudio evaluó la efectividad del sistema de detección de residuos utilizando los algoritmos ORB y SIFT. Los resultados se categorizaron según el porcentaje de detección exitosa de descriptores en distintas categorías de residuos (ver Tabla 1).

En la Tabla 1 se puede observar que el número de imágenes por categoría es un factor crítico para la detección precisa de los residuos sólidos urbanos reciclables. Las categorías con un mayor número de imágenes (por ejemplo, *Corrugated Carton*) presentan mejores porcentajes de detección (53.6 %) respecto a aquellas menos representadas, como *Glass Jair* o *Six Pak rings*. Esto sugiere que aumentar el conjunto de datos por categoría mejoraría significativamente la eficacia del sistema.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Tabla 1. Resultados obtenidos.

	Categoría	ORB	SIFT	Coincidencias
1	Aluminium foil	30	13	36.7 %
2	Aluminium blister pack	4	2	25.0 %
3	Egg carton	6	3	33.3 %
4	Corrugated carton	28	13	53.6 %
5	Meal carton	18	9	16.7 %
6	Pizza box	2	1	50.0 %
7	Paper bag	16	7	18.8 %
8	Magazine paper	3	2	33.3 %
9	Normal paper	47	21	8.5 %
10	Wrapping paper	7	4	28.6 %
11	Toilet tube	2	1	0.0 %
12	Plastic film	217	93	11.1 %
13	Six pack rings	2	2	0.0 %
14	Paper straw	2	2	0.0 %
15	Glass bottle	51	23	19.6 %
16	Glass jar	2	2	0.0 %
17	Glass cup	2	2	50.0 %
18	Drink can	105	46	21.9 %
19	Metal bottle cap	39	17	5.3 %
20	Scrap metal	8	4	12.5 %
21	Clear plastic bottle	157	68	10.8 %
22	Plastic bottle cap	129	56	0.8 %
23	Plastic lid	48	21	4.2 %

Además, el umbral de aceptación del 30 % se muestra como una limitación del estudio, por lo que es recomendable reconsiderar este umbral para mejorar la confiabilidad de la detección.

SIFT mostró una variación en la eficacia dependiendo de la categoría y el número de imágenes disponibles. En general, logró identificar descriptores con un cierto nivel de precisión, observando que la eficacia se degradaba notablementeen categorías con menos imágenes.

En resumen, ambos algoritmos mostraron limitaciones significativas en la detección precisa de residuos urbanos, especialmente cuando se comparan con los resultados de estudios previos que utilizaron técnicas más avanzadas de aprendizaje profundo. En contraste, utilizando ORB y SIFT se logró un máximo de 53.6 % en la categoría de cartón corrugado, lo que abre una cantidad considerable de áreas de oportunidad frente a técnicas modernas de Aprendizaje Automático.

Conclusiones

Una medida de acción para enfrentar los problemas de contaminación ambiental es la conciencia de reutilización de residuos sólidos urbanos, de tal forma que lo que en un momento dado se considera "*basura*" pudiera tener una segunda vida. En este artículo se presentaron los resultados iniciales de una estrategia de detección de objetos en imágenes. Los resultados obtenidos en los experimentos de demostración realizados para este artículo muestran que, aunque los algoritmos convencionales de Visión Artificial como ORB y SIFT pueden ser utilizados para la detección de residuos, su efectividad se ve limitada cuando se comparan con métodos modernos de Aprendizaje Automático.

Las líneas abiertas de estudio apuntan a contar con conjuntos de datos mucho mayores que hayan sido adquiridos tanto en entornos urbanos como en plantas de selección y separación de residuos sólidos para todos los objetos posibles de ser reciclados. Lo anterior es necesario dado que la disponibilidad de datos tiene un impacto directo en la eficacia en el desempeño de modelos robustos de Aprendizaje Automático. Estas mejoras son esenciales para avanzar hacia un sistema automatizado de gestión de residuos más eficiente y contribuir a la reducción de la contaminación ambiental.✷

REFERENCIAS

1. ECOCE (2022). México: Cada vez más responsable. Reporte de actividades, 20 años de ECOCE. www.ecoce.mx.
2. Subramaniam, V., et. al. (2020). Waste classification for effective recycling. *Journal of Materials Research and Technology*,9(1), 1206-1215.
3. Paudel, S., Raj Pokharel, B. y Manandhar, B. (2019). A review on sources, composition, and characteristics of household solid waste in developing countries. *Journal of Environmental and Public Health*.
4. Moñus, M. G. (2022). Aplicación de técnicas de Aprendizaje Profundo en imágenes para el reconocimiento de objetos. Universidad Complutense de Madrid, I, 212.
5. Mayorga, M. S. (2020). Implementación de un sistema de detección de residuos reciclables basado en visión computacional. Repositorio académico de la Universidad de Chile, I, 66.
6. Rai, R., Tiwari, M. K., Ivanov, D. y Dolgui, A. (2021). Machine Learning in Manufacturing and Industry 4.0 Applications. *International Journal of Production Research*, 59(16), 4773–78.
7. Revrón, L. L. (2021). Detección de residuos mediante redes neuronales en entornos de costa y playa. Universidad de las Palmas, Gran Canaria.
8. Romero, S. M. (2023). Sistema de clasificación de residuos mediante redes convolucionales. Universidad Politécnica de Catalunya, I, 51.
9. Szeliski, R., (2022). Computer Vision: Algorithms and Applications, 2nd ed. Springer.
10. Muja, M. y Lowe, D. G. (2014): Scalable Nearest Neighbor Algorithms for High Dimensional Data. *Pattern Analysis and Machine Intelligence*, 36.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

SOBRE LOS AUTORES



Josue Garatachia Mejia es estudiante de Ingeniería en Computación participante en proyectos de investigación relacionados a tratamiento de imágenes digitales, ciencia y analítica de datos aplicada.



Erubiel Urbina Gómez es estudiante de Ingeniería en Computación participante en proyectos de investigación relacionados a tratamiento de imágenes digitales, ciencia y analítica de datos aplicada.



Héctor Alejandro Montes Benegas es Dr. en Ciencias Computacionales por la Universidad de Lincoln, UK. Cuenta con varios años de experiencia en empresas del sector público y privado como asesor externo en desarrollo de software, capacitación de personal y asesoría en sistemas de cómputo. Es Profesor de Tiempo Completo en la Facultad de Ingeniería en la UAEMex. Sus intereses de investigación incluyen Computo Evolutivo y Percepción Robótica que involucra la adquisición, procesamiento, análisis e interpretación de los datos que producen para tareas de robótica móvil y aprendizaje máquina. Ha dirigido y codirigido tesis de Licenciatura y de Maestría.



Rosa María Valdovinos Rosas es Dra. en Ciencias Computacionales, miembro del SNI nivel II. Con productos que inciden en varios PRONACES: proyectos de investigación, artículos en revistas indizadas y congresos, capítulos de libro, libros y 1 licencia de patente. Ha formado estudiantes de nivel licenciatura, Maestría y Doctorado. Colabora en la organización de eventos académicos para la divulgación de la ciencia, acciones orientadas a niñas en la computación y vocaciones tempranas. Con más de 1000 citas y un índice H14, el *International AD Scientific index 2025* la considera una las 10,000 científicas más influyentes a nivel Nacional.



Marcelo Romero Huertas obtuvo el doctorado en Ciencias Computacionales en 2011 por la Universidad de York, Inglaterra. Es Profesor-Investigador de Tiempo Completo del Departamento de Ingeniería de la UAEMex desde 2011. Es miembro del Sistema Nacional de Investigadores (CONACYT) y de la Academia Mexicana. de Computación. Sus intereses de investigación incluyen procesamiento de imágenes y reconocimiento de patrones, y protocolos de comunicación de redes de datos.

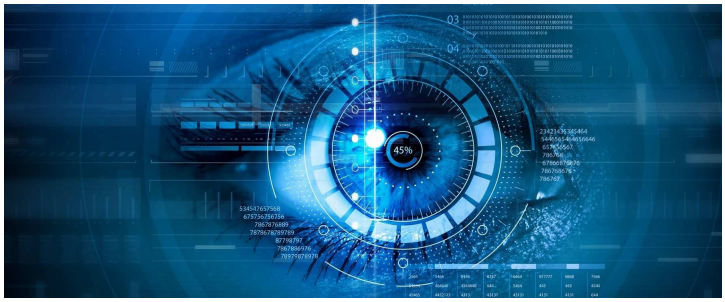


Imagen tomada de: https://futuroelectrico.com/vision-artificial-en-ia/.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO INVITADO

Aprendizaje Profundo para Localización Aérea

Aldrich Alfredo Cabrera-Ponce, Manuel Isidro Martin Ortíz y Jose Martínez-Carranza

Introducción

La determinación de la localización se ha erigido como un reto en diversos campos de la computación, abarcando desde la representación de entornos hasta las aplicaciones en el campo de la robótica. Con el propósito de obtener la localización dentro de un entorno específico, se han concebido diversas soluciones que integran en sus estructuras dispositivos sensoriales como las unidades de medida inercial (por sus siglas en inglés, IMU), sistemas láser y sistemas de posicionamiento global (por sus siglas en inglés, GPS). No obstante, destaca entre ellos el uso de las cámaras por su capacidad para capturar y procesar información del mundo real en imágenes y video. Esto ha motivado al desarrollo de métodos que procesan la información visual con el fin de obtener una coordenada o posición tridimensional de un entorno dado.

En el campo de la robótica aérea a través de drones, se ha afrontado el problema de la localización en situaciones en las cuales el sistema de posicionamiento global experimenta fallos debido a factores climatológicos o a la obstrucción de su señal. De esta manera, se ha adoptado una estrategia que aprovecha la cámara instalada en el dron para establecer una posición de referencia a través de la observación visual. Esto se lleva a cabo en lo que se espera la restauración de la señal del sistema GPS. En esencia, el procesamiento visual emerge como una alternativa para solucionar la problemática de la localización, permitiendo la extracción de información a partir de imágenes capturadas mientras el dron sigue su trayectoria de vuelo.

Bajo esta línea, se han aplicado diversos enfoques basados en visión por computadora, tales como el uso de descriptores visuales [1], sistemas de localización y mapeo simultáneo (o SLAM, acrónimo del vocablo inglés *Simultaneous Localization and Mapping*) [2] y sensores láser para la escaneo del entorno [3]. Estos métodos continúan siendo empleados en la actualidad en aplicaciones comerciales permitiendo procesar la información para adquirir múltiples resultados, uno de ellos la localización. La figura 1 muestra algunas de estas aplicaciones desarrolladas por empresas reconocidas como DJI [4], Parrot, ArduPilot, y Pix4D [5].

Redes Neuronales para Localización

En el campo de la inteligencia artificial se ha empleado el uso del aprendizaje profundo para obtener la localización utilizando las imágenes aéreas capturadas por la cámara de un dron. Así, estas imágenes pasan a través de redes neuronales para extraer información que permita a la computadora conocer el entorno y con ello determinar la posición y localización del dron. Dentro de la red neuronal se realizan procesos de extracción de características interpretables para la máquina, y existen múltiples estructuras arquitectónicas que mejoran el resultado final. PoseNet [6] es un ejemplo de una arquitectura de red neuronal diseñada para el procesamiento de imágenes con el objetivo de estimar la posición.



Figura 1. Aplicaciones para navegación y localización aérea.

Este proceso con redes neuronales se le conoce como entrenamiento, el cual utiliza muestras a partir de conjuntos de datos para que una máquina aprenda de ellos y otorgue un resultado. Sin embargo, realizar este proceso conlleva a un consumo considerable de recursos tanto a nivel de software como de hardware, sin mencionar la gran cantidad de datos requeridos para el entrenamiento y el tiempo de procesamiento.

Inspirados por esto último, mencionaremos tres trabajos de investigación enfocados en la localización aérea utilizando la cámara incorporada en los drones y redes neuronales. Dichas investigaciones son: "*Aerial Geolocalisation for MAVs using PoseNet*"[7], "*Convolutional Neural Networks for Geo-Localisation with a Single Aerial Image*"[8] y "*Multi-model Continual Learning for Camera Localisation from Aerial Images*"[9]. Estos trabajos muestran una metodología que utiliza las imágenes capturadas por los drones con una arquitectura de red neuronal. De esta manera, la red aprende de las imágenes con coordenadas asociadas a ellas para obtener la posición final. La metodología se compone de tres partes:

1. **Adquisición de datos:** Un operador humano guía un dron a lo largo de una zona específica, capturando imágenes que incluyen coordenadas GPS del sitio. Posteriormente, esta información se convierte

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

de un sistema de coordenadas en grados a uno en metros para facilitar la localización.

2. **Entrenamiento de la red neuronal:** En los trabajos considerados, las redes neurales reciben como entradas las imágenes aéreas con las coordenadas, permitiendo la extracción de información visual del entorno. Así se crea un modelo capaz de inferir la localización basada en los ejemplos de entrenamiento proporcionados.
3. **Evaluación del modelo:** Una vez se cuenta con el modelo de estimación, se somete a una prueba utilizando la imagen actual capturada por el dron. Como resultado, se obtiene una posición en metros que se asemeja a la posición real con la cual el modelo fue entrenado. De este modo, el modelo desarrolla la capacidad de obtener la localización a partir de una imagen aérea.

Las redes neuronales son modelos computacionales que pueden aprender patrones a través de neuronas interconectadas

En la figura 2 se presenta un diagrama que representa el proceso de recolección de datos de entrenamiento para la red neuronal, donde las imágenes actúan como entrada para crear un modelo de estimación capaz de inferir una posición en su resultado. Asimismo, en la figura 3 se muestra el diagrama para la localización aérea de un dron utilizando la imagen actual capturada en el lugar. Este diagrama representa la localización aérea utilizando la arquitectura PoseNet presentado Cabrera-Ponce y Martínez-Carranza [7].



Figura 2. Entrenamiento de una red neuronal.

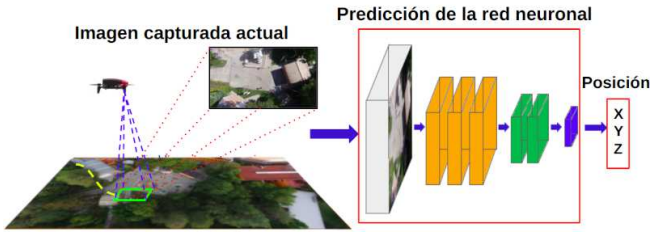


Figura 3. Diagrama general de la localización de un dron Enlace del video `https://youtu.be/tV9N6CIwpz0`.

El trabajo previo aplicó una metodología que permite a un dron localizarse en un escenario utilizando únicamente el modelo de estimación creado por la red neuronal e imágenes aéreas. Esto posibilita tener un sistema de localización de respaldo en casos donde el GPS falle. No obstante, emplear una arquitectura como PoseNet puede consumir recursos computacionales en su entrenamiento, sin mencionar el tiempo de procesamiento, siendo inadecuado para misiones de vuelo en tiempo real.

Debido a este problema, el artículo de Cabrera-Ponce y Martínez-Carranza [8] introduce una nueva arquitectura denominada *Compact PoseNet* (CompactPN). La finalidad de esta arquitectura es la reducción del tiempo para el entrenamiento y la aceleración del proceso de estimación. Bajo esta premisa, se eliminaron ciertos módulos y capas dentro de la red neuronal, conservando la estructura general sin afectar el resultado final. De esta manera, se logra una optimización que disminuye la complejidad de la red neuronal y los procesos internos.

La figura 4 presenta el diseño de la arquitectura CompactPN, concebida para optimizar la aceleración del proceso de estimación y mantener una alta eficiencia. Al mismo tiempo, se reduce el margen de error a menos de 1.68 cm, valor que representa un nivel de precisión aceptable para la localización aérea. También, en el artículo se presenta una evaluación comparativa con arquitecturas líderes en el campo, obteniendo un resultado favorable en términos de rendimiento y precisión.

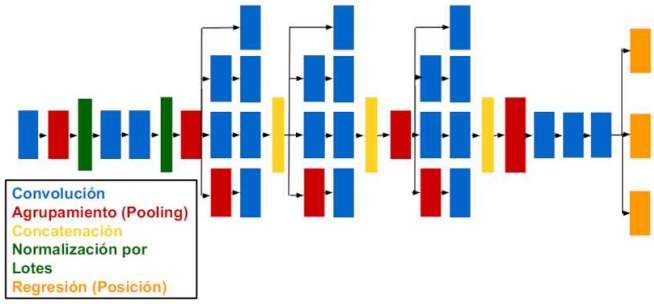


Figura 4. Arquitectura general de CompactPN.

A partir de los resultados obtenidos mediante un enfoque de entrenamiento tradicional, surge una pregunta de gran relevancia tanto en el campo de la robótica como dentro de la comunidad científica: ¿Es posible adquirir conocimiento de manera progresiva mientras un dron navega y captura imágenes con coordenadas GPS?. Esto con el propósito de generar un modelo capaz de proporcionar una posición estimada durante una misión de

vuelo. Teniendo en cuenta que cuando se realiza una misión de vuelo se tome información del escenario, de tal manera que cuando el posicionamiento tradicional del dron falle se da la posibilidad de utilizar un sistema de localización alterno con la información recabada.

Para responder a este cuestionamiento, se han investigado métodos de entrenamiento utilizando recursos limitados con una cantidad reducida de datos. Estos métodos tienen como propósito aprender la información como llega sin olvidar el conocimiento previamente adquirido. En consecuencia, se logra la construcción de un modelo con la capacidad de incorporar nueva información aprendida sin dejar de lado los conocimientos previos derivados de los datos anteriores. Esta reciente área de investigación se le conoce como aprendizaje continuo, aprendizaje incremental o aprendizaje permanente, utilizado principalmente en tareas de clasificación y segmentación.

El aprendizaje continuo permite aprender información nueva sin olvidar el conocimiento previo

Aprendizaje Continuo para Localización

Motivados por esta reciente área se estudian dos de los trabajados de Cabrera-Ponce et al. [9,10] que implementan el método del aprendizaje continuo en tareas de localización aérea. En estos trabajos se introduce una iniciativa que adopta un enfoque de aprendizaje continuo mediante un método llamado reproducción latente. Este enfoque permite la reproducción de características o patrones esenciales fusionando la información actual con los conocimientos previos en una memoria dentro de la red. Como resultado, el modelo se actualiza al incorporar nuevos datos durante el proceso de entrenamiento. La figura 5 ilustra el esquema general del sistema de aprendizaje continuo presentado en el artículo [9].

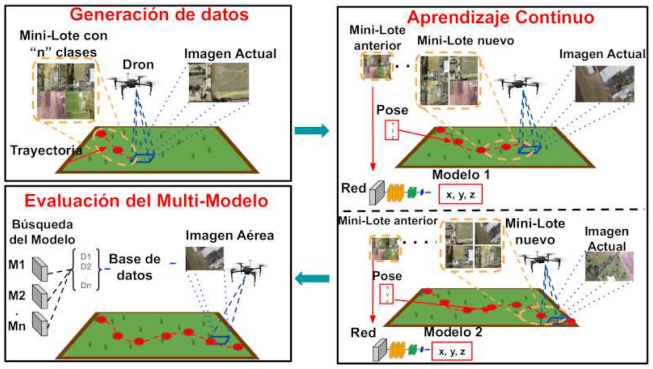


Figura 5. Entrenamiento continuo para localización aérea. Video `https://youtu.be/ZWCB3aQthOE`.

La metodología propuesta por los autores consiste en entrenar un modelo de aprendizaje individual para cada segmento de la trayectoria recorrida por el dron. En otras palabras, cuando el dron se desplaza por un entorno específico, se seleccionan secciones de aproximadamente 50 a 100 metros de distancia en las cuales se capturan imágenes y se entrena un modelo para cada una de ellas. Sin embargo, se presenta un desafío al momento de determinar a qué región pertenece una imagen en particular, teniendo como problema la selección del modelo correspondiente. Para superar este desafío, los autores proponen un método basado en la búsqueda de imágenes utilizando un histograma de color.

Este método selecciona el modelo correspondiente a la sección mientras el dron recorre una determinada área, en lugar de depender exclusivamente de un solo modelo para toda la trayectoria. Para lograr esto, se lleva a cabo una comparación entre la imagen actual capturada por el dron y una base de datos de imágenes previamente entrenadas. Esto permite calcular la similitud de la imagen utilizando un descriptor con información visual de los colores de los píxeles, indicando la región y el modelo correspondiente. La figura 6 ilustra el proceso de búsqueda del modelo por medio de la comparación del histograma de color de la imagen y la base de datos.

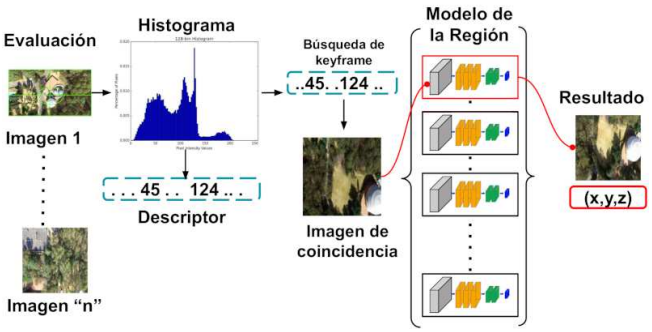


Figura 6. Búsqueda del modelo correspondiente utilizando el histograma de color de la imagen.

Este enfoque de la figura 6 se basa en la conversión de la información visual en datos cuantificables, seguida de una comparación con imágenes previamente almacenadas. Posteriormente, se carga el modelo correspondiente para estimar la posición aérea del dron a partir de la imagen capturada. Como resultado, se obtiene la localización aproximada con la proporcionada por el GPS.

El proceso anterior compara la imagen con una base de datos permitiendo identificar el área correspondiente de la trayectoria para cargar su modelo. Después este modelo se somete al proceso de evaluación para estimar

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

la posición de la imagen de entrada, lo que a su vez brinda la localización actual del dron. Esto nos da una iniciativa para el uso del método de aprendizaje continuo, haciendo que el entrenamiento sea más rápido al dividir la trayectoria en múltiples secciones. La idea es utilizar el método mientras se realiza una misión de vuelo, teniendo la ventaja de identificar al dron cargando el modelo correspondiente a partir de la imagen actual.

Un nuevo trabajo introduce una metodología fundamentada en los principios previamente discutidos para lograr una localización aérea de manera eficaz para los drones. En Cabrera-Ponce et al. [10] se propone incorporar el aprendizaje continuo como una red jerárquica con el propósito de identificar las diferentes regiones de la trayectoria, en lugar de depender de la comparación de histogramas. Esto debido a los efectos desfavorables que podría ocasionar el uso exclusivo del color de las imágenes al intentar identificar la región correcta para la localización.

El aprendizaje continuo permite la localización de un dron en escenarios con cambios constantes

La metodología jerárquica está compuesta de dos redes neuronales llamadas InceptionV4 y MobileNetV1. Este proceso simula el aprendizaje humano, donde una parte de nuestro cerebro retiene información sobre ciertos aspectos y otra parte características significativas. Bajo esta premisa, una red neuronal, InceptionV4, aprende ejemplos de las regiones de la trayectoria total para identificar que modelo carga para la localización final. Es decir, la primera red neuronal se encarga de identificar la región que concuerda con la imagen de entrada dentro de un conjunto de 10 regiones. Esto resulta más adecuado para encontrar la similitud de la imagen en la base de datos, con el fin de obtener un mejor rendimiento.

La figura 7 presenta una ilustración que visualiza la generación de las secciones a lo largo de la trayectoria, donde los rectángulos amarillos representan estas áreas llamadas submapas. La línea roja traza la trayectoria seguida por el dron, mientras que los puntos rojos señalan los lugares donde se tomaron las imágenes, cada uno de ellos con sus coordenadas GPS en metros. Esta representación proporciona una visión del entorno en el cual el dron realiza su navegación y captura de imágenes.

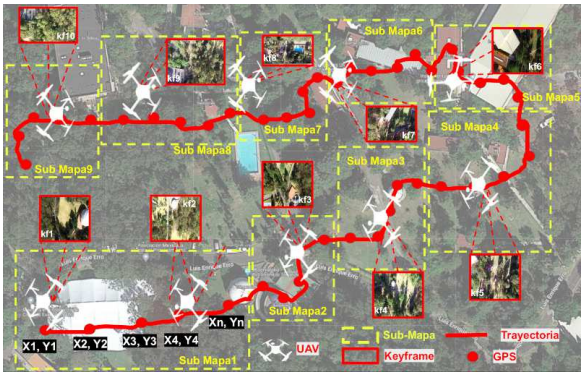


Figura 7. Secciones a lo largo de la trayectoria.

Finalmente, los autores respaldan la idea de que una metodología que emplea múltiples modelos en un sistema jerárquico, obtiene mejores resultados para obtener la localización. Bajo este argumento el dron con un sistema de este estilo será capaz de identificar la sesión en la que se encuentra para luego estimar la posición utilizando la imagen actual. Esta metodología tiene como objetivo reducir el error en la posicion obtenida y previene confusiones de similitud con el conjunto de datos que podría surgir con el uso del histograma de color.

Para evaluar el método de aprendizaje continuo de manera jerárquica, se utilizó una red neuronal entrenada de manera tradicional, la búsqueda de la región utilizando el histograma de color y un método de mapeo con SLAM. En la figura 8 se presentan las trayectorias de entrenamiento donde se realizaron las evaluaciones del aprendizaje jerárquico. En ellas se muestran, como círculos, las posiciones correctamente recuperadas de la región correspondiente a la imagen actual del dron.

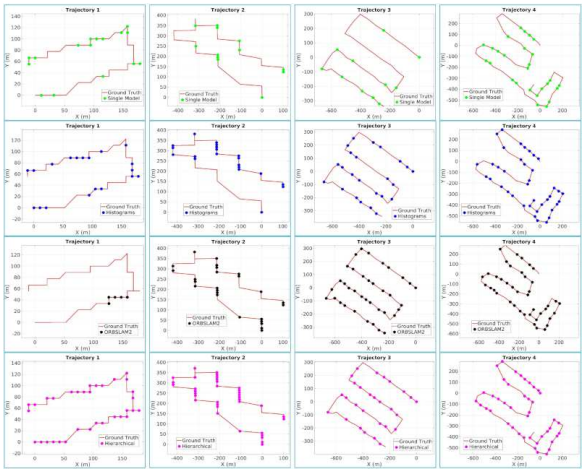


Figura 8. Localización recuperada jerárquicamente.

La primera fila corresponde a las evaluaciones con un único modelo utilizando todas las imágenes de evaluación. La segunda fila representa la localizacion utilizando la busqueda de la región con el histograma de color, la

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

tercera corresponde a la localización con un sistema de mapeo SLAM. Finalmente en la última fila se muestra las posiciones recuperadas utilizando el método jerárquico con el aprendizaje continuo, siendo este la metodología que mayores resultados obtuvo al obtener correctamente la localización lo que permite al dron continuar con su navegación en el escenario.

Conclusiones

A través de estos artículos, destacamos la importancia de los métodos de aprendizaje profundo y continuo para obtener la localización aérea utilizando imágenes capturada con drones. Estos métodos representan avances en el área de la robótica aérea y la visión por computadora, aprovechando la información visual para crear modelos de estimación de posición utilizando redes neuronales. Estos enfoques abren nuevas posibilidades para tener sistemas de localización alternos cuando los sistemas GPS fallen, de tal manera que los sistemas basados en imágenes puedan ser utilizados para recuperar la posición del dron en la zona donde la señal fue perdida. Finalmente, la localización es importante en cualquier campo de la robótica, ofreciéndole al robot conocimiento de dónde se encuentra mientras desarrolla una misión.✷

REFERENCIAS

1. Dusmanu, M., Miksik, O., Schönberger, J. L. y Pollefeys, M. (2021). Cross-descriptor visual localization and mapping. En *Proceedings of the IEEE/CVF International Conference on Computer Visionm¿*, 6058-6067.
2. Mur-Artal, R. y Tardós, J. D. (2017). Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5), 1255-1262.
3. Carrasco, P., Cuesta, F., Caballero, R., Pérez-Grau, F. J. y Viguria, A. (2021). Monte-Carlo Localization for Aerial Robots using 3D LiDAR and UWB sensing. En *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, 354-360
4. DJI C. (2024) DJI. Matrice 300 RTK. Recuperado el 4 de Junio de 2024, de `https://www.dji.com/mx/matrice-300`.
5. Pix4D, (2024). Professional photogrammetry and drone mapping. Recuperado el 4 de Junio de 2024, de `https://www.pix4d.com/`.
6. Kendall, A., Grimes, M. y Cipolla, R. (2015). Posenet: A convolutional network for real-time 6-dof camera relocalization. En `ProceedingsoftheIEEEinternationalconferenceoncomputervision`, 2938-2946.
7. Cabrera-Ponce, A. A. y Martinez-Carranza, J. (2019). Aerial geo-localisation for MAVs using PoseNet. En *2019 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED UAS)*, 192-198.
8. Cabrera-Ponce, A. A. y Martinez-Carranza, J. (2022). Convolutional Neural Networks for Geo-Localisation with a Single Aerial Image. *Journal of Real-Time Image Processing*, 19(3), 565-575.
9. Cabrera-Ponce, A. A., Martin-Ortiz, M. y Martinez-Carranza, J. (2022). Multi-Model Continual Learning for Camera Localisation from Aerial Images. En *Proceedings of the 13th Annual International Micro Air Vehicle Conference and Competition (IMAV)*, 103–109.
10. Cabrera-Ponce A. A., Martin-Ortiz, M. y Martinez-Carranza, J. (2023). Hierarchical continual learning for single image aerial localisation. En *Proceedings of the 14th Annual International Micro Air Vehicle Conference and Competition (IMAV)*, 39-47.

SOBRE LOS AUTORES



Aldrich Alfredo Cabrera Ponce es estudiante de doctorado en la Coordinación de Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla. Obtuvo su grado de maestro en Ciencias en Ciencias de la Computación en el Instituto Nacional de Astrofísica Óptica y Electrónica en 2021. Sus intereses de investigación incluyen robótica e inteligencia artificial, actualmente esta trabajando en localización aérea utilizando aprendizaje continuo con imágenes capturada con drones.



Manuel Isidro Martin Ortíz es Investigador Titular C en la Coordinación de Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla. En 2015 fue distinguido como Coordinador Académico del Laboratorio Nacional de Supercómputo del Sureste de México. Su principal tema de investigación es sobre procesamiento digital de imágenes y procesamiento paralelo y distribuido, para la cual desarrolla aplicaciones en sistemas especializados para ciencias naturales y exactas.



Jose Martínez Carranza es Investigador Senior C (Equivalente a Profesor Titular) en la Coordinación de Ciencias Computacionales del INAOE. En 2015 recibió la distinción Royal Society-Newton Advanced Fellowship, otorgada por el Fondo Newton. Es fundador y líder del grupo de investigación de Sistemas Aéreos No Tripulados, y su equipo, QuetzalC++, ha participado en competencias internacionales de gran prestigio, obteniendo diversos reconocimientos, destacando el 1er Lugar en la competencia IROS 2017 Autonomous Drone Racing y el 1er Lugar Regional (Latinoamérica) en la OpenCV AI Competition 2021. Se desempeñó como Presidente General de la 12ª Conferencia Internacional de Micro Vehículos Aéreos (IMAV) 2021 y como Editor Asociado de la conferencia internacional IEEE IROS en 2022-2023. Es Miembro Senior del IEEE y Nivel 2 en el Sistema Nacional de Investigadores de México.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

IA & Educación

Yasmín Hernández, Julieta Noguez y Lucía Barrón

Aplicaciones de visión computacional en la educación

El sentido de la vista y el cerebro identifican objetos a partir de la información que se recibe a través de los ojos. La visión computacional es una rama de la Inteligencia Artificial que trata de imitar esta capacidad en las computadoras, de forma que mediante la interpretación de las imágenes se puedan reconocer los diversos objetos en el ambiente y su posición en el espacio [1]. La visión computacional incluye métodos para adquirir, procesar, analizar y comprender las imágenes del mundo real con el fin de producir datos que puedan ser tratados por una computadora [2]. La Fig. 1 muestra el esquema general de la visión computacional. La imagen de entrada se procesa para extraer diversos atributos, y se obtiene como salida una descripción de la imagen. Esta descripción se usa para tomar decisiones y llevar a cabo acciones.

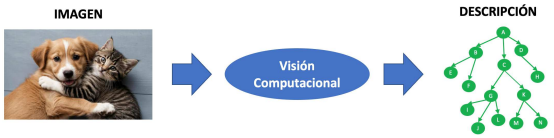


Figura 1. Esquema general de visión computacional [1].

La visión computacional necesita una gran cantidad de datos. Ejecuta análisis de datos hasta que logra distinguir en las imágenes y reconocer objetos o personas. Utiliza dos tecnologías esenciales: aprendizaje automático del tipo aprendizaje profundo y una red neuronal convolucional. El aprendizaje automático permite que la computadora se enseñe a sí misma sobre el contexto de los datos visuales. Si se alimentan suficientes datos a través del modelo, la computadora "mirará"los datos y aprenderá a distinguir una imagen de otra. Las redes neuronales convolucionales se usan para comprender imágenes, mientras que las redes neuronales recurrentes se utilizan de manera similar para aplicaciones de video y ayudar a las computadoras a comprender cómo se relacionan las imágenes en una serie de cuadros [2].

Las aplicaciones de la visión computacional se han extendido a muchos campos. En la educación ha demostrado ser útil para automatizar tareas administrativas y para apoyo educativo. Por ejemplo: facilitar experiencias de aprendizaje personalizadas, apoyar el aprendizaje remoto, automatizar la evaluación de tareas y exámenes, dar seguimiento a la atención y participación de los estudiantes, mejorar la gestión del aula y en general en tareas que consumen tiempo [3].

La motivación principal para aplicar la visión computacional en la educación es apoyar el aprendizaje del estudiante. Esta tecnología puede realizar un seguimiento de la atención de los estudiantes durante el proceso de aprendizaje y ofrecer información en tiempo real sobre su estado emocional, postura y contacto visual [3]. Estos datos son útiles para los profesores, por ejemplo, para cambiar la dinámica o la velocidad de la enseñanza y ser más sensibles a las necesidades individuales. Por otro lado, el software educativo puede afectar la forma en que los estudiantes utilizan los materiales educativos y se enganchan con el proceso de aprendizaje. La visión computacional ayuda a los estudiantes a avanzar en el aprendizaje, así como a lograr la adaptación requerida por diferentes estudiantes con diversas capacidades de aprendizaje. Por ejemplo, si un estudiante tiene dificultades para leer un texto, el sistema podría ofrecer definiciones más elaboradas o lecturas adicionales sobre el mismo tema.

Las evaluaciones son una parte integral de la educación; sin embargo, es una tarea tediosa. La visión computacional puede utilizarse para evaluar exámenes, en donde las preguntas de opción múltiple o respuesta corta pueden calificar de manera automática. De igual manera, la asistencia y la supervisión de exámenes son fundamentales para las clases presenciales y en línea. La visión computacional simplifica la verificación la identidad del estudiante y la exactitud del registro de asistencia. Además, en las clases en línea, particularmente durante los exámenes, puede observar a los estudiantes e informar comportamientos anormales, tales como apartar la mirada de la pantalla con frecuencia, utilizar dispositivos no autorizados, entre otros.

Por otra parte, a nivel mundial han aumentado los problemas de seguridad al interior de las escuelas. La visión computacional puede verificar la ocupación del aula, el distanciamiento social y cualquier otro peligro en clase y enviar alertas en tiempo real, lo que puede ayudar a que las escuelas sean más seguras.

Muchas de las aplicaciones anteriores se basan en el reconocimiento facial, lo que permite llevar a cabo las tareas en el ámbito educativo. Hernández y colegas dis-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

cuten la importancia de contar a los estudiantes y la identificación de sus sentimientos [4]. El reconocimiento facial se lleva a cabo mediante el uso de YOLOv3, DeepFace y un repositorio de imágenes de aulas virtuales [5]. YOLO es un sistema para la detección de objetos mientras que Deepface es un sistema de reconocimiento facial. La Fig. 2. muestra el reconocimiento el conteo y reconocimiento de estudiantes en un aula. Esta aplicación es el inicio para el desarrollo de un sistema que apoye en la seguridad de las aulas de una Universidad Pública en México.



Figura 2. El reconocimiento el conteo y reconocimiento de estudiantes en un aula [4].

Si bien los beneficios de la visión computacional en la educación son muchos, su implementación conlleva ciertos desafíos. Un aspecto es la ética y el uso adecuado de estas herramientas. Tal vez el problema más crítico sea la privacidad; se debe proteger la información de los estudiantes, incluidos los datos faciales y de comportamiento para generar cierta confianza entre los padres, los profesores y los propios estudiantes.✳

REFERENCIAS

1. Sucar, L. E. y Gómez G. (Sin fecha). Visión Computacional, https://ccc.inaoep.mx/~esucar/Libros/vision-sucar-gomez.pdf.
2. Shapiro, L. G. y Stockman G. C. (2001). Computer Vision, Prentice Hall.
3. Svet, O. (2024). Computer vision in education: benefits that accelerate the industry, https://computools.com/computer-vision-in-education/.
4. Hernández, J. A., Hernández, Y., Rivera, L., y Bonilla, J. C. (2023). A New Approach for Counting and Identification of Students Sentiments. En *proceedings of WILE 2023* , 29-40. Cham: Springer Nature Switzerland.
5. Roboflow. (2024). Bitirme-Son Comúyer Vision Project, https://universe.roboflow.com/bitirmeprojesi-4ecb2/bitirme-son.



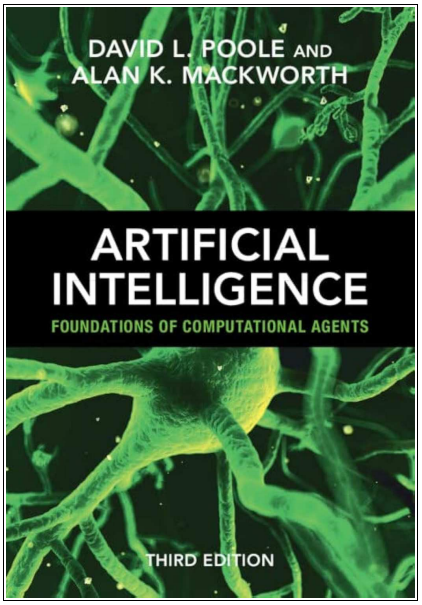
Imagen tomada de: https://megaprofe.es/impacto-de-la-inteligencia-artificial-en-el-aprendizaje/

Deskubriendo Konocimiento

Alejandro Guerra Hernández y **Gildardo Sánchez Ante**
deskubriendokonocimiento@komputersapiens.org

Revisión de "Artificial Intelligence: Foundations of Computational Agents" *

Dr. Alejandro Guerra Hernández
Universidad Veracruzana
Instituto de Investigaciones en Inteligencia Artificial



Portada del libro*

Para quienes nos dedicamos a la enseñanza y el estudio de la Inteligencia Artificial (IA), un libro de texto nuevo siempre es de interés y pertinencia dados los avances continuos en el área y el tiempo que toma llevar a buen puerto tal empresa. En esta ocasión les propongo una revisión de la tercera edición del libro de David L. Poole y Alan K. Mackworth, que actualiza la edición previa del año 2010. Ambos autores son profesores eméritos del Departamento de Ciencias de la Computación en la Universidad de la Columbia Británica. Su amplia trayectoria ha sido reconocida con el premio a una vida de logros de la Asociación Canadiense de IA.

La idea central del libro es similar a la del texto de Russell y Norvig ** que revisamos en nuestra columna del año XV, No. 1 (enero-abril 2023): Estudiar agentes inteligentes que actúan en un ambiente. La propuesta es partir de agentes simples que actúan en ambientes determinista simples, e incrementalmente ampliar sus capacidades para contender con ambientes cada vez más complejos. Para guiar esta aproximación incremental, los autores definen un espacio de diseños posibles con base en diez dimensiones: modularidad, horizonte al planear, representación, límites computacionales, aprendizaje, percepción y modelado de la incertidumbre, preferencias, número de agentes e interactividad. Este espacio de diseño provee un marco conceptual donde ubicar las diferentes técnicas propuestas por la IA en relación con el agente que las utiliza en un ambiente determinado.

Para ejemplificar tal aproximación, los autores proponen cinco aplicaciones prototípicas que serán utilizadas a lo largo del libro: un robot de ayuda y envío de objetos en un laboratorio con cubículos, pasillos, etc; un asistente de diagnóstico de fallos en una red eléctrica; un tutor de estudiantes de física elemental; un agente de compras; y una casa inteligente.

La estructura del libro refleja claramente la aproximación adoptada por los autores. La parte uno introduce los conceptos de IA y agente inteligente, presenta las aplicaciones seleccionadas para los ejemplos y define el espacio de diseño de los agentes. También considera diferentes arquitecturas de agente y las técnicas de control jerárquico. La segunda parte del libro contiende con técnicas de razonamiento y planificación bajo certidumbre, por ejemplo, búsquedas en espacios de soluciones, razonamiento con restricciones, inferencia con lógica proposicional y planificación determinista. La tercera parte cubre el aprendizaje y el razonamiento bajo incertidumbre, incluyendo el aprendizaje supervisado, las redes neuronales, el aprendizaje profundo, las redes de creencias, el aprendizaje bayesiano y la causalidad. La cuarta parte aborda la planificación y actuación bajo incertidumbre, considerando el uso de preferencias y utilidades en la toma

*D. L. Poole y A. K. Mackworth. *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press, Cambridge, UK, tercera ed., 2023.

**S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Pearson, New York, NY, USA, fourth edition, 2020.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

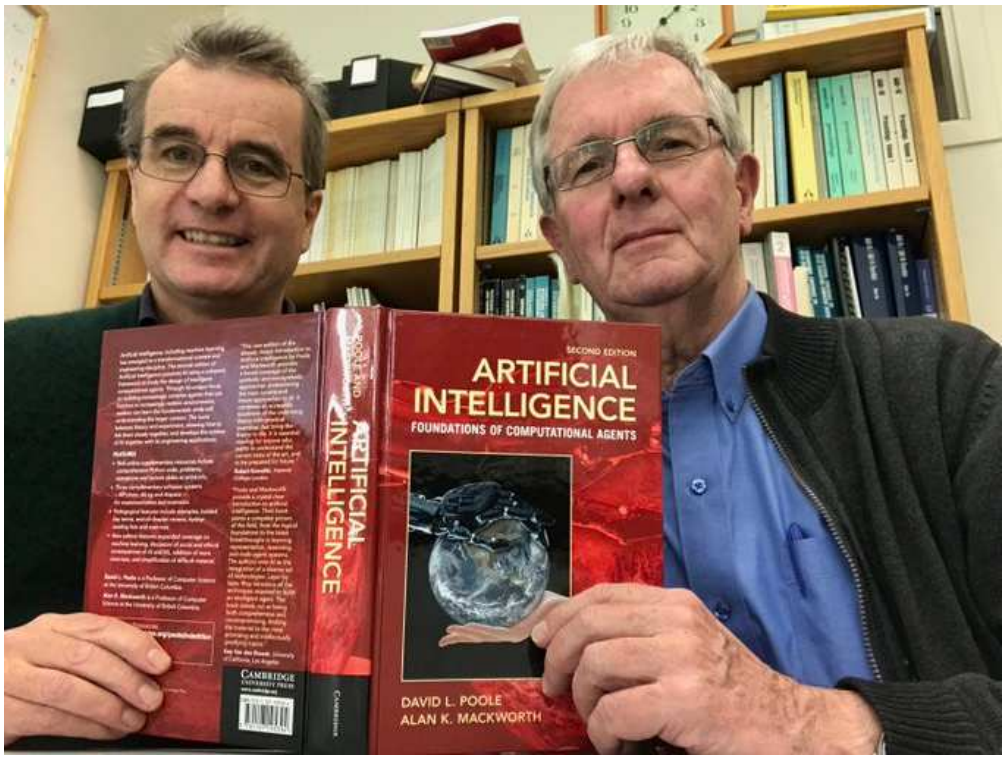
de decisiones, el aprendizaje por refuerzo y los sistemas multi-agente. La quinta parte del libro aborda las representaciones de individuos y relaciones considerando la lógica de primer orden, los grafos de conocimiento, las ontologías, el aprendizaje relacional y el razonamiento lógico probabilista. La sexta parte del libro provee una visión panorámica del área considerando su impacto social, una retrospectiva y su visión del futuro. Dos apéndices proveen nociones matemáticas preliminares y referencias algunos paquetes de software abiertos utilizados.

Las consideraciones en torno al impacto social de la IA son una de las novedades en esta edición. De hecho, además de la perspectiva ofrecida en el capítulo trece sobre estos temas, cada capítulo del libro discute los beneficios y el posible impacto negativo de las técnicas propuestas. Las otras novedades incluyen un nuevo capítulo sobre redes neuronales y el aprendizaje profundo; y otro más sobre causalidad. Por supuesto, la bibliografía propuesta ha sido revisada a profundidad en todo el libro.

En mi opinión hay tres decisiones en el diseño del libro que el lector agradecerá. El espacio de diseño de los agentes provee un marco coherente donde ubicar las técnicas abordadas, resultando en un texto menos fraccionado que otros; El énfasis en los fundamentos de las técnicas presentadas, así como el pseudo-código y ejemplos que los acompañan; Y la organización misma del libro, que aborda primero las representaciones basada en atributos, luego las proposicionales y finalmente las relacionales, lo cual refleja mejor la propuesta de diseños incrementalmente más complejos y coincide con el orden de los avances en IA de las últimas décadas.

El libro completo está disponible gratuitamente en línea en la página `https://artint.info`. También puede adquirirse como libro electrónico o físico. La implementación de los algoritmos presentados en Python se puede descargar en la página `https://AIPython.org`.

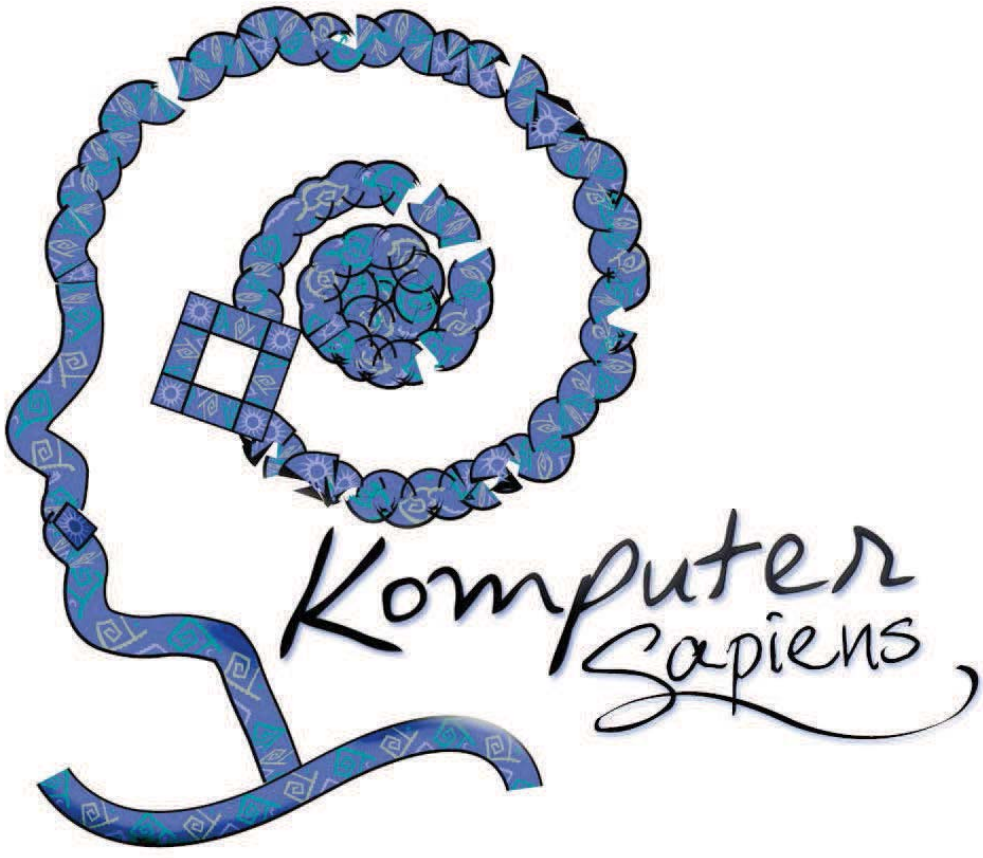
La comparación con el libro de texto de Russell y Norvig parece obligada. El texto que hoy nos ocupa es menos enciclopédico en el sentido que opta por la profundidad en los temas presentados, más que por cubrir todos los últimos avances en la IA. La propuesta del espacio de diseños y los casos de estudio van un paso más allá en la dirección marcada por Russell y Norvig. La oportunidad de evaluar el libro gratuitamente es de agradecer (y su costo también, por si deciden adquirirlo). Me parece que ambos textos son adecuados para un curso de introducción a la IA, pero éste es más amigable para un lector informal interesado en estos temas. ✳



Autores del Libro. *Imagen tomada del sitio:*
https://x.com/AlanMackworth/status/925467655158472705?lang=ar.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Indizada en el IRMDCT de CONACYT y en Latindex



¡Publique en Komputer Sapiens!

CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS

Komputer Sapiens solicita artículos de divulgación en todos los temas de Inteligencia Artificial, dirigidos a un amplio público conformado por estudiantes, académicos, empresarios, tomadores de decisiones y consultores. Komputer Sapiens es patrocinada por la **SMIA, la Sociedad Mexicana de Inteligencia Artificial**

SMIA
www.smia.org.mx

Instrucciones para autores e información general: http://www.komputersapiens.org
Síguenos en las redes sociales: www.facebook.com/Komputer.Sapiens, twitter.com/KomputerSapiens